Advanced Statistical Learning Approaches to Healthcare

by

Serhat Simsek

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 3, 2019

Keywords: Statistical and Machine Learning, Healthcare Analytics, Data mining, Genetic Algorithm, Simulated Annealing, Particle Swarm Optimization

Approved by

Mark Carpenter, Chair, Professor of Mathematics and Statistics
Philippe Gaillard, Associate Professor of Mathematics and Statistics
Erkan Nane, Associate Professor of Mathematics and Statistics
Hans Werner Van Wyk, Assistant Professor of Mathematics and Statistics

Abstract

In this dissertation, we propose a novel statistical modeling methodology that effectively analyzes and extracts information from large datasets. Two real datasets were employed to validate the proposed methodology: breast cancer patient dataset and a no-show patient dataset. The objective of this dissertation is then to extract novel and useful information from these large and complex datasets by applying the proposed methodology that involves statistical and machine learning, heuristic optimization, and advanced data resampling techniques. Particularly, in the first study (breast cancer), we developed prediction models for breast cancer 1-, 5- and 10-year survivability and studied the variables whose importance for survival change over time. The obtained results revealed that certain variables lose their importance over time, while others gain importance. This information can assist medical practitioners in identifying specific subsets of variables to focus on in different periods, which will in turn lead to more effective and efficient cancer care. In the second study, we employ several statistical and machine learning algorithms to accurately predict no-show patients and to obtain patient-specific risk scores. Also, we developed a prediction tool that enables practitioners to use the proposed model without having any knowledge in statistics, optimization, and programming. Therefore, the overarching goal of the study is to develop a parsimonious model embedded within the prediction tool so as to enable healthcare professionals to improve clinic utilization, and improve patient outcomes, at the same time, decrease the costs that originate from patient no-shows.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Mark Carpenter, for providing me with constructive comments, tireless support, and wise directions during my Ph.D. education. Without having his guidance, I would not be able to go through this challenging Ph.D. process. I would also like to thank my committee members Dr. Philippe Gaillard, Dr. Erkan Nane and Dr. Hans Werner Van Wyk for being a part of my study and my academic life.

I would like to express my deepest gratitude and appreciation to my beautiful wife, Sibel Simsek. Thank you for all the times you have supported me. Without your patient and true love, nothing would have been possible. Your being drives my success.

To my mother, Nefise Simsek, after becoming a father I realize that I always take for granted your unconditional love and endless support that most deserve my gratitude. Thank you so much for being my mother. I will always remember, with happiness, the moments that we have spent together.

Also, I would like to thank my father, Ramazan Simsek, mother-in-law, Sultan Solmaz, and father-in-law, Erhan Solmaz, thank you all so much for always being with us whenever we need you the most. Last but not least, I would like to thank my lovely kids: my son, Yusuf, and my daughter, Afife.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The advent of efficient and cost-effective technology for recording and storing data has led to an exponential growth in data volume, which in turn provides plenty of opportunities for statisticians to effectively analyze and to extract necessary information from large datasets. In this dissertation, we propose a novel statistical modeling approach that involves several statistical and machine learning techniques to extract novel, hidden and useful information from large data. In the proposed approach, we attempt to obtain the most parsimonious model possible. Therefore, optimization algorithms such as Particle Swarm Optimization (PSO), Simulated Annealing (SA), and Genetic Algorithms (GA) are deployed in addition to statistical Least Absolute Shrinkage and Selection Operator (LASSO) technique, in the variable selection process. We employ two different real datasets to validate the proposed methodology those of breast cancer patients and of no-show patients. Since the distribution of the class size was highly imbalanced for both datasets, several data balancing techniques, random under-sampling (RUS), random over-sampling (ROS), and synthetic minority over-sampling (SMOTE), are employed. A large number of statistical and machine learning algorithms as well as their combinations (ensemble learner) are used along with a k-fold cross-validation technique. Following the model building process, we study the importance of the variables that were used in the best performing models via Sensitivity Analysis (SA) and later combined by employing an Information Fusion (IF) technique.

The dissertation is organized as follows. In the second chapter, we apply our methodology to model breast cancer survivability, where the changing effect of the variables on breast cancer survivability is studied. In Chapter 3, the prediction of no-show patients is studied and patient-specific risk score provided. In Chapter 4, we employ a tree augmented Bayesian belief network to both predict patients who will not keep their appointments and to explore conditional relations among the predictors. Moreover, a prediction tool utilizing the Bayesian model is developed, allowing practitioners to make use of our study.

Chapter 2

A Hybrid Data Mining Approach for Identifying the Temporal Effects of Variables Associated with Breast Cancer Survival

2.1 Abstract

Predicting breast cancer survival is crucial for practitioners to determine possible outcomes and make better treatment plans for the patients. In this study, a hybrid data mining based methodology was constructed to differentiate the variables whose importance for survival change over time. Therefore, the importance of variables was determined for three different time periods (i.e. one, five, and ten years). To conduct such an analysis, the most parsimonious models were constructed by employing one regression analysis method—Least Absolute Shrinkage and Selection Operator (LASSO), and one metaheuristic optimization method, namely a Genetic Algorithm (GA). Due to the high imbalance between the number of survivals (majority) and deaths (minority), two well-known resampling procedures—Random Under-sampling (RUS) and Synthetic Minority Over-sampling Technique (SMOTE)—were applied to increase the performance of the classification models. In the final stage, two data mining models, namely Artificial Neural Networks (ANNs) and Logistic Regression (LR), were utilized along with 10-fold cross-validation. Sensitivity analysis (SA) was conducted for each model to identify the importance of each variable for a certain model and time period. The obtained results revealed that certain variables lose their importance over time, while others gain importance. This information can assist medical practitioners in identifying specific subsets of variables to focus on in different periods, which will in turn lead to a more effective and efficient cancer care. Moreover, the study's findings indicate that extremely parsimonious models can be developed by adopting a purely data-driven approach, rather than eliminating the variables manually. Such methodology can also be applied in treating other types of cancer.

2.2 Introduction

Breast cancer is the second most common cause of death after lung cancer among women in the U.S (Allegra et al., 2010). While the number of new breast cancer cases reported each year has been stable over the last ten years, in 2015, the American Cancer Society estimated that there would be 267,267 new cases and 40,610 deaths due to breast cancer in 2020 in the U.S. (Weir et al., 2015). With constant improvements in technology and patient screening, the 5-year survival rate of breast cancer patients in the U.S. has increased from 75.2% to 98.9% over the 1975−2010 period (Howlader et al., 2013).

Predicting the survival of cancer patients is important. Patients facing a cancer diagnosis want to know what their future may entail and what prognosis they can expect. This can help patients and their loved ones to plan treatment options, think about any possible lifestyle changes, and make financial and other important decisions. It also helps doctors identify possible treatment paths, understand possible outcomes of different prognoses, and make more data-informed decisions, in addition to being better able to guide their patients. On the other hand, making a medical prognosis and determining survival factors is very challenging. These factors can be grouped into two categories, comprising of chronological and biological factors (Bundred, 2001). The chronological factors include those that change with time and directly affect the prognosis, such as lymph node status, tumor size, and histological stage. As lymph nodes filter harmful substances, their increased involvement signifies a potentially wider spreading of cancerous cells and adversely impacts the prognosis (Bundred, 2001; Rampaul et al., 2001). Tumor size enlargement also adversely affects the prognosis and reduces the probability of survival. Moreover, the histological stage depicts whether a tumor is local or has spread to healthy breast tissues, or to other parts of the body. Biological factors, such as histological grade, on the other hand, pertain to the behavioral status of the tumor. Histological grade is indicative of tumor aggressiveness, whereby “low grade” cancers are less aggressive than “high grade” tumors (Bundred, 2001; Rampaul et al., 2001).

Determining the importance of these and other variables is as crucial as predicting patient survival, as it enables medical practitioners to better plan the treatment (Desforges et al., 1992). However, the impact of certain variables on the survival may vary across time. Quantin et al. (1999) provided empirical evidence for time-dependent effects of prognostic factors in colon cancer. The authors demonstrated that multivariate analysis can capture a variety of patterns in determining prognostic factors in different time ranges. Some variables may not be important during the first few years following cancer diagnosis, but might gain importance in the later stages of the illness. Being cognizant of this dynamic effect of certain variables enables practitioners to optimize the treatment.

On the other hand, cancer diagnosis and treatment methods require extensive medical research and analysis, which requires resources. Additionally, large and information-rich datasets are being collected and retained by medical practitioners and hospitals for all types of patients. In an effort to capitalize on such data and gain useful knowledge that can potentially contribute to cancer research efforts, as well as to the practitioners' and caregivers' decision-making process, statistical and data mining methodologies are increasingly being applied (Pendharkar et al., 1999; Thongkam et al., 2009; Kulkarni et al., 2011; Ting et al., 2018; Lu et al., 2019). Machine Learning (ML) algorithms are widely utilized for this purpose, as they have been proven effective in extracting hidden patterns from large and complex datasets (Fayyad et al., 1996). They can be broadly classified into three main categories, namely supervised, unsupervised, and semi-supervised algorithms. In supervised algorithms, such as ANNs, Bayesian Belief Networks (BBNs), and Decision Trees (DTs), a training set is used to supervise the specifications of input data to the ML algorithm in order to estimate the desired output. In contrast, in unsupervised algorithms, such as hierarchical clustering, patterns in data are discovered by the models without a learning process. Semi-supervised algorithms are, on the other hand, a combination of supervised and unsupervised algorithms (Dougherty et al., 1995; Huang et al., 2014).

Various machine learning and data analytics techniques have been used extensively to process and analyze data (Rouyendegh et al., 2016; Rouyendegh et al., 2018; Simsek et al., 2018), such as medical data (Dag et al., 2017; Nasir et al., 2019; South-Winter et al., 2018; Topuz et al., 2018), including breast cancer datasets, for survival and variable selection analysis (Gunasundari et al., 2016; Gupta & Sharma, 2011; Ryu et al., 2007; Walczak & Velanovich, 2018; West et al., 2005; Zolbanin et al., 2015). In breast cancer research, the focus is primarily given to identification of important predictors, predominantly from groups of genes, that assists to detect the cause of cancer. Authors of such studies mostly used unsupervised learning algorithms to detect the important genes by applying clustering methods. Other authors have applied supervised learning algorithms to clinical and sociodemographic predictors to find important factors and predict expected survival times based on these predictors. In the following subsections, a brief overview of the studies most pertinent for the current investigation is given.

As noted above, authors of most extant studies in this field used datasets comprised of various genes, as genetic data is particularly well suited for unsupervised methods like clustering. The aim of adopting clustering methods in these studies was identifying the most important genes for each cancer type. For example, Li et al. (2009) developed an unsupervised algorithm to detect previously unrecognized prognostic groups and features for gliomas, which is currently the most common brain tumor type in adults. In adopting this strategy, the authors aimed to overcome the bias generated by *a priori* gene selection that is subsequently adopted in classification (Alizadeh et al., 2000; Beer et al., 2002; Lapointe et al., 2004). The model developed by Li et al. (2009) determines two groups of gliomas, which are then clustered into six nested subgroups. Different sets of classifiers are identified for each group, which is then validated using different datasets. Although unsupervised learning allowed these authors to identify, using genomic information, relevant cancer subtypes that may coexist within a tumor, its main drawback is potential identification of cancer subtypes that are not pertinent to patient survival. In an earlier study, Lapointe et al. (2004) developed a semi-supervised algorithm in which gene data was combined with clinical data to determine patient survival times. The idea behind this approach was to use clinical data to identify relevant genes that will be employed in the subsequent unsupervised clustering. The dataset used by Lapointe et al. consisted of 7399 genes from 240 breast cancer patients. The authors used Cox proportional hazard scores of 7399 genes based on 160 training observations and ranked the genes based on their Cox scores, accordingly. Next, the test group was clustered using the 25 top scoring genes to predict the patients' survival time and associated probabilities.

While these studies are highly important in the context of breast cancer survival analysis and early prognosis discussions, they may not be as useful from the perspective of caregivers, as they may not be pertinent to the treatment decisions related to sustaining patient's quality of life (Bradbury-Huang, 2010). Moreover, since the aforementioned studies mostly focused on genetic data, the importance changes in various other patient factors, such as cancer stage or treatment type, which have to be considered separately by decision makers.

Authors of extant studies based on clinical data tend to use supervised ML algorithms for cancer survival prediction and variable analysis. For example, Lundin et al. (1999) used ANNs to predict 5-, 10-, and 15-year breast cancer survivability based on the data pertaining to 951 patients. Their results demonstrate that stage, nodal status, tumor size, age, mitotic count, nuclear pleomorphism, tubule formulation, and tumor necrosis are the most important features determining cancer survivability. In a later study, Delen et al. (2005) used ANNs, as well as decision trees and logistic regression, which they incorporated into a hybrid model for predicting breast cancer survivability over a 5-year period. The authors utilized the Surveillance, Epidemiology, and End Results (SEER) dataset, which includes 433,272 patient records spanning the 1973−2000 period, in order to determine the impact of certain variables on cancer survival. The study findings indicate that cancer grade, number of primary tumors, cancer stage, radiation treatment, lymph node involvement, and tumor size are the most important predictors of breast cancer survival.

Extensive research has also been conducted in order to improve the predictive capacity of already developed models. For example, Thongkam et al. (2009) proposed a hybrid breast cancer survival model that incorporates Support Vector Machines (SVMs), along with an outlier-filtering approach, followed by simple over-sampling. The results reported by these authors demonstrate that the hybrid scheme significantly improved the performance of SVM. Khan et al. (2008) also evaluated a hybrid data mining method for cancer prognosis by utilizing interference techniques together with fuzzy decision tree models. Their results demonstrated that an independently applied crisp classification underperforms a fuzzy decision tree model. In an earlier study, Pendharkar et al. (1999) utilized association analysis as a variable selection tool to uncover the relations between breast cancer occurrence and cancer related factors. The authors used data envelopment analysis (DEA), ANN and discriminant analysis (DA) to predict cancer diagnosis. Their results reveal that while increasing the training sample size improves the prediction accuracy of DEA and DA, the accuracy of ANN remains relatively unchanged. Zupan et al. (2000) used classification methods for prostate cancer survival analysis, whereas Churilov et al. (2004) subsequently employed clustering techniques to assign patients into homogeneous risk groups that would facilitate treatment decisions. The clusters were generated by examining patient's age, tumor size, prostate-specific antigen concentration in the blood, and pathology scores. Recently, Kate and Nadig (2017) created stage-specific breast cancer survival prediction models by employing ANN, SVM, and DT, as well as statistical methods, such as LR and Naïve Bayes, to compare the survival predictions based on the stage-wise subsets and the entire dataset. The survival rate based on the full dataset was 92.04% (36.17−99.42%), whereby survival rates declined with progressive cancer stages. The authors further demonstrated that stage-wise stratification increased the AUC (area under the ROC curve) of the models in all different breast cancer stages, and ranked the variables based on stages, rather than time.

The aforementioned studies focused either on predicting cancer patient survival or finding important survival predictors. On the other hand, to the best of our knowledge, the varying impact of predictors on short- and long-term survival rates of cancer patients (irrespective of the cancer type) has never been explored. Furthermore, a comprehensive purely data-driven variable selection approach has never been employed in this field, even though this would allow the researchers to eliminate the variables that do not contribute to the prediction power, thereby producing highly parsimonious models. In addition, with the exception of Thongkam et al. (2009) who used simple over-sampling, advanced sampling approaches have not been used to prevent class imbalance problems that occur due to significant differences in survival classes in most survival datasets.

These shortcomings must be addressed, as being cognizant of the changes in the predictor importance in short- and long-time periods would be informative for better estimation of the impacts of different predictors on cancer development. Thus, combining the balancing methods with variable selection algorithms should ensure that the resulting variable importance findings are not affected by the inherent class imbalance, which also changes over time, as the models are applied to longer time-frames (Chawla, 2005).

Given the limitations revealed by the review of pertinent literature, the overarching objective of the present study is to determine the varying impacts of the cancer related variables obtained from the SEER dataset on predicting breast cancer survivability over 1-, 5-, and 10-year periods. In constructing the models, the increasing class imbalance over shorter time periods and removing noisy variables on a per-time-frame basis were considered to ensure near-optimal model performance and gain accurate variable importance information. To achieve this goal, a hybrid data mining methodology was employed. More specifically, a purely data-driven (not manual) variable selection process based on using GA and LASSO methods was adopted in order to obtain the most parsimonious models possible. In order to increase the sensitivity of the prediction models, RUS and SMOTE methods were employed in the data balancing phase. Finally, prediction results that were obtained through ANNs and LR were combined through the Information Fusion (IF) technique, as this allowed us to identify the variables whose impact changes (increases/decreases) over time. Therefore, the proposed hybrid data-driven methodology contributes to the existing body of knowledge on cancer survival literature by (1) utilizing data balancing algorithms (i.e., SMOTE and RUS) to overcome the class imbalance problem, (2) adopting a comprehensive variable selection process by using GA and LASSO methods in order to obtain parsimonious models, (3) analyzing the varying impacts of variables on survivability prediction accuracy over short-, mid- and long-term periods, and finally (4) employing a hybrid method (by using IF) that combines the predictors obtained through different predictive models (i.e., ANN and LR) to reduce model uncertainty and increase robustness of the results. It is the authors' view that the contribution and the value of the current study do not stem from the standard machine learning and statistical models and/or better classification results employed in this study. Rather, the findings presented in this work augment the existing body of knowledge by identifying the variables whose impact can change over time. This was achieved by employing the most parsimonious models that were built by overcoming the class imbalance problems in most survival datasets. Therefore, the same method can be adopted not only by the practitioners in the breast cancer field, but is also applicable to all cancer types.

The remainder of this study is organized as follows. The dataset, data cleaning methodology, sampling methods, variable selection process, and the predictive models utilized in the current study are presented in Section 2.2. Section 2.3 is designated for the results and insights obtained via the hybrid data analytic methodology adopted in the current study. Finally, the study closes with the main study conclusions and suggestions for potential future research directions, which are given in Section 2.4.

2.3 Methodology

In the present study, a hybrid data-analytical method consisting of five consecutive phases was adopted, as illustrated in Figure 2.1. The first phase was divided into three steps, whereby the dataset was cleaned in Step 1 in order to eliminate duplicate records and variables that do not contribute to the dependent variable prediction. In Step 2, different datasets were created from the original data to evaluate temporal effects of each variable over 1-, 5-, and 10-year time intervals. In Step 3, GA and LASSO were employed for variable selection. In the second phase, the data was input into the prediction algorithms, along with 10-fold cross-validation technique, to train the data mining models, whereby the balancing methods (i.e., RUS and SMOTE) were applied to each fold to prepare them for the model training process. At the completion of these stages, classification results yielded by the individual prediction models were obtained, along with the relative importance of each of the variables, as revealed by each model. After comparing the models using a set of performance metrics, those that exhibited inferior performance were eliminated in Phase 3. In Phase 4, an information fusion technique was used to combine important variable sets obtained through the individual models in the previous phase. This process was repeated three times (for each time interval), resulting in three sets of important variables (i.e., for the 1-, 5- and 10-year period). In the final phase, the aforementioned variable sets were compared to identify the variables whose importance/impact for determining the survival after breast cancer diagnosis changes over time. More details about the study phases and individual steps are provided in the following subsections.

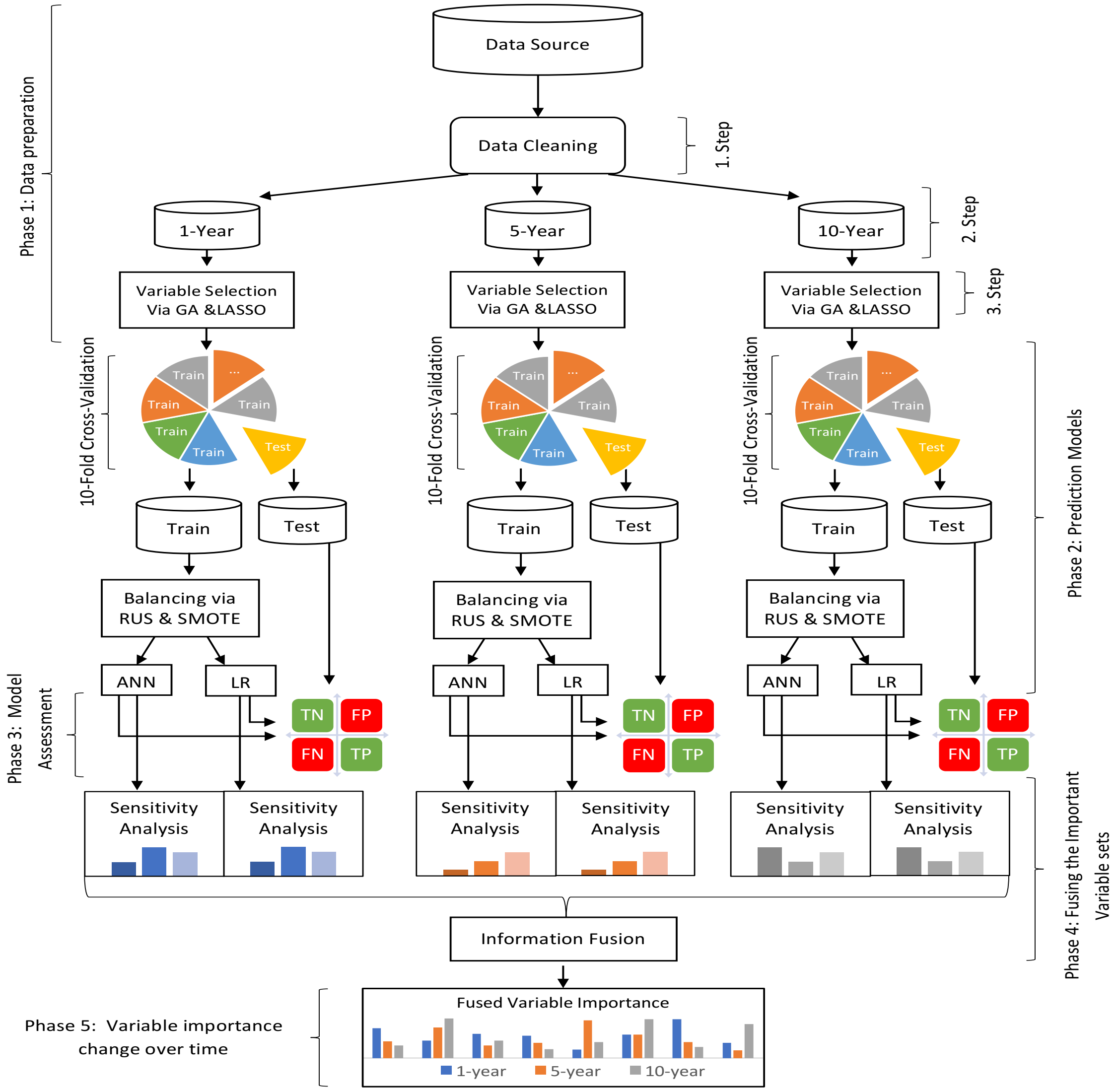

Phase 1: Data preparation
Data Source
Data Cleaning
1. Step
1-Year
5-Year
10-Year
2. Step
Variable Selection Via GA &LASSO
3. Step
10-Fold Cross-Validation
Train
...
Test
Phase 2: Prediction Models
Balancing via RUS & SMOTE
ANN
LR
TN
FP
FN
TP
Phase 3: Model Assessment
Sensitivity Analysis
Phase 4: Fusing the Important Variable sets
Information Fusion
Phase 5: Variable importance change over time
Fused Variable Importance
1-year
5-year
10-year

Figure 2.1: Proposed hybrid methodology.

2.3.1 Data Acquisition and Preparation

In the present study, the (SEER) dataset, which combines patient-specific predictors for any cancer type into a unique database, was utilized. The dataset contains records (173 variables) pertaining to 789,284 cancer patients covering the 1973−2013 period. The database provides comprehensive information for several cancer types, such as breast, lung, genital, and rectum cancer, including cancer-specific characteristics, socioeconomic conditions, and temporal information, such as the time of diagnosis and whether the patient died.

For the present investigation, the data cleaning process commenced by removing redundant information, and eliminating patient-specific information that was irrelevant for model training (e.g., patient ID, registry ID, etc.). Next, records of patients that died of causes other than breast cancer, such as an accident, natural disaster, etc., were discarded. Finally, all variables pertaining to other cancer types were eliminated. After the aforementioned data cleaning process, 53,732 breast cancer patient records and 17 variables that related solely to breast cancer remained. The descriptions of the variables are given in Table 2.1.

Table 2.1: Description of variables

Variable Name	Description
Cancer stage	EOD 3rd Edition and Collaborative Stage disease information
Nodes positive	The exact number of regional lymph nodes examined by the pathologist that were found to contain metastases
Grade	The measurement of how closely the tumor cells resemble the parent tissue, organ of origin
Age	Patient’s age at diagnosis (single-year ages)
Extension	The farthest documented extension of tumor away from the primary site
Primary site	The site from which the primary tumor originated
PR status	Includes indicators for estrogen receptors in cancer cells
ER status	Includes indicators for progesterone receptors in cancer cells
Tumor marker	Cancer or other cells of the body produce tumor markers as a reaction to cancer or certain benign conditions
Surgery	A surgical procedure aimed at removing and/or destroying the tissue of the primary cancer site performed as part of the initial work-up or the first course of therapy
Radiation	An indicator that shows whether radiation treatment has been applied or not
Race	Patient’s race

Tumor size	The largest dimension of the primary tumor in millimeters
Histology	Histologic cell type based on the microscopic composition of cells and/or tissue for a specific primary site
Marital status	Indicates whether a patient is married or not
Behavior	Collection of the malignancies, benign, uncertain whether benign or malignant, carcinoma *in situ*, malignant
Lymph node	The highest specific lymph node chain that is involved by the tumor

2.3.2 Temporal Dataset Derivation (Data Inclusion)

Since the current analysis is based on three temporal intervals, the original dataset was segregated into three new datasets, each pertaining to the temporally delineated dependent variable. Each dataset was derived using the survival month variable to determine a survival cutoff period. The survival month variable indicates the number of months a patient has survived after the initial diagnosis. For each of the three time intervals (1, 5, and 10 years), a new binary dependent variable was created as follows:

$$U_i = \begin{cases} 1(deceased), \text{ } if\, survival\, month \le i \times 12 \\ 0(survived), otherwise \end{cases} \quad (1)$$

where *i* = 1, 5 and 10 years. Therefore, the output variable is a binary categorical variable where 1 denotes *deceased* and 0 denotes *survived* patients. The number of patients in each class for each of the three time intervals is shown in Table 2.2.

Table 2.2: Distribution of survival cases

Cutoff Point	Deaths	Survivors
1 Year	866	52,886
5 Years	7,028	46,724
10 Years	10,787	42,965

2.3.3 Variable Selection

Machine learning algorithms are designed to learn the relationships between independent and dependent variables. As such, when training a model, all available variables in a given dataset can be used. In practice, however, this approach might bring certain disadvantages. First, some variables may not have any predictive power due to the absence of correlation between the outputs. Second, some variables may suffer from collinearity, which, depending on the machine learning algorithm, can have adverse effects on the predictive power of the model (Langley, 1994). Due to these reasons, including all variables does not guarantee the best possible predictive model, and in fact it may result in a model with lower than optimal predictive capability. In addition, the complexity of the machine learning model also increases with the number of predictors, both in terms of computational complexity and the explanatory power of the model. Consequently, in practice, variable selection methods are used to select only the features (predictors or variables) that are relevant to the model training process. This results in a model with increased predictive performance, better comprehensibility for users, and a reduced computational cost for training the model. The currently available variable selection methods can be broadly categorized into filter and wrapper methods (John et al., 1994).

Filter methods are univariate statistical methods that are mostly used in the data preparation phase. Many filter methods have been described in the literature (Saeys et al., 2007), and mostly consist of covariance tests, such as t-test, which can be used to determine if an independent variable is correlated to the output variable.

Wrapper methods are algorithmically or heuristically defined approaches to selecting features for a machine learning model. These methods are used in conjunction with the machine learning algorithm, and essentially "wrap" the model in the training process. In these methods, an algorithm or heuristic is employed to add or remove features from the dataset and retrain the model, with the goal of increasing the model's predictive performance, whereby the process is iterated until a stopping rule is reached. There are many wrapper methods, such as RFE (recursive feature elimination) (Guyon et al., 2002), simulated annealing (Aarts & Korst, 1988), genetic algorithms, etc. (Yang & Honavar, 1998).

In the present study, a wrapper (GA) and a filter method (LASSO) were adopted to conduct variable selection due to their high performance in the preliminary data analysis phases as well as their well-known high performance, as demonstrated in the pertinent literature (Dag et al., 2016). Variable selection was performed to obtain models that are better at predicting cancer survivability, and to gain a better understanding of the factors that are most important in predicting the survival interval. In previous studies (Delen et al., 2005; Kibis et al., 2017) in which the authors utilized the same dataset, the number of variables was manually reduced by applying expert knowledge. However, the goal of the present study was to use a pure data-driven methodology instead of human intuition and domain knowledge, each of which can be limited in its scope or is implicitly bound by assumptions. Additionally, having fewer variables reduces model complexity and renders models easier to understand, thus making them more useful for practitioners and decision makers. A brief description of each method is provided in the following paragraph.

Genetic algorithm mimics the gene selection process in biological organisms to select features that maximize (or minimize) predefined criteria. The algorithm is initialized with a predefined number of "individuals" (a solution) called a population. Each individual defines a solution to the problem under consideration that, in the case of variable selection, is a list of variables included for training. The model is trained for each individual, and the best performing individuals are selected to create the next generation of individuals by performing crossover operations between them. A certain amount of mutation can also be allowed, as defined in the input parameters for the GA. After a predefined number of generations has been created, an individual that yields the best prediction performance is identified. On the other hand, LASSO is a statistical regularization-based filter method that utilizes a linear regression fitting procedure in which estimation of coefficients is subject to a restriction. The purpose of this restriction in estimating coefficients is to overcome the issue of overfitting the data and to converge some coefficients towards exactly zero, which allows LASSO to be used as a variable selection tool. The LASSO (Tibshirani,1996) is given by Equation 2 below:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{2}$$

where y is the response, and the number of predictors and observations is represented by p and n, respectively. RSS is the residual sum of squares and λ is the penalization parameter added to the coefficients. The 10-fold cross-validation is used to find the optimal λ value. It should be noted that, for $\lambda = 0$, this procedure reduces to ordinary least squares regression.

Again, one of the study aims was to develop parsimonious models that have strong predictive power. Such parsimonious models can be used by practitioners and their patients to better understand the choices they have and the decisions that need to be made with regards to treatments, resource allocation, and patient care.

2.3.4 Data Imbalance and Resampling

The dataset used in the present study is a survival dataset, and is thus implicitly affected by a considerable imbalance across output classes. As seen in Table 2.2, the difference between the two classes of interest for the present study (surviving and deceased individuals) decreases monotonically for each increasing time interval. Several strategies for addressing the imbalanced data issue have been proposed (Chawla, 2005; Guo et al., 2008; Klement et al., 2011). In the present study, two well-known techniques— RUS and SMOTE—were adopted (due to their performance in the preliminary data analysis phase). They both have their benefits and drawbacks (Batista et al., 2004; Chawla, 2005; Drummond et al., 2003; Guo et al., 2008; Kotsiantis et al., 2006).

When RUS is adopted, cases from the majority class are randomly sampled to include a similar number of cases as in the minority class, thereby producing a balanced dataset (Chawla, 2005; Guo et al., 2008). The advantage of adopting RUS is that it includes actual data only; however, data that could have positively contributed to the model training process can be omitted. Conversely, SMOTE over samples the minority class by creating synthetic cases from the existing cases in the minority class. The synthetic cases are created by applying a *k-nearest neighbors algorithm* that imputes values for the synthetic cases from the actual nearby data points. More information on these methods can be found in the work by Chawla et al. (2002).

2.3.5 Prediction Models

In this study, ANN and LR models were employed, since these algorithms exhibited superior performance in preliminary analyses, as a part of which a collection of several classification algorithms was evaluated. We provide a brief summary of the algorithms, ANN and LR, in the following paragraphs.

ANNs are inspired by their biological counterparts, whereby a highly interconnected network of simplistic neurons generates a learning model that can learn arbitrarily complex non-linear functions. As such, these algorithms have been widely used in machine learning contexts, such as classification (Cetinic et al., 2018; Melin et al., 2018), regression (Genc & Dag, 2016; Zhou et al, 2019), optimization (Liu et al., 2009; Yazdi et al., 2011), and pattern recognition (El-Midany et al., 2010; Patterson, 1996). The artificial neurons mimic the signal integration and activation of their biological counterparts using mathematical functions, such as sigmoidal activation functions. These neurons are organized into input, middle, and output layers. There can be multiple middle layers, depending upon the complexity of the problem. Each neuron in a given layer is connected to all neurons in the next layer. The behavior of the neuron is governed by the training mechanism, which sets values for the weights of each neural connection, as well as values for the activation (or firing) of each neuron (Han & Kamber, 2006). In the present study, the Multi-layer Perceptron (MLP) algorithm, along with one hidden layer, was employed. The number of units in the hidden layer was tuned internally using a 10-fold cross-validation algorithm (thus, 360 models were developed). In addition, the *decay* parameter was tuned in order to prevent the possible overfitting issues in each model.

Logistic Regression is a probabilistic classification model that predicts probabilities of different class values based on the relationships between independent variables and the dependent variable. While the dependent variable values are continuous in linear regression, in many contexts, it is necessary to predict a discrete categorical class value, along with independent variables with discrete values. A logistic regression model transforms the discrete values into continuous form, and creates a linear combination of the variables to compute the probability of each class value of the dependent variable. As in linear regression, it predicts an outcome for the dependent variable, whereas logistic regression predicts a categorical class value, rather than a numerical value. For prediction, it computes the conditional probability of each categorical class value, and assigns the class value with the highest probability. The algorithm takes real values as input and outputs normalized values within the [0, 1] interval. The model can be utilized for binary and multi-class classification. The standard logistic function is a sigmoid function the value of which approaches 1 when the regression expression tends to ∞, whereas it approaches 0 when the regression expression tends to -∞. It takes the form given below:

$$F\left(x\right)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}} \tag{3}$$

where F denotes the probability of a class value when the dependent variable equals x, and p is the number of explanatory variables. In the present study, the 0.5 cutoff was chosen in order to assign the prediction results to the classes.

2.3.6 Performance Measures

In the present study, four metrics were adopted to measure model performance, namely AUC, accuracy, sensitivity, and specificity, with the latter three defined as follows:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{4}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$specificity = \frac{TN}{TN + FP} \tag{6}$$

where *TP*, *TN*, *FP*, and *FN* denote the true positives, true negatives, false positives, and false negatives, respectively. Because the datasets examined in the present study were imbalanced, AUC was considered as the main performance criterion, due to the fact that sensitivity and specificity have an inherent tradeoff among them, and can also be misleading for imbalanced datasets. However, all aforementioned criteria were measured and are presented here to provide a comprehensive picture of the model quality. The AUC metric is the area under the curve of all combinations of sensitivity and (1 – specificity), thus giving a more accurate measure regarding the models' performance, irrespective of the tradeoff or cutoff point chosen between sensitivity and specificity.

2.3.7 Sensitivity Analysis of Predictor Variables

One of the objectives of the present study was to develop a methodology that can be used to identify the most important cancer survivability predictors and establish how their importance changes over time. Findings yielded by such analyses can provide insight into how different factors affect the survival time for cancer patients. This information can be used by medical practitioners and decision makers to make better-informed data-driven treatment choices for their patients. SA was used for each of the models to understand what effect each predictor has on the model output and thus establish the relative importance of the examined variables for each examined period.

SA provides indication of the relative importance of each input for predicting the model output. Once the models have been trained, sensitivity analysis is conducted by assigning each variable in the model its nominal value. In the subsequent steps, one by one, each variable is changed gradually in both positive and negative direction, while the values of the remaining variables are held constant (equal to their respective means). As can be observed in Eq. 7, the sensitivity of a specific predictor variable is calculated by taking the proportion of the error of the model that includes this variable to the error of the model when it does not include this specific variable. Therefore, the importance of a variable is in direct proportion to variance of predictive error of the classification model in the absence of that specific variable. The effect these changes have on the output is observed and the same process is repeated with the next variable until the effect of all variables has been evaluated (Saltelli et al., 2002). At the end of the process, the effect of each variable on the output can be tabulated, and those that have the least significant effect on the output can be identified. In the present study, the relative importance of each classification model was established using the sensitivity measure introduced by Saltelli (2002), given by:

$$S_i = \frac{V_i}{V(y)} = \frac{V(E(y \mid x_i))}{V(y)} \tag{7}$$

where y is the dichotomous response variable, $V(y)$ is the unconditional response variable, E is the expectation operator that integrates all predictors except x_i. The normalized sensitivity of a variable is then calculated as described by Saltelli et al. (2002).

2.3.8 Information Fusion

IF is an umbrella term used for techniques that are applied to combine information from many different sources to increase some quality, such as knowledge discovery or information accuracy. Thus, IF techniques can be used for preprocessing raw data to increase its quality before being used for data mining, or for creating data mining models, to for example fuse partial results. Similarly IF can be used to extract and aggregate information from multiple sources (e.g., multiple models) to create an aggregated value using some function (Vicenc Torra, 2003). Such techniques have been shown to increase the robustness and accuracy of available information or knowledge discovery (Cang & Yu, 2014). At the same time, IF can decrease the uncertainty and bias of singular sources by combining many sources to obtain a better forecast estimate (Clemen, 1989). IF techniques have therefore become increasingly popular in the data mining field in recent years. While many such techniques have been described in the data mining literature, none can be considered the best at combining multiple information streams, whether applied before, during, or after the modeling process. In practice, an IF method should be chosen by considering the problem and the utility of several relevant models in a trial and error fashion (Ruiz & Nieto, 2000).

Any particular data mining model or information source can be formulated as:

$$\hat{y} = g\left(x_1, x_2, x_3, \ldots, x_m\right) \tag{8}$$

where g denotes the prediction function or information source, $\hat{y}$ is the predicted output, and the $x_1 \ldots x_m$ are the m input features. With such models or information sources, a fused output can be formulated as:

$$\hat{y}_{fused} = \Psi\left(g_1(x), g_2(x), \ldots, g_r(x)\right) \tag{9}$$

where $\boldsymbol{x}$ denotes vectors of the input features for each input model, and Ψ specifies a combination function. If this function is linear, the IF model can be represented as:

$$\hat{y}_{fused} = \sum_{i=1}^{r} \lambda_i g_i(x);\ \ where \sum_{i=1}^{r} \lambda_i = 1 \tag{10}$$

where λ_i represent weights, which can be based on the confidence level or accuracy measure of each information source or input model. For a prediction model, this can be the AUC measure, whereby a model with a higher AUC would have a greater impact on the final fused value obtained by the IF technique. As indicated from the above formulation, there are countless ways to combine the information from multiple sources, given the number of combination functions, model weights, source aggregations, etc. Thus, following the IF technique application, the final fused information will reduce the uncertainty, noise, or error contributed by any single source, thereby increasing the accuracy or usefulness of the final, fused data.

2.4 Results and Discussion

In this section, we discuss the variable selection results and the performance of the predictive models that were used to predict cancer patient survivability. The statistical programming language, R, was employed throughout the modeling process—from variable selection to model training and validation (R Core Team, 2014). In particular, GA and LASSO—the algorithms used for performing the variable selection—were executed using the R packages *gaselect* (Kepplinger, 2015) and *glmnet* (Friedman et al., 2010), respectively.

2.4.1 Variable Selection Results

In the present study, models that separate the output variable on three time-based intervals were developed, due to which three sets of variables were selected by each of the variable selection algorithms. Table 2.3 shows the variables selected for each of these time intervals.

Table 2.3: Variables selected for each time interval

Variable Name	1 YEAR			5 YEAR			10 YEAR		
	LASSO	GA	LASSO U GA	LASSO	GA	LASSO U GA	LASSO	GA	LASSO U GA
Cancer stage	✓		✓	✓	✓	✓	✓	✓	✓
Nodes positive	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tumor size									
Age		✓	✓	✓	✓	✓	✓	✓	✓
Marital status									
Lymph node	✓	✓	✓	✓	✓	✓	✓		✓
Primary site									
PR status				✓	✓	✓	✓	✓	✓
ER status				✓	✓	✓	✓		✓
Behavior									
Surgery				✓		✓		✓	✓
Extension	✓	✓	✓	✓	✓	✓	✓	✓	✓
Histology									
Grade		✓	✓	✓	✓	✓	✓	✓	✓
Radiation							✓	✓	✓
Race		✓	✓						
Tumor marker		✓	✓		✓	✓		✓	✓

As can be seen from Table 2.3, application of GA and LASSO resulted in different variable sets, most of which overlap for the 5- and 10-year intervals. For instance, *nodes positive* and *extension* were selected for all three time periods by both GA and LASSO. In the 1-year interval, only three variables were selected by both GA and LASSO, namely *nodes positive*, *extension,* and *lymph node*. On the other hand, for the 5-year interval, GA and LASSO selected the same variables, with the exception of *surgery,* and *tumor marker*. It should be noted that *tumor size, marital status, primary site, radiation, behavior, histology* and *race* were not selected as important variables for 5-year survival prediction in this study whereas these attributes have been selected for analysis in Delen et al. (2005). There are two possible reasons for the difference although the same dataset has been utilized. First, we used variable selection algorithms to select the important variables whereas expert opinion has been used in Delen et al. (2005). Secondly, while we used observations from 1973 to 2010, the comparative study has used dataset from 1973 to 2000. The extended time frame with new attributes such as ER- and PR status along with variable selection algorithms resulted in different set of attributes in this study compared to Delen et al. (2005). In the 10-year time period, *Lymph node*, *ER status, Surgery,* and *Tumor marker* were the only variables that were not selected by both methods. In the final stage, based on the union set, 8, 10, and 11 variables were selected for 1-, 5- and 10-year time points, respectively. It should be noted that the variables that were selected by GA and LASSO do not represent the final set of important variables. Rather, they are the candidates selected for inclusion in the final prediction models. The final set of important variables, along with their importance levels, is provided in Section 2.4.2, after applying SA and IF.

2.4.2 Classification Results

Two different classification models (ANN and LR) were employed in the present study for each of the three time intervals due to their high performance in the preliminary data analysis stage. Each model was trained using SMOTE and RUS sampling methods, along with 10-fold cross-validation. Therefore, 360 models were produced, as there were three variable combinations (GA, LASSO, GA **U** LASSO) × two learning algorithms (ANN and LR) × two sampling algorithms (RUS and SMOTE) × 3 time intervals (1-, 5- and 10-year periods) × 10 folds. The aggregated form of the classification results yielded by these models is shown in Table 2.4, where the second column represents the variable selection and balancing algorithms that are used for the associated year. To exemplify, *LASSO (RUS)* represents the model that is sampled through *RUS* and uses the variables selected through LASSO only. On the other hand, *LASSO* ∪ *GA (SMOTE)* represents the model that is sampled through *SMOTE* and uses the variables selected through the union set of LASSO and GA.

Some interesting patterns can be identified in the prediction results, as shown in Table 2.4. As previously noted, AUC was considered as the primary criterion, since it considers both sensitivity and specificity. It is thus more informative, especially when there is a significant imbalance between the minority and majority class. In such cases, the accuracy rate can be extremely high, while model sensitivity can be extremely low.

It is evident from Table 2.4 that the AUC level decreases as the prediction period increases. For example, the AUC rate for both ANN and LR is the lowest in the 10-year prediction. Furthermore, the power of AUC—as well as accuracy, sensitivity, and specificity—decreases from 1- to 10-year survival prediction. This might be due to a higher bias in predicting survival for longer time periods (Steyerberg, 2008).

Table 2.4: Classification results using SMOTE and RUS with different models

	Logistic Regression	Number of Variables	Accuracy	Sensitivity	Specificity	AUC
1-year	LASSO (RUS)	4	0.840 (0.003)	0.692 (0.034)	0.842 (0.003)	0.842 (0.024)
	LASSO (SMOTE)		0.838 (0.003)	0.693 (0.038)	0.841 (0.004)	0.845 (0.026)
	GA (RUS)	7	0.741 (0.006)	0.748 (0.045)	0.741 (0.007)	0.818 (0.020)
	GA (SMOTE)		0.819 (0.011)	0.756 (0.060)	0.820 (0.012)	0.861 (0.024)
	LASSO ∪ GA (RUS)	8	0.737 (0.005)	0.788 (0.46)	0.736 (0.005)	0.829 (0.018)
	LASSO ∪ GA (SMOTE)		0.807 (0.007)	0.789 (0.050)	0.808 (0.008)	0.870 (0.022)
5-year	LASSO (RUS)	9	0.726 (0.006)	0.751 (0.010)	0.722 (0.007)	0.815 (0.006)
	LASSO (SMOTE)		0.749 (0.005)	0.731 (0.006)	0.752 (0.006)	0.820 (0.006)
	GA (RUS)	9	0.746 (0.006)	0.740 (0.014)	0.747 (0.007)	0.824 (0.006)
	GA (SMOTE)		0.757 (0.006)	0.737 (0.011)	0.760 (0.006)	0.828 (0.005)
	LASSO ∪ GA (RUS)	10	0.747 (0.006)	0.752 (0.013)	0.734 (0.007)	0.824 (0.005)
	LASSO ∪ GA (SMOTE)		0.755 (0.006)	0.738 (0.012)	0.757 (0.007)	0.829 (0.005)
10-year	LASSO (RUS)	9	0.724 (0.007)	0.692 (0.011)	0.732 (0.009)	0.785 (0.007)
	LASSO (SMOTE)		0.736 (0.006)	0.694 (0.015)	0.747 (0.007)	0.791 (0.007)
	GA (RUS)	9	0.728 (0.007)	0.702 (0.012)	0.735 (0.009)	0.789 (0.007)
	GA (SMOTE)		0.737 (0.007)	0.693 (0.014)	0.749 (0.009)	0.794 (0.007)
	LASSO ∪ GA (RUS)	11	0.726 (0.007)	0.702 (0.013)	0.732 (0.008)	0.790 (0.007)
	LASSO ∪ GA (SMOTE)		0.738 (0.006)	0.695 (0.015)	0.749 (0.007)	0.796 (0.007)

	Artificial Neural Networks	Number of Variables	Accuracy	Sensitivity	Specificity	AUC
1-year	LASSO (RUS)	4	0.828 (0.011)	0.694 (0.034)	0.830 (0.012)	0.840 (0.021)
	LASSO (SMOTE)		0.822 (0.013)	0.714 (0.040)	0.824 (0.013)	0.845 (0.021)
	GA(RUS)	7	0.736 (0.006)	0.777 (0.005)	0.735 (0.006)	0.826 (0.025)
	GA (SMOTE)		0.819 (0.010)	0.777 (0.051)	0.820 (0.010)	0.869 (0.019)
	LASSO ∪ GA(RUS)	8	0.722 (0.013)	0.801 (0.049)	0.721 (0.012)	0.829 (0.026)
	LASSO ∪ GA(SMOTE)		0.822 (0.009)	0.772 (0.046)	0.823 (0.009)	0.871 (0.020)
5 -year	LASSO (RUS)	9	0.744 (0.006)	0.745 (0.010)	0.744 (0.007)	0.823 (0.006)
	LASSO (SMOTE)		0.761 (0.008)	0.743 (0.015)	0.764 (0.010)	0.835 (0.006)
	GA (RUS)	9	0.744 (0.007)	0.748 (0.012)	0.743 (0.008)	0.824 (0.006)
	GA (SMOTE)		0.755 (0.005)	0.758 (0.014)	0.754 (0.007)	0.835 (0.005)
	LASSO ∪ GA(RUS)	10	0.737 (0.007)	0.758 (0.017)	0.734 (0.009)	0.824 (0.006)
	LASSO ∪ GA(SMOTE)		0.756 (0.003)	0.755 (0.013)	0.759 (0.005)	0.836 (0.006)
10-year	LASSO (RUS)	9	0.726 (0.007)	0.706 (0.016)	0.731 (0.010)	0.792 (0.008)
	LASSO (SMOTE)		0.745 (0.006)	0.695 (0.015)	0.758 (0.008)	0.803 (0.007)
	GA (RUS)	9	0.721 (0.007)	0.710 (0.013)	0.723 (0.010)	0.791 (0.007)
	GA (SMOTE)		0.741 (0.007)	0.706 (0.011)	0.751 (0.008)	0.802 (0.007)
	LASSO ∪ GA (RUS)	11	0.724 (0.006)	0.705 (0.012)	0.729 (0.008)	0.790 (0.008)
	LASSO ∪ GA (SMOTE)		0.742 (0.011)	0.697 (0.016)	0.754 (0.013)	0.801 (0.007)

Another interesting observation is that the ANN models always outperform the LR models in all time periods (.871 and .870 in 1-year, .836 and .829 in 5-year and .803 and .796 in 10-year time periods). Irrespective of the statistical significance of the observed differences, it is not possible to identify models that would work the best on a certain type of data, due to which it is necessary to resort to the *trial & error* option. It is likely that the inability to differentiate models based on the aforementioned measures stems from the potential non-linear structure of the dataset.

When the balancing (sampling) procedures are considered, the results reported in Table 2.4 indicate that SMOTE always outperformed the RUS procedure for both ANN and LR models, as well as for both GA and LASSO selection models. This outcome is expected when a dataset includes a large number of variables. As the number of variables increases, the number of cases (i.e., the number of observations) should also increase exponentially in order to train the model well (Han & Kamber, 2006). Hence, instead of decreasing the number of cases (as in RUS), increasing the number of cases (as in SMOTE) might increase the model predictive power, since there is no information loss. It should be noted that similar results might not be obtained from a dataset with the same number of cases but with fewer variables. These arguments are in line with the findings reported by Kumar et al. (2012), who demonstrated that sampling methods can be used interchangeably for a given model without diminishing its predictive power.

Finally, the results reported in Table 2.4 demonstrate that neither variable selection method consistently outperforms the other alternative. However, for both ANN and LR, and for all three time periods, the models provide better results when the union set of LASSO and GA with SMOTE is considered with the exception of 10-year period with ANN prediction.

Each prediction/classification method provides different outputs for a given dataset, as it was also provided in the related literature (Delen et al., 2005). Therefore, in the present study, information fusion was employed to combine multiple predictions made by different models and thus ensure more reliable and robust results. After applying the sensitivity analysis procedure, which was explained in detail in Section 2.3.7, the variable importance for both models and all three time periods was obtained. Therefore, six variable importance reports have been generated (i.e., 1-year ANN, 1-year LR, 5-year ANN, 5-year LR, 10-year ANN, and 10-year LR), each focusing on the importance level of a variable for a given period and a given model. Subsequently, the two reports that belong to the same time period were combined via the IF model, thus yielding three importance reports for their respective time periods. In the final stage, the results were presented graphically, as shown in Figure 2.2, to visualize model outputs for three survival time periods after information fusion.

The results shown in Figure 2.2 demonstrate that the prediction power of *grade, extension*, *lymph nodes* and *nodes positive* does not decrease from 1- to 5-year, but increases in the 10-year survival prediction. Similarly, although there is 1% reduction in the 5-year survival prediction power of *historic stage*, it also increases while predicting 10-year survival rate. While such reduction could have occurred due to overfitting, the increase in the predictor importance from 1- to 10-year survival rate can be attributed to the information about the potential spread rates of the tumor cells, which is provided by these five predictors (Bunting et al., 1976; Carter et al., 1989; Kennecke et al., 2010). Therefore, while impacts of the cancer characteristics might diminish in the short term due to treatment, these tumor characteristics are highly informative in determining the long-term patient survival and cancer recurrence, and thus gain importance as the illness progresses (Delen, 2009).

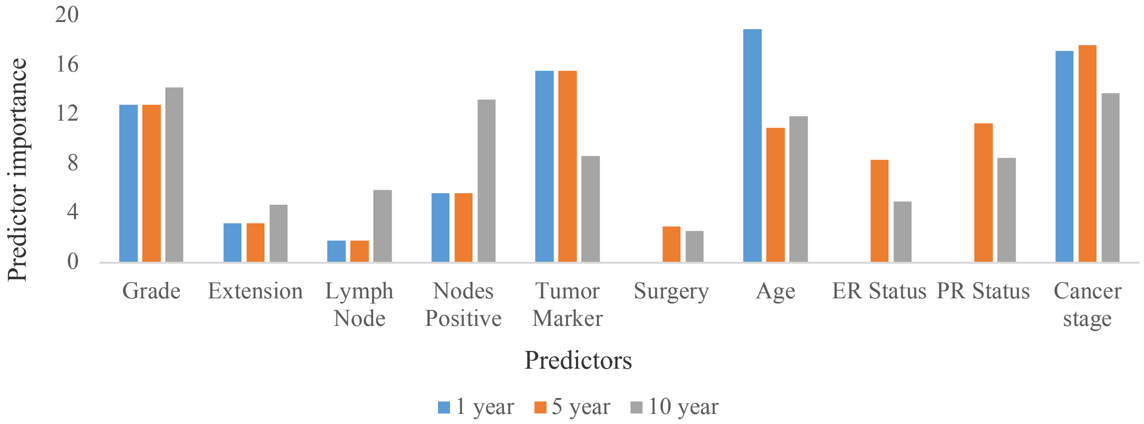


Figure 2.2: Variable importance for 1-, 5- and 10-year survival prediction via IF model

Furthermore, *tumor marker* provides information about cancer behavior and the treatment types that can be effective for a specific patient (McShane et al., 2005). Although one type of treatment can benefit the patient and reduce the recurrence risk in the first five years, it may lose its effect in the long term (Clarke et al., 2005). Therefore, while *tumor marker* is of similar relevance for predicting 1- and 5-year survival, its power in predicting 10-year survival is much lower. *Cancer stage* revealed a similar trend in predicting short- and long-term survival. Owing to the reliance on state-of-the-art technologies, 5-year survival rate is 100% for *Stage 0* and *Stage I* breast cancer. However, it decreases to 93%, 72% and 22% for *Stage II, III* and *IV* breast cancer patients, respectively (Howlader et al., 2013). Considering the fact that the median survival time of patients with metastatic breast cancer is four years (Bafford et al., 2009), and patients diagnosed with *Stage II* and *III* cancer often die for reasons other than breast cancer such as flu, cold, etc. (Howlader et al., 2013), the prediction model loses its power to predict 10-year survival.

In addition, *ER* and *PR status* indicate whether cancer cells have estrogen or progesterone receptors that promote cancer cell growth in the presence of estrogen or progesterone hormones, respectively. Results yielded by the current analyses demonstrate that *ER* and *PR status* do not have an impact on 1-year survivability, as both are also related with prognostic factors, such as tumor size, lymph node involvement, cancer stage, and histology type (Moise et al., 2013). On the other hand, both *ER* and *PR status* follow a similar trend in predicting 5- and 10-year survival. While a patient diagnosed *ER* positive may have higher chance of 5-year survival (Berry et al., 2006), both *ER*- and *PR-status* have diminishing effect for predicting 10-year survival. The results support the findings reported by Kasami et al. (2008) and Hirata et al. (2009), who demonstrated that adjuvant systematic treatment can change *ER* and *PR status*, which would in turn reduce their long-term predictive power.

Moreover, while *age* is highly relevant in predicting cancer patient survival, its power diminishes with time. There are conflicting opinions on the importance of age for breast cancer survival. While Adami et al. (1986) and Nixon et al. (1994) demonstrated that younger breast cancer patients are more vulnerable to distant metastases and are more likely to experience cancer recurrence, Yancik et al. (1989) found that older patients do as well as young ones in the short-term. Moreover, Lundin et al. (1999) showed that *age* does not provide any statistical power to the prediction model for predicting 5- and 10- year survival whereas it becomes statistically significant in analyzing 15-year survival. Therefore, available evidence supports the importance of *age* in predicting 1-year survival. On the other hand, its predictive power reduces from 19% to 12% from 1- to 10-year survival period, likely due to the higher significance of other predictors, such as *grade* and *lymph node involvement*. Finally, while undergoing *surgery* provides no predictive importance for 1-year survival, it has almost equal power in predicting 5- and 10-year survival. Surgery, however, has a relatively low importance compared to other predictors because *mastectomy* is performed on almost all patients to prevent potential cancer growth and spread. Therefore, since most of the patients included in the analyzed dataset opted to undergo *mastectomy*, the prediction model assigned low importance rate to *surgery*.

2.5 Conclusions and Future Research Plan

In the present work, a novel approach for predicting the survival time of breast cancer patients after diagnosis was proposed and was applied to three time periods. The analyses reported in the preceding sections indicate that certain predictive attributes contained within existing medical databases can be used to develop models that can accurately predict cancer outcomes and patient survival rates. To achieve this aim, a hybrid methodology was formulated to elucidate whether and how the importance of various factors changes over time, rather than focusing on cancer stages. The methodology was developed using the SEER dataset pertaining to the 1973−2013 period obtained from NIH. Since the dataset was imbalanced, RUS and SMOTE methods were applied. In addition, before performing classification, two well-known methods GA and LASSO were utilized in order to identify the best set of predictor candidates for inclusion into the prediction models. In the classification process, the importance of each variable was determined using two classification algorithms namely ANN and LR, before applying an IF technique. The aim of this work was to ascertain (1) which factors contribute the most to the prediction of survival of breast cancer patients in various time intervals, and (2) whether and how the importance of those factors changes over time. The study findings revealed that some factors typically considered relevant for cancer treatment do not equally contribute to patient survival. As a result of identifying the key factors for patients' short- and long-term survival, the most parsimonious models were developed for each time period, by utilizing purely data-driven algorithms. Moreover, the importance of these variables and their changes over time was established, revealing the underlying dynamics within and among factors. Thus, the contribution of the current study to the extant body of knowledge stems from mapping out the variables whose impact changes over time. As this was achieved by employing the most parsimonious models, these strategies can have much broader implications for cancer treatment and research. In particular, advanced resampling techniques were built by preventing the class imbalance problems that occur due to significant differences in survival classes in most survival datasets. Therefore, the same method can be adopted not only for the practitioners in breast cancer field but also in all cancer types. Moreover, the information provided here will be beneficial to medical practitioners, cancer specialists, in conducting their treatments in a more efficient, reliable, and productive manner.

The methodology presented in this study can be used to augment decision-making processes based on data-driven predictions in order to determine better prognosis and treatment courses. However, not all patterns identified by machine learning models may be useful, or interesting to practitioners. Any predictions made by such models, in order to be useful, must identify patterns that can be actionable and logical. Therefore, application of these models in practice requires a careful consideration of these model characteristics by medical practitioners.

We plan to continue on the same research path and build a graphical user interface (GUI) that can be used by medical practitioners without any machine learning experience or statistical background. We plan to design the GUI to allow the practitioners to enter all relevant patient data and obtain the survival probability for that patient for three different time periods. As this tool can quickly identify "high risk" patients, medical practitioners can use this information to formulate the best course of treatment. We hope that such tool would remove some personal bias from already challenging treatment decisions. Finally, with minor adjustments, the tool can be applied to any cancer type.

Chapter 3

A Data-Driven Statistical Modeling Approach for Identifying No-show Patients

3.1 Abstract

Accurate prediction of no-show patients plays a crucial role in that it enables researchers to increase the efficiency of their scheduling systems. The purpose of the current study is to formulate a novel hybrid data mining based methodology to a) accurately predict patients no-shows, b) build the most parsimonious model possible by employing a comprehensive variable selection procedure and, c) provide healthcare agencies with a patient-specific risk level. Our study suggests that an Artificial Neural Network (ANN) model should be employed as a classification algorithm in predicting patient no-shows by using the variable set that is commonly selected by a Genetic Algorithm (GA) and Simulated Annealing (SA). In addition, the Random Under-sampling (RUS) balancing method should be employed to improve the performance of the model in predicting the minority group (no-show) patients. The patient-specific risk scores obtained were justified by applying a threshold sensitivity analysis. It has been shown that the medical experts can purely and confidently rely on the probabilistic scores provided, while their intuition/incentive should collaborate with prediction models to make the final decision on the cases where the model is not confident enough. Those insights enable health care professionals to improve clinic utilization, and improve patient outcomes. The output of the model can be applied to the appointment scheduling system for a robust overbooking strategy.

3.2 Introduction

When patients miss appointments, they impose a substantial burden on healthcare providers. Such patients, called “no-shows,” are an expensive problem because they reduce revenue and increase costs for healthcare providers (DuMontier et al., 2013; Moore et al., 2001). No-show patients harm themselves, as they have poor healthcare outcomes (Fortin et al., 2016). Due to these problems, the decision science literature has explored several strategies to mitigate both skipped appointments as well as their effects such as overbooking to reduce the number of open appointments (Reid et al., 2016; Zeng et al., 2010), modifying patient behavior through incentives or deterrents (Vikander et al., 1986), requiring patient prepayments (Garuda et al., 1998), and imposing fines on patients who have skipped appointments (Bech, 2005) etc.

Predicting patient no-shows patients plays a crucial role in that it enables researchers to increase the efficiency of the prescription (optimization, simulation, scheduling etc.) for the problem. In other words "*a good predictive model leads to a good prescriptive model*". The enormous amount of complex data that are being made available to researchers offer a huge potential increase in the amount of useful knowledge to be gained, with the help of today's computing power that can host efficient but computationally expensive data analytical models (Olson & Delen, 2008).

In the current study, we develop a data-driven method that classifies the "*No-show*" patients into five categories (i.e. (1) *very low risk* (of *no-show*), (2) *low risk*, *(3) moderate risk, (4) high risk,* and *(5) very high risk*) by employing efficient predictive methods in addition to heuristic-based optimization models and data balancing algorithms. Our study fills a significant gap in the existing body of knowledge in that it 1) develops a very parsimonious model by using wrapper-based *variable selection* models since "*the simpler is the better*", 2) employs cutting edge data balancing algorithms to increase the ability of the predictive model in detecting the minority class, as the no-show datasets are mostly imbalanced (i.e. the number of the no-shows are much less than the number of show-ups), 3) determines the patients who have *very low-, low-, moderate-, high- and very high-risk* of being a no-show patient by using the probabilities that can be achieved by some of these probabilistic machine learning models and 4) analyzing publicly available patient data that were collected over a significant period of time. To the best of our knowledge, the outcome of the current study is significant to the no-show problem, as none of these points were addressed in the related literature.

It should be clearly indicated that our study targets only the predictive side of the problem as it does not attempt to come up with a prescriptive solution (i.e. scheduling). Having said that, the scheduling systems can be designed based on these insights to optimize the utilization of healthcare resources and come up with more efficient no-show management systems as our study provides the *patient-specific (No-show) risk levels* that are obtained through a highly parsimonious model. In addition, we examine the effects of the primary factors that lead to the no-show issues. Our approach is to identify and pinpoint the critical factors involved in the no-show problem such as which patient types regularly miss appointments, whether there are discernable patterns at different time scales. That is, whether some times of the day, month or year have a higher probability of no-shows, etc. Such insights can help healthcare providers to implement mitigation strategies based on root causes by deploying prediction models and intervention systems and offer healthcare providers an accurate decision support system that can help them understand their patients' behaviors and manage their resource deployment accordingly.

3.3 Methodology

To provide practitioners with better information for developing a scheduling system, we propose a technique that can be used by clinics to minimize their no-show patient cost. Figure 3.1 shows the proposed four-step methodology. In the first step, we begin with a healthcare provider dataset from the state of Espírito Santo, Brazil. We downloaded the patient data from the *Kaggle* website and preprocessed it to prepare for the machine learning algorithms. The second step was to select the relevant predictors. To choose the predictors, we used two optimization techniques: GA and SA. In the third step, we balanced the dataset because the number of no-show patients in the dataset is significantly less than patients who kept their appointments. We used RUS, ROS, and SMOTE. Balancing applied solely to the training set. To split the dataset into training and test sets, we implemented a *10*-fold cross-validation technique. We trained three classification models LR, RF, and ANN along with the variable sets obtained through the optimization algorithms (SA and GA) as well as combinations of these variable sets. In addition, an ensemble learning model is developed by combining the aforementioned three models. In the fourth step, we analyzed the individual performance of the classification/prediction models. After finding the best performing variable set with the best classification algorithm and the best balancing technique, we explored the importance of these predictors using sensitivity analysis. An Information Fusion method is then used to combine essential variables. Lastly, patients are classified into five risk level with regard to patient-specific no-show risk scores. The details of each of the steps of the methodology are given the following subsections.

3.3.1 Data Acquisition and Preparation

We acquired the dataset from the data science competition platform Kaggle (*https://www.kaggle.com/joniarroba/noshowappointments*) ("Medical Appointment No-shows," 2016). The dataset has 110,528 observations with 14 variables. The dataset has two categories: (1) appointment characteristics (*appointment date, appointment scheduling time, appointment ID, patient ID*); and (2) demographic information (*age, gender, financial status, alcoholism,*

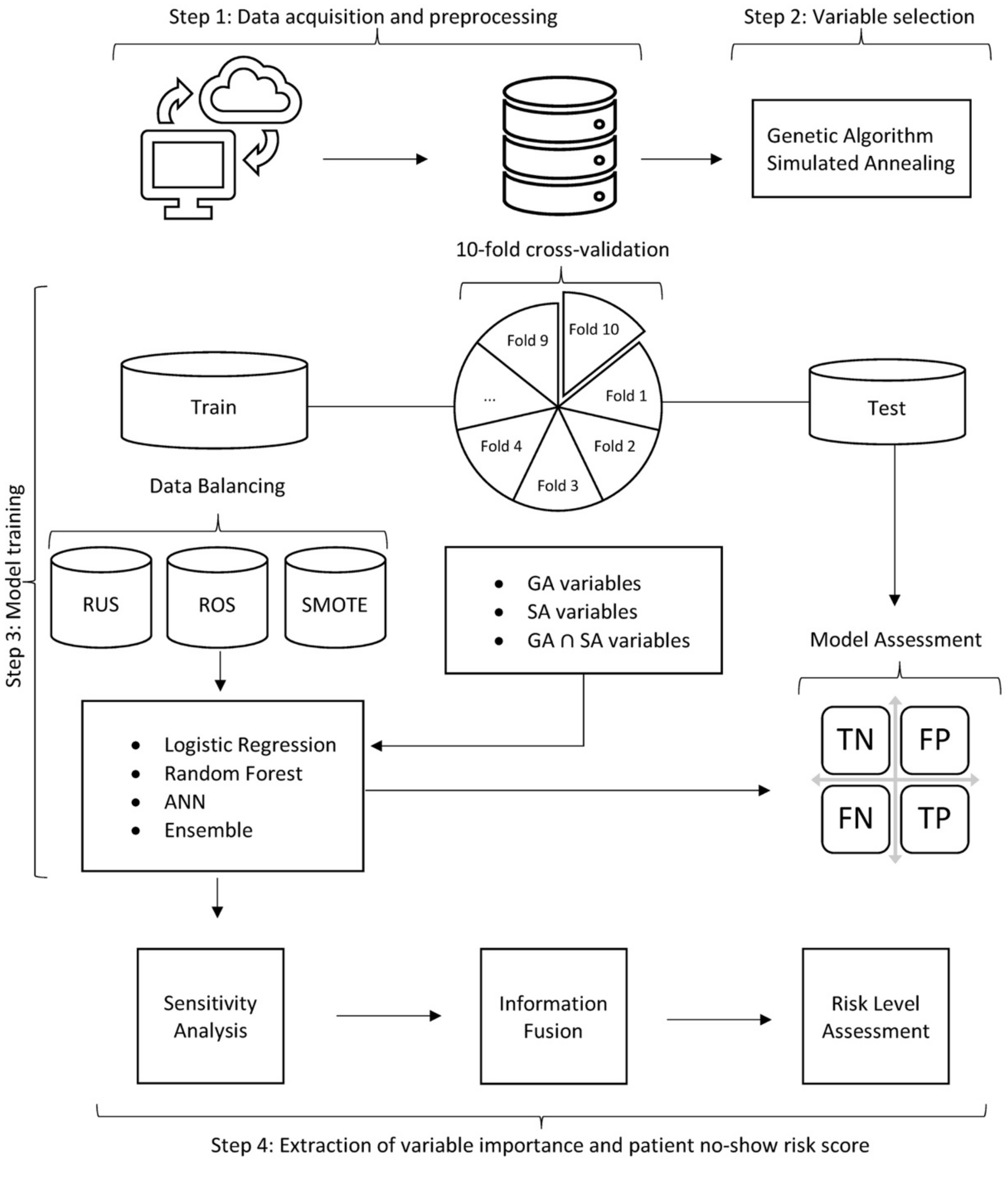


Figure 3.1: An outline of the proposed data analytics methodology

handicap, and patient health). In addition to these variables, we derived new variables from the originals. We created a *lead time* variable using the number of days between when the appointment day was scheduled and the appointment day itself.

Similarly, we constructed *calling time*, *appointment day, appointment time, and appointment month* variables from *appointment date*. We used the patient appointment records to create *prior no-show* (proportion of prior missed appointments) and *time between appointments* (the time elapsed between two consecutive appointments) variables. Lastly, we pruned outliers and irrelevant variables like Patient ID and Appointment ID from the dataset. After completing the variable extraction and elimination process, we obtained a dataset with 72,602 observations and 17 variables. The definitions of the predictors are given in Table 3.1.

Table 3.1: Description of variables

Variable Name	Definition of Variable
Age	Chronological patient age
Gender	Patient sex
Season	The month of the appointment
Appointment Day	Day (1-31) of the appointment
Scheduling Day	The day an appointment was scheduled to take place
Lead Time	Waiting time for the up-coming appointment
Calling Time	The time patients called to schedule their appointments
Appointment Reminder	Whether a reminder message or call is received
Alcoholism	Whether the patient is an alcoholic
Financial Aid	Whether the patient has financial support
Handicap	Whether the patient has a permanent physical impediment
Hypertension	Whether the patient has high blood pressure
Diabetes	Whether the patient has diabetes
Neighborhood	Hospital location
Time between Appointments	The elapsed time between two consecutive appointments
Prior No-shows	The proportion of no-shows

3.3.2 Variable Selection

Variable selection is the extraction of the most relevant subset of variables from the original set. Performing variable selection enables researchers to find the simplest model that predicts the target variable well. However, finding an optimal subset of variables is a challenging task. An exhaustive search, evaluating all combinations of candidate subsets requires computational effort that grows exponentially (2^n where *n* is the variable count); therefore, it may not execute in real time, especially when data size is big. On the other hand, computationally efficient algorithms can possess different shortcomings. The traditional statistical variable selection techniques (as employed in the related no-show literature) such as ridge regression, backward selection, forward selection, and stepwise selection, perform poorly when the collinearity among the predictors exists

(James et al., 2013). Balancing these competing interests requires careful consideration of the tradeoffs, and implementing a solution accordingly. In this study, we sought an accurate, parsimonious model, without the computationally intense demands of the exhaustive search. Therefore, we employed two metaheuristic algorithms, SA and GA; we explain the optimization techniques and advantages of these methods in the sub-sections below.

3.3.2.1 Simulated Annealing

Simulated annealing, inspired by the annealing process, seeks a global optimum for a given function (Kirkpatrick et al., 1983). The SA algorithm mimics the annealing process using a temperature variable that initially has a large (hot) value and then cools down (decreases) at each iteration step. SA always accepts better solutions, yet it also randomly accepts worse solutions as long as the system temperature is high. The system temperature cools down gradually, thereby restricting the search space.

```
GENERATE: An initial random best-performing variable subset, X.
INITIALIZE:
            The number of iterations (move attempt): Γ;
            The current system temperature: $\varphi$;
For $\alpha = 1\ to\ \Gamma$ {
            Initiate a random move, i.e., perturb the current variable set $\boldsymbol{X}$,
            Calculate the performance of the model $\Phi(\boldsymbol{X}_{\alpha})$,
            Compute the change in performance, $\Delta\Phi = \Phi(\boldsymbol{X}_{best}) - \Phi(\boldsymbol{X}_{\alpha})$
      if $\Delta\Phi < 0$ then
            Accept the downhill move and update the current predictor set
            Set $\Phi(\boldsymbol{X}_{best}) = \Phi(\boldsymbol{X}_{\alpha})$
      else
            Calculate the acceptance probability of the uphill move: Θ*
            Set $\Theta_{\alpha}^{*} = \exp\left(\frac{\Phi(X_{\alpha}) - \Phi(X_{best})}{\varphi(\alpha)}\right)$
            Sample a number, $\zeta$, from a uniformly distributed population between [0,1]
            If $\Theta_{\alpha}^{*} \leq \zeta$ then
                  Set $\Phi(\boldsymbol{X}_{best}) = \Phi(\boldsymbol{X}_{\alpha})$
            else
                  Set $\Phi(\boldsymbol{X}_{best}) = \Phi(\boldsymbol{X}_{best})$
            end
      end
end
Report the model with the predictor set having the highest $\Phi(\boldsymbol{X}_{\alpha})$, for $\forall\ \alpha \in [1, \Gamma]$
```

Figure 3.2: The proposed simulated annealing variable selection algorithm

The central premise of simulated annealing algorithms is to “jump” out of local minima that the algorithm otherwise cannot escape, thereby gaining the ability to pursue the search in other regions. In the optimal case, the algorithm finds the global maximum. In this study, we propose a simulated annealing algorithm utilizing Random Forest model as it is invulnerable to both the issue of the multicollinearity and the overfitting, contrary to many other machine learning models (Breiman, 2001). Figure 3.2 show pseudocode of the proposed algorithm where the algorithm is run 10 times with a different initial random variable subset in the attempt to find the best possible solution. For detailed information about SA, see Rutenbar 1989 (Rutenbar, 1989).

3.3.2.2 Genetic Algorithms

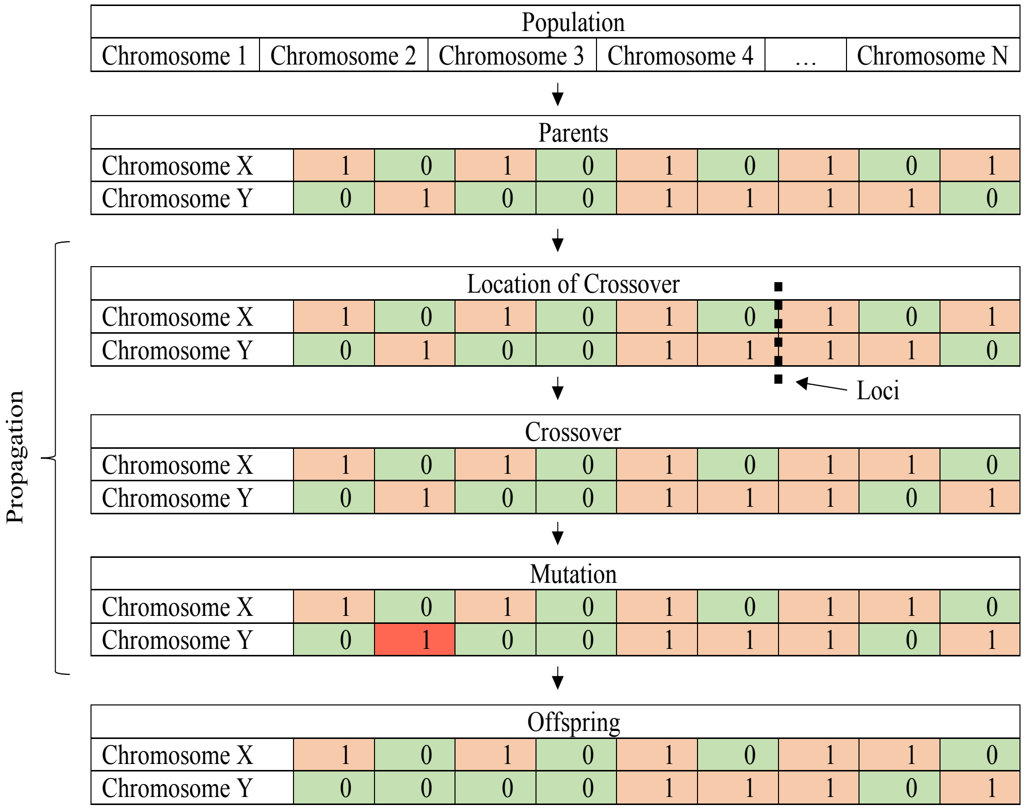
Genetic Algorithms, inspired by biological evolution, seek an optimal solution both for continuous (differentiable or not) and discrete functions (Goldberg & Holland, 1988). The effectiveness of GAs as variable selection tools has been proven in many fields (Bhanu & Lin, 2003; Drake & Marks, 2002; Lavine et al., 2002).

```
GENERATE: An initial population: $\Theta = \{\theta_1, \theta_2, \theta_3, \ldots, \theta_N\}$
INITIALIZE:
        Rate of mutation: $\gamma$
        Rate of elitism: $\eta$
        Number of iteration: $K$
for $k$ =1 to $K$
        Number of elitism $\tau = \eta * N$
        Compute $f(\theta_i^{(k)})$ where function, $f$, is produced using RF model
        Select the best $\tau$ chromosomes in $\Theta$ and save them in $\Theta_1$
        Set $\Theta_2 = \{\theta \in \Theta : \theta \notin \Theta_1\}$ , and set number of crossover $\delta = \frac{|\Theta_2|}{2}$
        for $j$=1 to $\delta$
                Select two chromosomes $\theta_x$ and $\theta_y$ from $\Theta$ based on fitness criterion
                Randomly select a split position (loci) to exchange the genes of the chromosomes
                beyond the loci and store $\theta_a$ and $\theta_b$ in $\Theta_2$
        end
        for $j$=1 to $\delta$
                mutate binary values of each gene of the chromosome $\theta_j$ in $\Theta_2$ with the rate $\gamma$
        end
        Update $\Theta = \Theta_1 + \Theta_2$
end
Return $\theta^* = argmax_{\theta_i^{(k)}} f(\theta_i^{(k)})$.
```

Figure 3.3: The proposed genetic variable selection algorithm

Therefore, in this study, we adopt GA as one of the variable selection tools. We propose a genetic algorithm, shown in Figure 3.3, consisting of the following initial parameters: number of maximum generations, 250, number of population per generation, 50, crossover probability, 0.8, and mutation probability: 0.1.

The intuition behind the choice of the crossover and mutation probability can be explained as follows. We keep the probability of crossover between pairs of chromosomes high, at 0.8, so as to promote the algorithm to move the next generations towards the space defined by the fittest chromosomes. Moreover, even though it is a common practice to keep the mutation probability small, say < 0.05, to reduce convergence time of the algorithm to an optimal solution, we set this probability to 0.1 in an effort to provide the proposed algorithm with the ability to escape local optima, with the tradeoff of having longer execution time. The proposed GA is also run 10 times as in SA to obtain the best possible solutions. Figure 3.4 illustrates the next generation of offspring, where two chromosomes propagate via the processes and the loci is the split position, beyond which the chromosomes exchange the genes.



Population

Chromosome 1	Chromosome 2	Chromosome 3	Chromosome 4	…	Chromosome N

Parents

Chromosome X	1	0	1	0	1	0	1	0	1
Chromosome Y	0	1	0	0	1	1	1	1	0

Location of Crossover

Chromosome X	1	0	1	0	1	0	1	0	1
Chromosome Y	0	1	0	0	1	1	1	1	0

Crossover

Chromosome X	1	0	1	0	1	0	1	1	0
Chromosome Y	0	1	0	0	1	1	1	0	1

Mutation

Chromosome X	1	0	1	0	1	0	1	1	0
Chromosome Y	0	1	0	0	1	1	1	0	1

Offspring

Chromosome X	1	0	1	0	1	0	1	1	0
Chromosome Y	0	0	0	0	1	1	1	0	1

Figure 3.4: The propagation step of a genetic algorithm

3.3.3 K- fold Cross-validation

Cross-validation is a model validation technique to estimate the error rate on unseen data (Kohavi & Kohavi, 1995). k-fold Cross-Validation (CV) splits the dataset into k disjoint folds (subsets); k – 1 of these disjoint folds are used to fit a predictive model, and the fitted predictive model's performance is evaluated on the fold left out. The procedure is repeated $k - 1$ times, so that each fold takes its turn as the left-out fold, exhausting the entire dataset. The mathematical definition of the k-fold cross-validation is

$$CV = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{1}$$

where k is the number of folds and A_i is the model accuracy on the ith fold. We used 10-fold CV to evaluate the performance of the predictive models. We chose $k = 10$ since it provides a good trade-off between the bias and variance, as well as the time required to run models (Dougherty et al., 1995).

3.3.4 Sampling Techniques

The issue of having an uneven class size in the target variable encourages machine learning algorithms to take a naïve approach to minimize the loss function by classifying all the instances as the majority class (show-ups in our case). In doing so, the machine learning algorithms maximize overall accuracy by generating high predictive accuracy for the majority class and low predictive accuracy for the minority class (see Table 3.5) (J.-S. Lee & Zhu, 2011). Table 3.2 illustrates the number of no-show and show-up patients and their percentages in the dataset. Classifying all patients as show-up patients would be 79.07% accurate. However, such a model is useless, because it would not predict any no-shows.

Table 3.2: Summary of the target variable

	No-show Patients	Show-up Patients
# of instances	15196	57406
Percentage	20.93%	79.07%

Two commonly used techniques that compensate the issue of imbalance are resampling the target variable, and cost-sensitive learning (Kotsiantis et al., 2006). However, in cost-sensitive learning, misclassification costs may not be available, and if costs are available, overfitting the training set can result unless the costs are specified precisely (Bansal et al., 2008). Researchers use resampling methods to decrease or eliminate the imbalance. Approaches include minority class oversampling, majority class under-sampling, or a blend of both methods. (Dag et al., 2017; Kibis et al., 2017). However, it should be noted that there is no absolute winner among the techniques as different modeling techniques react differently to sampling. Thus, in this study, all the resampling techniques have been utilized during the model training process.

The sampling techniques we used are RUS, ROS and SMOTE. RUS randomly eliminates some majority-class observations until the majority class count is approximately equal to the minority-class observation count. Similarly, ROS increases the number of minority-class observations by randomly including existing samples as new samples until the number of observations in the minority class equals the majority class. SMOTE, on the other hand, uses both random under-sampling and over-sampling depending class distribution in the data. SMOTE increases the count of minority class by interpolating pairs of close neighbors in the minority class randomly, whereby generating synthetic instances. At the same time, SMOTE is capable of decreasing the majority class through under-sampling. The operational parameters of SMOTE—the number nearest neighbors (KNNs), the rate of under-sampling and over-sampling—need to be specified by researchers before the execution of the algorithm. After our extensive preliminary analysis, we use 5-nearest neighbors to have new cases synthesized, and equal rates for up- and down-sampling.

3.3.5 Predictive Modeling

Predictive modeling is the process of creating a statistical or machine learning model to predict a future event. We used an ANN, RF, and LR to predict whether a given patient will be a no-show. We also created Ensemble Learners (EL) by using combinations of ANN, RF, and LR since empirical studies have shown that EL-based algorithms often produce more accurate results compared to single algorithms (Opitz & Maclin, 1999; West et al., 2005). However, counterexamples are reflected (Orimoloye, 2017); therefore, we included both ensemble and single learning algorithms. We describe our models in the following sub-sections.

3.3.5.1 Artificial Neural Networks

Artificial Neural Networks, inspired by the work theory of human brains, are complex analytical techniques that can model complicated, nonlinear functions (Patterson, 1996). In ANNs, the outcome is modeled by a set of latent variables that are a non-linear combination of the original variables, named as hidden units. These units, promoting the learning process by increasing the flexibility of the model, are located in the middle of the network and are connected to the outcomes through the output function.

Once the numbers of hidden units and predictors are defined, initial random values (weights) are assigned to the parameters of the network. The weight of the parameters, then, are updated (i.e., the new weights are estimated) through optimization algorithms—attempting to minimize the loss function.

In this study, we deploy the most commonly used neural network, namely the single hidden layer back-propagation network (vanilla). The vanilla uses the back-propagation algorithm to find the solution to the optimal parameters. The back-propagation algorithm can be described as follows:

Let $\mathbf{\Theta}$ denote the weight vector consisting of

$$\begin{gathered} \{\alpha_{0q}, \alpha_q; q = 1,2,\dots,Q\}, \text{where Q is the number of hidden units} \\ \{\gamma_{0k}, \gamma_k; t = 1,2,\dots,K\}, \text{ where K is the number of level in the target variable.} \end{gathered} \tag{2}$$

Define $\boldsymbol{X}$ be the input vector (predictor variables), $\boldsymbol{Z}$ be the hidden units (latent variables), and $\boldsymbol{Y}$ be the categorical target variable, then

$$\begin{gathered} Z_q = \sigma\left(\alpha_{0q} + \alpha_q^T \boldsymbol{X}\right), \text{ where } \sigma \text{ is the sigmoid function: } \sigma(v) = \frac{1}{1-e^{-v}}, \\ T_k = \gamma_{0k} + \gamma_k \boldsymbol{Z}, \text{ where } \boldsymbol{Z} = \left(Z_1, Z_{2,\dots,}Z_q\right),\ T = (\,T_1, T_{2,\dots,}T_k), \\ f_k(\boldsymbol{X}) = \ \pi_k(T). \end{gathered} \tag{3}$$

Compute $\hat{f}_k(x_i)$ using (3), and calculate the errors via (4)

$$\begin{aligned} L(\mathbf{\Theta}) &\equiv \sum_{j=1}^{N} L_j \\ &= \sum_{j=1}^{N} \sum_{k=1}^{K} \left(y_{jk} - f_k(x_j)\right)^2, \end{aligned} \tag{4}$$

where L is the loss function. Then back propagate the errors by means of (5)

$$\begin{aligned}\frac{\partial L_j}{\partial \gamma_{kq}} &= -2\left(y_{jk} - f_k(x_j)\right)\pi_k'(\gamma_k^T z_j) z_{qj}, \\ \frac{\partial L_j}{\partial \alpha_{qp}} &= -\sum_{k=1}^{K} 2\left(y_{jk} - f_k(x_j)\right)\pi_k'(\gamma_k^T z_j)\gamma_{kq}\sigma'(\alpha_q^T x_i) x_{jp}.\end{aligned} \tag{5}$$

Where $z_{qj} = \sigma(\alpha_{0q} + \alpha_q^T x_j)$, and $z_j = (z_{1j}, z_{2j}, \dots, z_{Qj})$.

Now update the weights using the following derivatives

$$\begin{aligned}\gamma_{kq}^{(l+1)} &= \gamma_{kq}^{(l)} - \psi \sum_{j=1}^{N} \frac{\partial L_j}{\partial \gamma_{kq}^{(l)}}, \\ \alpha_{qp}^{(l+1)} &= \alpha_{qp}^{(l)} - \psi \sum_{j=1}^{N} \frac{\partial L_j}{\partial \alpha_{qp}^{(l)}},\end{aligned} \tag{6}$$

Where ψ is the learning rate. The process is repated until either the maximum number of iteration number is reached or estimates of the error rate start increasing.

The neural networks involve a great number of parameters to be estimated, thereby being prone to overfitting the data. To mitigate this problem, we regularize the model using the penalized loss function:

$$\sum_{j=1}^{N}\sum_{k=1}^{K}\left(y_{jk} - f_k(x_j)\right)^2 + \xi\left(\sum_{kq}\gamma_{kq}^2 + \sum_{qp}\alpha_{qp}^2\right). \tag{7}$$

Where ξ is the penalty term added on the parameter estimations known as weight decay.

It should be noted that tuning the parameters of ANN is a challenging task as there is no constraint and explicit way to select optimal parameters. In this study, therefore, all parameters of the ANN model, such as weight decay, learning rate, and the number of units in the hidden layer are tuned using the cross-validation technique.

3.3.5.2 Random Forests

Random Forests, introduced by Breiman (2001), is a tree-based machine learning algorithm that recursively partitions data using a plurality of decision trees. The RF algorithm uses the bootstrap sampling technique to grow unique trees, which in turn not only overcomes the issue of overfitting but also renders the algorithm robust against noise in the data set. Moreover, RF randomly samples a portion of the predictors at each tree split but uses just one predictor to divide the data into two partitions. Limiting the predictors at each node decorrelates the trees produced.

Otherwise, a strong predictor, or a few strong predictors, will dominate and grow strongly correlated trees. The tree-growing process yields a forest of multiple trees. Then, the RF tallies each tree's vote and chooses the class by majority vote. In this study, we tune the parameters of the RF using the cross-validation technique, and intentionally dictate the model to grow an odd number of trees to avoid possible ties in the decision-making process. The detailed information regarding the RF and its parameters can be found in (Breiman, 2001) and (Breiman, 2002), respectively.

3.3.5.2 Logistic Regression

Logistic Regression is a member of the generalized linear models where the distribution of the response variable is assumed to belong to the exponential family of distributions (Hunter et al., 2008). Researchers use LR mainly for predicting dichotomous dependent variables. LR uses the logit function to relate the probability of the occurrence of events to the predictor variables. In a two-class problem (as in our case), when the odds of the occurrence exceed 1, the instance can be classified as one, and zero otherwise. The mathematical definition of the standard logistic function can be defined as follows:

$$logit(p) = \ ln\left(\frac{p}{1-p}\right) = \boldsymbol{\beta X} \tag{8}$$

where $\boldsymbol{\beta}$ and **X** are the coefficient and the input vector, respectively, and p is the probability of the target variable being equal one. The ultimate classification, then, is made as follows.

$$Prediction = \begin{cases} "1", if \ \dfrac{p}{1-p} > 1 \\ "0", if \ \dfrac{p}{1-p} < 1 \end{cases} \tag{9}$$

where $\frac{p}{1-p}$ is called the odds.

The researchers can manipulate the threshold value of the odds to get the desired level of sensitivity and specificity. For many of the programming languages, the default threshold for the odds is 1, or in other words, 0.5 for the p.

3.3.6 Sensitivity Analysis

The relative importance of predictor variables can be obtained and interpreted in theory-based statistical models such as generalized linear models, lasso, ridge regression, lars, and others.

(James et al., 2013). Conversely, black box models, (models with no closed mathematical form), such as ANNs do not provide direct access to the mechanisms in the underlying process. Consequently, alternatives are required to provide insight into black box models. One path to insights is sensitivity analysis (Davis, 1989).

Sensitivity analysis determines the influence of independent variables in predictive models. The algorithm includes each of the predictors, calculates the model error, then excludes one of the predictors and recalculates the model error. The ratio of the model error with the variable to the model error without the variable is the sensitivity measure, indicating how sensitive the model is to the predictor (Principe et al., 2000). This process is repeated recursively. Saltelli introduced a sensitivity measure used to rank the relative importance of the variables for each predictive model. Its form is (Saltelli, 2002):

$$S_i = \frac{V_i}{V(y)} = \frac{V(E(y \mid x_i))}{V(y)} \tag{10}$$

The binary response variable is *y,* and *V*(*y*) is the unconditional output variance. *E* is the expectation operator, integrated over x_i. We applied the sensitivity analysis on the best performing model and calculated the normalized sensitivity of the variable importance as described by Saltelli et al. (2002) (see Figure 3.5).

3.3.7 Information Fusion

Information Fusion (IF) gathers information obtained through different algorithms to reduce the model uncertainty, and improve robustness and information completeness. Studies have shown that combining multiple predictive models produces more information and greater accuracy when compared to single models (Elder, 2003).

IF techniques have three primary use cases: 1) preprocessing to improve data quality before using data mining techniques, 2) model building, e.g., using a combination of predictive models to improve results, and 3) information extraction – calculating results that generate knowledge (Vicenc Torra, 2003).

In the final step of our study (see Figure 3.1), we used IF to aggregate the sensitivity measures from each of the predictors into an overall sensitivity measure. The overall sensitivity measure enables us to discover predictor importance in target variable prediction despite the black box nature of the models.

IF aggregation can take numerous forms. We adopted the method demonstrated in Dag et al. (2017) to combine the information gained out of each model.

3.4 Results & Discussion

3.4.1 Variable Selection Results

As can be seen from Table 3.3, the optimization algorithms found different sub-optimal solutions, due to their differing algorithmic nature. For example, the GA algorithm classified *age* and *season* as principal predictors in explaining the variation in the dataset while SA ignored them. Similarly, the SA algorithm considers *handicap* as a crucial predictor, but the GA algorithm did not. However, the algorithms arrived at the same conclusion for some variables, such as *appointment day, lead time, appointment reminder, alcoholism, financial aid, diabetes, prior no-show,* and *time between appointments*. Lastly, neither of the methods considered *gender, scheduling day, calling time, hypertension and neighborhood* as essential variables.

Table 3.3: Variable selections results

Variable Name	Simulated annealing	Genetic algorithm
Age		✓
Gender		
Scheduling Day		
Appointment Day	✓	✓
Season		✓
Lead time	✓	✓
Calling time		
Appointment reminder	✓	✓
Alcoholism	✓	✓
Financial Aid	✓	✓
Handicap	✓	
Hypertension		
Diabetes	✓	✓
Neighborhood		
Prior no-show	✓	✓
Time between appointments	✓	✓

The optimization algorithms used in this study, GA and SA, are metaheuristic techniques in that they do not explore all search space, and thus cannot guarantee to provide the global best solution. In an attempt to compensate this shortcoming, besides using the variable sets provided by GA and SA, we considered using their possible combinations, namely the intersection set and union set of the variable sets given by the two algorithms. In our preliminary analysis, the union set of the predictor variables has not performed as well as the intersection set, presumably because of the issue of overfitting the data. Therefore, it is not included in Table 3.4. On the other hand, the intersection set not only provides promising results, but also advances the parsimony (simplicity) of the model, as it is one of the essential products of the current study. More details about the variables and prediction/classification results are discussed in the subsequent sections.

3.4.2 Classification Results

To begin with, to evaluate and compare the model performances we used four well-known performance evaluation metrics: accuracy, sensitivity, specificity, and AUC (a detailed explanation of the measures can be found in Powers 2011 (Powers, 2011)).

We trained our models with each of the variable subsets, which were defined in the previous sections (GA, SA and the intersection of these two) along with the balancing techniques. We evaluated the results using 10-fold cross-validation. We summarized the four measures of the performance of the predictive models numerically (i.e. accuracy, sensitivity, specificity, and AUC). Table 3.4 presents the classification/prediction results, with the standard deviation of each measure delimited by parentheses. The best results are set in boldface for each of the measures. We found the best AUC results (0.844) in the EL and ANN models, using RUS as the balancing technique and GA as the variable set provider. Similarly, the EL model using the predictor variables suggested by GA and balance via SMOTE produced the best accuracy (0.786) and specificity (0.814). Finally, the ANN model using the predictors obtained from the intersection of the variable set of GA and SA, and balanced by RUS, had the best sensitivity result (0.792).

Our study used the sensitivity metric as a primary criterion as it is crucially important to be able to correctly detect no-show patients so that necessary actions can be taken to mitigate any possible cost caused by no-show patients. Therefore, the ANN model, when trained with the variables that are commonly selected by GA and SA (intersection) and when the data is balanced with RUS algorithm, outperformed all of the other options in Table 3.4.

On the other hand, Table 3.5 shows the performance of the predictive models that are developed without implementing the proposed data analytics technique. As expected, all of the

Table 3.4: Classification results

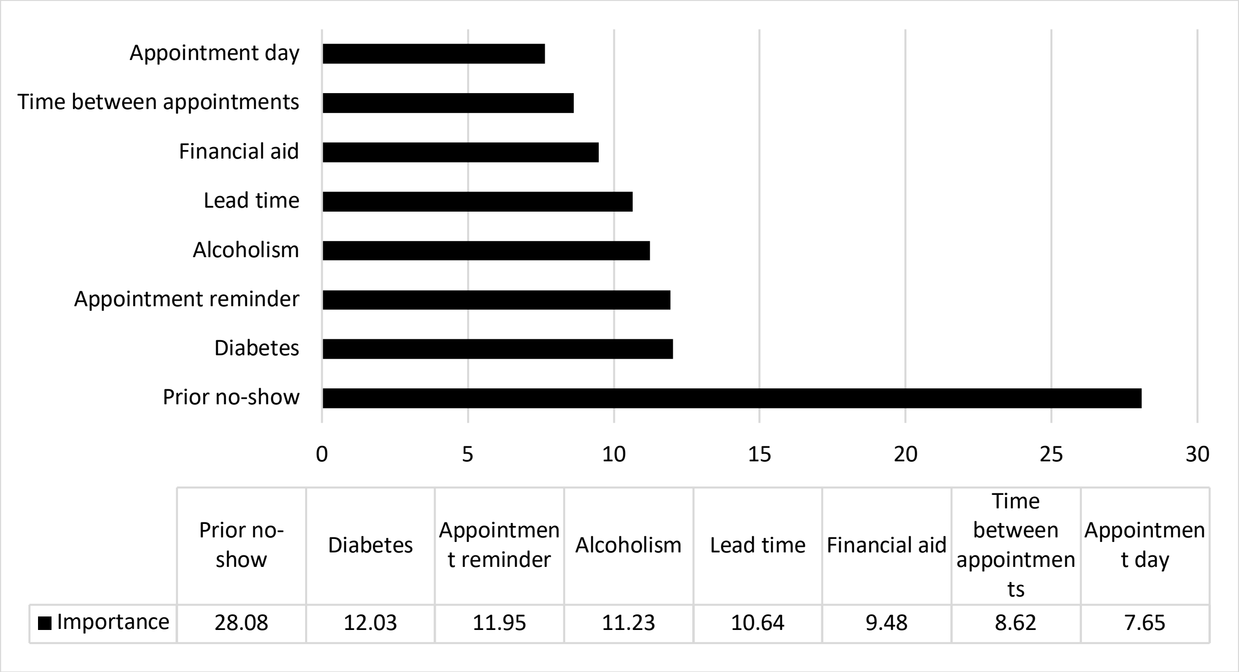
Variable Set		Model	AUC	Accuracy	Sensitivity	Specificity
SIMULATED ANNEALING	SMOTE	LR	0.805(0.004)	0.735(0.006)	0.706(0.011)	0.743(0.008)
		RF	0.818(0.006)	0.766(0.006)	0.703(0.018)	0.782(0.010)
		ANN	0.838(0.006)	0.757(0.009)	0.734(0.030)	0.763(0.018)
		EL	0.837(0.004)	0.770(0.005)	0.702(0.017)	0.788(0.008)
	ROS	LR	0.805(0.004)	0.727(0.006)	0.723(0.012)	0.728(0.008)
		RF	0.817(0.005)	0.756(0.004)	0.730(0.017)	0.763(0.008)
		ANN	0.839(0.005)	0.749(0.008)	0.756(0.013)	0.747(0.009)
		EL	0.837(0.004)	0.758(0.005)	0.733(0.012)	0.765(0.007)
	RUS	LR	0.805(0.004)	0.751(0.005)	0.664(0.013)	0.775(0.007)
		RF	0.824(0.006)	0.771(0.004)	0.701(0.017)	0.789(0.006)
		ANN	0.838(0.004)	0.768(0.006)	0.707(0.013)	0.784(0.009)
		EL	0.837(0.004)	0.777(0.004)	0.688(0.013)	0.800(0.005)
GENETIC ALGORITHM	SMOTE	LR	0.806(0.004)	0.734(0.007)	0.714(0.012)	0.739(0.008)
		RF	0.831(0.006)	0.782(0.005)	0.686(0.013)	0.807(0.005)
		ANN	0.837(0.004)	0.758(0.010)	0.728(0.030)	0.766(0.020)
		EL	0.843(0.005)	**0.786(0.006)**	0.680(0.009)	**0.814(0.009)**
	ROS	LR	0.806(0.004)	0.727(0.007)	0.732(0.012)	0.726(0.008)
		RF	0.827(0.006)	0.771(0.007)	0.696(0.012)	0.791(0.009)
		ANN	0.837(0.007)	0.761(0.011)	0.725(0.028)	0.770(0.019)
		EL	0.842(0.005)	0.776(0.005)	0.701(0.010)	0.796(0.007)
	RUS	LR	0.806(0.004)	0.729(0.006)	0.727(0.011)	0.730(0.007)
		RF	0.833(0.006)	0.745(0.006)	0.777(0.015)	0.737(0.008)
		ANN	**0.844(0.005)**	0.742(0.007)	0.781(0.012)	0.732(0.010)
		EL	**0.844(0.005)**	0.757(0.005)	0.750(0.012)	0.759(0.007)
INTERSECTION	SMOTE	LR	0.805(0.004)	0.735(0.006)	0.706(0.011)	0.743(0.008)
		RF	0.827(0.005)	0.765(0.006)	0.715(0.027)	0.778(0.014)
		ANN	0.838(0.004)	0.757(0.005)	0.735(0.024)	0.763(0.011)
		EL	0.837(0.004)	0.771(0.004)	0.703(0.021)	0.789(0.008)
	ROS	LR	0.805(0.004)	0.727(0.006)	0.723(0.012)	0.728(0.008)
		RF	0.830(0.005)	0.752(0.005)	0.752(0.017)	0.752(0.008)
		ANN	0.838(0.004)	0.746(0.010)	0.756(0.017)	0.744(0.015)
		EL	0.837(0.004)	0.757(0.005)	0.736(0.015)	0.762(0.009)
	RUS	LR	0.805(0.004)	0.729(0.006)	0.718(0.010)	0.732(0.008)
		RF	0.831(0.005)	0.740(0.005)	0.780(0.015)	0.729(0.008)
		ANN	0.838(0.005)	0.730(0.005)	**0.792(0.019)**	0.714(0.008)
		EL	0.838(0.004)	0.747(0.004)	0.765(0.014)	0.742(0.008)

models produced very low sensitivity scores, as low as 0.378 and as high as 0.426, while producing relatively high specificity scores. Such an outcome clearly shows that these models have very low power in correctly classifying no-show patients, emphasizing the crucial necessity of performing data balancing.

Table 3.5: Classification results without use of the proposed data analytics technique

Variable Set	Data Balancing	Model	AUC	Accuracy	Sensitivity	Specificity
NONE	NONE	LR	0.826(0.005)	0.822(0.004)	0.378(0.008)	0.940(0.003)
		RF	0.847(0.005)	0.837(0.004)	0.426(0.009)	0.945(0.003)
		ANN	0.825(0.010)	0.826(0.005)	0.416(0.059)	0.935(0.016)
		EL	0.842(0.005)	0.835(0.004)	0.384(0.018)	0.954(0.005)

After selecting the ideal model (ANN with RUS with the intersection of SA and GA variables), we investigated the importance of the variables used in building the ideal (ANN) model via sensitivity analysis. Namely, 10 different variable importance reports are produced, as we ran 10 different models because of 10-fold cross-validation. After obtaining the individual importance of the variables for these models, we combined the reports obtained from the models via using the IF technique.



	Prior no-show	Diabetes	Appointment reminder	Alcoholism	Lead time	Financial aid	Time between appointments	Appointment day
■ Importance	28.08	12.03	11.95	11.23	10.64	9.48	8.62	7.65

Figure 3.5: Sensitivity analysis for predictor variables

Figure 3.5 shows the contribution (importance) of each variable in predicting the outcome. Note that these numbers present the relative importance of each variable, when compared with each other. For example, the contribution of *prior no-show* is 28.08% and *diabetes* is 12.03%. Collectively, the two comprise 40.11% of the total contribution.

Our findings confirm most of what has been published in the related literature. To exemplify, many studies have reported that patients with a high prior no-show history are more likely to miss their next appointments, thereby making prior no-show one of the most important predictors cited in the literature (Cronin et al., 2013; Daggy et al., 2010; Dove & Schneider, 1981; Farid & Alapont, 1993; Y. Huang & Hanauer, 2014; Kempny et al., 2016; V. J. Lee et al. , 2005; Torres et al., 2015).

Similarly, the literature reported that patients with diabetes have no-show rates as high as 40% (Nuti et al., 2012). Bindman et al. (1995) have reported that chronic medical conditions such as *diabetes*, can require acute care service. Therefore, it might be the reason that a patient with diabetes who fails to show up for their appointments, bears a high chance of being hospitalized due to the worsening health condition, thereby missing the next appointments.

Moreover, the relative contributions of *appointment reminder, alcoholism, lead time,* and *financial aid* are 11.95%, 11.23%, 10.64%, and 9.48% respectively. The importance of these variables is also well articulated in the literature. For example, researches have argued that appointment reminders such as SMS, phone call, and email reminders, are effective at reducing the no-show rate of patients (Hasvold & Wootton, 2011; Junod Perron et al., 2010; Milne et al., 2006; Vodopivec-Jamsekr et al., 2012). The reduction has been reported to be as high as 75% (A. Tibble et al., 2000). Similarly, Dantas et al. (2018) have shown that patients with addiction to alcohol are more likely to miss their appointments. Further, researchers comparing four significant factors (age, payer, no-show history and lead-time) found that lead-time was the strongest predictor of no-show patients (Norris et al., 2014) Another study, using data from an ophthalmology clinic, confirmed that longer lead times increased no-show rates (McMullen & Netland, 2015). On the other hand, Norris et al. have shown that financial aid is one of the greatest reasons associated with patient no-shows. This claim supported by Horsley et al. (2007) asserting that Medicaid vs. non-Medicaid use was a significant predictor of missed appointments, with Medicaid patients having a much higher incidence of no-shows.

Lastly, we find that *time between appointments* and *appointment day* are other two important variables in the prediction of the no-show patients, having relatively lower contributions: 8.62%, and 7.65 %, respectively.

Different than the published relevant literature, we have not found any studies on the time between two consecutive appointments that was created in the data processing step. The variable simply measures the elapsed time between the last appointment and the appointment scheduled before the last appointment, thereby providing insights regarding patient's appointment pattern. For example, if the length of the time between the appointments are relatively low, or in other words, if the appointments are scheduled frequently, then it might indicate a serious health condition requiring the patient to schedule appointments on regular basis, and thus the patient can be expected to show up for his/her appointments. On the other hand, a long time period between the two consecutive appointments might imply a regular check-up appointment or sickness that the patient might choose to not keep his/her appointment depending on the urgency of the disease.

Finally, studies reported appointment day as an important predictor (Kheirkhah et al., 2015; Torres et al., 2015). Monday and Friday are deemed as two specific days on which the highest no-show rates occur (Y.-L. Huang & Hanauer, 2016; Y. Huang & Hanauer, 2014; Kheirkhah et al., 2015; Torres et al., 2015).

The (ideal) ANN model assigns (estimates) probabilistic scores for each instance (patients), ranging between 0 and 1, as presented in Table 3.6. It should be noted that these results are obtained

Table 3.6: Classification results for different probabilistic scores

Probability Score	Dropped Cases	# Show-up	# No-show	TP	FN	AUC	Accuracy	Sensitivity	Specificity
				FP	TN				
0.5	0	5757	1470	1208	1660	0.846	0.734	0.822	0.712
				262	4097				
$\notin [0.4, 0.6]$	1945	4276	1006	828	599	0.892	0.853	0.823	0.860
				178	3677				
$\notin [0.3, 0.7]$	2602	3780	845	712	441	0.904	0.876	0.843	0.883
				133	3339				
$\notin [0.2, 0.8]$	3568	3003	656	584	302	0.919	0.898	0.890	0.899
				72	2701				
$\notin [0.1, 0.9]$	5722	1248	257	234	65	0.950	0.942	0.890	0.948
				23	1183				
$\notin [0.05, 0.95]$	6538	627	62	54	4	0.960	0.983	0.871	0.994
				8	633				

from the 7th fold as the best sensitivity score was obtained from it. The estimate of a probability score close to *1* indicates a high chance of being a show-up for the patient, while a probability score of *0* indicates otherwise. Any numbers in-between the two are classified as one way or the other, according to the specified threshold value. For example, If a patient is assigned a probability score of 0.9, then it can be considered that there is a 90% chance that the patient will be a *no-show*. On the other hand, a probability score of *0.05* for a patient indicates that there is a 5% chance of being a *no-show*. In other words, there is a 95% chance that the patient will show up for her/his appointment. With that in mind, Table 3.6 illustrates the changing performance of the proposed ANN model when the different probabilistic thresholds are considered. For example, when the instances with the assigned prediction probability scores on the interval [*0.4, 0.6*] are dropped, the performance of the model increases in each metric.

It should be noted that the interval of *0.4* and *0.6* is an interval that the model does not feel very confident in predicting the outcome (the closer the probability is to *0.5*, the less confident the model is in its predictions). In other words, the model performance increases when the instances that the model is not confident are dropped. As expected, the same pattern can be seen as we widen the interval and drop the cases in between the updated probability borders. As such, the performance of the model reaches an accuracy score of 0.983, an AUC score of 0.960, a sensitivity score of 0.871, and a specificity score of 0.994 in the classification of the patients with the estimated probability scores less than *0.05* and greater than *0.95*, attesting to the reliability of the probabilities estimated by the proposed model. Therefore, expert help can be used for the cases that the model probability is close to *0.5* in order to aggressively increase the accuracy of the model. This is another common practice in the modern machine learning world, which artificial intelligence and humans collaboration becomes inevitable (H. James & Paul R., 2018).

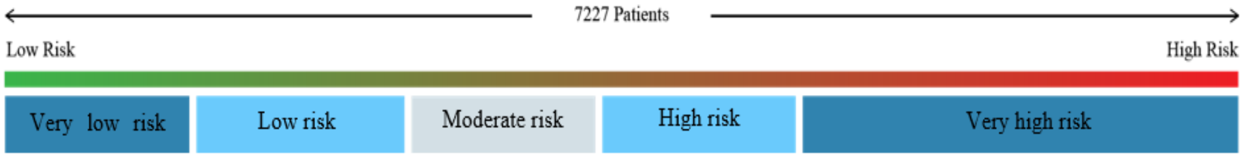


Figure 3.6: The risk level classification through k-means clustering algorithm

In addition, we have stratified the patients into five categories (i.e. *Very low risk(1)*, *Low risk(2), Moderate risk(3), High risk(4) and Very high risk(5)*) by employing k-means clustering algorithm on the estimated probabilities (Figure 3.6). As discussed earlier, the cases associated with these groups were incrementally dropped starting with the middle group and moving outward in both directions on the probability spectrum, and the remaining cases were run through the ANN (ideal) model. Such risk level stratification can be used by medical decision makers to augment their decision-making process when assessing a specific patient's situation for no-shows. Such a decision support mechanism can be used to lighten the burden of a large portion of the cases associated with no-shows, depending upon a specific threshold probability between available resources, risks, and costs. Healthcare agencies can therefore adjust their risk scoring metric to achieve an ideal balance between type I and type II errors based on the cost-benefit analysis.

3.5 Conclusion

In this study, a hybrid data mining based methodology was formulated to provide healthcare agencies and medical decision makers with a patient-specific risk level of no-shows. The overarching goal was to provide patients better care by utilizing available resources more efficiently and, at the same time, decrease the costs that originate from patient no-shows.

Our study fills a significant gap in the existing body of knowledge in that it; 1) develops a very parsimonious model by using wrapper-based *variable selection* models as "*the simpler is the better*", 2) employs cutting edge data balancing algorithms to increase the ability of the predictive model in detecting the minority class, as the no-show datasets are mostly imbalanced (i.e. the number of the no-shows are much less than the number of show-ups), 3) determines the patients who have *very low-, low, moderate-, high- and very high risk* score of being a no-show and, 4) analyzing a publicly available patient data that were collected over a significant period of time. To the best of our knowledge, the outcome of the current study is significant to the no-show problem, as none of these points were addressed in the related literature.

It should be clearly indicated that the contribution of the current study is not the application of popular classification models. It is rather; a) the simplicity of the model built, b) a high sensitivity (low Type 1 error) score that was achieved and c) the probability risk level that was provided as an end-product to the medical decision makers.

Our study suggests that the ANN model should be employed as a classification model in predicting the no-show patients by using the variable set that is commonly selected by GA and SA algorithms. In addition, the RUS balancing method should be employed to improve the performance of the model in predicting the minority group of no-show patients.

The patient-specific risk scores obtained were justified by applying a threshold sensitivity analysis and it has been shown that model performance consistently increases when the patients, about whom the model is not very confident, are dropped. The medical experts will be able to confidently rely on the probabilistic score provided, while the data analytical models and medical experts' intuition/incentive should collaborate to make the final decision on the cases where the model is not confident enough. Healthcare agencies can take advantage of the risk assessment to augment their decision-making procedures.

Those insights enable health care professionals to improve clinic utilization, and improve patient outcomes. Healthcare delivery systems can adopt measures targeting patients that are no-show candidates effectively by systematizing intervention tactics.

Chapter 4

Predicting No-show Patients with Individual Probabilistic Risk Scores: A Bayesian Belief Network Approach

4.1 Abstract

Patients who miss their appointments (no-shows) reduce revenues and impair the delivery of quality healthcare. Much research has been devoted to identifying “no-show patients”. However, they have had limitations in that: a) publicly available datasets have not been employed; b) parsimonious models based on data-driven variable selection models have not been produced; c) data balancing methods have not been employed even though imbalance is inherent in no-show datasets; d) individualized patient risk scores of being a “*no-show”* have not been devised, and; e) a web-based decision support system tool, which can be adapted by medical practitioners/facilities to minimize the no-show risk of a patient has not been built for practical use. In this study, we build a probabilistic data-driven methodology that consists of five steps to overcome these limitations. After data acquisition and preparation in the first step, the second step is dedicated to selecting important variables through pure data-driven wrapper methods such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA) and Extreme Gradient Boosting (XGB). Then, in the third step, the Synthetic Minority Oversampling Technique (SMOTE) and Random Under-sampling (RUS) are employed along with 10-fold cross-validation to overcome the data imbalance issue that exists in the dataset. In the fourth step, the patient-specific probabilistic risk scores along with the conditional interrelations among the predictors are obtained via the Tree Augmented Bayesian Belief Network (TAN) model. Finally, the probabilistic risk scores are justified, patients are clustered into 5 risk groups and a web-based decision support system (GUI: graphical user interface) is built. Results show that an overall AUC score of 0.828 can be achieved with a sensitivity score of 0.785 if the data is balanced through RUS and the variable selection is made via GA and PSO. More importantly, results show that an extremely parsimonious model can be built (by only employing 7 variables), and interesting conditional inter-relations exist between these variables. In addition, a web-based GUI is developed, which can be adapted by healthcare

facilities, which would enable them to obtain individualized risk scores and come up with better scheduling scenarios. We believe that such an automated tool can easily be adapted by medical clinics to decrease the number of no-shows.

4.2 Introduction

No-shows, i.e. patients who do not attend scheduled appointments, create an expensive problem for medical providers. Skipped appointments reduce revenues for healthcare institutions, and increase the cost of healthcare (DuMontier et al., 2013; Moore et al., 2001). Missing appointments is also associated with poor healthcare outcomes (Fortin et al., 2016). Several strategies to mitigate skipped appointments have been proposed. For example, reducing the number of open appointments through overbooking (Reid et al., 2016; Zeng et al., 2010), using incentives and disincentives to modify patient behavior, including moving no-show patients to the end of the waiting list (Vikander et al., 1986), having patients prepay a portion of the cost (Garuda et al., 1998), or using fines to penalize patients who miss appointments (Bech, 2005) have all been shown to reduce incidences of no-shows. An alternative strategy is to use patient interventions to reduce non-attendance (Macharia et al., 1992; Shah et al., 2016). To use interventions effectively requires the identification of essential predictors. However, the efficacy of these predictors can vary (Junod Perron et al., 2010). Selecting effective predictors will produce better results by allowing judicious use of interventions when compared to indiscriminate targeting.

The cost of no-shows is substantial. One study estimated the daily loss for no-shows at $725.42 per US provider (Berg et al., 2013). Applying that estimate to the 968,743 active US physicians, plus 87,835 active US physician assistants (Kaiser, J, 2018), indicates a US revenue loss of at least $766 million per year. The actual cost is likely higher. The 2008 Veterans Health Administration estimate of unused appointments was $564 million ("Office of inspector general audit of veterans health administration's effort to reduced unused," 2008). A 2006 estimate for England's NHS missed appointments cost was about £780 million a year (Atun & Sittampalam, 2006). Attempts to reschedule other patients in the available time slot, even using overbooking, still leaves resources underutilized (Moore et al., 2001; Reid et al., 2016). Missed appointments are associated with lower preventive health service utilization and inadequately controlled hypertension and diabetes (Nguyen, DeJesus, & Wieland, 2011).

Predictive analytics has been used in several studies to incorporate the probabilities of no-shows into a scheduling system (Daggy et al., 2010; Glowacka et al. , 2009; Samorani & LaGanga, 2015). Once the probabilities are found, they inform the schedule. Glowacka et al. (2009) predicted no-shows at a free clinic in the southeastern United States using association rule mining. They used their results as a scheduling policy input to reduce no-shows.

Daggy et al. (2010) used data from a Veterans Affairs hospital in the United States to develop a logistic regression model to estimate the number of no-shows. In the regression model, patients' age, marital status, travel distance, diagnoses, insurance, the days since the last visit, appointment lead time, prior no-show rate, the total number of previous visits and the season were all used as independent variables. The model's purpose was to find a regression model that could be used to optimally schedule patients while using an overbooking policy. Once the no-show probabilities were generated using the regression, physician utilization and overtime were estimated and analyzed using simulations. In a similar study, Samorani & LaGanga (2015) used individual appointment characteristics and appointment day to predict no-shows. These predictions were then used in the heuristic scheduling procedure.

In a study at a pediatric medical facility, Topuz et al. (2018) identified significant contributors for no-shows and discovered relationships in predictors such as demographics, current appointment information, socioeconomic status and appointment attendance history of the patient and the family. Although they have used feature selection models, none of these selection models was wrapper based method. The advantages of using wrapper methods will be discussed in detail in the subsequent sections.

Although the aforementioned studies focus either on predicting or finding important predictors etc., to our best knowledge, none of them have; a) used publicly available data in their analysis, b) employed a comprehensive data-driven feature selection approach that consists of wrapper methods; c) employed data balancing methods even though imbalance is inherent in no-show datasets (i.e. the number of "*no-show"* patients is very low when compared to patients who are not "*no-show*"); d) obtained individualized patient risk scores and; e) built a web-based decision support tool that can be easily adapted by medical practitioners/facilities to predict the risk of being a "*no-show*" patient, therefore allowing for a chance to decrease the amount of "*no-shows"* using appropriate interventions.

Given these limitations of the previous literature, our study contributes to the existing body of knowledge by; a) using a publicly available large dataset, b) removing noisy variables to ensure near-optimal model performance and accurate feature importance information, which in turn enables us to obtain the most parsimonious model possible; c) taking into account the class imbalance problem, which prevents the model from "cherry-picking" potential no-show patients; d) determining the patient-specific risk score of being a no-show; and finally e) building a web-based decision support systems tool for the practical use of medical staff & facilities to prevent/decrease possible future *no-shows*. To achieve this goal, a hybrid data mining methodology was employed. More specifically, we built a probabilistic data-driven methodology that consists of five steps. After data acquisition and preparation in the first step, our second step was dedicated to select important variables through data-driven models such as PSO, GA, and XGB. Then, in the third step, SMOTE and RUS were employed along with 10-fold cross-validation to overcome the imbalance problem, which exists in most of the no-show datasets. The combination of using balancing methods along with feature selection algorithms should ensure that the resulting feature importance findings are not affected by the inherent class imbalance. In the fourth step, the individualized patient (probabilistic) risk scores along with the conditional interrelations were obtained via TAN model. In the final step, the probabilistic risk scores were justified and a web-based decision support system (GUI) is built for the medical practitioners' use. This tool is designed in a way that it does not require the end user to have any background knowledge in machine learning, statistics, or optimization, etc. fields.

We believe that the contribution and the value of the current study do not come from the classical machine learning, statistical and/or optimization models, which have been employed in many health- and medical informatics studies previously. Rather, it is contributing to the existing body of knowledge because of its holistic probabilistic data analytic approach, which obtains patient-specific risk scores (of being a no-show) by employing the most parsimonious models that were built by preventing the class imbalance problems that occur due to significant differences in most no-show datasets. It should also be noted that the web-based tool also serves as a practical contribution of our study.

4.3 Methodology

Our data analytics approach consists of five steps as shown in Figure 4.1. In the first step, we prepared the dataset obtained from the *kaggle.com* website ("Medical Appointment No Shows," 2016). In the second step, a machine learning algorithm, namely XGB, and two optimization algorithms, namely PSO and GA were employed to select important variables. All possible combinations (unions/intersections) of the variables that are obtained through these algorithms were created in addition to SMOTE and RUS balancing techniques in the third step. In the fourth step, the TAN model for each variable set was employed using 10-fold cross-validation to discover relationships among the variables and to obtain a patient-specific (individual) probabilistic risk score. In the final step, a web-based DSS was built for practitioners' use.

4.3.1 Data Preparation

In this study, we obtained the dataset from a well-known data science competition platform; *kaggle.com* ("Medical Appointment No-shows," 2016). The dataset was recorded by public healthcare providers in the state of Espírito Santo, Brazil, and consists of 110,528 observations with 14 variables. Using these, we also created a few more variables. For example, *lead time, appointment day, appointment month and appointment scheduling day* were created using the variables *scheduled date* and *appointment date*. Similarly, using the patient ID, two patient-specific variables such as *prior no-shows and time between appointments* were created. The rationale behind creating such variables is the potential contribution of these variables in predicting the patient's no-show. During data exploration, we discovered that the *lead time* for some patients was less than zero (the number of days between the appointment scheduled day and the appointment day), which we believe is due to the erroneous data entries. Therefore, we eliminated these records from the dataset. Although some variables did not play any role in predicting the response variable (*Patient ID* and *Appointment ID*), they were kept in the dataset since they provided information about the patient's history. However, they were excluded from the analysis part. After creating additional variables and the variable elimination process, we were left with a dataset that has 72,602 observations and 17 variables. The summaries of the predictor variables are shown in Table 4.1.

Table 4.1: Predictor variables

Categorical Variable Name	Number of levels		
Gender	2		
Appointment month	3		
Appointment day	6		
Scheduling day	6		
SMS reminder	2		
Alcoholism	2		
Financial aid	2		
Handicap	2		
Hypertension	2		
Diabetes	2		
Neighborhood	79		
Continuous Variable Name	Mean	Median	Range
Age	37.24	37	0-115
Lead time	9.25	4	0-179
Calling time	10.88	10	6-20
Time between appointments	5.62	1	0-40
Prior no-shows	0.13	0	0-1

4.3.2 Variable Selection

Variable selection is the task of searching for optimal variable subsets from the original dataset by evaluating the relationship between the explanatory variables and response variable with some metrics such as correlation, entropy, error-probability, etc. Variable selection is necessary for several reasons. For example, it reduces dimensionality by eliminating redundant variables, which in turn decreases model training time (Saptarsi & Amlan, 2014). It also enables learning algorithms to generalize existing relationships between the explanatory variables and the response variable, thus prevents overfitting (Bermingham et al., 2015), and eases the model interpretability by simplifying the model complexity (James et al., 2013). There are various types of feature selection methods, which can be categorized into two types: filter methods and wrapper methods (John et al., 1994).

Filter methods are univariate statistical methods that are mostly used in the data preparation phase of
analysis. Many filter methods have been described in the literature (Saeys et al., 2007), and mostly consist of covariance tests such as t-test, ridge regression (Hoerl & Kennard, 1970), least absolute

shrinkage and selection operator (LASSO) (Tibshirani, 1996), etc. which can be used to determine if an independent variable is correlated to the output variable.

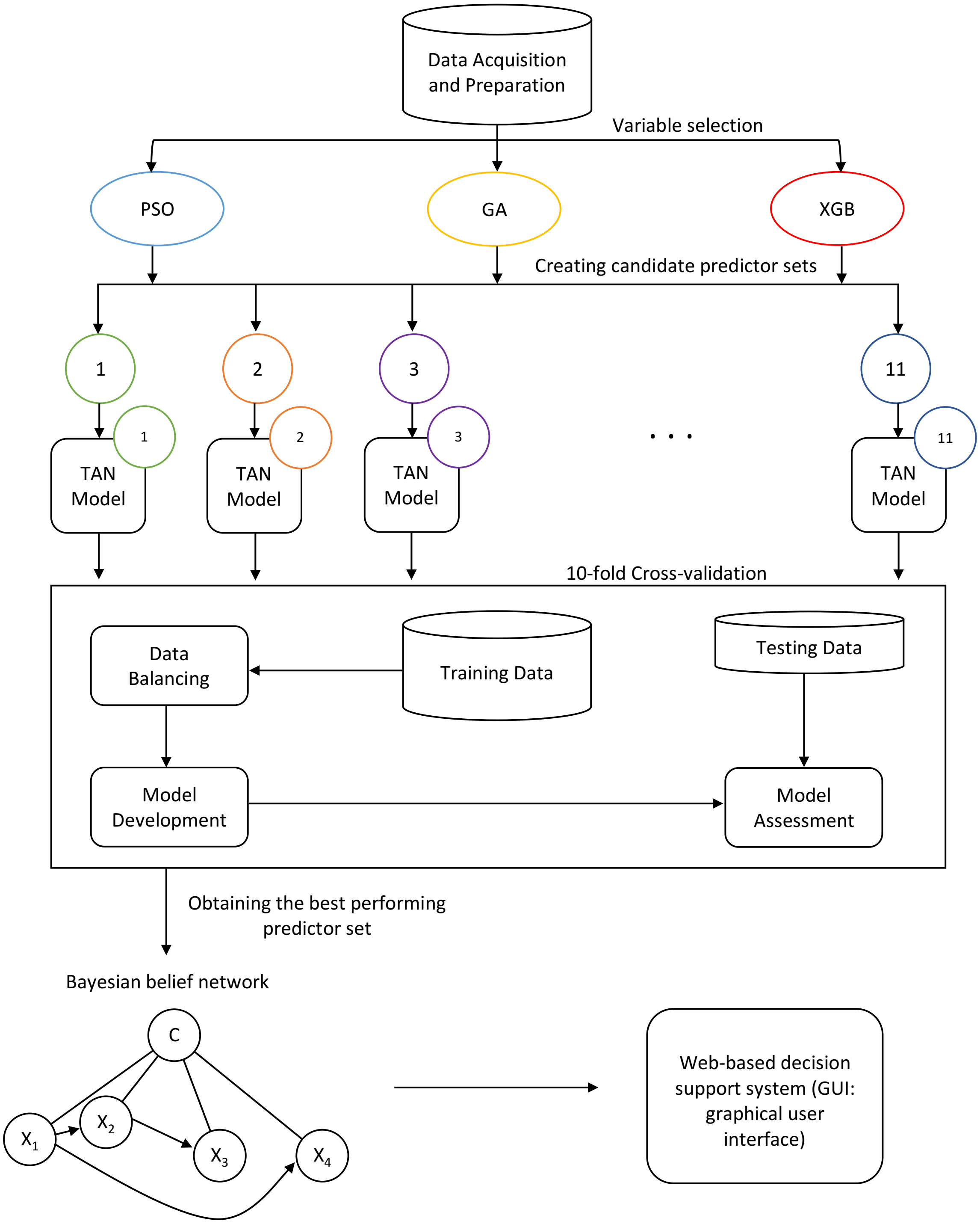


Figure 4.1: An Outline of the proposed methodology

Wrapper methods, instead, are algorithmically or heuristically defined approaches to select features for a machine learning model. The methods are used in conjunction with a machine learning algorithm, and essentially "wrap" the model in the training process. These methods use an algorithm or heuristic to add or remove features from the dataset and retrain the model, with the goal of increasing the models' predictive performance, iterating the process until a stopping rule is reached. There are many wrapper methods, such as Recursive Feature Elimination (Guyon et al., 2002), Simulated Annealing (Aarts & Korst, 1988), Genetic Algorithms (J. Yang & Honavar, 1998) and Particle Swarm Optimization (Kennedy, 2011), etc.

In this study, we used two wrapper methods (i.e. GA and PSO) and a machine learning algorithm (i.e. XBG) to conduct feature selection due to their high performance in our preliminary data analysis stage. The detailed information regarding the algorithms is given in the following subsections.

4.3.2.1 Genetic Algorithm

Only organisms who adapt themselves to the environment, can thrive and get ahead of the ones who are less adaptive. This phenomenon is referred to as "natural selection". Genetic algorithms, inspired by natural selection, are adaptive heuristic search algorithms that are used to solve optimization problems. They are preferred in situations where exhaustive optimization techniques are computationally expensive. Genetic algorithms initially create a random set of individuals (chromosomes) that explore the whole solution space. These chromosomes are tested at each iteration according to their fitness functions. The chromosomes considered to be superior in terms of the fitness function are kept in the population and are mated with other well-performing chromosomes to create new chromosomes. The new chromosomes, members of a new generation, can also have random mutations. The chromosomes that do not perform well are wiped out from the population so that the population size of chromosomes is controlled.

Genetic algorithms have also been used for variable selection purposes in the literature (Dag et al., 2017; J. Yang & Honavar, 1998). In this study, the genetic algorithm uses the *Partial Least Squares* (PLS) regression model to evaluate the prediction performance of the chromosomes (i.e. variable subsets in our case).

All parameters of the PLS model are optimized using the cross-validation technique. For detailed information about the genetic algorithm and the PLS regression model, the reader is referred to (Goldberg & Holland, 1988) and (Wold, 2006), respectively.

4.3.2.2 Gradient Boosting

Gradient tree boosting is a machine learning technique that creates a predictive model by combining many weak learners. The idea behind gradient tree boosting is to grow a sequence of simple trees where each tree is grown to explain the prediction residuals of the preceding tree. A direct implementation of the boosting tree algorithm, however, might be computationally expensive since the time required to train a model increases exponentially, depending on the sample size. In an effort to improve the computational efficiency of the boosted tree algorithm, Chen and Guestrin et al. (2016) introduced a scalable tree boosting system called Extreme Gradient Boosting (XGB). The gradient tree boosting algorithm predicts the output by summing K independent regression and classification tree functions:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{1}$$

where K stands for the number of trees and $\mathcal{F}$ is the space of regression trees in which f is a tree function that maps each observation vector to a certain output. To find the optimal tree structure, one must minimize the objective function:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

where l is the loss function calculating the difference between the actual response y_i and the predicted response $\hat{y}_i$. The second term $\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j-1}^{T} w_j^2$ is the penalization parameter that tunes model complexity, where γ and λ are the constants that control regularization degree, w is the scores vector on leaves, and the number of leaves is represented by T. It should be noted that when Ω is set to be 0, the optimization objective function boils down to regulating the tree boosting algorithm. Moreover, in the optimization process, the XGB algorithm uses shrinkage and column subsampling techniques (Chen & Guestrin, 2016) to prevent overfitting. The optimization task of the objective function cannot be accomplished by traditional optimization methods since it contains functions as parameters. Therefore, the model needs to be trained iteratively, as follows:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

where the prediction of $\hat{y}_i^{(t)}$ for i^{th} instance at the t^{th} iteration, and $\hat{y}_i^{(t)} = 0$ when t is equal to 0 (i.e. the prediction value at first step starts with 0).

Therefore, Equation 2 takes the form:

$$L^{(t)}(\phi) = \sum_i l\left(\hat{y}_i{}^{(t-1)} + f_t(x_i), y_i\right) + \Omega(f_t) \tag{4}$$

The second-order Taylor approximation can be applied to Equation 4 in order to hasten the optimization of the objective function (J. Friedman et al., 2000):

$$L^{(t)}(\phi) \approx \sum_i \left[l\left(\hat{y}_i{}^{(t-1)}, y_i\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t{}^2(x_i) \right] + \Omega(f_t) \tag{5}$$

where the first and second order statistics on the loss function are represented by $g_i = \partial_{\hat{y}^{(t-1)}} l\left(\hat{y}_i{}^{(t-1)}, y_i\right)$ and $h_i = \partial^2{}_{\hat{y}^{(t-1)}} l\left(\hat{y}_i{}^{(t-1)}, y_i\right)$, respectively. Since the loss function $l\left(\hat{y}_i{}^{(t-1)}, y_i\right)$ is constant at step t, it can be removed to simplify the objective function and thus can be written as

$$L^{(t)}(\phi) \approx \sum_i \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t{}^2(x_i) \right] + \Omega(f_t) \tag{6}$$

After extending the regularization term Ω, given that the instance set of leaf j is $I_j = \{i | q(x_i) = j\}$, the objective function takes the form:

$$\begin{aligned} \tilde{L}^{(t)}(\phi) &\approx \sum_i \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t{}^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2) \\ &= \sum_i \left[\left(\sum_{i \in I_j} g_i \right) w_i + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \end{aligned} \tag{7}$$

The optimal weight w_j^* of leaf j with a fixed structure q(x) and the optimal objective function can be calculated as in Equations 8 and 9, respectively.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{8}$$

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{9}$$

After finding a way to measure the goodness of the tree, the algorithm attempts to optimize the tree one level at a time by assigning a score to a leaf node in the splitting process:

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{10}$$

where the first term is the score for the new left leaf, the second term is the score for the new right leaf and the third term is the score for the original leaf from which the left and right leaves grow.

4.3.2.3 Particle Swarm Optimization

Particle swarm optimization is a metaheuristic search technique that mimics the social behavior of swarming organisms to accomplish an optimization task (Kennedy, 2011). By virtue of its metaheuristic nature, PSO manages to obtain a solution for irregular, time-dependent or even noisy functions, which in turn, leads PSO to be successfully applied in various scientific fields, such as machine learning (Lin et al., 2008), image processing (Dhanalakshmi et al., 2016), business optimization (X.-S. Yang et al. , 2011), operations research (B. Liu et al., 2008). PSO algorithms create a population of candidate solutions i.e. a swarm. Each member of this swarm is called a particle and has its own position and fitness value calculated by the fitness function. The particles fly (iterate) across the search space in accordance with their velocity vector. The velocity vector has two main components called personal best and global best. The personal best information belongs to an individual particle and is the best point experienced by the particle in the search space. On the other hand, the global best is the best point obtained by the swarm. Each particle uses the personal best and the global best information to update its current position to a presumably better position. In this process, the extent to which personal best and the global best will be given an ear to update the velocity vector is managed by weights called the learning factor. Here, we denoted the number of particles by *N* and i refers to the specific ith particle in the swarm. The position vector of the particle i at time t is denoted by $X_i^t = (x_{i1}^t, x_{i2}^t, \dots, x_{ik}^t, \dots, x_{iD}^t)$ where *D* is the number of variables and $x_{ij}^{t} \in \{0,1\}$. The personal best for the ith particle and the global best at time t are represented by $P_i^t = (p_{i1}^t, p_{i2}^t, \dots, p_{ik}^t, \dots, p_{iD}^t)$ and $G^t = (g_1^t, g_2^t, \dots, g_2^t, \dots, g_2^t)$, respectively where $P_i^t \; and \;\; G^t \in \{0,1\}$. The global best G^t does not have the index i as it belongs to the swarm, not the individual particle. After each iteration, the velocity vector for the particle i is updated as follows:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (P_i^t - X_i^t) + c_2 r_2 (G^t - X_i^t) \tag{11}$$

where, $V_i^t = (v_{i1}^t, v_{i2}^t, \dots, v_{ik}^t, \dots, v_{iD}^t)$ is the previous velocity vector and, c_1 and c_2 are the weight factors for the personal best solution and global best solution, respectively. For particles to move stochastically in the search space, random numbers, uniformly distributed in [0,1], r_1 and r_2 are included in Equation 11. The inertia weight ω controls the extent to which the previous velocity will affect the new velocity. It should be noted that the particles are kept in the search space by bounding the velocity vector, as shown in Equation 18.

In continuous PSO, the position vector is updated according to:

$$X_i^{t+1} = X_i^t + V_i^{t+1}. \tag{12}$$

On the other hand, in discrete PSO, the velocity vector is transformed into the probability vector via a sigmoid function and takes the form of

$$S_{ij}^t = \frac{1}{1 + e^{V_{ij}^t}} \tag{13}$$

where S_{ij}^t represents the probability that jth bit X_i^t is 1. Therefore, the position of the particle is updated as follows:

$$X_{ij}^t = \begin{cases} 1, & \Omega < S_{ij}^t \\ 0, & otherwise \end{cases} \quad \text{j=1, …, K,} \tag{14}$$

where Ω is a uniform random number between 1 and 0.

The maximization of the function f is equal to the minimization of function $-f$. For this reason, without loss of generality, maximization of the fitness function in a D-dimensional space can be expressed as:

$$Given\, f\colon \Re^D \to \Re$$
$$Find\;\; \boldsymbol{X}_{opt} | f(\boldsymbol{X}_{opt}) \geq f(\boldsymbol{X}) \quad \forall \boldsymbol{X} \in \Re^D \tag{15}$$

where $\boldsymbol{X}_{opt}$ refers to the maximum value of the fitness function. However, the maximization of the fitness function can be challenging if the function is not differentiable at all points. In such cases, possible candidate solutions are iteratively investigated by evaluating their fitness value until the maximum solution is found or stopping criterion is met. In an effort to find the best solution, PSO uses the following algorithm (Weerasinghe, Chi, & Cao, 2016):

1.) Randomly assign the positions and velocities of the N particles in the search space $\Re^D$.

Start the loop

2.) Calculate the fitness function value for each particle.

3.) If the current fitness function value of the particle (X_i^t) is better than its previous personal best fitness value (P_i^{t-1}) then set the new position as the personal best i.e.

$$P_i^t = \begin{cases} P_i^{t-1}, \; if \; f(X_i^t) < f(P_i^{t-1}) \\ X_i^t \quad , \; iff(X_i^t) \geq f(P_i^{t-1}) \end{cases} \tag{16}$$

4.) Find the best performing particle in the swarm in terms of the fitness value and set its personal best as a global best.

Select the P_s^t such that $f(P_i^t) < f(P_s^t)$ for $s \; and \; i \in \{1, 2, ..., \mathrm{N}\}$

$$f(G^t) = \begin{cases} G^t, & if \; f(P_s^t) < f(G^t) \\ P_s^t, & if \; f(P_s^t) \geq f(G^t) \end{cases} \tag{17}$$

5.) Update the particle velocity based on Equation 1:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (P_i^t - X_i^t) + c_2 r_2 (G^t - X_i^t).$$

In order for the particle to be kept in the search space, make the following modification:

$$V_i^{t+1} \begin{cases} V_{min}, & if \;\; V_i^{t+1} < \; V_{min}. \\ V_i^{t+1}, & if \;\; V_{min} < V_i^{t+1} < V_{max}. \\ V_{max}, & if \;\; V_{max} < \; V_i^{t+1}. \end{cases} \tag{18}$$

6.) Update the particle position based on Equation 2:

$$X_i^{t+1} = X_i^t + V_i^{t+1}.$$

End of the loop

7.) Exit if the termination criterion is met.

4.3.3 K-fold Cross-validation

In data mining literature, a commonly identified problem is overfitting of machine learning algorithms (Hawkins, 2003). Therefore, it is important to validate the algorithm on an unseen dataset. Finding an unseen dataset, however, might be challenging in terms of budget and time. In such situations, one solution is to randomly separate the dataset into two subsets so one can be used as a training set and another can be used as a test set (i.e. unseen dataset). After splitting the dataset, the machine learning algorithm can be fitted on the training set and then applied on the test set. However, a random separation of the dataset can introduce bias into the model's performance evaluation (Kohavi, 1995). To mitigate this problem, researchers commonly use the *k*-fold cross-validation technique in which the whole dataset is randomly split into *k* approximately equal subsets (folds). The algorithm is then trained on all folds except one, which will be used as the test set (Kohavi, 1995). This process is repeated *k* times to exhaust all the folds (Figure 4.2).

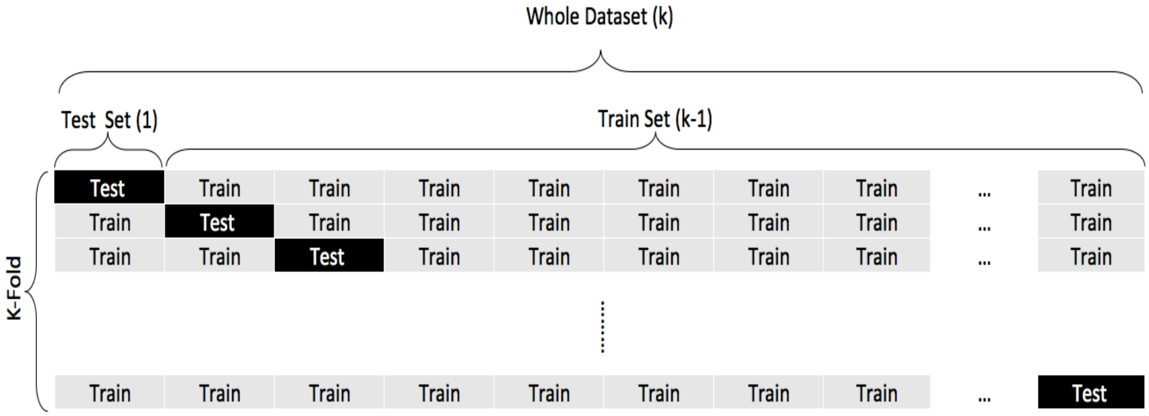


Figure 4.2: *K*-fold Cross-validation

The cross-validation estimate of the overall performance of the algorithm can be obtained by simply taking the average of the *k* individual performance, such that:

$$CVP = \frac{1}{k}\sum_{i=1}^{k} \mathcal{F}_i \tag{19}$$

where CVP is the *cross-validation performance*, *k* is the number of folds and $\mathcal{F}_i$ stands for the performance measure of each fold. In this study, we use 10-fold cross-validation (i.e. *k*=10) since it provides a good trade-off between the time required to train models, and minimization of the bias and variance associated with the validation process (Kohavi, 1995; Olson & Delen, 2008).

4.3.4 Data Balancing Techniques

It is common to come across an imbalanced dataset in many real-world applications such as health informatics (Dag et al., 2017), fraud detection (Fawcett & Provost, 1997), manufacturing process (Riddle et al., 1994), etc. An imbalanced dataset is a dataset where one or more of the response classes are less represented (i.e. minority) than the other class(es), which in turn can raise difficulties for the machine learning algorithms in the optimization process (Krawczyk, 2016). To cope with this imbalance, many approaches have been proposed in the literature (Chawla, 2005; Guo et al., 2008; Ling & Sheng, 2009). We have used RUS and SMOTE in this analysis. RUS is a simple technique that reduces the majority class by randomly eliminating several instances until having an approximately equal number of instances with the minority class. SMOTE, on the other hand, seeks to increase the minority class by creating synthetic instances (Chawla et al., 2002). The synthetics instances are created as follows:

- Select a random sample x_i from the dataset
- Find its k-nearest neighbors in the feature space and randomly select one of them, say, x_j
- Calculate the Euclidian difference between x_i and x_j then multiply by uniform random number range between 0 and 1
- Add this difference to the x_i to create a new synthetic instance x along the line segment connecting x_i and x_j.

4.3.5 Tree Augmented Naïve Bayesian Belief Network

A Bayesian Belief network is a directed acyclic graph that attempts to encode a joint distribution over a random vector $\boldsymbol{X} = \{X_1, \ldots, X_k\}$ and it consists of two main components $\mathcal{B} =< \mathfrak{H}, \Theta >$ (Pearl & Judea, 1997). The component $\mathfrak{H}$ forms the structure of the Bayesian Belief Network where each node of $\mathfrak{H}$ represents one of the random variables in $\boldsymbol{X}$ and the arcs connecting the nodes indicate a probabilistic dependency among the variables. Each variable is independent of its non-descendant given that its parent is $\mathfrak{H}$. The other component, Θ, is a parameter vector which calibrates the network. Given that the set of parents of X_i is $P_\gamma(X_i)$, the joint probability distribution defined by the Bayesian Belief Network $\mathcal{B}$ over $\boldsymbol{X}$ is:

$$P(\boldsymbol{X}) = \prod_{j=1}^{k} P\left(X_i \middle| P_\gamma(X_i)\right), \tag{20}$$

where k is the number of variables.

The simplest form of the BBN is the Naïve Bayes (NB) which does not require any structure to be learned and assumes, given the target variable, all variables to be independent. Namely, the nodes are disconnected and have no parents but the target node (Figure 4.3 (a)). However, the assumption of independence of the NB is not realistic and rarely holds true.

Therefore, in order to circumvent the independence assumption of NB, Friedman *et al.* developed a new algorithm called Tree Augmented Naïve Bayes (TAN) (N. Friedman, Geiger, & Goldszmidt, 1997). Unlike Naïve Bayes, TAN allows each variable to be dependent on, at most, one non-target variable in addition to the target variable (Figure 4.3 (b)). In TAN, the target node is connected to all predictor nodes through arcs i.e. the target node is the parent of all variables. Also, arcs connect the predictor nodes to the other predictor nodes. These connections show the dependency between the predictor nodes in that the node having the starting point of the arc is the parent node of the pointed node, and the contribution of the child node in predicting the target variable is dependent on its parent node.

Formally, the Bayesian Belief Network $\mathcal{B}$ is considered as a TAN network if there exists a tree function δ over **X** where the following equation holds, given that C is the target variable and $P_\gamma(C) = \emptyset$:

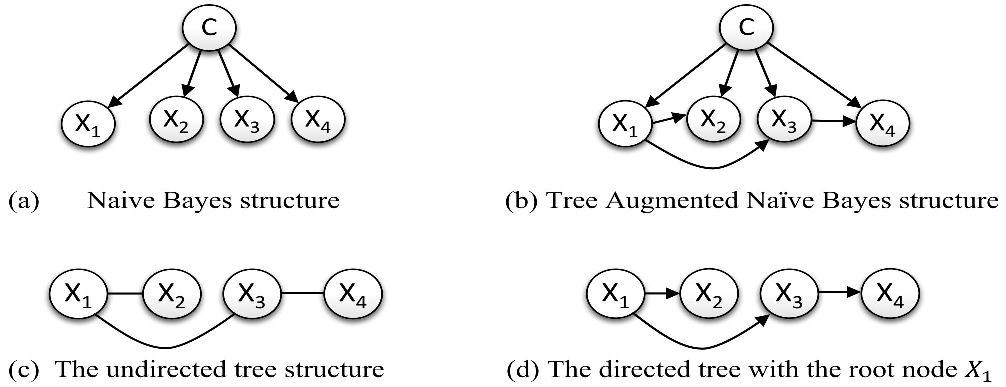
$$P_\gamma(X_i)=\begin{cases} \{\mathrm{C}, X_{\delta(i)}\}, & if\ \delta(i) > 0. \\ \{C\}, & if\ \delta(i) = 0. \end{cases} \tag{21}$$

Finding the optimal tree configuration which has maximum spanning weight is an optimization problem where the log likelihood of the δ function needs to be maximized. The TAN network algorithm uses the general outline of Chow-Liu's algorithm for one-dependence estimator (Chow & Liu, 1968) to maximize the δ function and thus follows the following steps (N. Friedman et al., 1997):

1. Calculate $I_{\hat{P}_D}(X_i; X_j|C)$ between each pair of variables, i≠j where

   $$I_P(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) = \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} P(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \log \frac{P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}{P(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{y}|\boldsymbol{z})} \tag{22}$$

   is the conditional mutual information and $D = \{C, X_1, \dots, X_k\}$ is a training dataset.
2. Construct a complete undirected graph denoting variable $X_1, \dots, X_k$ by the nodes. Use the results obtained through conditional mutual information function in order to place the weights of arcs connecting $X_i\ to\ X_j$.
3. Construct a maximum weighted spanning tree as in Figure 4.3 (c).
4. Build a directed graph by transforming the undirected tree structure. To do so, choose a root variable from the undirected tree and set the direction of all arcs to be outward from it, as shown in Figure 4.3 (d)
5. Build a TAN model by adding a node labeled C and adding an arc from C to each X_i



(a) Naive Bayes structure

(b) Tree Augmented Naïve Bayes structure

(c) The undirected tree structure

(d) The directed tree with the root node X_1

Figure 4.3: An augmented naïve Bayesian model.

4.3.5.1 Parameter Tuning of the TAN Model

As discussed in the above section, there are two steps to construct the Tree Augmented Naïve Bayesian network. In the first step, the structure learning algorithm attempts to find an optimal tree configuration by maximizing the log-likelihood (in Equation 22) under the condition that *n–1* edges need to be constructed. However, this condition can lead to overfitting the model since some edges can be unnecessarily included (N. Friedman et al., 1997). Therefore, to avoid overfitting, a penalization term can be involved in the maximization process of the likelihood (Equation 22).

In this study, the penalty terms *Akaike's information criterion* (AIC) (Akaike, 1998) and *Bayesian Information Criterion* (BIC) (Wit et al., 2012) are used to find the optimal tree structure. It should be noted that maximizing penalized log-likelihood may result in a Forest Augmented Naïve Bayesian (FAN) model instead of a TAN model (N. Friedman et al., 1997). That is, the algorithm may choose to ignore some of the edges between the nodes, which in turn, destroys the structure of TAN by having less than *n–1* arcs. After obtaining the tree configuration, the second step is to learn the parameters of the tree structure from the data. Sometimes, since the variables are conditioned on one another, TAN can suffer from lack of instances and hence make an unreliable estimate. To mitigate this difficulty the parameter estimation process is biased by adding the smoothing parameter (Laplace correction) (Kohavi & Kohavi, 1996). The smoothing parameter can control the effect of lack of instances by fixing zero probabilities. Note that introducing a smoothing parameter does not change the existing tree structure. Namely, the tree structure remains untouched, while the weights of the arcs are tuned. In this study, we cross validate both the learning process of the structure and the parameters to find an optimal model. The data specific behavior of the ROC (AUC) metric with the different smoothing parameters and the score functions is illustrated in Figure 4.4.

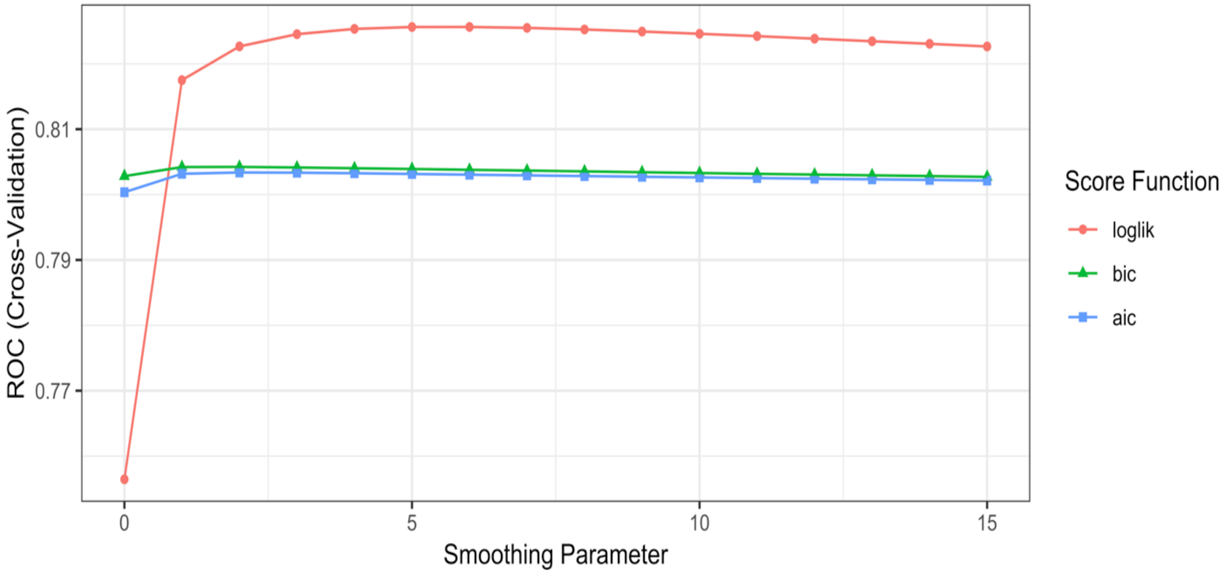


Figure 4.4: Parameter tuning of a TAN model

4.3.6 Performance Metrics

The model performances are compared by using 4 different evaluation metrics: accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Detailed information about the performance evaluation metrics can be found in (Ling et al., 2003).

4.4 Results and Discussion

4.4.1 Variable Selection Results

To better interpret the causal relationship among the dependent variables, we seek to reduce the data dimensionality by employing three different variable selection methods explained in Section 4.3.2.1, 4.3.2.2 and 4.3.2.3. Variable selection methods, however, use their own way of judging the importance of the variables and thus have different shortcomings (Hall, 1999). For example, correlation-based feature selection techniques are sensitive to multicollinearity and might fail to detect all the relevant variables (Hall, 1999). Similarly, exhaustive variable selection techniques can be computationally expensive and may not terminate in a realistic time. Time-efficient algorithms, on the other hand, cannot guarantee the optimal subset solution (Dash et al., 2000). For this reason, we consider all possible combinations of the three variable sets selected by the previously mentioned machine learning and optimization algorithms to reduce the bias associated with each algorithm. The variables selected through these variable selection methods are shown in Table 4.2.

Table 4.2: Variable selection results

Variable Name	PSO	GA	XGB
Age	✓	✓	✓
Gender		✓	✓
Scheduling day			✓
Appointment day	✓	✓	✓
Appointment month			✓
Lead time	✓	✓	✓
Calling time	✓	✓	✓
SMS reminder	✓	✓	✓
Alcoholism	✓		
Financial aid	✓		
Handicap			
Hypertension			
Diabetes			
Neighborhood		✓	✓
Time between appointments	✓	✓	✓
Prior no-shows	✓	✓	✓

The variables; *Age, Appointment day Lead time, Calling time, Prior no-show, Time between appointments,* and *SMS reminder* are selected by all the algorithms while only two algorithms find *Neighborhood and Gender* be important. Moreover, *Alcoholism* and *Financial aid* are chosen only by particle swarm optimization. Similarly, *Scheduled day* and *Appointment Month* are regarded as important variables only by XGB. It should be noted that none of the algorithms consider *Diabetes, Hypertension, Handicap* as important variables.

4.4.2 Creating Predictor Sets and Classification Results

Three different variable sets have been extracted from the dataset through a machine learning algorithm (i.e. XGB) and the two optimization techniques (i.e. PSO and GA). In our preliminary analysis phase, we considered all possible combinations of these variable sets (i.e. intersections and unions of the all variable sets) to find the optimal set that results in maximizing the TAN models predictive performance with a minimum number of variables. The number of possible scenarios/combinations is 11 (i.e. GA, PSO, XGB, GA ∩ PSO, GA ∩ XGB, PSO ∩ XGB, GA ∩ PSO ∩ XGB, GA ∪ PSO, GA ∪ XGB, PSO ∪ XGB, and GA ∪ PSO ∪ XGB). To exemplify, GA ∩ PSO ∩ XGB represents the set of variables that are common in all of these three models employed, while GA ∪ PSO ∪ XGB represents the union of the three sets variables selected through GA, PSO, and XGB. Recall that we employed two algorithms to handle the data imbalance (i.e. RUS and SMOTE). Table 4.3 originally had 11 scenarios for each balancing algorithm. However, four of these 11 scenarios produced the same set of variables (For example, the intersection of GA and XGB (i.e. GA ∩ XGB) was exactly the same with GA (only)). Therefore, those scenarios were eliminated for the sake of simplicity, which left us with seven unique scenarios.

As can be seen in Table 4.3, the best performing scenario is bolded for each metric. For example, the model with variable sets given by the intersection set of GA and PSO (GA∩PSO) achieved the best performance in terms of AUC (0.828) on the test set balanced through RUS, while the best sensitivity (0.785) was again achieved with the intersection set of the GA and PSO using SMOTE. Moreover, the best specificity (0.752) is obtained with the intersection set of all algorithms (GA∩PSO∩XGB) with RUS, GA led to reach the best accuracy (0.742) using the RUS balancing technique. Even though the accuracy number obtained through GA ∪ PSO was the same, this scenario had a higher number of variables when compared to GA only (i.e.11 > 9). It should be noted that our goal is to maximize the performance while minimizing the number of the variables to be selected (to obtain the most parsimonious model possible).

The best performing variable set was selected to further investigate the causal relationships among the variables. To do so, the AUC metric is considered as the main criterion since it is not sensitive to a cutoff value, and it is a better way to judge a model performance when the dataset is imbalanced (Ling et al., 2003). With that said, the variable set giving the best AUC has fewer variables compare to the other variable sets, which also appeals to us since one of our major goals is to obtain the most parsimonious model possible. Therefore, we chose to investigate the variable set of GA ∩ PSO when sampled with RUS, as underlined in Table 4.3.

Table 4.3: Model performance with different scenarios and sampling techniques

Sampling technic: Random Under-sampling (RUS)

Variable Set	Number of Variable	Accuracy	Sensitivity	Specificity	AUC
GA	9	**0.742(0.003)**	0.746(0.011)	0.742(0.005)	0.822(0.006)
PSO	9	0.729(0.004)	0.783(0.012)	0.715(0.006)	0.827 (0.006)
XGB	11	0.718(0.005)	0.739(0.010)	0.712(0.007)	0.809(0.004)
GA∪PSO	11	**0.742(0.003)**	0.744(0.013)	0.741(0.005)	0.821(0.006)
GA∩PSO	7	0.728(0.004)	**0.785(0.012)**	0.713(0.006)	**0.828(0.006)**
PSO∪XGB	13	0.717(0.003)	0.741(0.010)	0.711(0.006)	0.809(0.004)
GA∩PSO∩XGB	7	0.737(0.007)	0.682(0.018)	**0.752(0.013)**	0.796(0.005)

Sampling technic: Synthetic Minority Over-sampling (SMOTE)

Variable Set	Number of Variable	Accuracy	Sensitivity	Specificity	AUC
GA	9	0.696(0.006)	0.748(0.022)	0.682(0.011)	0.791(0.007)
PSO	9	**0.726(0.006)**	0.709(0.022)	**0.731(0.011)**	0.8(0.007)
XGB	11	0.683(0.008)	0.734(0.014)	0.669(0.010)	0.77(0.007)
GA∪PSO	11	0.719(0.004)	0.712(0.016)	0.721(0.005)	0.795(0.007)
GA∩PSO	7	0.703(0.007)	0.759(0.023)	0.689(0.014)	**0.803(0.007)**
PSO∪XGB	13	0.708(0.005)	0.685(0.013)	0.714(0.008)	0.768(0.006)
GA∩PSO∩XGB	7	0.669(0.005)	**0.767(0.006)**	0.643(0.007)	0.785(0.005)

4.4.3 Bayesian Network Results

The Bayesian belief network structure of the best performing model (GA ∩ PSO) is illustrated in Figure 4.5 with seven independent variables and the outcome (i.e. whether the patient is no show or not). As explained in Section 4.3.5, the Tree Augmented Naïve Bayesian (TAN) algorithm requires variables to have parent-child relations. Understanding the inter-relations among the predictors and, the association of each predictor with the outcome can be observed from the TAN model. An arc from a predictor (parent) to another (child) indicates that the contribution of the child node on the model outcome depends on the value of the parent node.

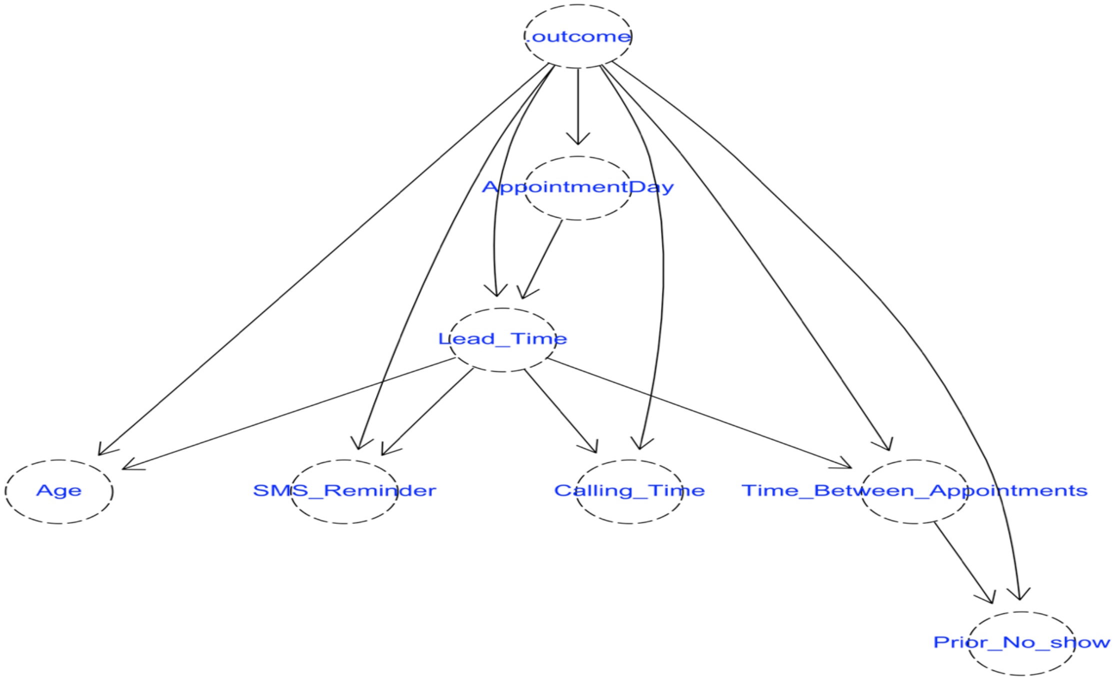


Figure 4.5: Bayesian belief network

Figure 4.5 demonstrates that the most influential variable is *Lead time* since it is the parent of four variables (*Age, SMS remainder, Calling time* and *Time between appointments*), and the impacts of all these variables on the target variable are directly associated with the value of the *Lead time*. For example, the *age* seems to be an important factor for the outcome. However, the effect of *age* might decrease with the value of *lead time*. For example, if the lead time is too high, even if the *age* of the patient is between 50 and 75 (the lowest risk *age* frame), there is a high chance that this patient might be a *no-show* patient. This suggests that clinics need to be careful about the lead time when scheduling an appointment for a patient in order to increase the show-up rate. With that said, the effect of *Lead time* depends on the *Appointment Day*, which can be an indication of a general tendency among patients towards specific appointment days. Namely, some days can be more popular and more attractive to the patients in order to schedule an appointment, which in turn might increase the lead time (Y.-L. Huang & Hanauer, 2016; Norris et al., 2014).

Additionally, the individual effects of the variables can be investigated. For instance, age-specific show-up rate demonstrates that the no-show rate steadily increases from an early age to around 18, as seen in Figure 4.6. The fact that the likelihood of the *show up* decreases with child's age has been supported by many studies in the literature (Barron, 1980; Y. Huang & Hanauer, 2014; McLeod et al., 2015; Pang et al., 1995). Moreover, the *show up* rate hits the lowest rate of attendance at ages between 18 and 25. One possible reason why young adults miss their appointment is that they are less likely to understand the purpose of the appointments (Frankel et al., 1989).

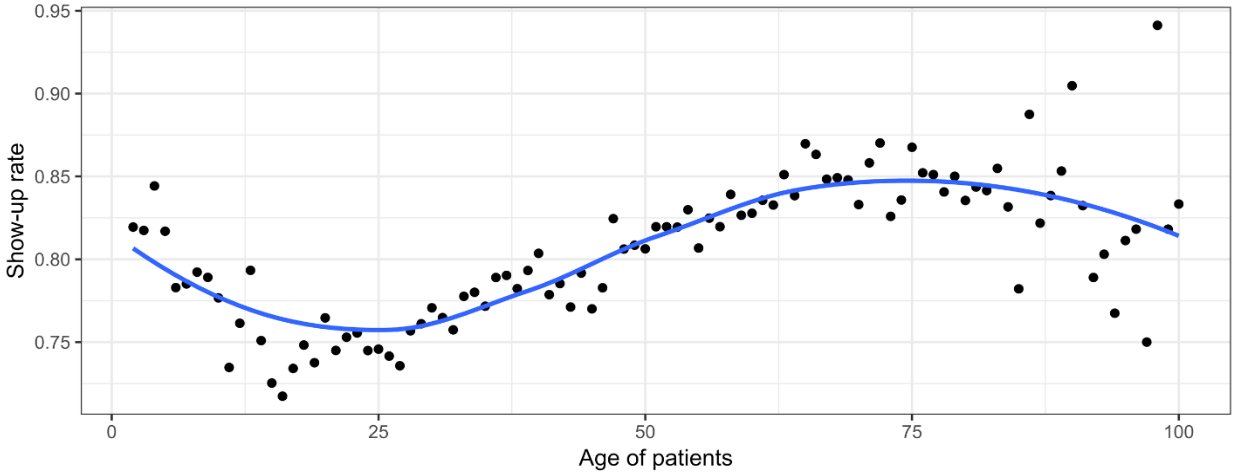


Figure 4.6 The show-up probability of patients by age

The show-up trend starts going up consistently until the age of 75 after which the trend gets nosier by experiencing some sudden decreases. The reasons for this decrease for patients older than 75 can be due to transportation difficulties (Bean & Talaga, 1992; Frankel et al., 1989; Pang et al., 1995), hospitalization (Kaplan-Lewis & Percac-Lima, 2013) or deteriorated health conditions (Cosgrove, 1990). However, it should be noted that the effect of age should be considered along with the *lead time* according to the BNN network.

Similarly, *calling time* is an important factor whose effect on the outcome also depends on the *lead time*. This might happen since the patients calling early in the morning may be experiencing an unexpected health problem, and thus will attend the appointment regardless of any other factors. However, if the *lead time* is too long for an early-calling patient, it potentially indicates a non-emergency situation and so it might not decrease the *no-show* risk.

As discussed in the earlier sections, the variable *SMS received* has an impact on the outcome, which again depends on its parent node, *Lead time*. Therefore, one can say that sending a text reminder might become even more important when the *lead time* is too long (Barron, 1980; Bean & Talaga, 1992; Hardy et al., 2001; O'Brien & Lazebnik, 1998; Smith & Yawn, 1994). Or adversely, *SMS reminder* might have less impact on the outcome if the *lead time* is too low. The rationale behind it is that *SMS reminder* might become unnecessary (obsolete) for very short lead times. Having said that, Figure 4.5 indicates that appointment day has a strong effect on the model outcome i.e. some days can be challenging or easy for the patients to make it to the appointment.

Prior no-shows is another variable that has an effect on the model outcome depending on *the time between appointments* (i.e. the elapsed time between consecutive appointments). It has been shown that prior no-show history is one of the most important predictors for predicting the no-show patients (Collins et al., 2003; Cronin et al., 2013; Y.-L. Huang & Hanauer, 2016; Kempny et al., 2016; Torres et al., 2015). However, counter-arguments have also been made (Dantas et al., 2018). The reason behind this can be the failure of ignoring other factors i.e. the relation of the prior no-show history with other indicators. With that said, the BBN suggest that the impact of *prior no-show* is effected by the elapsed time between consecutive appointments. The rationale behind this might be; if the time between the most recent appointments is too short, this might indicate that the health issue that the patient might not be a minor problem. Therefore, the patient had to go to the clinic again for follow-up. In such situation, if there is some prior show in the patient's history, there is a high chance that the patient will show up because of the seriousness of the issue that she/he has.

4.4.4 Evaluation of the Predictive Power of the TAN model

Monte Carlo Cross Validation (MCCV) technique is used to evaluate the ability of the model prediction and its robustness. Namely, the dataset was split into two parts: one part is used to train model, and the other part is left outside of the training process to validate the model performance. This process repeated 1000 times. Each model was fitted on the training dataset, and predictions were obtained using the test set. The predictive power of the model is then summarized by the aforementioned performance metrics: the area under the curve, accuracy, sensitivity, and specificity. In addition to these metrics, another metric called Brier's score is used to validate the probabilistic accuracy of the predictive model. Brier score reflects how close the model probabilistic forecast is to the truth. The mathematical definition of the Brier score can be defined as follow

$$\beta = \frac{1}{N}\sum_{j=1}^{N}(f_j - \partial_j)^2$$

where f_j is the probabilistic prediction and ∂_j is the actual outcome for the j-th instance, and N is the number of instances. Simply, the metric calculates the square of the difference between the assigned probability for each patient and his/her actual status; therefore, the lower Brier's score refers to the better probabilistic accuracy.

```
for j = $1, 2 \dots, 1000$ do
    threshold$_1$ = cutoff + i/100
    threshold$_2$ = cutoff – i/100, where i ∈ $\{0, 1 \dots, 50\}$ and cutoff = 0.5
   for k = $1, 2 \dots, N$, where N is the number of patients in the test set do
        if prediction $_k$ ≥ threshold$_1$ then
            classify the patient as 1(no-show)
        if else prediction $_k$ ≤ threshold$_2$ then
           classify the patient as 0 (show-up)
        else threshold$_1$ ≤ prediction $_k$ ≤ threshold$_2$ then
        do not make any classification drop the instance.
   end for
  accuracy$_{[j]}$ = calculate accuracy
  sensitivity$_{[j]}$ = calculate sensitivity
  specificity$_{[j]}$ = calculate specificity
  ROC$_{[j]}$ = calculate ROC
  Brier's score$_{[j]}$ = calculate Brier's score
end for
```

Figure 4.7: The pseudo code

The model fitting and validation process, repeated 1000 times, are done for different cutoff points to better evaluate the consistency and reliability of the fitted model. Namely, the model performance investigated when the probabilistic prediction of the model for each instance goes away from the default cutoff value 0.5 with the expectation of getting better results. For example, for a prediction probability of 0.49, the patient would be classified as show-up while the same patient would be classified as no-show when the predicted probability is 0.51. In this study, in addition to the regular model performance investigation—where the cutoff value is generally set to 0.5, we explore the model performance when the model is not allowed to make a prediction between specified cutoff values. For example, the lower bound cutoff value is set at most to be 0.4 and the upper bound cutoff value is set at least to 0.6, then the model simply is prevented from making a decision for patients for whom the assigned probabilities are between the lower (0.4) and upper (0.6) cutoff values. These patients are not classified as either "no-show" or "show-up", and they are dropped from the confusion matrix—where we obtain the evaluation metrics: accuracy, sensitivity, specificity, and ROC. Therefore, the model performance is summarized by the evaluation metrics without taking those patients into account. In doing so, the model consistency and reliability is well studied for the expansive cutoff values. For each cutoff value, the TAN model is validated 1000 times. In other words, 50x1000 = 50,000 models are fitted and tested. The behavior of the five metrics is shown in Figure 4.8, and the pseudo code for this procedure is provided in Figure 4.7.

Figure 4.8 illustrates the mean behavior of each of the metrics as a line, and the shaded areas around the lines are the one standard deviation intervals. It can be observed from Figure 4.8 that when the gap between the lower and the upper bound cutoff values are expanded, the performance of the model improves steadily. In other words, when the lower cutoff value is pushed towards 0 and at the same time the upper cutoff value is pushed towards 1, the probabilistic predictions of the model get more precise. This suggests that the model credibility increases when a prediction for the patient is close to the endpoints. Moreover, the marginal distributions of the four cutoff values are investigated. The cutoff values are chosen to be both lower and upper 0.5, lower 0.3 and upper 0.7, lower 0.1 and upper 0.9, and lower 0.05 and upper 0.95. The distributions of the evaluation metrics for these cutoff points are illustrated in Figure 4.10.

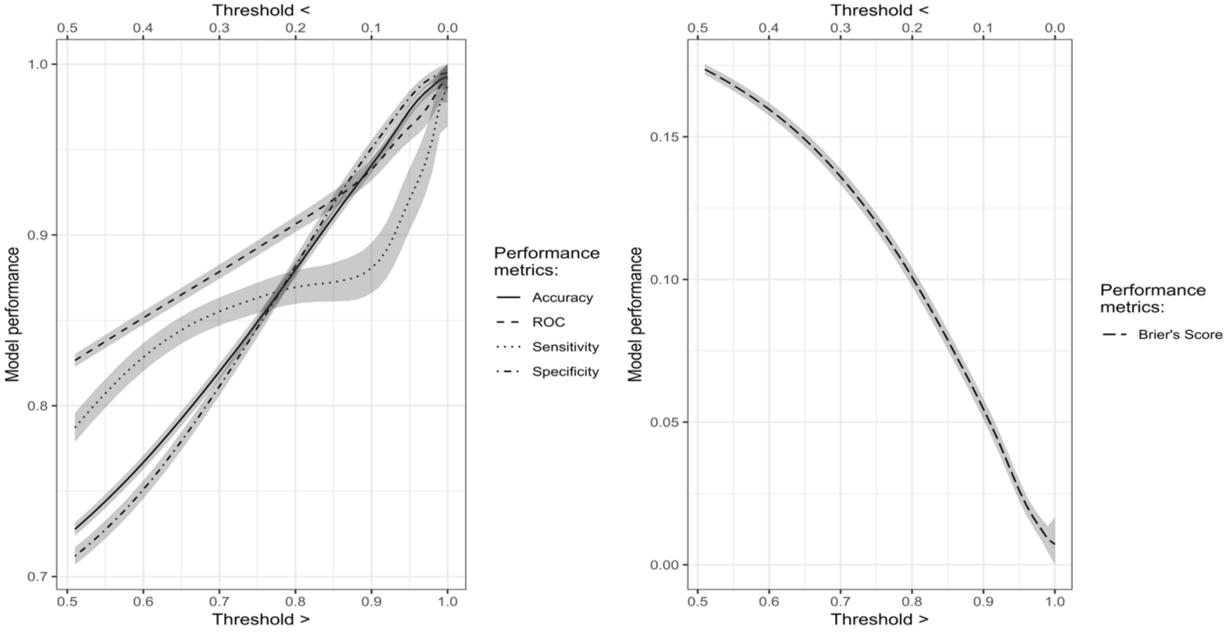


Figure 4.8: Behavior of the evaluation metrics in 1000 validation samples

Figure 4.9 suggests that the distributions of the evaluation metrics are approximately normal for each of the chosen cutoff points. Therefore, in Table 4.4, the mean value of each of the metrics is shown as a point estimate.

Table 4.4: Model results with different cutoffs

Cutoff	ROC	Accuracy	Sensitivity	Specificity	Brier's Score
0.5	.827	.728	.787	.712	.174
0.3 < | > 0.7	.881	.826	.857	.818	.133
0.1 < | > 0.9	.943	.947	.885	.957	.049
0.05 < | > 0.95	.968	.979	.930	.985	.021

When the cutoff value is 0.5, the mean ROC curve area is .827. For the second cutoff values (0.3 < | > 0.7), the mean ROC curve area becomes .881, where any probabilistic prediction between 0.7 and 0.3 is exempted from the model evaluation process. Similarly, for the third cutoff values (0.1 < | > 0.9), the mean ROC curve area becomes .943. And lastly, for the fourth cutoff values (0.05 < | > 0.95), the mean ROC curve area hits as high as .968.

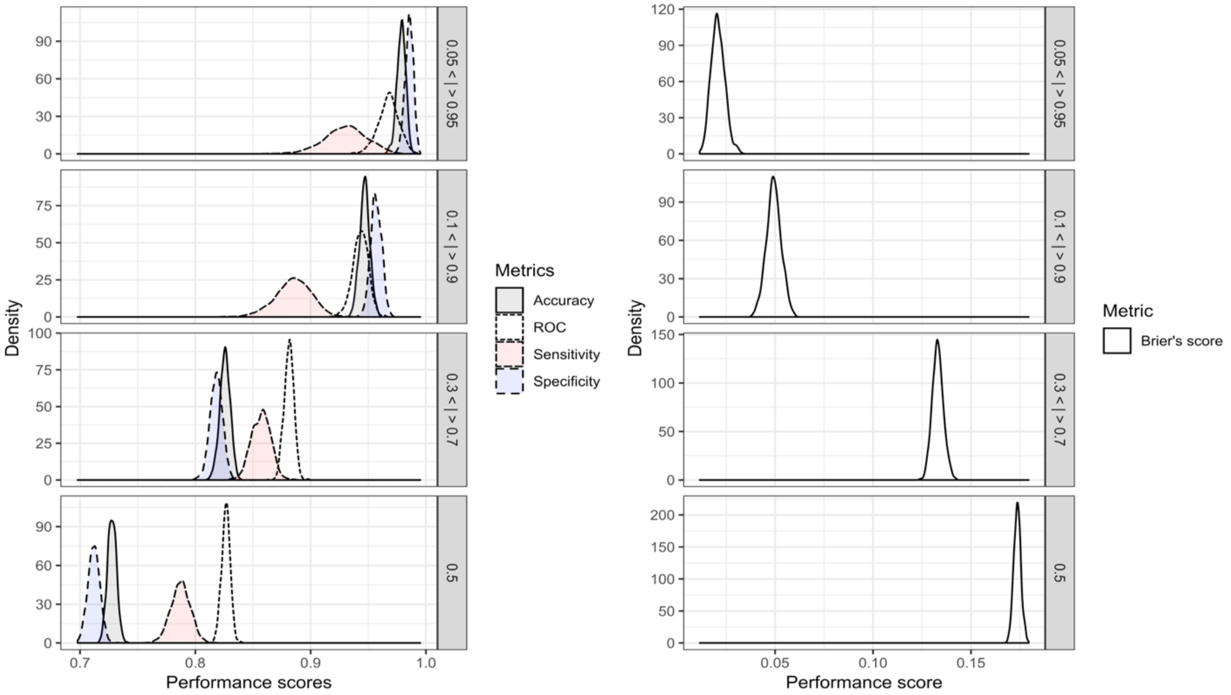
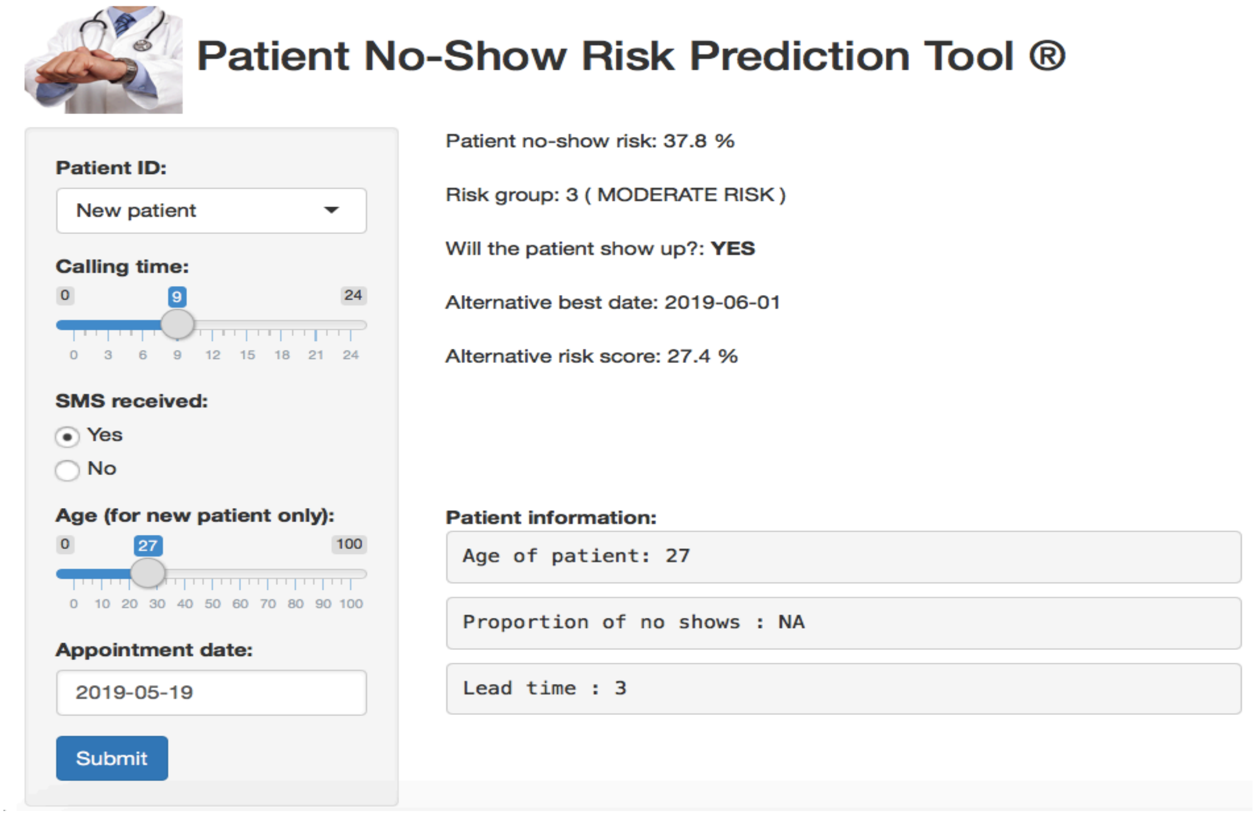


Figure 4.9: Distribution of the metrics for certain cutoff points in 1000 validation samples

The same consistent patterns are observed for the other evaluation metrics: accuracy, sensitivity, and specificity. On the other hand, the mean value of the Brier's score decreases from .174 to .021 when the cutoff value is changed from 0.5 to 0.05 < | > 0.95, which demonstrates the assigned probabilities for each patient merely deviate from the truth; therefore, again the probabilistic prediction of the model can be trusted especially if the assigned probability is close zero or one.

4.5 Patient No-Show Risk Prediction Tool

We have developed an interactive prediction tool that can be used by clinics to cope well with no-show patients (a screenshot of the tool is shown in Figure 4.10). The tool requires the appointment setter (clerk) to enter the necessary information of the patient who will be scheduled an appointment. The tool then not only provides patient-specific information such as age, the proportion of prior no-shows, and the lead time but also provides a patient-specific risk score of being no-show. It also suggests an alternative appointment date with an updated risk score, which



Patient No-Show Risk Prediction Tool ®

Patient ID: New patient

Calling time: 9

SMS received:
● Yes
○ No

Age (for new patient only): 27

Appointment date: 2019-05-19

Patient no-show risk: 37.8 %

Risk group: 3 (MODERATE RISK)

Will the patient show up?: **YES**

Alternative best date: 2019-06-01

Alternative risk score: 27.4 %

Patient information:

```
Age of patient: 27
```

```
Proportion of no shows : NA
```

```
Lead time : 3
```

Figure 4.10: The interface of the risk prediction tool

is less than the appointment date that was offered by the patient herself/himself. In order to provide the clerk with an alternative appointment date and its risk score, the tool iteratively searches the best possible neighbor appointment days, on which the patient is most likely to come, (i.e., the risk score for the patient hits its lowest value). To exemplify, the possible neighbor appointment days found by the tool is illustrated in Figure 4.11 for five patients selected at random

The tool has been developed in R using the shiny package, and it can be utilized through computers (with any operating systems such as Microsoft Windows, Linux, Mac OS, etc.) and any smartphone.

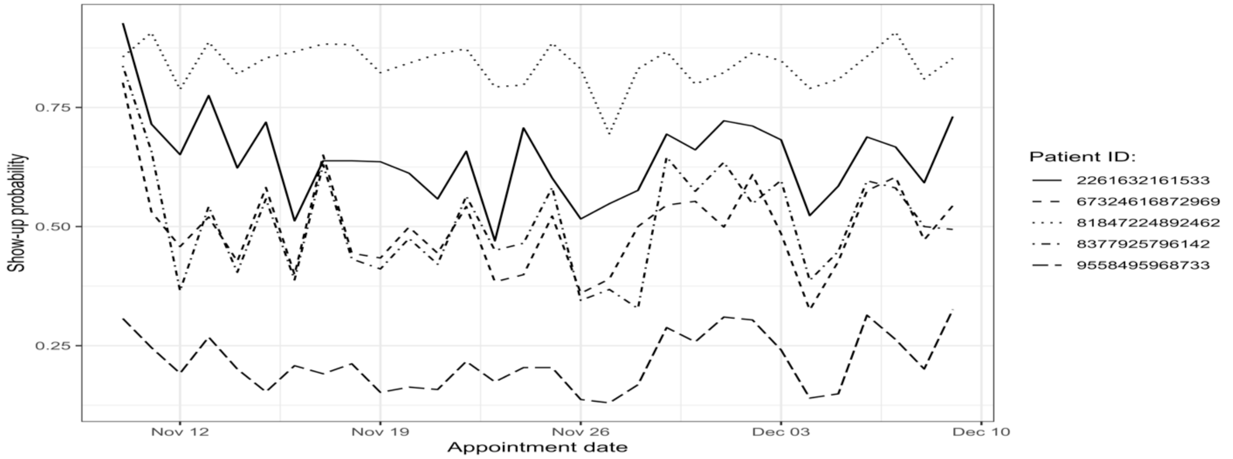


Figure 4.11: Patient-specific show-up probability structure

4.6 Conclusions

Given that no-shows present a big problem for medical practitioners, wasting resources and creating an undue burden on the system, understanding the reasons behind no-shows and predicting patients likely to be no-shows is a valuable problem that has been covered in the related literature both in terms of descriptive and predictive analytical studies. The main targets of this study were to create a parsimonious Bayesian belief network model to identify the likelihood of patients to be no-shows and to find inter-relations among the explanatory variables.

The first contribution of the study was the use of machine learning techniques that could potentially find the most parsimonious models by eliminating irrelevant or redundant features and in the process increasing the accuracy of the resulting predictions. The second contribution of this analysis was a data-driven exploration of the relationships and interactions between the selected features and how these could be used to understand patient behavior and design better problem mitigation and decision support solutions. To have a parsimonious model, three machine learning algorithms (GA, PSO, and XGB) were used to perform variable selection. Our approach was applied to the dataset acquired from the kaggle website. Since the dataset was imbalanced, two sampling techniques; random under-sampling (RUS) and synthetic minority over-sampling (SMOTE) were used while the model building process along with 10-fold cross-validation. We emphasized the following questions:

1) What are the most important variables in detecting no-show patients?

2) Can the interactions among the variables be visualized and analyzed?

3) How can clinics reduce their cost by using data-driven methods?

The proposed framework results show that performing comprehensive variable selection can lead to a better performing predictive model with fewer variables. Furthermore, the information obtained using a probabilistic approach in the final step has provided relations among these important predictors as well as the patient-specific no-show probabilities. These predictions and insights can be useful to clinic managers, decision makers, and policy makers since they provide a picture of their patient's expected behavior, which should enable them to improve their clinical utilization, and in the process, improve patient outcomes.

5 Conclusion

The statistical modeling methodology and the tool introduced in this dissertation play the critical roles in filling the gaps in the relevant literature of healthcare. In the first study, we proposed a novel statistical modeling approach that is used to model the survival time of breast cancer patients. Therefore, the contribution of the first study to the breast cancer literature can be summarized as follows: 1) we investigated breast cancer patient survivability from three-different-year perspective, thereby revealing the factors that contribute the most to the prediction of survival of breast cancer patients in various time intervals, and 2) we explored how the importance of those factors change over time. As this was achieved by employing the most parsimonious models, these strategies have a high potential to have much broader implications for cancer treatment and research. In the second study, the similar modeling approach was formulated and applied to the no-show patient dataset, where the study contributes the extant patient no-show prediction literature by 1) finding the most parsimonious model, 2) exploring the conditional relation among the predictors, and 3) developing a prediction tool providing an individual patient no-show risk score and an alternative appointment day on which the patient is most likely to come. The main goal of the study is to reduce the costs of clinics introduced by patient no-shows and to provide patients better care by utilizing available resources more efficiently.

References

A. Tibble, J., Forgacs, I., Bjarnason, I., & Przemioslo, R. (2000). The Effects of a Preassessment Clinic on Nonattendance Rates for Day-Case Colonoscopy. *Endoscopy*, *32*(12), 963–965. https://doi.org/10.1055/s-2000-9629

Aarts, E., & Korst, J. (1988, January). Simulated Annealing and Boltzmann Machines. New York, NY; John Wiley and Sons Inc.

Adami, H.-O., Malker, B., Holmberg, L., Persson, I., & Stone, B. (1986). The Relation between Survival and Age at Diagnosis in Breast Cancer. *New England Journal of Medicine*, *315*(9), 559–563. https://doi.org/10.1056/NEJM198608283150906

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle (pp. 199–213). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_15

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., … Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511. https://doi.org/10.1038/35000501

Allegra, C. J., Aberle, D. R., Ganschow, P., Hahn, S. M., Lee, C. N., Millon-Underwood, S., … Zon, R. (2010). National Institutes of Health State-of-the-Science Conference Statement: Diagnosis and Management of Ductal Carcinoma In Situ September 22-24, 2009. *JNCI Journal of the National Cancer Institute*, *102*(3), 161–169. https://doi.org/10.1093/jnci/djp485

Atun, R., & Sittampalam, S. (2006). The role of Mobile Phones in Increasing Accessibility and Efficiency in Healthcare. Retrieved August 6, 2018, from https://www.vodafone.com/content/dam/vodafone/about/public_policy/policy_papers/public_policy_series_4.pdf

Bafford, A. C., Burstein, H. J., Barkley, C. R., Smith, B. L., Lipsitz, S., Iglehart, J. D., … Golshan, M. (2009). Breast surgery in stage IV breast cancer: impact of staging and patient selection on overall survival. *Breast Cancer Research and Treatment*, *115*(1), 7–12. https://doi.org/10.1007/s10549-008-0101-7Bansal, G., Sinha, A. P., & Zhao, H. (2008). Tuning Data Mining Methods for Cost-Sensitive Regression: A Study in Loan Charge-Off Forecasting. *Journal of Management Information Systems*, *25*(3), 315–336. https://doi.org/10.2753/MIS0742-1222250309

Barron, W. M. (1980). Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care*, *7*(4), 563–574. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7010402

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20. https://doi.org/10.1145/1007730.1007735

Bean, A. G., & Talaga, J. (1992). Appointment breaking: causes and solutions. *Journal of Health Care Marketing*, *12*(4), 14–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10123581

Bech, M. (2005). The economics of non-attendance and the expected effect of charging a fine on non-attendees. *Health Policy*, *74*(2), 181–191. https://doi.org/10.1016/j.healthpol.2005.01.001

Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., … Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, *8*(8), 816–824. https://doi.org/10.1038/nm733

Berg, B. P., Murr, M., Chermak, D., Woodall, J., Pignone, M., Sandler, R. S., & Denton, B. T. (2013). Estimating the Cost of No-Shows and Evaluating the Effects of Mitigation Strategies. *Medical Decision Making*, *33*(8), 976–985. https://doi.org/10.1177/0272989X13478194

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., … Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, *5*(1), 10312. https://doi.org/10.1038/srep10312

Berry, D. A., Cirrincione, C., Henderson, I. C., Citron, M. L., Budman, D. R., Goldstein, L. J., … Winer, E. P. (2006). Estrogen-Receptor Status and Outcomes of Modern Chemotherapy for Patients With Node-Positive Breast Cancer. *JAMA*, *295*(14), 1658. https://doi.org/10.1001/jama.295.14.1658

Bhanu, B., & Lin, Y. (2003). Genetic algorithm based feature selection for target detection in SAR images. *Image and Vision Computing*, *21*(7), 591–608. https://doi.org/10.1016/S0262-8856(03)00057-X

Bindman, A. B., Grumbach, K., Osmond, D., Komaromy, M., Vranizan, K., Lurie, N., … Stewart, A. (1995). Preventable Hospitalizations and Access to Health Care. *JAMA: The Journal of the American Medical Association*, *274*(4), 305. https://doi.org/10.1001/jama.1995.03530040033037

Borovicka, T., Jirina, M., Kordik, P., & Jiri, M. (2012). Selecting Representative Data Sets. In *Advances in Data Mining Knowledge Discovery and Applications*. InTech. https://doi.org/10.5772/50787

Bradbury-Huang, H. (2010). What is good action research? *Action Research*, *8*(1), 93–109. https://doi.org/10.1177/1476750310362435

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2002). Manual On Setting Up, Using, And Understanding Random Forests V3.1. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.

Bundred, N. J. (2001). Prognostic and predictive factors in breast cancer. *Cancer Treatment Reviews*, *27*(3), 137–142. https://doi.org/10.1053/ctrv.2000.0207

Bunting, J. S., Hemsted, E. H., & Kremer, J. K. (1976). The pattern of spread and survival in 596 cases of breast cancer related to clinical staging and histological grade. *Clinical Radiology*, *27*(1), 9–15. https://doi.org/10.1016/S0009-9260(76)80004-9

Cang, S., & Yu, H. (2014). A combination selection algorithm on forecasting. *European Journal of Operational Research*, *234*(1), 127–139. https://doi.org/10.1016/j.ejor.2013.08.045

Carter, C. L., Allen, C., & Henson, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, *63*(1), 181–187. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2910416

Cetinic, E., Lipic, T., & Grgic, S. (2018). Fine-tuning Convolutional Neural Networks for fine art classification. *Expert Systems with Applications*, *114*, 107–118.

https://doi.org/10.1016/J.ESWA.2018.07.026

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 853–867). New York: Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_40

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. Retrieved from https://www.jair.org/media/953/live-953-2037-jair.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794). New York, New York, USA: ACM Press. https://doi.org/10.1145/2939672.2939785

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *14*(3), 462–467. https://doi.org/10.1109/TIT.1968.1054142

Churilov, L., Bagirov, A. M., Schwartz, D., Smith, K., & Dally, M. (2004). Improving risk grouping rules for prostate cancer patients with optimization. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (p. 9 pp.). IEEE. https://doi.org/10.1109/HICSS.2004.1265355

Clarke, M., Collins, R., Darby, S., Davies, C., Elphinstone, P., Evans, V., … Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (2005). Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *The Lancet*, *366*(9503), 2087–2106. https://doi.org/10.1016/S0140-6736(05)67887-7

Clemen, R. T. (1989). Combining forecasts: A review and annotated. *International Journal of Forecasting*, *5*, 559–583. Retrieved from https://pdfs.semanticscholar.org/7117/9279738b91df0520061b351cb3e0124a411c.pdf

Collins, J., Santamaria, N., & Clayton, L. (2003). Why outpatients fail to attend their scheduled appointments: a prospective comparison of differences between attenders and non-attenders. *Australian Health Review*, *26*(1), 52. https://doi.org/10.1071/AH030052

Cosgrove, M. P. (1990). Defaulters in general practice: reasons for default and patterns of attendance. *The British Journal of General Practice : The Journal of the Royal College of General Practitioners*, *40*(331), 50–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2107849

Cronin, P. R., DeCoste, L., & Kimball, A. B. (2013). A Multivariate Analysis of Dermatology Missed Appointment Predictors. *JAMA Dermatology*, *149*(12), 1435. https://doi.org/10.1001/jamadermatol.2013.5771

Dag, A., Oztekin, A., Yucel, A., Bulur, S., & Megahed, F. M. (2017). Predicting heart transplantation outcomes through data analytics. *Decision Support Systems*, *94*, 42–52. https://doi.org/10.1016/j.dss.2016.10.005

Dag, A., Topuz, K., Oztekin, A., Bulur, S., & Megahed, F. M. (2016). A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decision Support Systems*, *86*, 1–12. https://doi.org/10.1016/J.DSS.2016.02.007

Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P.-C., … Sands, L. (2010). Using no-show modeling to improve clinic performance. *Health Informatics Journal*, *16*(4), 246–259. https://doi.org/10.1177/1460458210380521

Dantas, L. F., Fleck, J. L., Cyrino Oliveira, F. L., & Hamacher, S. (2018). No-shows in appointment scheduling – a systematic literature review. *Health Policy*, *122*(4), 412–421. https://doi.org/10.1016/j.healthpol.2018.02.002

Dash, M., Liu, H., & Motoda, H. (2000). Consistency Based Feature Selection (pp. 98–109). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45571-X_12

Davis, G. W. (1989). Sensitivity analysis in neural net solutions. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(5), 1078–1082. https://doi.org/10.1109/21.44023

Delen, D. (2009). Analysis of cancer data: a data mining approach. *Expert Systems*, *26*(1), 100–112. https://doi.org/10.1111/j.1468-0394.2008.00480.x

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Desforges, J. F., McGuire, W. L., & Clark, G. M. (1992). Prognostic Factors and Treatment Decisions in Axillary-Node-Negative Breast Cancer. *New England Journal of Medicine*, *326*(26), 1756–1761. https://doi.org/10.1056/NEJM199206253262607

Dhanalakshmi, L., Ranjitha, S., & Suresh, H. N. (2016). A novel method for image processing using Particle Swarm Optimization technique. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 3357–3363). IEEE. https://doi.org/10.1109/ICEEOT.2016.7755326

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features. *MACHINE LEARNING: PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE*, 194--202. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.6141

Dove, H. G., & Schneider, K. C. (1981). The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. *Medical Care*, *19*(7), 734–740. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7266121

Drake, A. E., & Marks, R. E. (2002). Genetic Algorithms In Economics and Finance: Forecasting Stock Market Prices And Foreign Exchange — A Review. In *Genetic Algorithms and Genetic Programming in Computational Finance* (pp. 29–54). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-0835-9_2

Drummond, C., Drummond, C., & Holte, R. C. (2003). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling, 1--8. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.9672

DuMontier, C., Rindfleisch, K., Pruszynski, J., & Frey, J. J. (2013). A multi-method intervention to reduce no-shows in an urban residency clinic. *Family Medicine*, *45*(9), 634–641. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24136694

El-Midany, T. T., El-Baz, M. A., & Abd-Elwahed, M. S. (2010). A proposed framework for control chart pattern recognition in multivariate process using artificial neural networks. *Expert Systems with Applications*, *37*(2), 1035–1042. https://doi.org/10.1016/J.ESWA.2009.05.092

Elder, J. F. (2003). The Generalization Paradox of Ensembles. *Journal of Computational and Graphical Statistics*, *12*(4), 853–864. https://doi.org/10.1198/1061860032733

Farid, B. T., & Alapont, E. (1993). Patients who fail to attend their first psychiatric outpatient appointment: Non-attendance or inappropriate referral? *Journal of Mental Health*, *2*(1), 81–83. https://doi.org/10.3109/09638239309016957

Fawcett, T., & Provost, F. (1997). Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, *1*(3), 291–316. https://doi.org/10.1023/A:1009700419189

Fortin, K., Pries, E., & Kwon, S. (2016). Missed Medical Appointments and Disease Control in

Children With Type 1 Diabetes. *Journal of Pediatric Health Care*, *30*(4), 381–389. https://doi.org/10.1016/j.pedhc.2015.09.012

Frankel, S., Farrow, A., & West, R. (1989). Non-attendance or non-invitation? A case-control study of failed outpatient appointments. *BMJ (Clinical Research Ed.)*, *298*(6684), 1343–1345. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2502248

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, *29*, 131–163. Retrieved from http://www.cs.technion.ac.il/~dang/journal_papers/friedman1997Bayesian.pdf

Gao Huang, G., Shiji Song, S., Gupta, J. N. D., & Cheng Wu, C. (2014). Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE Transactions on Cybernetics*, *44*(12), 2405–2417. https://doi.org/10.1109/TCYB.2014.2307349

Garuda, S. R., Javalgi, R. G., & Talluri, V. S. (1998). Tackling No-Show Behavior. *Health Marketing Quarterly*, *15*(4), 25–44. https://doi.org/10.1300/J026v15n04_02

Genc, O., & Dag, A. (2016). A machine learning-based approach to predict the velocity profiles in small streams. *Water Resources Management*, *30*(1), 43–61. https://doi.org/10.1007/s11269-015-1123-7

Glowacka, K. J., Henry, R. M., & May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society*, *60*(8), 1056–1068. https://doi.org/10.1057/jors.2008.177

Goldberg, D. E., & Holland, J. H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, *3*(2/3), 95–99. https://doi.org/10.1023/A:1022602019183

Gunasundari, S., Janakiraman, S., & Meenambal, S. (2016). Velocity Bounded Boolean Particle Swarm Optimization for improved feature selection in liver and kidney disease diagnosis. *Expert Systems with Applications*, *56*, 28–47. https://doi.org/10.1016/J.ESWA.2016.02.042

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. In *2008 Fourth International Conference on Natural Computation* (pp. 192–201). IEEE. https://doi.org/10.1109/ICNC.2008.871

Gupta, S., & Sharma, A. (2011). Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. *Indian Journal Of Computer Science and Engineering*, *2*(2), 188–195.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1/3), 389–422. https://doi.org/10.1023/A:1012487302797

H. James, W., & Paul R., D. (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. Retrieved from https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces

Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.4643

Han, J., & Kamber, M. (2006). *Data mining : concepts and techniques*. Elsevier. Retrieved from https://dl.acm.org/citation.cfm?id=1076797

Hardy, K. J., O'Brien, S. V, & Furlong, N. J. (2001). Information given to patients before appointments and its effect on non-attendance rate. *BMJ (Clinical Research Ed.)*, *323*(7324), 1298–1300. https://doi.org/10.1136/BMJ.323.7324.1298

Hasvold, P. E., & Wootton, R. (2011). Use of telephone and SMS reminders to improve attendance at hospital appointments: a systematic review. *Journal of Telemedicine and Telecare*, *17*(7), 358–364. https://doi.org/10.1258/jtt.2011.110707

Hawkins, D. M. (2003). The Problem of Overfitting. https://doi.org/10.1021/CI0342472

Hirata, T., Shimizu, C., Yonemori, K., Hirakawa, A., Kouno, T., Tamura, K., … Fujiwara, Y. (2009). Change in the hormone receptor status following administration of neoadjuvant chemotherapy and its impact on the long-term outcome in patients with primary breast cancer. *British Journal of Cancer*, *101*(9), 1529–1536. https://doi.org/10.1038/sj.bjc.6605360

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Horsley, B. P., Lindauer, S. J., Shroff, B., Tüfekçi, E., Abubaker, A. O., Fowler, C. E., & Maxfield, B. J. (2007). Appointment keeping behavior of Medicaid vs non-Medicaid orthodontic patients. *American Journal of Orthodontics and Dentofacial Orthopedics*, *132*(1), 49–53.

Howlader, N., Noone, A. M., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., … Tatalovich, Z. (2013). SEER Cancer Statistics Review, 1975–2010, National Cancer Institute. Bethesda, MD, based on November 2012 SEER data submission, posted to the SEER web site, 2013. *Http://Seer.Cancer.Gov/Csr/1975_2010 (Accessed on June 08, 2013)*.

Huang, Y.-L., & Hanauer, D. A. (2016). Time dependent patient no-show predictive modelling development. *International Journal of Health Care Quality Assurance*, *29*(4), 475–488. https://doi.org/10.1108/IJHCQA-06-2015-0077

Huang, Y., & Hanauer, D. A. (2014). Patient no-show predictive model development using multiple data sources for an effective overbooking approach. *Applied Clinical Informatics*, *5*(3), 836–860. https://doi.org/10.4338/ACI-2014-04-RA-0026

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, *24*(3), nihpa54860. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19756229

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

John, G. H., Kohavi, R., & Ppeger, K. (1994). Irrelevant Features and the Subset Selection Problem, 121–129. Retrieved from https://pdfs.semanticscholar.org/a83b/ddb34618cc68f1014ca12eef7f537825d104.pdf

Junod Perron, N., Dominicé Dao, M., Kossovsky, M. P., Miserez, V., Chuard, C., Calmy, A., & Gaspoz, J.-M. (2010). Reduction of missed appointments at an urban primary care clinic: a randomised controlled study. *BMC Family Practice*, *11*(1), 79. https://doi.org/10.1186/1471-2296-11-79

Kaiser, J, H. (2018). Professionally Active Physicians | The Henry J. Kaiser Family Foundation. Retrieved August 6, 2018, from https://www.kff.org/other/state-indicator/total-active-physicians

Kaplan-Lewis, E., & Percac-Lima, S. (2013). No-Show to Primary Care Appointments. *Journal of Primary Care & Community Health*, *4*(4), 251–255. https://doi.org/10.1177/2150131913498513

Kasami, M., Uematsu, T., Honda, M., Yabuzaki, T., Sanuki, J., Uchida, Y., & Sugimura, H. (2008). Comparison of estrogen receptor, progesterone receptor and Her-2 status in breast cancer pre- and post-neoadjuvant chemotherapy. *The Breast*, *17*(5), 523–527. https://doi.org/10.1016/j.breast.2008.04.002

Kate, R. J., & Nadig, R. (2017). Stage-specific predictive models for breast cancer survivability. *International Journal of Medical Informatics*, *97*, 304–311.

Kempny, A., Diller, G.-P., Dimopoulos, K., Alonso-Gonzalez, R., Uebing, A., Li, W., … Gatzoulis, M. A. (2016). Determinants of outpatient clinic attendance amongst adults with congenital heart disease and outcome. *International Journal of Cardiology*, *203*, 245–250. https://doi.org/10.1016/J.IJCARD.2015.10.081

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M. C. U., Voduc, D., Speers, C. H., … Gelmon, K. (2010). Metastatic Behavior of Breast Cancer Subtypes. *Journal of Clinical Oncology*, *28*(20), 3271–3277. https://doi.org/10.1200/JCO.2009.25.9820

Kennedy, J. (2011). Particle Swarm Optimization. In *Encyclopedia of Machine Learning* (pp. 760–766). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_630

Kepplinger, D. (2015). gaselect: Genetic Algorithm (GA) for Variable Selection from High-Dimensional Data. *R Package Version 1.0.5. Https://CRAN.R-Project.Org/Package=gaselect*.

Kheirkhah, P., Feng, Q., Travis, L. M., Tavakoli-Tabasi, S., & Sharafkhaneh, A. (2015). Prevalence, predictors and economic consequences of no-shows. *BMC Health Services Research*, *16*(1), 13.

Kibis, E., Buyuktahtakin, E., & Dag, A. (2017). Data analytics approaches for breast cancer survivability: comparison of data mining methods. *Proceedings of the 2017 Industrial and Systems Engineering Conference*.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science (New York, N.Y.)*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Klement, W., Wilk, S., Michalowski, W., & Matwin, S. (2011). Classifying Severely Imbalanced Data (pp. 258–264). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21043-3_31

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Appears in the International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1–7). https://doi.org/10.1067/mod.2000.109031

Kohavi, R., & Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, 1137--1143. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.529

Kohavi, R., & Kohavi, R. (1996). Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 202--207. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4952

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 25–36.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kulkarni, A., Naveen Kumar, B. S. C., Ravi, V., & Murthy, U. S. (2011). Colon cancer prediction with genetics profiles using evolutionary techniques. *Expert Systems with Applications*, *38*(3), 2752–2757. https://doi.org/10.1016/J.ESWA.2010.08.065

Kumar, S., Mohri, M., & Talwalkar, A. (2012). Sampling Methods for the Nyström Method. *Journal of Machine Learning Research*, *13*, 981–1006. Retrieved from http://www.jmlr.org/papers/volume13/kumar12a/kumar12a.pdf

Langley, P. (1994). Selection of Relevant Features in Machine Learning. *In Proceedings of the AAAI Fall Symposium on Relevance*, 140–144.

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., … Pollack, J. R. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences*, *101*(3), 811–816. https://doi.org/10.1073/pnas.0304146101

Lavine, B. K., Davidson, C. E., & Moores, A. J. (2002). Genetic algorithms for spectral pattern recognition. *Vibrational Spectroscopy*, *28*(1), 83–95. https://doi.org/10.1016/S0924-2031(01)00147-3

Lee, J.-S., & Zhu, D. (2011). When Costs Are Unequal and Unknown: A Subtree Grafting Approach for Unbalanced Data Classification*. *Decision Sciences*, *42*(4), 803–829. https://doi.org/10.1111/j.1540-5915.2011.00332.x

Lee, V. J., Earnest, A., Chen, M. I., & Krishnan, B. (2005). Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Research*, *5*(1), 51. https://doi.org/10.1186/1472-6963-5-51

Li, A., Walling, J., Ahn, S., Kotliarov, Y., Su, Q., Quezado, M., … Fine, H. A. (2009). Unsupervised Analysis of Transcriptomic Profiles Reveals Six Glioma Subtypes. *Cancer Research*, *69*(5), 2091–2099. https://doi.org/10.1158/0008-5472.CAN-08-2100

Lin, S.-W., Ying, K.-C., Chen, S.-C., & Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, *35*(4), 1817–1824. https://doi.org/10.1016/J.ESWA.2007.08.088

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A Better Measure than Accuracy in Comparing Learning Algorithms (pp. 329–341). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44886-1_25

Ling, C. X., & Sheng, V. S. (2009). Cost-Sensitive Learning and the Class Imbalance Problem. Retrieved from https://www.semanticscholar.org/paper/Cost-Sensitive-Learning-and-the-Class-Imbalance-Ling-Sheng/9c4a953ed2cfc999eef0901d43097f9d2933005c

Liu, B., Wang, L., & Jin, Y.-H. (2008). An effective hybrid PSO-based algorithm for flow shop scheduling with limited buffers. *Computers & Operations Research*, *35*(9), 2791–2806. https://doi.org/10.1016/J.COR.2006.12.013

Liu, D., Yuan, Y., & Liao, S. (2009). Artificial neural networks for optimization of gold-bearing slime smelting. *Expert Systems with Applications*, *36*(9), 11671–11674. https://doi.org/10.1016/J.ESWA.2009.03.016

Lu, H., Wang, H., & Yoon, S. W. (2019). A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. *Expert Systems with Applications*, *116*, 340–350. https://doi.org/10.1016/J.ESWA.2018.08.040

Lundin, M., Lundin, J., Burke, H. B., Toikkanen, S., Pylkkänen, L., & Joensuu, H. (1999). Artificial Neural Networks Applied to Survival Prediction in Breast Cancer. *Oncology*, *57*(4), 281–286. https://doi.org/10.1159/000012061

Macharia, W. M., Leon, G., Rowe, B. H., Stephenson, B. J., & Haynes, R. B. (1992). An overview of interventions to improve compliance with appointment keeping for medical services. *JAMA*, *267*(13), 1813–1817. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1532036

McLeod, H., Heath, G., Cameron, E., Debelle, G., & Cummins, C. (2015). Introducing consultant outpatient clinics to community settings to improve access to paediatrics: an observational impact study. *BMJ Quality & Safety*, *24*(6), 377–384. https://doi.org/10.1136/bmjqs-2014-003687

McMullen, M. J., & Netland, P. A. (2015). Lead time for appointment and the no-show rate in an ophthalmology clinic. *Clinical Ophthalmology (Auckland, NZ)*, *9*, 513.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., Clark, G. M., & Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. (2005). Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK). *JNCI: Journal of the National Cancer Institute*, *97*(16), 1180–1184. https://doi.org/10.1093/jnci/dji237

Medical Appointment No-shows. (2016). Retrieved from https://www.kaggle.com/joniarroba/noshowappointments

Melin, P., Miramontes, I., & Prado-Arechiga, G. (2018). A hybrid model based on modular neural networks and fuzzy systems for classification of blood pressure and hypertension risk diagnosis. *Expert Systems with Applications*, *107*, 146–164. https://doi.org/10.1016/J.ESWA.2018.04.023

Milne, R. G., Horne, M., & Torsney, B. (2006). SMS reminders in the UK national health service: an evaluation of its impact on" no-shows" at hospital out-patient clinics. *Health Care Management Review*, *31*(2), 130–136.

Moise, M., Buruian, M. M., Ilie, C., Zamfir, C. L., Folescu, R., & Motoc, A. G. M. (2013). Estrogen and progesterone receptor expression in the mammary gland tumors. *Rom J Morphol Embryol*, *54*(4), 961–968.

Moore, C. G., Wilson-Witherspoon, P., & Probst, J. C. (2001). Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine*, *33*(7), 522–527. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11456244

Muhammad Umer Khan, M. U., Jong Pill Choi, J. P., Hyunjung Shin, H., & Minkoo Kim, M. (2008). Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Vol. 2008, pp. 5148–5151). IEEE. https://doi.org/10.1109/IEMBS.2008.4650373

Nasir, M., South-Winter, C., Ragothaman, S., & Dag, A. (2019). A comparative data analytic approach to construct a risk trade-off for cardiac patients' re-admissions. *Industrial Management & Data Systems*, *119*(1), 189–209. https://doi.org/10.1108/IMDS-12-2017-0579

Nguyen, D. L., DeJesus, R. S., & Wieland, M. L. (2011). Missed Appointments in Resident Continuity Clinic: Patient Characteristics and Health Care Outcomes. *Journal of Graduate Medical Education*, *3*(3), 350–355. https://doi.org/10.4300/JGME-D-10-00199.1

Nixon, A. J., Neuberg, D., Hayes, D. F., Gelman, R., Connolly, J. L., Schnitt, S., … Harris, J. R. (1994). Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage I or II breast cancer. *Journal of Clinical Oncology*, *12*(5), 888–894. https://doi.org/10.1200/JCO.1994.12.5.888

Norris, J. B., Kumar, C., Chand, S., Moskowitz, H., Shade, S. A., & Willis, D. R. (2014). An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decision Support Systems*, *57*, 428–443. https://doi.org/10.1016/J.DSS.2012.10.048

Nuti, L. A., Lawley, M., Turkcan, A., Tian, Z., Zhang, L., Chang, K., … Sands, L. P. (2012). No-shows to primary care appointments: subsequent acute care utilization among diabetic patients. *BMC Health Services Research*, *12*, 304. https://doi.org/10.1186/1472-6963-12-304

O’Brien, G., & Lazebnik, R. (1998). Telephone call reminders and attendance in an adolescent clinic. *Pediatrics*, *101*(6), E6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9606248

Oeffinger, K. C., Fontham, E. T. H., Etzioni, R., Herzig, A., Michaelson, J. S., Shih, Y.-C. T., … American Cancer Society. (2015). Breast Cancer Screening for Women at Average Risk. *JAMA*, *314*(15), 1599. https://doi.org/10.1001/jama.2015.12783

Office of inspector general audit of veterans health administration’s effort to reduced unused. (2008). Retrieved August 6, 2018, from https://www.google.com/search?ei=gqloW-aCDa-1ggfuirXYBg&q=office+of+inspector+general+audit+of+veterans+health+administration%27s+effor+to+reduced+unused&oq=office+of+inspector+general+audit+of+veterans+health+administration%27s+effor+to+reduced+unused&gs

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Publishing Company, Incorporated. https://doi.org/10.1007/978-3-540-76917-0

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Orimoloye, L. (2017). Are ensemble classifiers always better than single classifiers? - SAS Users. Retrieved December 28, 2018, from https://blogs.sas.com/content/sgf/2017/03/10/are-ensemble-classifiers-always-better-than-single-classifiers/

Pang, A. H. T., Tso, S., Ungvari, G. S., Chiu, H., & Leung, T. (1995). An Audit Study of Defaulters of Regular Psychiatric Outpatient Appointments in Hong Kong. *International Journal of Social Psychiatry*, *41*(2), 103–107. https://doi.org/10.1177/002076409504100203

Patterson, D. W. (1996). *Artificial neural networks : theory and applications*. Prentice Hall. Retrieved from https://dl.acm.org/citation.cfm?id=521611

Pearl, J., & Judea. (1997). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers. Retrieved from https://dl.acm.org/citation.cfm?id=534975

Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., & Benner, M. (1999). Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems with Applications*, *17*(3), 223–232. https://doi.org/10.1016/S0957-4174(99)00036-6

Powers, D. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. Retrieved from https://dspace2.flinders.edu.au/xmlui/handle/2328/27165

Principe, J. C., Euliano, N. R., & Curt Lefebvre, W. (2000). Innovating adaptive and neural systems instruction with interactive electronic books. *Proceedings of the IEEE*, *88*(1), 81–94. https://doi.org/10.1109/5.811604

Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett, G., MacKenzie, T., Adnane Tazi, M., … Faivre, J. (1999). Variation Over Time of the Effects of Prognostic Factors in a Population-based Study of Colon Cancer: Comparison of Statistical Models. *American Journal of Epidemiology*, *150*(11), 1188–1200. https://doi.org/10.1093/oxfordjournals.aje.a009945

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/.

Rampaul, R. S., Pinder, S. E., Elston, C. W., & Ellis, I. O. (2001). Prognostic and predictive factors in primary breast cancer and their role in patient management: The Nottingham Breast Team. *European Journal of Surgical Oncology (EJSO)*, *27*(3), 229–238. https://doi.org/10.1053/EJSO.2001.1114

Reid, M. W., May, F. P., Martinez, B., Cohen, S., Wang, H., Williams Jr, D. L., & Spiegel, B. M.

R. (2016). Preventing Endoscopy Clinic No-Shows: Prospective Validation of a Predictive Overbooking Model. *The American Journal of Gastroenterology*, *111*(9), 1267–1273. https://doi.org/10.1038/ajg.2016.269

Riddle, P., Segal, R., & Etzioni, O. (1994). Representation Design and Brute-force Induction in a Boeing Manufacturing Domain. *Applied Artificial Intelligence*, *8*(1), 125–147. https://doi.org/10.1080/08839519408945435

Rouyendegh, B. D., Oztekin, A., Ekong, J., & Dag, A. (2016). Measuring the efficiency of hospitals: a fully-ranking DEA–FAHP approach. *Annals of Operations Research*, 1–18. https://doi.org/10.1007/s10479-016-2330-1

Rouyendegh, B. D., Topuz, K., Dag, A., & Oztekin, A. (2018). An AHP-IFT Integrated Model for Performance Evaluation of E-Commerce Web Sites. *Information Systems Frontiers*, 1–11. https://doi.org/10.1007/s10796-018-9825-z

Ruiz, E., & Nieto, F. H. (2000). A note on linear combination of predictors. *Statistics & Probability Letters*, *47*(4), 351–356. https://doi.org/10.1016/S0167-7152(99)00177-7

Rutenbar, R. A. (1989). Simulated annealing algorithms: an overview. *IEEE Circuits and Devices Magazine*, *5*(1), 19–26. https://doi.org/10.1109/101.17235

Ryu, Y. U., Chandrasekaran, R., & Jacob, V. S. (2007). Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research*, *181*(2), 842–854. https://doi.org/10.1016/J.EJOR.2006.06.031

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2002). *Sensitivity Analysis in Practice*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470870958

Samorani, M., & LaGanga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*, *240*(1), 245–257. https://doi.org/10.1016/j.ejor.2014.06.034

Saptarsi, G., & Amlan, C. (2014). Feature Selection: A Practitioner View. *I.J. Information Technology and Computer Science Information Technology and Computer Science*, *11*(11), 66–77. https://doi.org/10.5815/ijitcs.2014.11.10

Shah, S. J., Cronin, P., Hong, C. S., Hwang, A. S., Ashburner, J. M., Bearnot, B. I., … Kimball, A. B. (2016). Targeted Reminder Phone Calls to Patients at High Risk of No-Show for Primary Care Appointment: A Randomized Trial. *Journal of General Internal Medicine*, *31*(12), 1460–1466. https://doi.org/10.1007/s11606-016-3813-0

Simsek, S., Bayraktar, E., Ragothaman, S., & Dag, A. (2018). A Bayesian Approach to Detect the Firms with Material Weaknesses in Internal Control. In *Proceedings of the 2018 Industrial and Systems Engineering Conference* (pp. 1247–1253). Orlando.

Smith, C. M., & Yawn, B. P. (1994). Factors associated with appointment keeping in a family practice residency clinic. *The Journal of Family Practice*, *38*(1), 25–29. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8289047

South-Winter, C. A., Dag, A., & Ragothaman, S. (2018). Factors Associated with Readmission of Cardiac Patients. *International Journal of Health Sciences*, *6*(4), 2372–5079. https://doi.org/10.15640/ijhs.v6n4a3

Steyerberg, E. W. (2008). *Clinical prediction models : a practical approach to development,* *validation, and updating*. Springer Science & Business Media. Retrieved from https://books.google.com/books/about/Clinical_Prediction_Models.html?id=kHGK58cLsMIC

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. https://doi.org/10.1016/j.eswa.2009.04.067

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. WileyRoyal Statistical Society. https://doi.org/10.2307/2346178

Ting, F. F., Tan, Y. J., & Sim, K. S. (2018). Convolutional Neural Network Improvement for Breast Cancer Classification. *Expert Systems with Applications*. https://doi.org/10.1016/J.ESWA.2018.11.008

Topuz, K., Uner, H., Oztekin, A., & Yildirim, M. B. (2018). Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network. *Annals of Operations Research*, *263*(1–2), 479–499. https://doi.org/10.1007/s10479-017-2489-0

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems*, *106*, 97–109. https://doi.org/10.1016/J.DSS.2017.12.004

Torra, Vicenc. (2003). Trends in Information fusion in Data Mining (pp. 1–6). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-36519-8_1

Torra, Vicenç. (2003). Trends in Information fusion in Data Mining (pp. 1–6). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-36519-8_1

Torres, O., Rothberg, M. B., Garb, J., Ogunneye, O., Onyema, J., & Higgins, T. (2015). Risk Factor Model to Predict a Missed Clinic Appointment in an Urban, Academic, and Underserved Setting. *Population Health Management*, *18*(2), 131–136. https://doi.org/10.1089/pop.2014.0047

Vikander, T., Parnicky, K., Demers, R., Frisof, K., Demers, P., & Chase, N. (1986). New-patient no-shows in an urban family practice center: analysis and intervention. *The Journal of Family Practice*, *22*(3), 263–268. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3950555

Vodopivec-Jamsek, V., de Jongh, T., Gurol-Urganci, I., Atun, R., & Car, J. (2012). Mobile phone messaging for preventive health care. *Cochrane Database of Systematic Reviews*, *12*, CD007457. https://doi.org/10.1002/14651858.CD007457.pub2

Walczak, S., & Velanovich, V. (2018). Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks. *Decision Support Systems*, *106*, 110–118. https://doi.org/10.1016/J.DSS.2017.12.007

Weerasinghe, G., Chi, H., & Cao, Y. (2016). Particle Swarm Optimization Simulation via Optimal Halton Sequences. *Procedia Computer Science*, *80*, 772–781. https://doi.org/10.1016/J.PROCS.2016.05.367

West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. *European Journal of Operational Research*, *162*(2), 532–551. https://doi.org/10.1016/J.EJOR.2003.10.013

Wit, E., Heuvel, E. van den, & Romeijn, J.-W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, *66*(3), 217–236. https://doi.org/10.1111/j.1467-9574.2012.00530.x

Wold, H. (2006). Partial Least Squares. In *Encyclopedia of Statistical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471667196.ess1914.pub2

Yancik, R., Ries, L. G., & Yates, J. W. (1989). Breast cancer in aging women. A population-based study of contrasts in stage, surgery, and survival. *Cancer*, *63*(5), 976–981. https://doi.org/10.1002/1097-0142(19890301)63:5<976::AID-CNCR2820630532>3.0.CO;2-A

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, *13*(2), 44–49. https://doi.org/10.1109/5254.671091

Yang, X.-S., Deb, S., & Fong, S. (2011). Accelerated Particle Swarm Optimization and Support Vector Machine for Business Optimization and Applications (pp. 53–66). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22185-9_6

Yazdi, M. R. S., Khorasani, A. M., & Faraji, M. (2011). Optimization of coating variables for hardness of industrial tools by using artificial neural networks. *Expert Systems with Applications*, *38*(10), 12116–12127. https://doi.org/10.1016/J.ESWA.2011.03.002

Zeng, B., Turkcan, A., Lin, J., & Lawley, M. (2010). Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research*, *178*(1), 121–144. https://doi.org/10.1007/s10479-009-0569-5

Zhou, F., Zhou, H., Yang, Z., & Yang, L. (2019). EMD2FNN: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. *Expert Systems with Applications*, *115*, 136–151. https://doi.org/10.1016/J.ESWA.2018.07.065

Zolbanin, H. M., Delen, D., & Hassan Zadeh, A. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, *74*, 150–161. https://doi.org/10.1016/J.DSS.2015.04.003

Zupan, B., Demsar, J., Kattan, M. W., Beck, J. R., & Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, *20*(1), 59–75.