A Machine Learning Approach to Transit Fraud Detection

by

Jerry Craig Claiborne

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
May 7, 2022

Keywords: fraud, transit, machine learning, public transportation

Approved by

Ashish Gupta, Chair, Professor & College of Business Advisory Council Fellow
David Paradice, Professor & Harbert Eminent Scholar
Dianne Hall, Torchmark Professor
Pei Xu, Associate Professor

Abstract

This research is a collection of 3 papers on the use of machine learning methods to detect and classify transit media fraud using passenger transaction data. Academically, this work is an extension of machine learning research into the largely unexplored area of transit media fraud. The implication for industry, is a series of tested and highly effective methods of fraud detection that can be implemented to mitigate the millions of dollars lost in transit fraud each year.

Paper 1 - Public transit systems lose millions of dollars each year to fraud. The primary types of transit fraud are fare evasion and fare media fraud. The lack of research around fare media fraud and the associated obstacles are presented, and a transit fraud framework is introduced. Using this framework as a foundation, an unsupervised machine learning approach to fraud is demonstrated utilizing principal component analysis and k-means clustering. The findings reveal that higher levels of bus activity, lower levels of rail use, and disparities between tapping in and out of the system appropriately are key variables in the detection of fraud. During testing, high concentration clusters (greater than 75% fraud or not-fraud) were achieved, accounting for as much as 46% percent of the total records.

Paper 2 - Globally, public transit provides billions of rides each year. The scale of transit related fraud is estimated to be in the hundreds of millions annually. In this research the significant challenges in transit fraud research of access to data, constantly evolving fraud techniques, and coping with the highly imbalanced nature of fraud data are presented and addressed. Using supervised machine learning methods (logistic regression, k-nearest neighbor, Naïve Bayes, & random forests) coupled with SMOTE to account for data imbalance, transit fraud classification model accuracy rates varied from approximately 80-97%. Corresponding ROC AUC scores ranged from approximately 75-87%. The highest performing models were KNN and random forests using bi-weekly and monthly data sets.

Paper 3 - The scale of both public transit and transit related financial losses are enormous. Billions of passenger trips are provided annually by public transit systems that are heavily dependent on revenue collections via passenger revenues. Transit media fraud costs authorities millions of dollars per year and has thus far been largely unexplored in academic research. This research describes the difficulties associated with transit research fraud and then addresses them via a demonstration of data techniques (SMOTE & ADASYN) and machine learning methods (deep learning). A series of 10 deep learning model variations, pretreated with SMOTE, are tested with the highest performing model achieving approximately 93% accuracy. These results represent compelling findings for both transit fraud researchers and public transit authorities

Acknowledgments

I would like to thank the various members of the Harbert College of Business for their invaluable advice and support throughout my time in the program. Specific thanks to Dr. Ashish Gupta for graciously agreeing to participate as the committee chair and for his constant guidance navigating conferences, research submissions, and career decisions. I'd like to extend my gratitude to Dr. David Paradice for his example of academic professionalism. The value added to papers and projects by Dr. Paradice's prompt and thorough responses cannot be overstated. Dr. Dianne Hall was an exceptional ambassador for the program and her support during the application process, and later navigating the program requirements, was critical to my success. Her friendly, and often very direct advice was incredibly helpful. Finally, Dr. Pei Xu introduced me to many of the concepts that eventually became central to this research. Her class has deeply influenced my approach to machine learning research, and I am grateful.

On a personal note, I would like to thank my parents, Theresa & Bill Thomas. They demonstrated daily that there is simply no substitute for hard work and sacrifice. They are a constant reminder to strive for more, but to always prioritize family above ambition.

Lastly, I'd like to thank my wife Melissa and our children Caleb, Noah, Kylie, & Aubrey. Melissa proved that a working professional and parent can pursue a PhD with poise and dignity. Her example was an inspiration and I appreciate the many sacrifices that she and our children made so that I could complete this program.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

PCA	Principal Component Analysis
KNN	K Nearest Neighbors
APTA	American Public Transportation Association
MTN	Mass Transit Network
AICPA	American Institute of Certified Public Accountants
SAS	Statement on Auditing Standards
CPA	Certified Public Account
PCAOB	Public Company Accounting Oversight Board
COSO	Committee of Sponsoring Organizations
UITP	International Association of Public Transport

Introduction

This dissertation follows the 3-paper style and is comprised of 3 independent but related studies. The overarching theme is the use of machine learning techniques to detect transit related fraud. The components and perspectives of the papers are as follows:

Paper	1	2	3
Research Question(s)	Can public transit transaction details be used to effectively identify transit media fraud?	Can machine learning models effectively classify transit fare media fraud?	Can a deep learning approach to transit fraud detection accurately categorize transactions?
Research Framework	Transit Fraud Framework		
Theoretical Support	Fraud Triangle Theory	General Deterrence Theory	Routine Activity Theory
Machine Learning Approach	Unsupervised	Supervised	Supervised
Method/ Algorithm	(1) Principal Component Analysis (PCA) (2) K-means clustering	(1) Logistic regression (2) KNN (3) Naïve Bayes (4) Random Forests	Deep Learning
Data	15-day samples of unlabeled transit transaction data	Transit record samples of daily, weekly, biweekly, and monthly time periods	14-day samples of labeled transit transaction data

Table 1: Overview of 3 Paper Dissertation

The goal of utilizing a variety of support theories, machine learning approaches, methods/algorithms, and data sets is to increase the generalizability and transferability of the work.

Paper 1: Utilizing Clustering Techniques to Classify Transit Fraud

Abstract – Public transit systems lose millions of dollars each year to fraud. The primary types of transit fraud are fare evasion and fare media fraud. The lack of research around fare media fraud and the associated obstacles are presented, and a transit fraud framework is introduced. Using this framework as a foundation, an unsupervised machine learning approach to fraud is demonstrated utilizing principal component analysis and k-means clustering. The findings reveal that higher levels of bus activity, lower levels of rail use, and disparities between tapping in and out of the system appropriately are key variables in the detection of fraud. During testing, high concentration clusters (greater than 75% fraud or not-fraud) were achieved, accounting for as much as 46% percent of the total records.

Keywords - Public transit, fraud, fare media, Fraud Triangle, machine learning, clustering

1.1 Introduction

Public Transit

Public transit is typically understood to be locally operated systems of buses and/or trains for public transportation use. A more formal definition provided by the American Public Transportation Authority (APTA) is "Transportation by bus, rail, or other conveyance, either publicly or privately owned, which provides to the public general or special service on a regular and continuing basis. Also known as 'mass transportation', 'mass transit', and 'transit'." (APTA, 2007). The public transit industry has evolved over centuries from simple ferry operations in ancient times, to public buses in Paris during the 1600's, to public trains in England during the 1800's, and finally to high-speed rail in modern times (Mass Transit Network, 2020).

Transit systems can be found in varying degrees of size, composition, and complexity in most modern cities. Globally, billions of trips are provided each year. APTA reports that average annual U.S. transit ridership exceeded 10 billion trips per year over the last decade and passenger fare revenues were approximately $16 billion in 2018 (APTA, 2020). However, transit companies lose millions of dollars annually to various forms of fraud. While pinpointing exact fraud totals is difficult, estimates at the agency level are often provided by industry publications, local/national news agencies, or law enforcement representatives as byproducts of stories/investigations. U.S. examples include estimates in Washington, D.C. of $40 million in 2019 (Wagner, 2019), New York, NY $300 million (Nguyen, 2019), and San Francisco, CA $25 million (May, 2019).

Transit Fraud

As evidenced by its scale, fraud is a significant issue for public transportation authorities. Two primary categories of transit fraud are fare evasion and fare media fraud. The Bureau of Transportation Statistics defines fare evasion as "*the unlawful use of transit facilities by riding without paying the applicable fare*" (BTS, 2017). Examples of fare evasion include forcing gates (prying open closed gates), jumping turnstiles, gate surfing (entering behind a paying customer before the gate closes), and simply sneaking aboard without paying.

Media fraud is the abuse or manipulation of fare media such that it negatively or illegally impacts one or more of the transaction participants. Typically, media fraud precludes the transit authority from receiving the full fare value of a provided trip, or the cost of the trip is illegally purchased or transferred to another party via media theft, media forgery, or credit card fraud. Examples of fare media fraud include the use or production of counterfeit or altered media, unauthorized card sharing, and buying/selling stolen or illegally purchased media.

While the two forms of fraud are sometimes commingled in the research, perhaps the easiest way to distinguish between them is by the nature of the interaction between a deviant customer and the transit system. Fare evasion can be best characterized as exhibiting a behavior that is intended to go undetected or unrecorded, such as forcing open a gate or boarding a train without purchasing a ticket. The actor in this scenario is attempting to cheat the system while simultaneously avoiding discovery. Alternatively, fare media fraud entails the user intentionally interacting with the system to create a transaction record under the false pretense of lawful payment and conduct. From an observer standpoint, a fraudulent fare media transaction would appear completely legitimate.

Because fare evasion and fare media fraud involve inherently different forms of interaction with transit systems, the associated detection and deterrence strategies are necessarily different. Since the goal of fare evasion is essentially to avoid interaction with the payment system, a typical mitigation strategy is centered around higher rates of customer monitoring. Customer monitoring levels are typically increased by employing additional personnel to check tickets and guard entry points, or by expanding the use of security camera systems.

Since fare media fraud involves interacting with the system directly, albeit under the false premise of legitimacy, transaction records are generated. The migration of transit authorities to automated fare collection and monitoring systems means details for every transaction are logged and archived regularly. This environment of large, frequently updated datasets requires a different fraud mitigation approach than fare evasion. In the case of fare media fraud, data-centric approaches are more viable and offer authorities the potential to implement solutions that rely less on increased personnel and more on data analysis. However, these data driven approaches present their own set of challenges.

Transit Fraud Research Challenges

Fraud research in a public transit setting presents a unique set of challenges. Chief among these challenges are the lack of a general framework to assist in the development and comparison of transit fraud research, access to data, and the evolving nature of fraud.

The first obstacle in transit fraud research is the lack of a unifying framework. While researchers can draw from a number of popular fraud frameworks, none of the current options are oriented specifically toward transit fraud. Typically, researchers draw from criminology, psychology, or sociology theories to provide context for fraud research. The use of a standardized framework augments the effort to contrast/compare research while providing perspective and context to readers outside of the public transit and/or fraud domains.

Once a research perspective is developed, the next step is to consider the data. Transit officials are often reluctant to disclose fraud levels for fear of encouraging additional criminal activity. High levels of reported fraud can also potentially damage bond ratings, negatively impact ridership, and/or damage political interest in expanding the system to neighboring communities. For these reasons and others, one of the most significant issues researchers face is access to appropriate data. While research on fare evasion exists, the majority of transit fare media fraud information must be constructed from piecemeal news and law enforcement reports. The data is typically restricted to the number of tickets seized, estimated value of the counterfeit tickets produced/sold, and arrest details. To gather fraud data independently, researchers are largely reduced to conducting general surveys or the use of the Delphi method to poll experts. These methods are limited in terms of the of volume, breadth, and precision of the data they provide. Given that transit agencies collect and store enormous amounts of detail-rich data, the best source is information derived directly from transit operations records.

Once the data has been acquired, researchers must account for the tendency of criminals to continually reshape fraud methods to avoid detection. Transit fraud has experienced an evolution consistent with other types of fraud. Beginning with relatively simple methods, techniques have gradually evolved into highly complex schemes to commit fraud.

An excellent example of the extreme ranges in complexity of fare media fraud is the progression from slugs to advanced counterfeit ticket operations. In early systems that used tokens, dishonest riders would utilize slugs to fool terminals and enter the system. Slugs are simple counterfeit coins or tokens, usually made from inexpensive materials, to simulate the size and weight of a legitimate coin/token. This is in stark contrast to modern transit systems, which use contactless smartcards to store customer information and digital media. Counterfeit smartcards can now be created via near-field communication (NFC) or radio-frequency identification (RFID) signal interception. This illustration of counterfeiting complexity growth, from simple token slugs to vastly more complicated smartcard duplication, is a prime example of the fraud evolution challenge faced by researchers.

The variety and complexity of transit fraud practices require the use of flexible detection solutions capable of identifying new methods. Machine learning techniques have been used to effectively classify some types of fraud with a high degree of accuracy. Of specific relevance, supervised learning methods have been used to identify factors of fare evasion. However, the inherent biases that often make supervised learning so effective at classifying known fraud, may also limit its effectiveness when encountering new types of fraud. By concentrating on deviations within the data set, unsupervised machine learning techniques offer the ability to account for the variations and new approaches introduced to frustrate fraud detection efforts.

Given the relative abundance of fare evasion research compared to the scarcity of fare media fraud studies, this paper concentrates on the latter. The objective is to establish a replicable research approach to fare media fraud that is academically and practically relevant using the best available data. As a potential solution to developing detection solutions for new/undocumented types of transit media fraud, the objective is to determine which available variables act as the best indicators for possible transit fraud. In doing so, this research can contribute to a standardized method of researching transit fraud and lay the foundation for a centralized repository for the inter-authority comparison of transit fraud.

Research question:

Can public transit transaction details be used to effectively identify transit media fraud?

The remainder of this paper presents a review of relevant literature and the proposed framework, a discussion on the proposed methods, an introduction to the data, an examination of the results, and the research conclusions.

1.2 Literature Review

Fraud

In academic settings, fraud is typically defined within the context of the occurrence. For example, telecommunications fraud deals with subscription scams wherein the perpetrator subscribes to services using a false identity with no intention of paying. This is radically different from medical prescription fraud where the perpetrator submits drug subscription claims for fictional or deceased patients (Bolton & Hand, 2002). Even when considering a specific subset of fraud such as financial fraud, a wide variety of fraud types exist. Examples of financial fraud include financial statement fraud, money laundering, credit card fraud, insurance fraud, auction fraud, and insider trading.

In terms of transit fraud, the major fraud categories are fare evasion and fare media fraud. Fare evasion has been and continues to be studied extensively. Barabino, Lai, & Olivo used 113 publications on fare evasion to produce a review of literature on the subject. The compiled body of work included 62 journal articles, 33 conference proceedings, 9 technical reports, 5 dissertations, 2 book chapters, and 2 working papers. In addition to the volume of research reported, they also showed that fare evasion research is a growing area of academic interest (Figure

1) (Barabino, Lai, & Olivo, 2020).

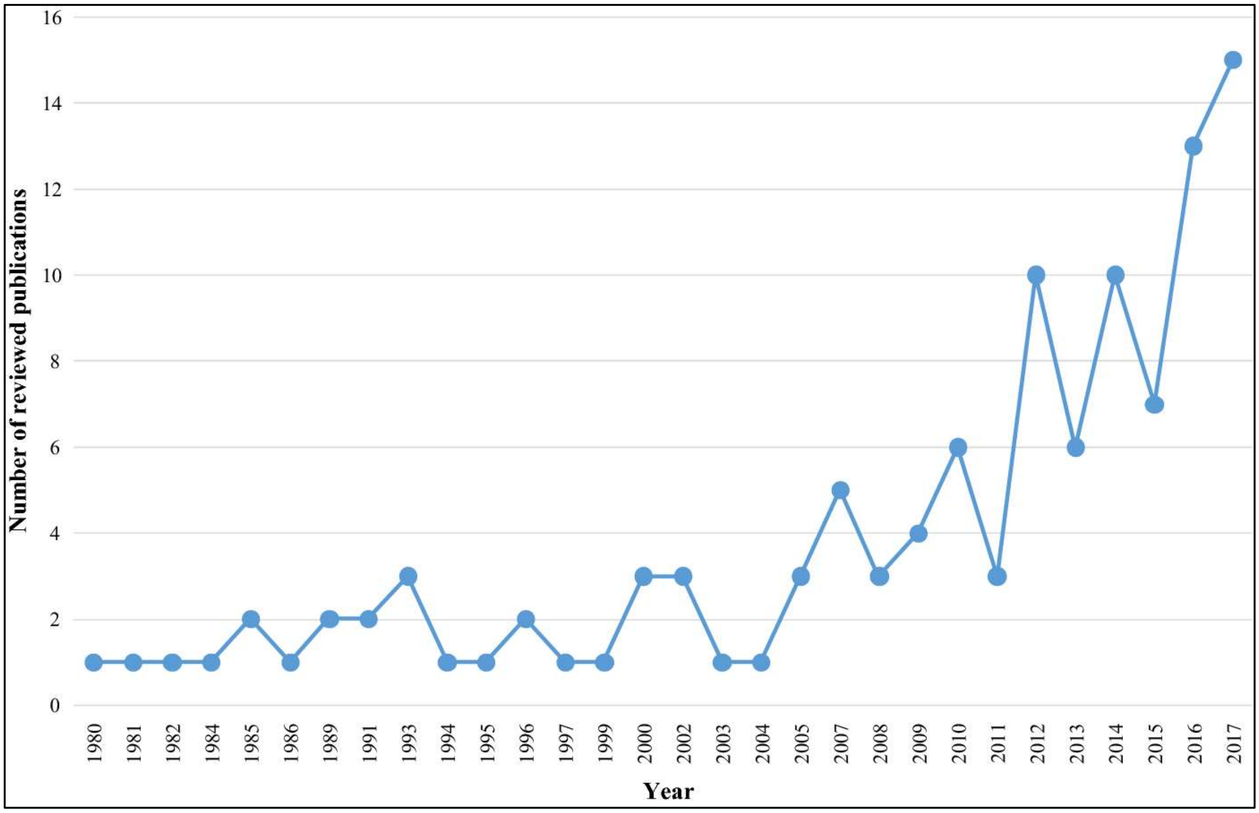


Figure 1: Fare Evasion Publications

(Adopted from Barabino, Lai, & Olivo, 2020)

Alternatively, fare media fraud research is scarce. This is attributable to the lack of researcher access to transaction records of transit authorities. Fürst & Herold (2018) comment repeatedly on of the lack of ticket forgery data and associated research. The limited examples of fare media fraud research that do exist include an application of neural network analysis to counter RFID transit card cloning efforts (DÜZENLİ, 2015), a tangentially related proposal to integrate fingerprint-based technology in the reservation / ticket booking system of the 2nd largest rail transit authority in India (Merja & Shah, 2013), and smart card attack and security research

(Markantonakis, Tunstall, Hancke, Askoxylakis, & Mayes, 2009). The scarcity of directly relevant transit fraud research clearly warrants attention from the academic research community.

Criminology Influences

As a theoretical underpinning for the proposed transit fraud framework, the fraud triangle theory is well represented in academic literature. The Fraud Triangle was originally introduced by Cressey (1953), who described the three contributing factors of fraud as pressure, rationalization, and opportunity (see Figure 1). In Cressey's words:

- Pressure – *"Trust violators, when they conceive of themselves as having a financial problem which is non-shareable…"*
- Opportunity – *"... and have knowledge or awareness that this problem can be secretly resolved by violation of the position of financial trust."*
- Rationalization – *"Also they are able to apply to their own conduct in that situation verbalizations which enable them to adjust their conceptions of themselves as trusted persons with their conceptions of themselves as users of the entrusted funds or property."* (Cressey, 1953).

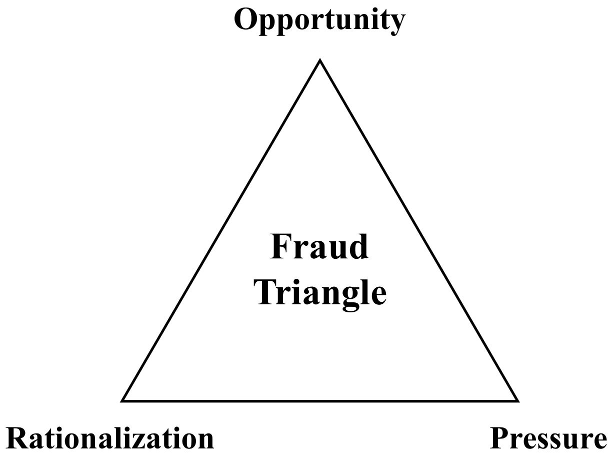


Figure 2: Fraud Triangle

In a 2009 international instructor survey of fraud examination and forensic accounting classes, the fraud triangle was found to be the most frequently taught fraud framework. A total of 111 faculty members at four-year institutions participated from the USA, UK, Australia, Hong Kong, and Lebanon (Smith & Crumbley, 2009).

A sample of notable works that modified the shape of the fraud triangle include Wolfe & Hermanson (2004) who expanded the triangle to a diamond by adding a fourth leg for "*capability*", Cieslewics (2010) who introduced a square by adding "*societal influences*", and Marks (2009) who added "*arrogance*" and "*competence*". These extensions of the fraud triangle focus primarily on customer features. Since the basis for this research is transaction data, these extensions are outside the scope of this paper.

In 2002, the fraud triangle became the foundation for the American Institute of Certified Public Accountants (AICPA) Statement on Auditing Standards (SAS) No. 99 (AICPA, 2002). The AICPA represents the interests of U.S. Certified Public Accounts (CPAs). The fraud triangle also appears in the auditing standards for the Public Company Accounting Oversight Board (PCAOB) and in the revised Internal Control—Integrated Framework (2013) and Risk Management Guide (2016) of the Committee of Sponsoring Organizations (COSO) of the Treadway Commission (Mintchik & Riley, 2019).

Cressey theorized that mitigating any one of the three central factors would lessen the total impact of fraud. Pressure, while originally proposed in financial terms, can also be rooted in time constraints (e.g., the customer is late and ticket lines are long) or social influence (e.g., peer pressure to disobey transit rules). Rationalization deals with a person's inner dialog and how he/she justifies deviant behavior. From a fraud prevention perspective, pressure and rationalization are difficult to influence due to their intangible natures. Conversely, and to varying extents, opportunity can be more directly impacted. By implementing system controls aimed at reducing fraud opportunities, transit authorities can proactively mitigate fraud. Examples of anti-fraud controls include inspection agents, surveillance camera systems, purchase limits, credit card verification protocols (e.g., use of PIN code or zip code), and transaction monitoring rules (e.g., velocity checking).

Transit Fraud Factors

As previously established, the majority of transit fraud research is centered around fare evasion. However, some of the contributing factors for fare evasion may also be applicable to fare media fraud. In 2007, a study of light rail fare evasion in 18 major European cities was published. The research considered the impacts of open vs. closed platforms, inspection factors (roving vs. permanent inspectors as well as overall inspection rates), flat vs. zone fare systems, fare media types, and the severity of fraud penalties. Findings included system recommendations of straight forward fare and use policies, closed system architectures (requires a ticket to enter), high customer contact levels (for fare inspections), use of smart cards, and an emphasis on system reliability (Dauby & Kovacs, 2007).

A study of Chilean bus fare evasion identified the level of inspection, proximity to a station, bus occupancy levels, time of day, location, and volume of passengers (boarding and alighting) at each bus stop to be significant factors in the likelihood of fare evasion occurrence (Guarda, Galilea, Handy, Muñoz & de Dios Ortúzar, 2016). A subsequent study of the same system analyzed fare levels, level of local employment, and the number of monthly fare inspections. Evasion rates were found to be positively correlated to fare levels and negatively correlated to unemployment. Inspection rates were found to be a poor indicator of fare evasion (Troncoso & de Grange, 2017).

In 2014, a qualitative fare evasion study of the Melbourne, Australia public transit system focused on consumer attitudes and motivations. Participant variables included gender, age, frequency of public transit utilization, and frequency of self-admitted fare evasion. The authors presented transit fraud as a spectrum of behavioral segments and intent levels. Findings showed a wide range of fraud occurrence rates, intentions, feelings, and fare evaders views based on participant rationale of fare evasion (Delbosc & Currie, 2016).

While these research examples are informative and cover a broad range of factors associated with fare evasion, fare media fraud and the associated transaction level details are noticeably absent. The proposed transit fraud framework addresses research perspectives (e.g., customer, system controls, and authority) that include these factors and more.

Transit Fraud Framework

Researchers have previously recognized the need for a central framework for specific categories of fraud research. Abbasi et al. (2012), developed a meta-learning framework of layered machine learning techniques to identify financial fraud (Abbasi, Albrecht, Vance, & Hansen, 2012). In 2018 a group of researchers from Belgium developed the SCARFF (Scalable Real-time Fraud Finder) framework for the detection of credit card fraud (Carcillo, Dal Pozzolo, Le Borgne, Caelen, Mazzer, & Bontempi, 2018).

The adoption of a comprehensive transit fraud framework encourages the development of research that can be readily compared and contrasted. By creating a defined research perspective, other academics and practitioners are encouraged to offer improvements, rebuttals, and/or additions. The proposed framework assumes a modern transit authority capable of detailed transaction logging. This is a reasonable standard given the percentage of U.S. transit systems utilizing smart cards has increased from 12% in 2009 to 48% as of 2019 (APTA, 2020). The essential elements of the framework are the data, a designated perspective, and the analytical approach that will be utilized. The proposed framework components are as follows:

Legitimate & Fraudulent Activity - As transactions are recorded there is little initial indication regarding their proclivity toward legitimacy or fraud. Transactions are stored by the transit authority for further analysis. Transactions are ultimately evaluated based on the transit fraud factors.

Government/External– Several potentially important factors exist beyond the span of control of the parent transit authority. Examples include the unemployment rate, community crime rates, law enforcement presence, and the severity of punishment and enforceability of anti-fraud statutes. Unemployment has been shown to be correlated with higher rates of fare evasion (Salis, Barabino, & Useli, 2017). Several criminology theories include elements that address how the opportunity to effectively commit a crime, the cost/benefit considerations of being caught, and the deterrence effects of punishment impact unlawful behavior.

Transactions– Access to transactional data is likely the prevailing reason behind the shortage of transit fraud research. Modern transit operations produce enormous collections of digital records that log every detail of electronic interactions between users and the system. The level of detail captured likely varies between systems, with dozens to hundreds of variables being available in a modern transit system's automated fare collection system. This study utilizes transactional data to demonstrate a viable transit fraud detection method.

System Controls – Enhanced ticketing, improved barriers and fare gates, installation of security cameras, uniformed and undercover enforcement agents, and enhanced security software are typical areas of concentration. An upgrade to these components is typically the first response of transit systems when fraud becomes an issue. APTA reports that 78% of buses now have security cameras (APTA, 2020). Cameras are also common aboard trains. Controls can by physical, digital, and/or policy based.

Customer– Consideration of customer factors covers a broad range of qualitative and quantitative variables including typical demographic data, income levels, value systems, cultural influences, access to private transportation, tolerance for risk, perception of public transit etc. Some research has been published on demographic fare evasion factors. For instance, young males with public transit ridership experience are a higher risk for fare evasion (Cools, Fabbro, & Bellemans, 2018).

Authority – Public transit systems take a variety of forms and offer a range of transportation options. The characteristics of transit systems warrant review when considering fraud. System settings (rural vs urban), location, size, complexity of use, brand, age, pricing structures, policy decisions on frequency of fare increases, customer service effectiveness, service levels, and technology adoption are all potential factors in transit fraud.

Media – Media considerations include the type of media (e.g., cash, tokens, magnetic tickets, smart cards, mobile phone apps etc.), the individual features of the media (smart cards use a variety of digital security protocols), and the various models of media. This media could also include the fare type (e.g., daily pass, weekly pass, monthly pass, single trip, multi-trip, stored value etc.).

Analysis – The analysis portion of the framework includes the examination of data using the tools, techniques, models, etc. selected by the researcher. The nature of each study will necessitate the type of analysis that is most effective and appropriate. The example demonstrated in this study utilizes an unsupervised machine learning approach consisting of principal component analysis and k-means clustering.

SME Evaluation - Use of experts to help interpret the findings is key (Cheeseman & Stutz, 1996). Subject matter experts must be involved in the transaction vetting process because decision makers may not be entirely comfortable with an analysis based solely on technical and statistical modeling. Subject matter experts provide a level of comfort and familiarity to the transit operator and can bridge the gap between the technical and the business sides of the fraud evaluation process. They may also be able to explain data anomalies and unexpected transaction phenomenon.

Business Decision – Ultimately transit operators must decide how to proceed with the recommendations provided by analysis and expert evaluations. Potential business decision outcomes include adopting an updated rule-based process, conducting case-by-case reviews, investing in system security upgrades (gates, cameras, security personnel, software etc.), deactivating the suspect media, updating policies regarding the involvement of law enforcement or direct punitive action (banning repeat offenders), and lobbying local, state, and/or federal lawmakers for stricter enforcement/punishment mandates.

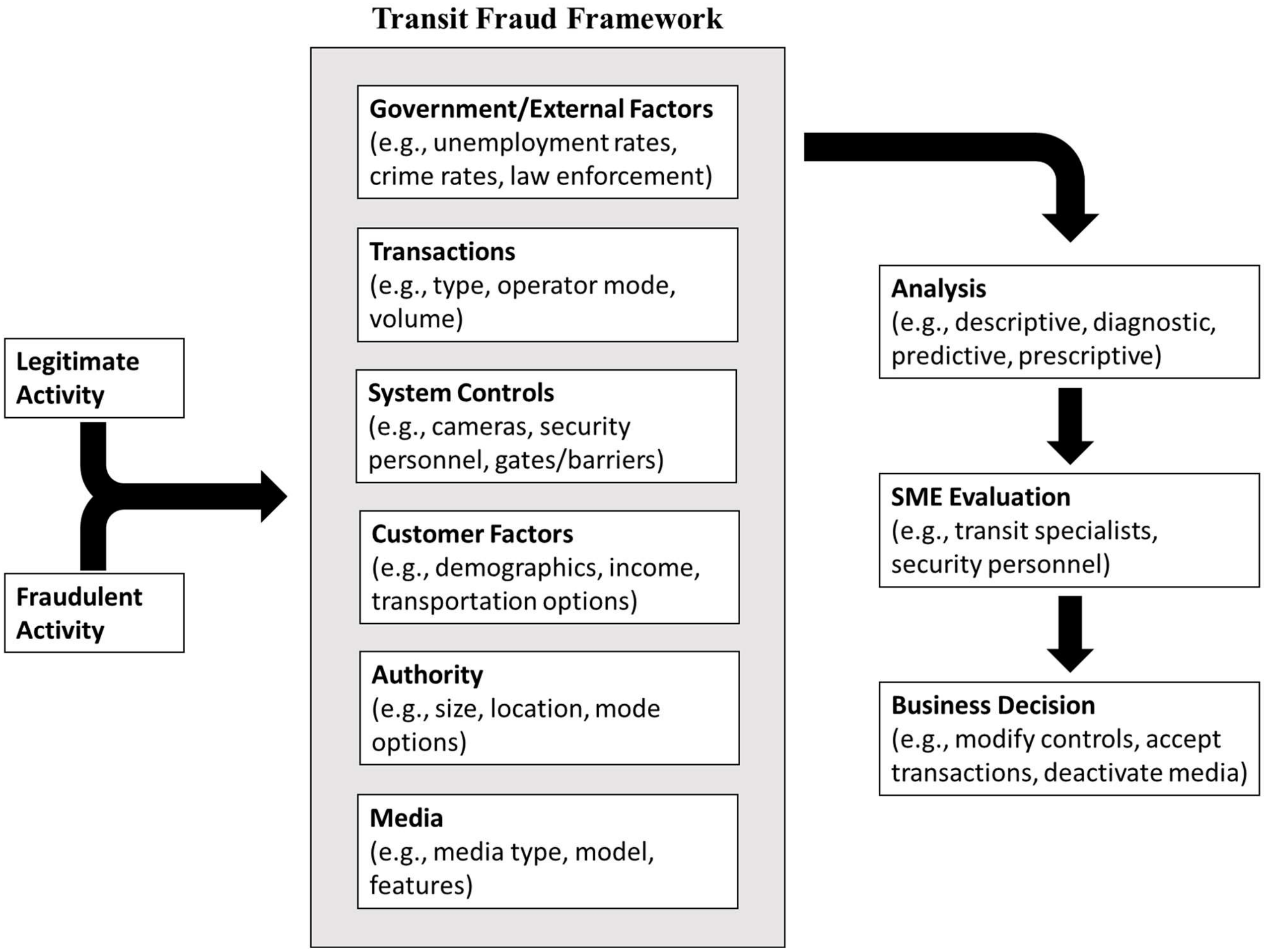

Transit Fraud Framework
Legitimate Activity
Fraudulent Activity
Government/External Factors
(e.g., unemployment rates, crime rates, law enforcement)
Transactions
(e.g., type, operator mode, volume)
System Controls
(e.g., cameras, security personnel, gates/barriers)
Customer Factors
(e.g., demographics, income, transportation options)
Authority
(e.g., size, location, mode options)
Media
(e.g., media type, model, features)
Analysis
(e.g., descriptive, diagnostic, predictive, prescriptive)
SME Evaluation
(e.g., transit specialists, security personnel)
Business Decision
(e.g., modify controls, accept transactions, deactivate media)

Figure 3: Transit Fraud Framework

1.3 Methodology

Model

The model for this study is derived from the Transit Fraud Framework introduced previously. The focus will be on transactions and specifically the transaction variable categories.

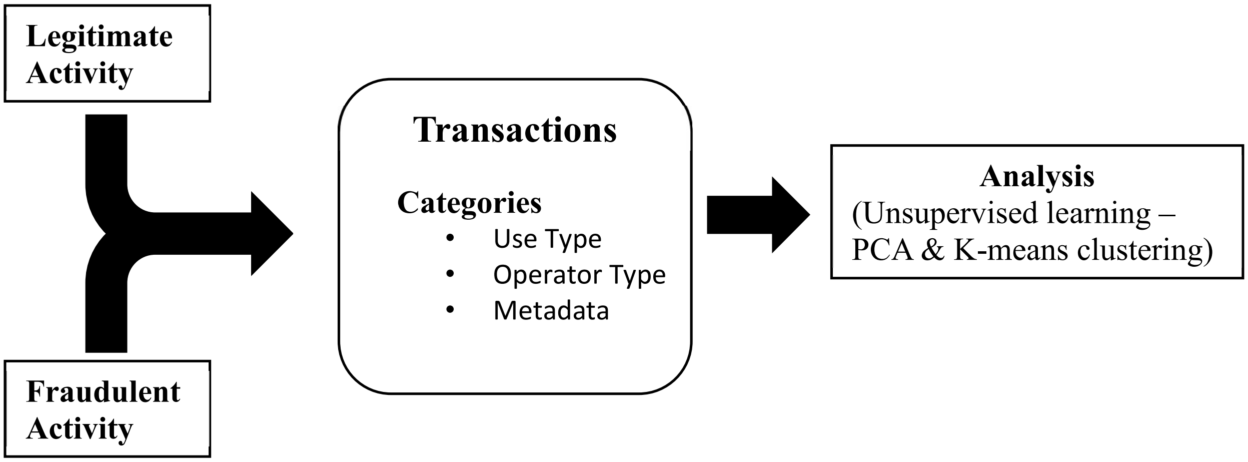


Figure 4: Research Methodology

Research Methodology

Supervised vs. Unsupervised Learning

Much of the general fraud research being conducted relies on the existence of labeled data to help researchers apply a variety of supervised machine learning techniques. A particularly relevant example is a study in 2018 using logistic regression to conduct fare evasion factor analysis. Factors identified included socio-demographics, transportation characteristics, perceptions of tariffs, presence/absence of fare inspectors, fines associated with fare evasion, behavior of acquaintances, nationality, weather, length of trip, and satisfaction with the public transportation service (Cools, Fabbro, & Bellemans, 2018). These factors were explored using a supervised learning technique.

While effective, the dependency on labeled data to identify factors may mitigate the effectiveness of these models at detecting new types of fraud. Unsupervised learning attempts to generalize data without the use of an initial underlying function. This notion is implemented by allowing the data to effectively self-organize into groups with high similarity for members and maximized dissimilarity for non-members. This study utilizes unsupervised learning to avoid any labeling bias and to maximize the potential for discovering new fraud methods.

Clustering

Because new approaches to fraud are always being introduced, unsupervised learning techniques that leave a degree of flexibility for pattern discovery are better suited for this research. With that principle in mind, clustering is a logical place to begin. Clustering techniques have been used in academic research to detect corporate fraud, credit card fraud, money laundering, financial reporting fraud, online auction fraud, stock market fraud and accounting fraud (Ngai et al, 2011), (Sabau, 2012). The intent of clustering techniques is to group data into natural groups such that homogeneity within clusters is maximized while simultaneously maximizing the heterogeneity between clusters (Thakare & Bagal, 2015).

K-means clustering is an unsupervised method in which data is assigned to *k* clusters based on the nearest cluster mean. Beginning with *k* clusters, each of which contains a single random point, an additional point is introduced and assigned to the cluster with the nearest mean (least squared Euclidian distance), and the centroid of the cluster is recalculated. This process repeats until the algorithm converges and cluster assignments stabilize (Wagstaff, Cardie, Rogers, & Schrödl, 2001). The iteration and convergence processes are demonstrated in the appendix (Figure 2).

Examples of k-means clustering success are prevalent in the literature. A comparison study using benchmark datasets from the UCI machine learning repository for Iris, Wine, Vowel, Ionosphere, and Crude Oil shows the flexibility of k-means clustering. See Table 8 in the appendix for additional details. Researchers found that across a variety of data sets and with varying cluster count parameters, k-means clustering had an average recognition rate of over 83%. Examples of k-means used specifically in fraud research can be found in Table 1.

Fraud Type	**Reference**
Medical claims fraud	(Wakoli, 2014)
Refund fraud /financial fraud	(Issa et al., 2011)
Healthcare insurance fraud	(Thiprungsri et al., 2011)
Money laundering	(Liu et al., 2011)
Credit card fraud	(Wu et al., 2010)
Money laundering	(Le Khac et al., 2010)
Online auction fraud	(Chang et al., 2010)
Insurance fraud	(Jurek et al., 2008)
Accounting fraud	(Virdhagriswaran et al., 2006)
Insider trading fraud	(Donoho, 2004)

Table 2: Financial Fraud Examples

The methodological approach employed is a logical progression of data collection, preparation, and modeling. The intent was to gather the most relevant features from the available data, treat the data to optimize the model performance, optimize and execute the models, and report the findings. Using principal component analysis (PCA) and k-means clustering techniques, public transit transaction records were analyzed for clues that certain categories of variables were more predictive than others.

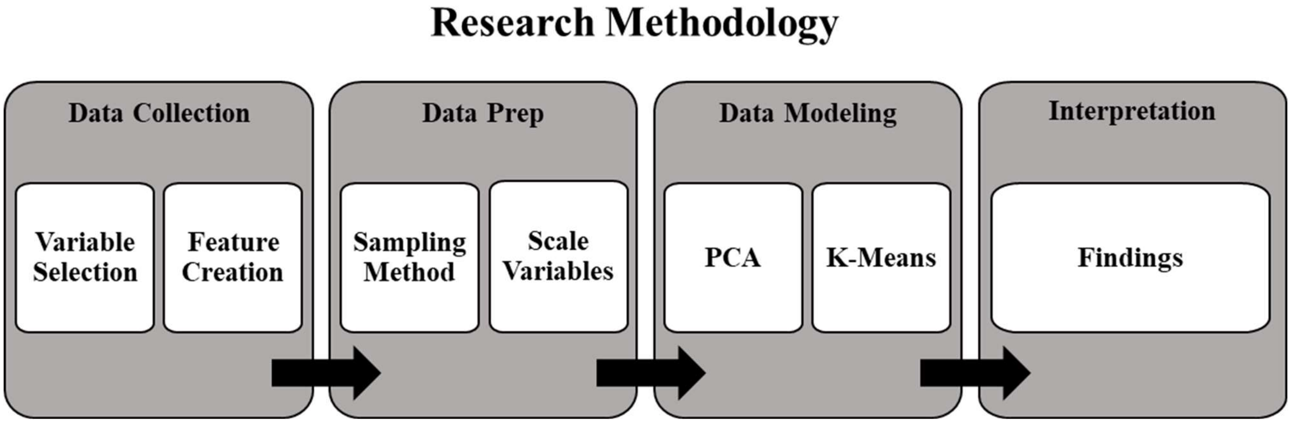


Figure 5: Research Methodology

1.4 Data

Source

The data for this research is based on transaction records from the Metropolitan Atlanta Rapid Transit Authority (MARTA) from 2018. MARTA is representative of modern transit systems and is a top 10 public transit authority in the U.S. The dataset includes 12,790 records of aggregated 15-day samples of transaction records for a random set of transit users. No customer identifying data was collected or reported. Each record represents the total number of transactions for the fare media across the various features. The data represents approximately 182,000 total transactions.

Variable Selection / Creation

Data features were extracted from several categorical types including use type, transaction status, operator, rider class, fare instrument category, and a variety of aggregated metadata features. Dummy variables were used to convert categorical data into numerical values. Dummy variables expanded the data to 160 features across 6 categories, but during the data preparation it was determined that many of the features were either partially or completely unpopulated in the dataset. Non-value-added features were removed, and the resulting dataset was reduced to 20 fully populated features across 3 categories.

Category	Original Features	Reduced Features
Use Type	19	5
Operator Provider	9	4
Metadata	6	11
Total	**160**	**20**

Table 3: Research Categories and Feature Summary

Category	Feature	Description
Use Type		
	ENT_EXT_RATIO	Entries minus exits
	ENTRIES_PER_DAY	Total entries divided by 15
	ENTRY_TAG_ON	Entries (gates or buses)
	EXIT_TAG_OFF	Exits (gates or buses)
	EXITS_PER_DAY	Total exits divided by 15
Operator Type		
	BUS_RATIO	Percentage of transactions that were bus related
	MARTA_BUS	Bus related transactions
	MARTA_RAIL	Rail related transactions
	RAIL_RATIO	Percentage of transactions that were rail related
Metadata		
	CATEGORIES_PER_DAY	Total categories divided by 15
	DEVICES	Total devices utilized
	DEVICES_PER_DAY	Total devices divided by 15
	FACILITIES	Total facilities visited
	FACILITIES_PER_DAY	Total facilities divided by 15
	FI_CATEGORIES	Total fare categories utilized
	MODES	Total modes utilized
	MODES_PER_DAY	Total modes divided by 15
	TRANSIT_DAYS	Total days with transactions
	TRANSIT_DAYS_RATIO	Total days with transactions divided by 15
	USES	Total uses

Table 4: Variable Details

Data Sampling & Scaling

Like most fraud research, the dataset being tested is highly imbalanced in its native form. For the transit dataset, the imbalance equates to approximately 400 legitimate transactions for each fraudulent transaction. To avoid constructing a classifier that simply selects the majority class to achieve an optimal accuracy rate of 99.75%, the classes need to be balanced. Several balancing methods exist including oversampling the minority class, undersampling the majority class, and use of a synthetic sample generator. Because this study is the first, or at the very least one of the first, to use actual transit transaction data, it is important to maintain as much data integrity as possible. For that reason, use of a synthetic generator was eliminated from consideration. If fraud samples were sparse and collecting more was not an option, the synthetic route would be more appealing. Use of a synthetic resampler like SMOTE does not create records outside of the bounds of the existing data. Instead, data density is increased to support the classifier. Rather than create additional uninformative samples, a combination of random majority undersampling and minority oversampling was employed. The minority class was resampled 6 times get a sufficiently high number of records to be representative of the population. This resulted in a total of 6,395 fraudulent transactions for consideration. For the majority class, a random number was assigned to each record, the records were shuffled thoroughly, and the first 6,395 records were selected for use. This provided a balanced dataset for the cluster analysis. No synthetic samples were utilized because of the abundance of available fraudulent records.

To mitigate the impact of scale variations, clusters variables were standardized by subtracting the mean and dividing by the standard deviation. The standard score of a sample x is calculated as:

$$\mathbf{z = (x - u) / s}$$

where **u** equals the mean of the training samples and **s** is the standard deviation of the training samples (Minitab, 2019).

Principal Component Analysis

Records were transformed using principal component analysis (PCA) to reduce dimensionality and provide legible plots. PCA combines existing high-dimension variables into new low-dimension combinations of linear variables. PC1 is plotted on the X-axis to exhibit direction of the highest variation in the data and PC2 is plotted along the Y-axis and indicates the direction of the next-highest level of variation. Component counts were determined by mapping the eigenvalues of the factors in a scree plot, then using the "elbow" method to identify a range of interest. This was further supported by the Kaiser criterion which limits the components based on eigenvalues of greater than 1. A PC count of 4 meets both criteria and accounts for approximately 90% of the cumulative variation in the dataset.

K-Means

Clustering was applied on both raw data and PCA scores using a range of k-means values. Based on the results, cluster values between 3 and 6 were explored.

1.5 Results

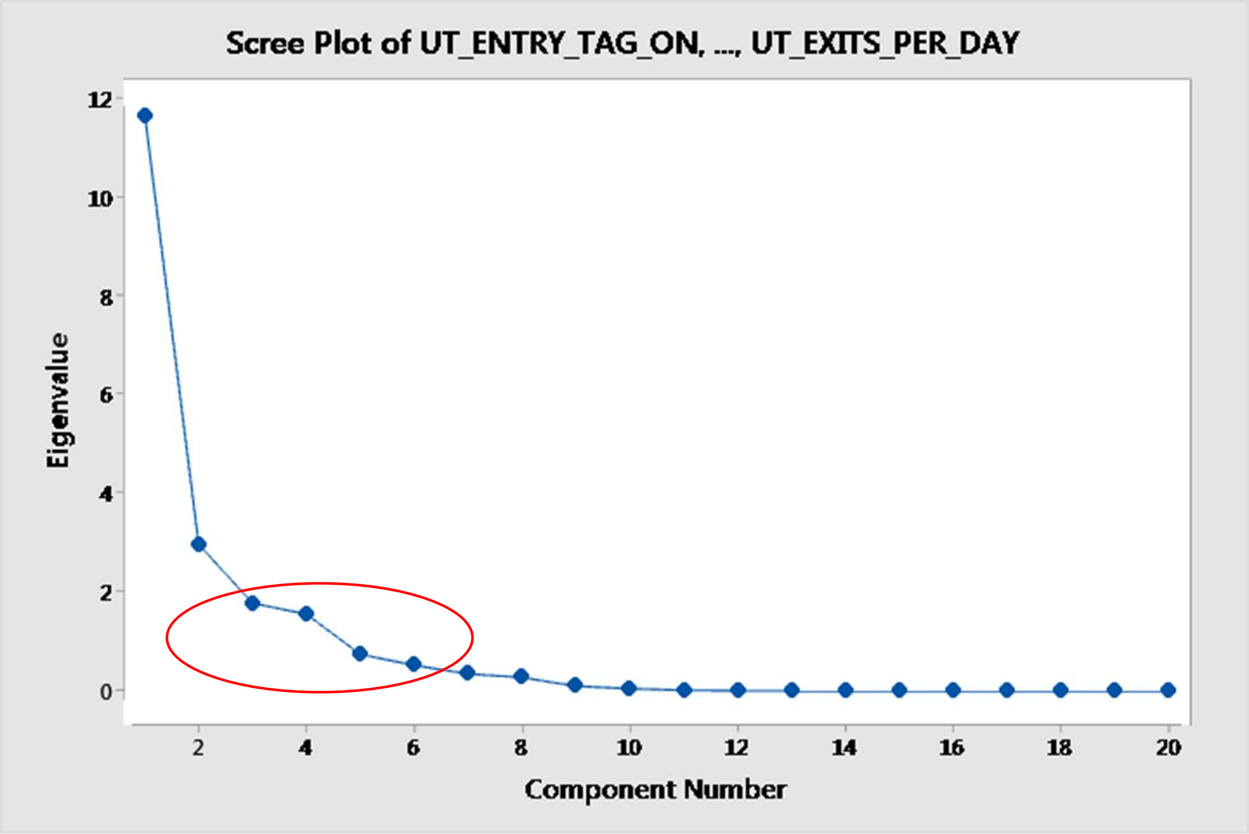


Figure 6: Scree Plot

Eigenanalysis of the Correlation Matrix

Eigenvalue	11.661	2.964	1.776	1.553	0.742	0.520	0.345	0.275	0.098	0.048
Proportion	0.583	0.148	0.089	0.078	0.037	0.026	0.017	0.014	0.005	0.002
Cumulative	0.583	0.731	0.820	0.898	0.935	0.961	0.978	0.992	0.997	0.999

Table 5: Eigenanalysis for PCA

Principal Component Analysis

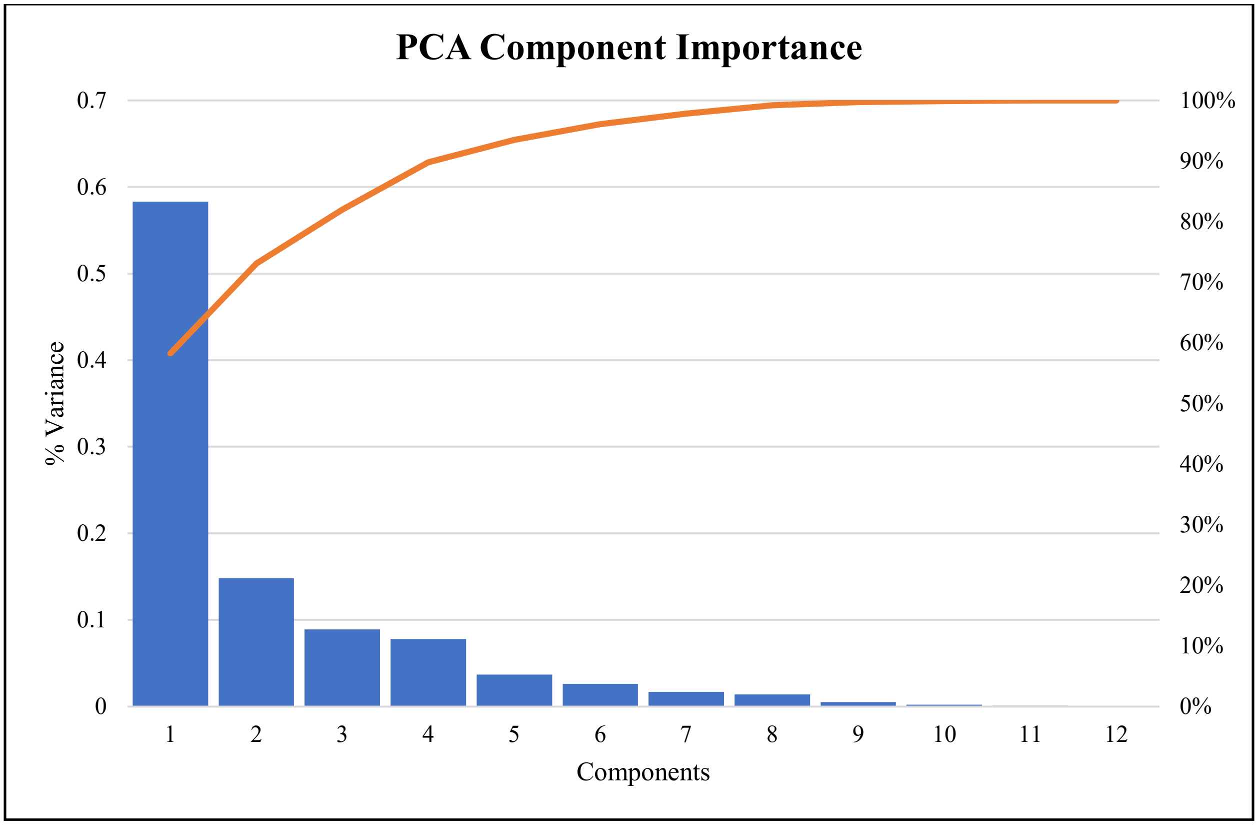


Figure 7: PCA Component Importance

Components	% Variance	Cumulative % Variance
1	0.583	0.583
2	0.148	0.731
3	0.089	0.82
4	0.078	0.898
5	0.037	0.935
6	0.026	0.961
7	0.017	0.978
8	0.014	0.992
9	0.005	0.997
10	0.002	0.999
11	0.001	1
12	0	1
13…20	…0	…1

Table 6: PCA Cumulative Variance

PCA Component Weights

Variable	PC1	PC2	PC3	PC4
UT_ENTRY_TAG_ON	0.268	0.132	-0.175	-0.055
UT_EXIT_TAG_OFF	0.252	-0.211	-0.154	-0.044
UT_ENT_EXT_RATIO	0.095	0.469	-0.075	-0.029
OP_MARTA_RAIL	0.26	-0.183	-0.16	-0.049
OP_MARTA_BUS	0.23	0.264	-0.029	0.041
OP_BUS_RATIO	-0.01	0.506	0.093	-0.005
OP_RAIL_RATIO	0.017	-0.516	-0.079	0.016
MD_MODES	0.179	-0.009	0.448	0.363
MD_DEVICES	0.279	0.032	-0.02	0.046
MD_FACILITIES	0.251	-0.088	0.164	0.199
MD_FI_CATEGORIES	0.136	-0.044	0.406	-0.559
MD_TRANSIT_DAYS	0.261	0.064	-0.147	-0.079
MD_TRANSIT_DAYS_RATIO	0.261	0.064	-0.147	-0.079
MD_USES	0.285	-0.001	-0.124	-0.016
MD_MODES_PER_DAY	0.179	-0.009	0.448	0.363
MD_DEVICES_PER_DAY	0.279	0.032	-0.02	0.046
MD_FACILITIES_PER_DAY	0.251	-0.088	0.164	0.199
MD_CATEGORIES_PER_DAY	0.136	-0.044	0.406	-0.559
UT_ENTRIES_PER_DAY	0.268	0.132	-0.175	-0.055
UT_EXITS_PER_DAY	0.252	-0.211	-0.154	-0.044

Table 7: PCA Component Weights

Fraud Plot

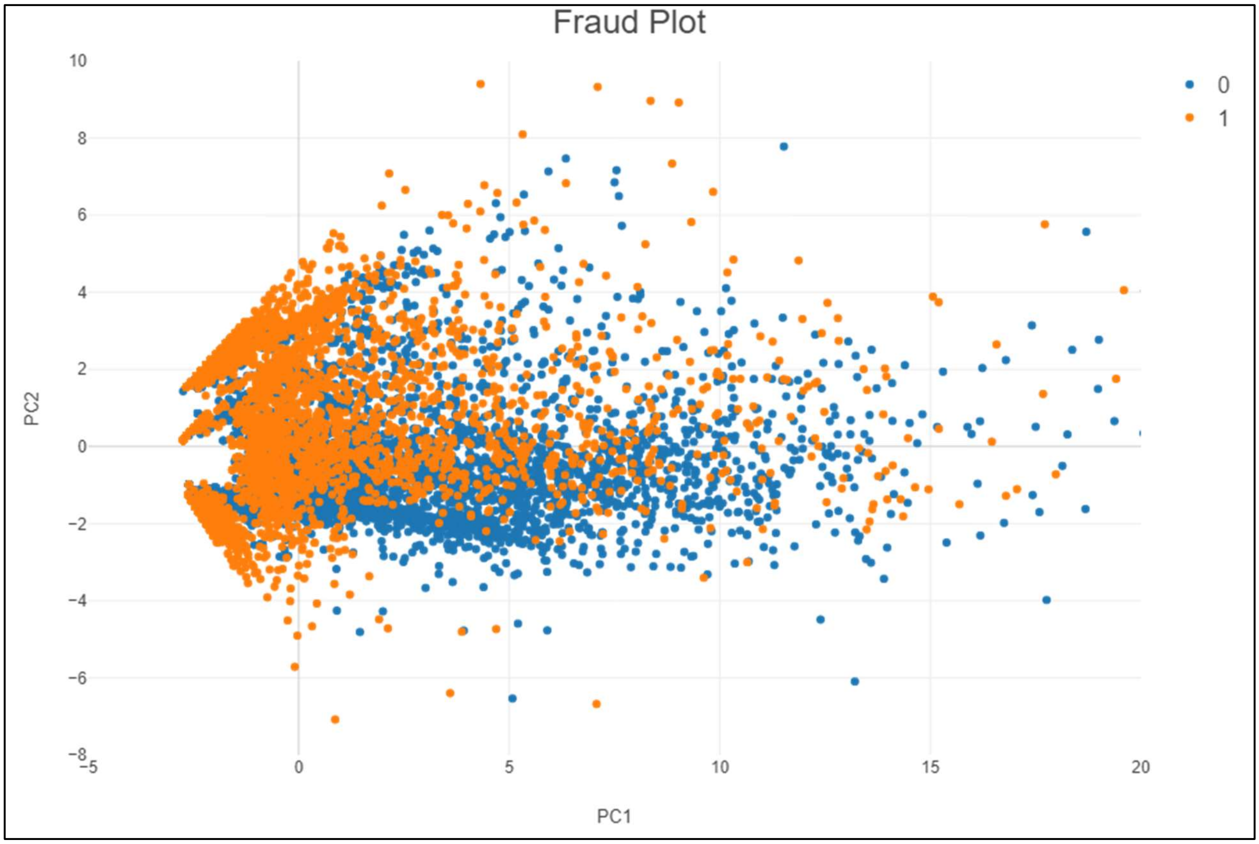
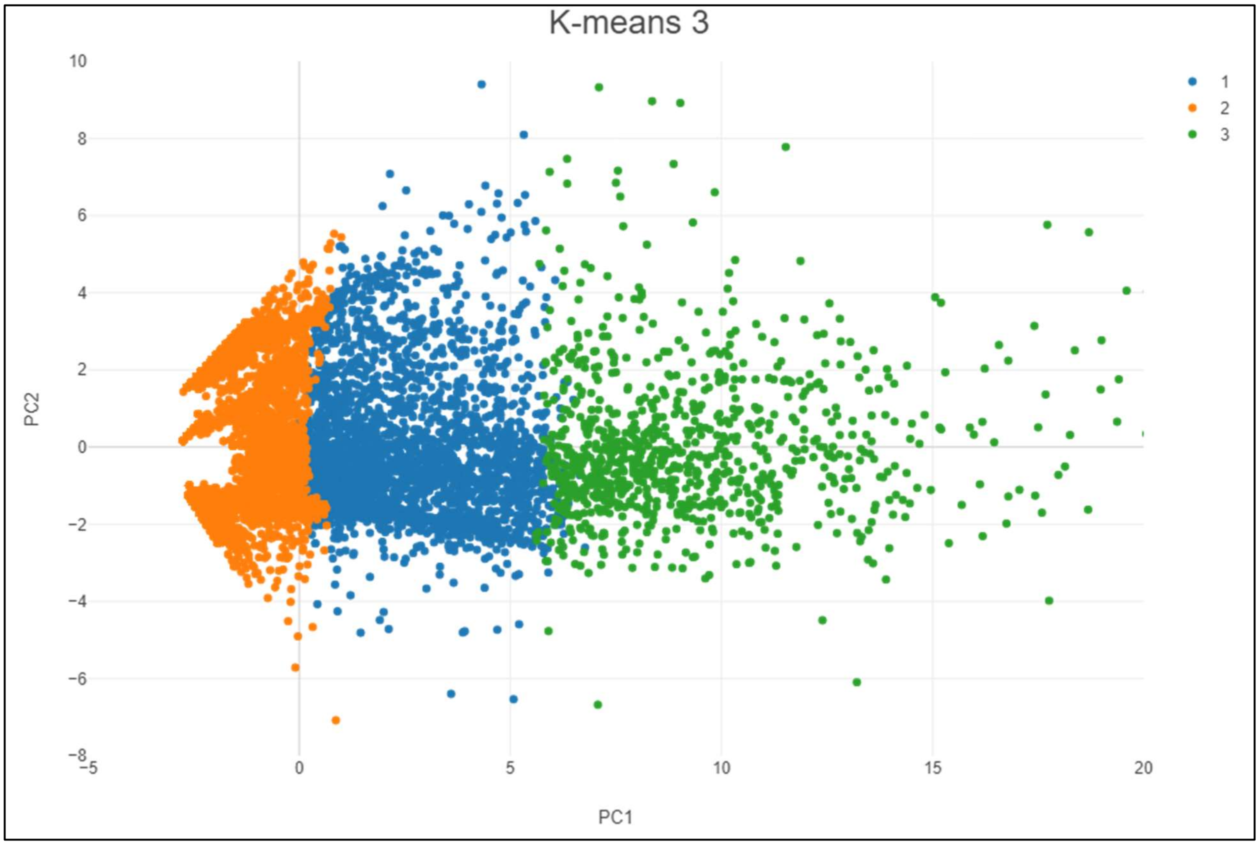


Figure 8: Fraud Plot

K-Means Plots



K-means = 3

Cluster by Count

Cluster	0	1	Total
1	2,097	825	2,922
2	3,633	5,313	8,946
3	665	257	922
Total	**6,395**	**6,395**	**12,790**

Cluster by Fraud Label %

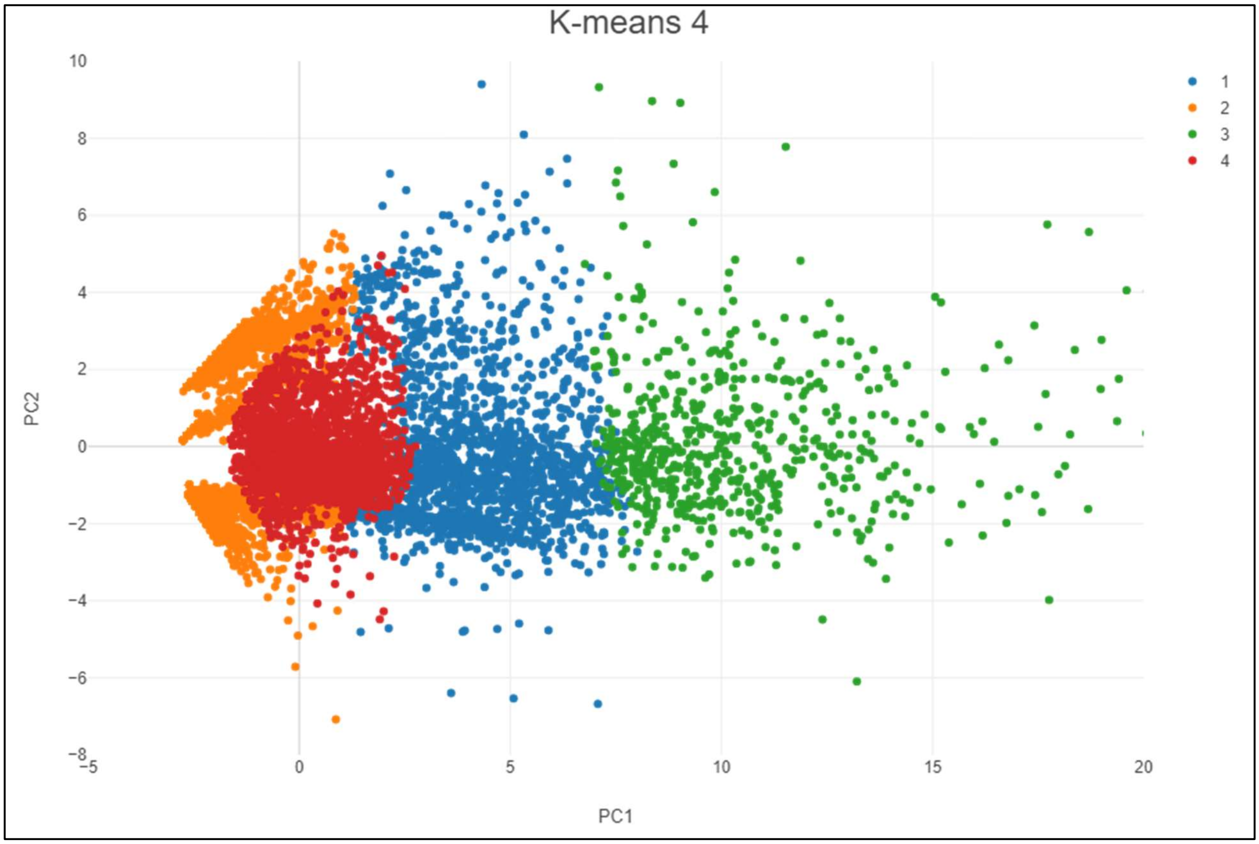
Cluster	0	1	Total
1	32.79%	12.90%	22.85%
2	56.81%	83.08%	69.95%
3	10.40%	4.02%	7.21%
Total	**100.00%**	**100.00%**	**100.00%**

Cluster by Row %

Cluster	0	1	Total
1	71.77%	28.23%	100.00%
2	40.61%	59.39%	100.00%
3	72.13%	27.87%	100.00%
Total	**50.00%**	**50.00%**	**100.00%**

Cluster by Total %

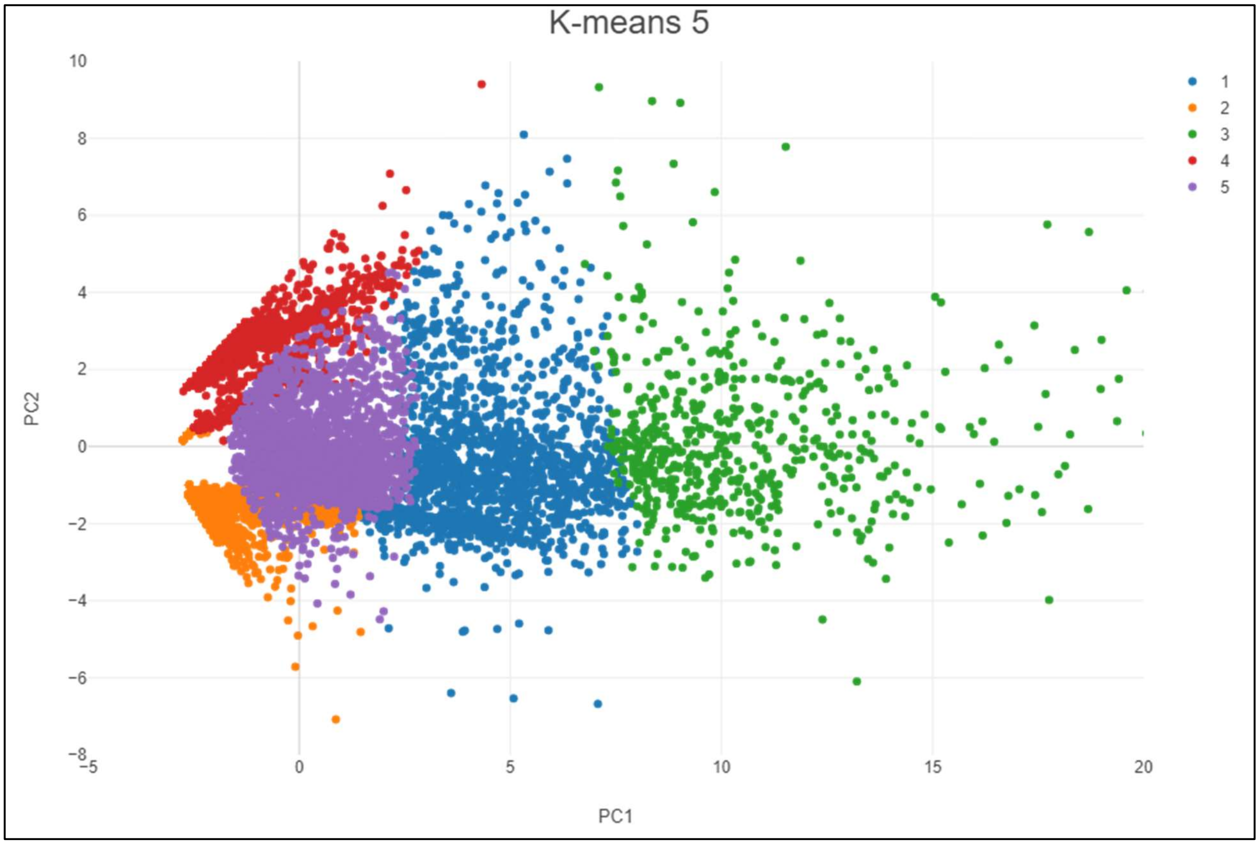
Cluster	0	1	Total
1	16.40%	6.45%	22.85%
2	28.41%	41.54%	69.95%
3	5.20%	2.01%	7.21%
Total	**50.00%**	**50.00%**	**100.00%**



K-means = 4

Cluster by Count				Cluster by Fraud Label %			
Cluster	0	1	Total	Cluster	0	1	Total
1	1,456	439	1,895	1	22.77%	6.86%	14.82%
2	3,038	4,450	7,488	2	47.51%	69.59%	58.55%
3	458	181	639	3	7.16%	2.83%	5.00%
4	1,443	1,325	2,768	4	22.56%	20.72%	21.64%
Total	**6,395**	**6,395**	**12,790**	**Total**	**100.00%**	**100.00%**	**100.00%**

Cluster by Row %				Cluster by Total %			
Cluster	0	1	Total	Cluster	0	1	Total
1	76.83%	23.17%	100.00%	1	11.38%	3.43%	14.82%
2	40.57%	59.43%	100.00%	2	23.75%	34.79%	58.55%
3	71.67%	28.33%	100.00%	3	3.58%	1.42%	5.00%
4	52.13%	47.87%	100.00%	4	11.28%	10.36%	21.64%
Total	**50.00%**	**50.00%**	**100.00%**	**Total**	**50.00%**	**50.00%**	**100.00%**



K-means = 5

Cluster by Count

Cluster	0	1	Total
1	1,381	404	1,785
2	2,457	1,875	4,332
3	433	174	607
4	654	2,605	3,259
5	1,470	1,337	2,807
Total	**6,395**	**6,395**	**12,790**

Cluster by Fraud Label %

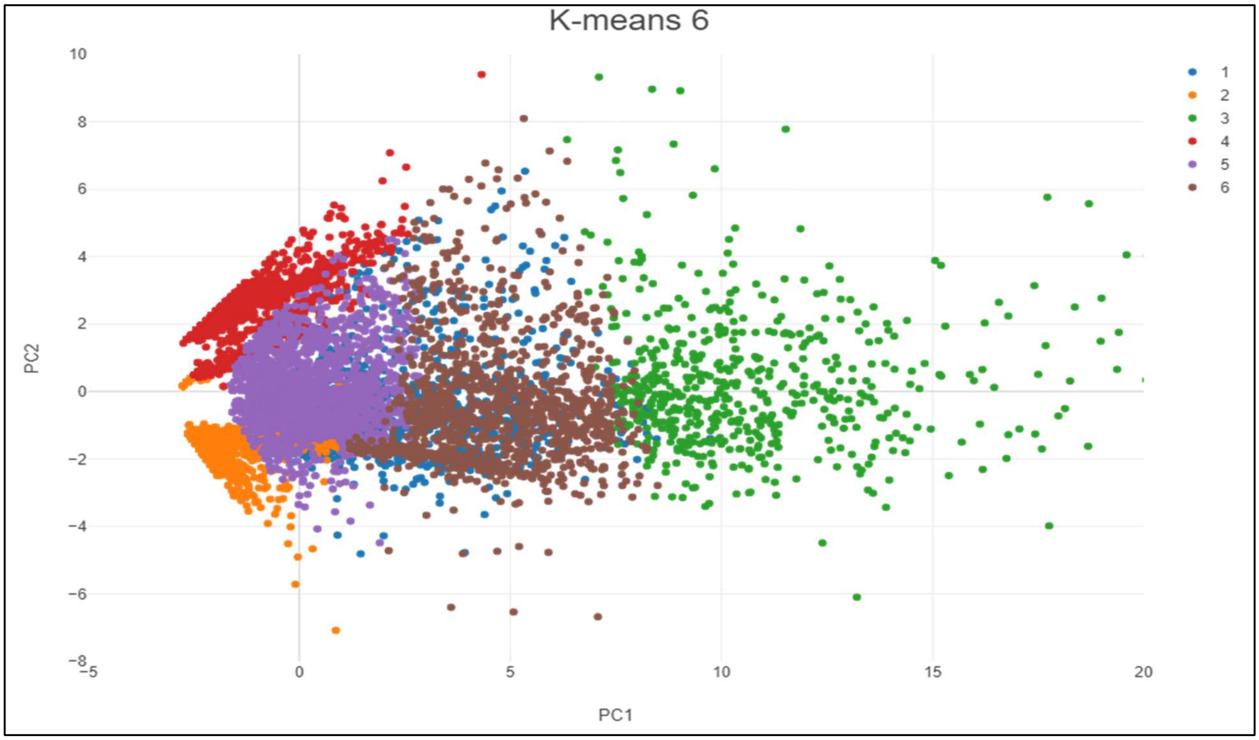
Cluster	0	1	Total
1	21.59%	6.32%	13.96%
2	38.42%	29.32%	33.87%
3	6.77%	2.72%	4.75%
4	10.23%	40.73%	25.48%
5	22.99%	20.91%	21.95%
Total	**100.00%**	**100.00%**	**100.00%**

Cluster by Row %

Cluster	0	1	Total
1	77.37%	22.63%	100.00%
2	56.72%	43.28%	100.00%
3	71.33%	28.67%	100.00%
4	20.07%	79.93%	100.00%
5	52.37%	47.63%	100.00%
Total	**50.00%**	**50.00%**	**100.00%**

Cluster by Total %

Cluster	0	1	Total
1	10.80%	3.16%	13.96%
2	19.21%	14.66%	33.87%
3	3.39%	1.36%	4.75%
4	5.11%	20.37%	25.48%
5	11.49%	10.45%	21.95%
Total	**50.00%**	**50.00%**	**100.00%**



K-means = 6

Cluster by Count

Cluster	0	1	Total
1	881	49	930
2	2,386	1,875	4,261
3	426	162	588
4	635	2,593	3,228
5	1,108	1,338	2,446
6	959	378	1,337
Total	**6,395**	**6,395**	**12,790**

Cluster by Fraud Label %

Cluster	0	1	Total
1	13.78%	0.77%	7.27%
2	37.31%	29.32%	33.32%
3	6.66%	2.53%	4.60%
4	9.93%	40.55%	25.24%
5	17.33%	20.92%	19.12%
6	15.00%	5.91%	10.45%
Total	**100.00%**	**100.00%**	**100.00%**

Cluster by Row %

Cluster	0	1	Total
1	94.73%	5.27%	100.00%
2	56.00%	44.00%	100.00%
3	72.45%	27.55%	100.00%
4	19.67%	80.33%	100.00%
5	45.30%	54.70%	100.00%
6	71.73%	28.27%	100.00%
Total	**50.00%**	**50.00%**	**100.00%**

Cluster by Total %

Cluster	0	1	Total
1	6.89%	0.38%	7.27%
2	18.66%	14.66%	33.32%
3	3.33%	1.27%	4.60%
4	4.96%	20.27%	25.24%
5	8.66%	10.46%	19.12%
6	7.50%	2.96%	10.45%
Total	**50.00%**	**50.00%**	**100.00%**

Cluster Findings

		Fraud % by Cluster						
	Categories	1	2	3	4	5	6	% Of Total*
K-means=3	UT, OP, & MD	28.23%	59.39%	27.87%	NA	NA	NA	NA
	OP & MD	41.53%	58.64%	25.74%	NA	NA	NA	NA
	UT & OP	74.80%	40.91%	21.64%	NA	NA	NA	13.51%
	UT & MD	30.63%	59.61%	26.07%	NA	NA	NA	NA
	UT	64.11%	53.12%	18.53%	NA	NA	NA	12.45%
	OP	78.11%	38.12%	28.06%	NA	NA	NA	32.15%
	MD	46.96%	56.62%	24.18%	NA	NA	NA	12.90%
K-means=4	UT, OP, & MD	23.17%	59.43%	28.33%	47.87%	NA	NA	14.82%
	OP & MD	22.44%	58.98%	29.17%	49.64%	NA	NA	14.57%
	UT & OP	75.26%	41.70%	33.81%	18.29%	NA	NA	46.39%
	UT & MD	24.18%	59.48%	27.51%	47.94%	NA	NA	15.17%
	UT	67.63%	54.44%	22.74%	20.75%	NA	NA	17.97%
	OP	80.83%	37.12%	28.26%	43.28%	NA	NA	28.14%
	MD	10.06%	57.94%	31.89%	53.67%	NA	NA	9.71%
K-means=5	UT, OP, & MD	22.63%	43.28%	28.67%	79.93%	47.63%	NA	39.44%
	OP & MD	27.65%	60.90%	29.40%	23.66%	49.64%	NA	8.33%
	UT & OP	62.02%	40.24%	30.90%	16.85%	76.31%	NA	42.70%
	UT & MD	23.96%	59.74%	29.57%	38.67%	48.05%	NA	11.49%
	UT	66.01%	44.02%	30.68%	16.84%	56.17%	NA	10.68%
	OP	48.79%	37.50%	23.47%	43.17%	82.98%	NA	33.63%
	MD	10.06%	60.27%	27.95%	35.00%	55.28%	NA	9.79%
K-means=6	UT, OP, & MD	5.27%	44.00%	27.55%	80.33%	54.70%	28.27%	32.51%
	OP & MD	6.49%	80.82%	27.20%	41.79%	55.09%	35.13%	34.88%
	UT & OP	64.92%	41.96%	14.30%	27.13%	76.87%	33.73%	38.20%
	UT & MD	11.30%	60.58%	25.98%	51.66%	54.14%	27.98%	8.16%
	UT	66.47%	57.89%	16.24%	22.16%	54.56%	39.81%	16.95%
	OP	48.00%	41.23%	32.15%	44.40%	83.16%	12.20%	34.89%
	MD	10.12%	61.88%	23.89%	28.24%	55.20%	43.43%	14.20%

Table 8: Fraud by Cluster

* Percentage of records that fall in 1st or 4th quartile

Category Findings

	Fraud % by Cluster						
Categories	1	2	3	4	5	6	% Of Total*
UT, OP, & MD	28.23%	59.39%	27.87%	NA	NA	NA	NA
UT, OP, & MD	23.17%	59.43%	28.33%	47.87%	NA	NA	14.82%
UT, OP, & MD	22.63%	43.28%	28.67%	79.93%	47.63%	NA	39.44%
UT, OP, & MD	5.27%	44.00%	27.55%	80.33%	54.70%	28.27%	32.51%
OP & MD	41.53%	58.64%	25.74%	NA	NA	NA	NA
OP & MD	22.44%	58.98%	29.17%	49.64%	NA	NA	14.57%
OP & MD	27.65%	60.90%	29.40%	23.66%	49.64%	NA	8.33%
OP & MD	6.49%	80.82%	27.20%	41.79%	55.09%	35.13%	34.88%
UT & OP	74.80%	40.91%	21.64%	NA	NA	NA	13.51%
UT & OP	75.26%	41.70%	33.81%	18.29%	NA	NA	46.39%
UT & OP	62.02%	40.24%	30.90%	16.85%	76.31%	NA	42.70%
UT & OP	64.92%	41.96%	14.30%	27.13%	76.87%	33.73%	38.20%
UT & MD	30.63%	59.61%	26.07%	NA	NA	NA	NA
UT & MD	24.18%	59.48%	27.51%	47.94%	NA	NA	15.17%
UT & MD	23.96%	59.74%	29.57%	38.67%	48.05%	NA	11.49%
UT & MD	11.30%	60.58%	25.98%	51.66%	54.14%	27.98%	8.16%
UT	64.11%	53.12%	18.53%	NA	NA	NA	12.45%
UT	67.63%	54.44%	22.74%	20.75%	NA	NA	17.97%
UT	66.01%	44.02%	30.68%	16.84%	56.17%	NA	10.68%
UT	66.47%	57.89%	16.24%	22.16%	54.56%	39.81%	16.95%
OP	78.11%	38.12%	28.06%	NA	NA	NA	32.15%
OP	80.83%	37.12%	28.26%	43.28%	NA	NA	28.14%
OP	48.79%	37.50%	23.47%	43.17%	82.98%	NA	33.63%
OP	48.00%	41.23%	32.15%	44.40%	83.16%	12.20%	34.89%
MD	46.96%	56.62%	24.18%	NA	NA	NA	12.90%
MD	10.06%	57.94%	31.89%	53.67%	NA	NA	9.71%
MD	10.06%	60.27%	27.95%	35.00%	55.28%	NA	9.79%
MD	10.12%	61.88%	23.89%	28.24%	55.20%	43.43%	14.20%

Table 9: Fraud by Category

* Percentage of total records that fall in 1st or 4th quartile

1.6 Discussion

PCA Findings

PCA was used to offset the high dimensionality impact of the dataset. By combining variables into a two-dimensional scatter plot using the first two principal components, the data was made suitable for visualization and interpretation. Based on an interest in examining a range of cluster values for comparison, k-means values between 3 and 6 were evaluated. The breakdown of principal components included variables from all three categories. Variables from Use Type and Operator Type groups were more uniform in terms of coefficient average and range. Metadata variable coefficients showed the greatest amount of variability and were notably lower in PC2. PC2 was heavily impacted by three variables in the first two categories. UT_ENT_EXT_RATIO and OP_BUS_RATIO had strong positive correlations of .469 and .506 respectively. OP_RAIL_RATIO had the strongest negative correlation and the highest absolute value of any variable in either of the first two principal components.

Plots and K-means Findings

To give a point of reference to the k-means clustering plots, an initial plot of Fraud vs. Not Fraud points was provided. While all modeling was performed without labels, the context of the research question requires that the labels be considered when determining the efficacy of the k-means clustering results. The fraudulent data points display a complex relationship with legitimate values when viewed as a 2-dimensional plot. It appears that from the perspective of PC1 and PC2 fraud does not form highly differentiated clusters, but rather overlaps the legitimate transactions with concentrated groups forming as the cluster count increases. There are plot areas that are clearly more likely to be fraud and areas that show clear tendencies toward non-fraudulent data points.

K-means =3 revealed a simple plot of ostensibly vertically banded data points. While the graph itself is of limited value, the concentration in cluster 2 is clear with 70% of the transactions and 83% of the total fraud being located there. This concentration is consistent across all 4 cluster quantities tested. Each cluster quantity shows a spread concentration around primarily negative values of PC1 where the majority of both overall data points and fraudulent data points are clustered. When k-means=4 the concentration shifts to 59% of the sample population and 70% of the fraud. For k=5 and k=6 the concentration splits into separate clusters but when combined still total 59% and 70% for population and fraud percentages respectively.

The area with the most ambiguity is centered at approximately the origin. For k=3 the clusters are insufficiently defined to capture a fraud differential in this region. However, when cluster values reach 4 and 5 the regions become more legible and show distributions of 52% for non-fraud. When 6 clusters are used this area is approximately 45% non-fraud. These clusters range from 20-22% of the total data point volume for k values of 4,5, & 6. This region offers the most difficulty in estimating whether a new data point should be labeled fraud or non-fraud.

The data also indicates clusters with higher concentrations of non-fraudulent data points. Mid-level values for PC1 were consistently clustered into groups that ranged between 72% for k=3 to 81% for k=6 for percentage of non-fraudulent transactions. While these points were clustered into a single group for k values 3-5, k=6 split this group into two clusters that when combined gave the approximately 81% non-fraud total. For data points with the highest values along the PC1 axis the 4 k-means variables consistently formed a cluster which was both the smallest cluster in terms of total data point volume and ranged from 71-72% non-fraudulent.

Excluding Use Type category variables caused the clusters to become less distinct when viewed from a fraud vs. non-fraud perspective. The k=3 cluster that previously accounted for 83% of fraud softened to 70%.

For PC1 values less than 0 with PC2 values ±1.5 units from the origin, the likelihood of fraud is very high. Plotted points closer to the origin are roughly evenly distributed in the sample data. As PC1 axis values increase, the likelihood of fraud diminishes. PC1 values are largely driven by total transaction levels over the test period with a lesser impact for daily transaction levels. As PC2 values increase so does the probability of fraud. Given that PC2 is largely driven by Operator Type, it can be inferred that greater OP_BUS_RATIO values result in elevated levels of fraud.

To address the impact of variable categories and their individual impacts on the effectiveness of clustering, each category/cluster combination was modeled and analyzed. Because of the binary nature of the fraud, effectiveness is signaled by substantially higher or lower values of fraud concentration while middling values indicate ambiguous clusters with less delineation between fraud/non-fraud records. A quartile approach was adopted to simplify the findings. K-means values and variable categories were screened for cluster fraud concentrations of less than 25% or greater than 75%. The results were consolidated into 2 tables (Tables 8 & 9). The highlighted cells indicate clusters that returned fraud percentages in the 1st or 4th quartiles (i.e., less than 25% or greater than 75%). These clusters present business stakeholders with more actionable data in terms of how to allocate resources for additional analysis and/or institute mitigation responses for factor combinations that are historically more fraudulent. The tables also provide the total percentage of records that fall into these high-risk clusters.

The highest percentages of population by k-means value were 46.39%, 42.70%, and 38.20% for k=4-6 respectively when considering Use Type and Operator Provider categories only. On average, k-means=4 provided the highest percentage of population across all category combinations. This is consistent with the PCA component weights which indicated that the top 3 factors of PCA2 were part of the Use Type and Operator Provider categories. To simplify, fare media with a history of high bus activity, low rail activity, and high entrance to exit ratios are the highest risk group for fraud.

1.7 Limitations & Future Research

The most significant limitation to this research is the imperfect nature of the dependent variable. Because Hotlisting is the only data marker provided for fraud in the transit data, researchers are forced to utilize it despite its shortcomings. The most troubling aspect of Hotlisting is that it is applied for effect as opposed to explanation. Hotlisting a card/ticket is the current method of disabling the media. Fare media can be disabled for several reasons including reported lost media, canceled promotional fare media, irregular transactions, fare stock control (i.e., sunsetting aging media), remote media replacement, and suspected/known fraud. Without a more directed approach to labeling fraud, any research centered on Hotlisted media will necessarily include some quantity of false positives.

A second concern of using Hotlisting is the impact on transactional data. During the course of this research, it was revealed that many of the fare media attributes available in the automated fare collection system are nullified when a card/ticket is Hotlisted. While meaningful data is still collected, a substantial amount of data is purged once the fare media is Hotlisted. This lost data could potentially be used to improve models to assist in the detection of current and future methods of fraud.

Other limitations of this research include that it was limited to a single transit authority for a specific window of time. With consideration of the transit fraud framework, multiple transit authorities are a prerequisite for exploring the varying impact of External Factors, Customer Factors, Transit Authority, and System Controls. New insights might be made using other public transit systems and/or for longer periods of time. Lastly there are additional unsupervised learning techniques that should be tested and compared to the k-means approach used in this research.

The difficulty of determining which transactions are legitimate vs fraudulent is compounded by the transactional overlap of the two record types. Transit fraud has a variety of forms such as counterfeit tickets, stolen tickets/cards, and altered media. Some of these forms may exhibit identical transactional behavior to legitimate fare media. This is illustrated in Figure 8, where both record types are shown to mingle and overlap across a broad spectrum of variable values. While this research focused on transaction records, some types of fraud may only be detectable by investigating the details of their creation, distribution, and/or purchase.

1.8 Conclusion

Unlike previous works, this study used actual transaction data from a U.S. top 10 transit authority to explore variable categories for determining transit fare media fraud. Additionally, a novel framework was introduced to provide a guide for future transit fraud research.

Based on a tolerance threshold of less than 25% or greater than 75%, this research showed that almost half of sample population could be effectively grouped into fraudulent vs. non-fraudulent clusters using k-means clustering. A k-mean value of 4 coupled with variable categories Use Type and Operator Provider generated the most fraud-differentiated clustering. By allowing investigators and analysts to quickly determine particularly high or low risk clusters, resources can be directed toward records with the highest probability of fraud.

These findings support the contention that unsupervised machine learning is suitable for producing meaningful transit fraud classification models based on transactional data. Transit fraud classification modeling offers transit systems a direct and effective means to mitigate the impacts of fare media fraud.

1.9 References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. Mis Quarterly, 1293-1327.

American Institute of Certified Public Accountants. Auditing Standards Board. (2002). Consideration of Fraud in a Financial Statement Audit:(supersedes Statement on Auditing Standards No. 82, AICPA, Professional Standards, Vol. 1, AU Sec. 316; and Amends SAS No. 1, Codification of Auditing Standards and Procedures, AICPA, Professional Standards, Vol. 1, AU Sec. 230," Due Professional Care in the Performance of Work," and SAS No. 85, Management Representations, AICPA, Professional Standard, Vol. 1, AU Sec 333). American Institute of Certified Public Accountants.

American Public Transportation Association. (2020). 2020 Public Transportation Fact Book.

APTA. (2007, September 27). Glossary of Transit Terminology. Retrieved January 6, 2020, from https://web.archive.org/web/20070927220938/http://www.apta.com/research/info/online/glossary.cfm

APTA. (2020, September 14). Ridership Report. Retrieved September 23, 2020, from https://www.apta.com/research-technical-resources/transit-statistics/ridership-report/

Barabino, B., Lai, C., & Olivo, A. (2020). Fare evasion in public transport systems: a review of the literature. Public Transport, 12(1), 27-88.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, 235-249.

BTS. (2017). Reports of Violent Crime, Property Crime, and Arrests by Transit Mode. Retrieved from https://www.bts.gov/archive/publications/national_transportation_statistics/2011/table_02_38

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). Scarff: a scalable framework for streaming credit card fraud detection with spark. Information fusion, 41, 182-194.

Chang, W. H., & Chang, J. S. (2010, June). Using clustering techniques to analyze fraudulent behavior changes in online auctions. In 2010 International Conference on Networking and Information Technology (pp. 34-38). IEEE.

Cheeseman, P. C., & Stutz, J. C. (1996). Bayesian classification (AutoClass): theory and results. Advances in knowledge discovery and data mining, 180, 153-180.

Cieslewicz, J. (2010), “The fraud square: societal influences on the risk of fraud”, paper presented at 2010 American Accounting Association Annual Meeting, San Francisco, CA, 31 July-4 August.

Cools, M., Fabbro, Y., & Bellemans, T. (2018). Identification of the determinants of fare evasion. Case studies on transport policy, 6(3), 348-352.

Cressey, D. R. (1953). Other people's money; a study of the social psychology of embezzlement.

Dauby, L., & Kovacs, Z. (2007). Fare evasion in light rail systems. Transportation Research Circular, (E-C112).

Delbosc, A., & Currie, G. (2016). Four types of fare evasion: A qualitative study from Melbourne, Australia. Transportation Research Part F: Traffic Psychology and Behaviour, 43, 254-264.

Donoho, S. (2004, August). Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 420-429). ACM.

Düzenlį, G. (2015). RFID card security for public transportation applications based on a novel neural network analysis of cardholder behavior characteristics. Turkish Journal of Electrical Engineering & Computer Sciences, 23(4), 1098-1110.

Fürst, E. W. M., & Herold, D. M. (2018). Fare Evasion and Ticket Forgery in Public Transport: Insights from Germany, Austria and Switzerland. Societies, 8(4), 98.

Guarda, P., Galilea, P., Handy, S., Muñoz, J. C., & de Dios Ortúzar, J. (2016). Decreasing fare evasion without fines? A microeconomic analysis. Research in Transportation Economics, 59, 151-158.

Issa, H., & Vasarhelyi, M. A. (2011). Application of Anomaly Detection Techniques to Identify Fraudulent Refunds. Available at SSRN 1910468.

Jurek, A., & Zakrzewska, D. (2008, October). Improving naïve Bayes models of insurance risk by unsupervised classification. In 2008 International Multiconference on Computer Science and Information Technology (pp. 137-144). IEEE.

Kalmár, A., Öllös, G., & Vida, R. (2011). Analysis of an Event Forecasting Method for Wireless Sensor Networks. Acta Universitatis Sapientiae-Electrical & Mechanical Engineering, 3.

Le Khac, N. A., & Kechadi, M. T. (2010, December). Application of data mining for anti-money laundering detection: A case study. In 2010 IEEE International Conference on Data Mining Workshops (pp. 577-584). IEEE.

Liu, R., Qian, X. L., Mao, S., & Zhu, S. Z. (2011, May). Research on anti-money laundering based on core decision tree algorithm. In 2011 Chinese Control and Decision Conference (CCDC) (pp. 4322-4325). IEEE.

Markantonakis, K., Tunstall, M., Hancke, G., Askoxylakis, I., & Mayes, K. (2009). Attacking smart card systems: Theory and practice. information security technical report, 14(2), 46-56.

Marks, J. (2009), Playing Offense in a High-Risk Environment, Crowe Horwath, NewYork, NY, available at: http://internalaudits.duke.edu/documents/articles_archive/PlayingOffense WhitePaper4_09.pdf (accessed 29 December 2014).

Mass Transit Network. (2020). About Mass Transit. Retrieved January 6, 2020, from https://masstransit.network/about-mass-transit

May, P. (2019, October 02). Video shows mass BART fare evasion in Oakland after 'Rolling Loud Festival'. Retrieved September 23, 2020, from https://www.eastbaytimes.com/2019/10/01/video-shows-mass-bart-fare-evasion-in-oakland-after-rolling-loud-festival/

Merja, J., & Shah, S. (2013, June). Simplified secure wireless railway for public transport. In 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks (pp. 77-82). IEEE.

Minitab, LLC (2019). "Cluster K-Means," Retrieved January 6, 2020, from https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/perform-the-analysis/enter-your-data/#standardize-variables

Mintchik, N., & Riley, J. (2019). Rationalizing Fraud: How Thinking Like a Crook Can Help Prevent Fraud. The CPA Journal, 89(3), 44-50.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), 559-569.

Nguyen, T. (2019, November 12). Fare evasion costs cities millions. But will cracking down on it solve anything? Retrieved September 23, 2020, from https://www.vox.com/the-goods/2019/11/12/20959914/fare-evasion-costs-cities-millions

Sabau, A. S. (2012). Survey of clustering based financial fraud detection research. Informatica Economica, 16(1), 110.

Salis, S., Barabino, B., & Useli, B. (2017). Segmenting fare evader groups by factor and cluster analysis. WIT Transactions on The Built Environment, 176, 503-515.

Smith, G. and Crumbley, D. (2009), "How divergent are pedagogical views toward the fraud/ forensic accounting curriculum?", Global Perspectives on Accounting Education, Vol. 6 No. 1, pp. 1-24.

Thakare, Y. S., & Bagal, S. B. (2015). Performance evaluation of K-means clustering algorithm with various distance metrics. International Journal of Computer Applications, 110(11), 12-16.

Thakare, Y. S., & Bagal, S. B. (2015). Performance evaluation of K-means clustering algorithm with various distance metrics. International Journal of Computer Applications, 110(11), 12-16.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. International Journal of Digital Accounting Research, 11.

Troncoso, R., & de Grange, L. (2017). Fare evasion in public transport: A time series approach. Transportation Research Part A: Policy and Practice, 100, 311-318.

Virdhagriswaran, S., & Dakin, G. (2006, August). Camouflaged fraud detection in domains with complex relationships. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 941-947). ACM.

Wagner, P. (2019, November 12). Metro, DC Council at odds over fare evasion: Order to not write tickets now open-ended. Retrieved September 23, 2020, from https://www.fox5dc.com/news/metro-dc-council-at-odds-over-fare-evasion-order-to-not-write-tickets-now-open-ended

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

Wakoli, L. W. (2014). Application of the k-means clustering algorithm in medical claims fraud/abuse detection (Doctoral dissertation).

Wolfe, D. and Hermanson, D. (2004), "The fraud diamond: considering the four elements of fraud", The CPA Journal, Vol. 74 No. 12, pp. 38-42.

Wu, J., Xiong, H., & Chen, J. (2010). COG: local decomposition for rare class analysis. Data Mining and Knowledge Discovery, 20(2), 191-220.

1.10 Appendix

K-means clustering convergence

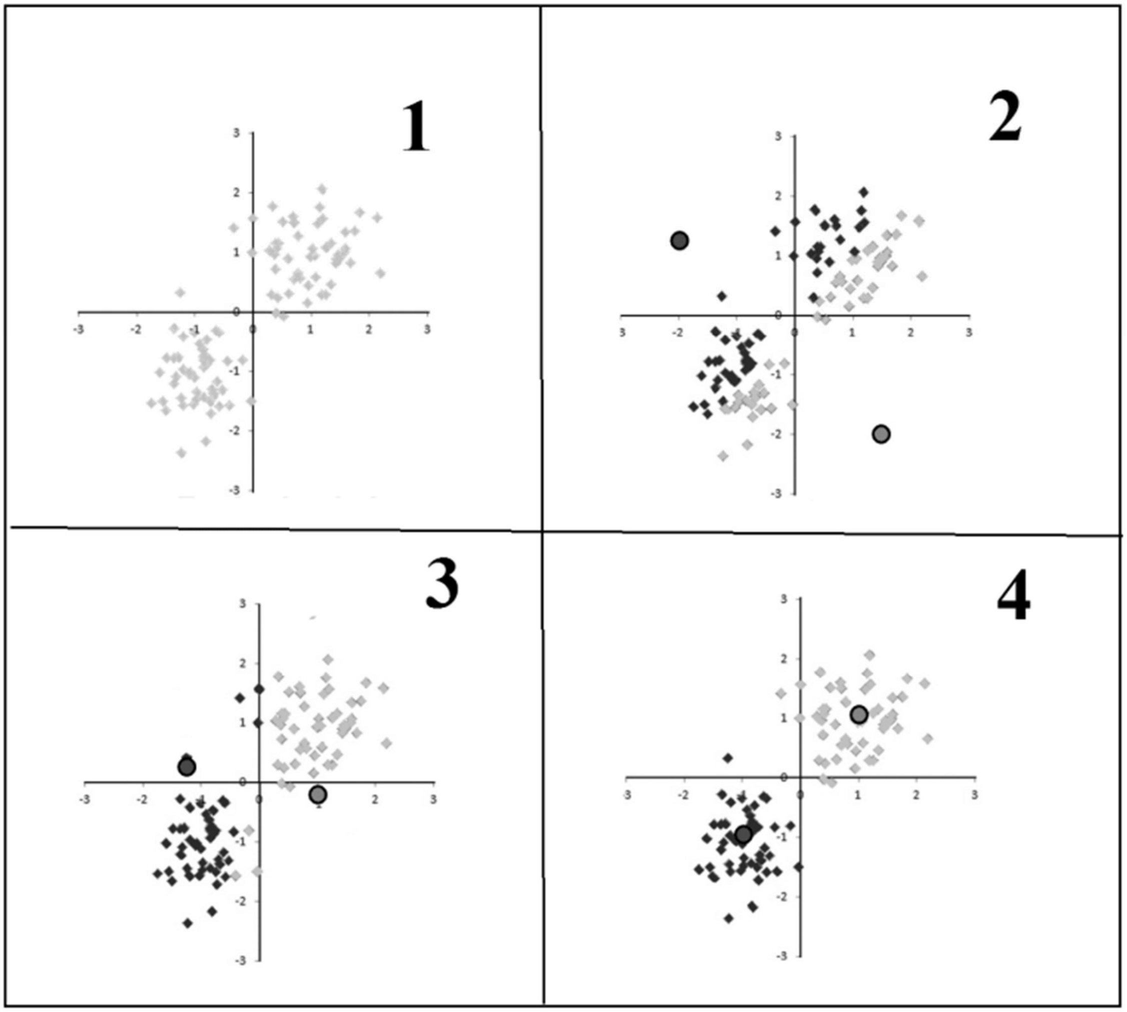


Figure 9: K-means Clustering Convergence

(Adopted from Kalmár, Öllös, & Vida, 2011)

K-means clustering results

Dataset	Recognition Rate (%) Number of clusters formed				
	3	5	10	14	16
Iris	89.33	98	98	99.33	99.33
Wine	69.6	72.64	75	75.87	79.74
Vowel	-	-	70.72	72.02	72.44
Ionosphere	81.58	84.41	88.07	88.34	90.01
Crude Oil	61.4	78.33	85.64	89.47	92.1

Table 10: K-means Clustering Results

(Adopted from Thakare & Bagal, 2015)

Paper 2: Classification of Transit Fraud via Supervised Machine Learning

Abstract –Globally, public transit provides billions of rides each year. The scale of transit related fraud is estimated to be in the hundreds of millions annually. In this research the significant challenges in transit fraud research of access to data, constantly evolving fraud techniques, and coping with the highly imbalanced nature of fraud data are presented and addressed. Using supervised machine learning methods (logistic regression, k-nearest neighbor, Naïve Bayes, & random forests) coupled with SMOTE to account for data imbalance, transit fraud classification model accuracy rates varied from approximately 80-97%. Corresponding ROC AUC scores ranged from approximately 75-87%. The highest performing models were KNN and random forests using bi-weekly and monthly data sets.

Keywords - Public transit, fraud, fare media, machine learning, supervised learning

2.1 Introduction

The transit industry has experienced explosive levels of growth during the last 50 years. During this period of growth, public transportation has evolved from simple bus lines to highly automated mass transit organizations. Like any industry, public transit suffers from the effects of those who seek to defraud the system. The scale, methods, and complexity of transit fraud vary widely both within and between systems. To mitigate criminal efforts, researchers have proposed and tested a variety of anti-fraud methods. Some of the most promising solutions involve machine learning, which is well-suited to address the unique challenges of transit fraud. To understand the motivation behind transit fraud research, it is important to establish the scale of public transit, the prevalence of transit fraud, and the estimated impacts of transit fraud. Once this context has been established, the data obstacles, relevant criminology considerations, and the roles of classification and machine learning can be more fully appreciated.

Scale of Public Transit

Public transit authorities consist of a variety of transportation modes. The largest segments of land-based systems are rail and bus. To better understand the definition and scope of the components, the International Association of Public Transport (UITP) defines a metro (rail) system as follows:

"Metropolitan railways are urban, electric transport systems with a high capacity and a high frequency of service. Metros are totally independent from other traffic, roads or pedestrians. They are consequently designed for operations in tunnels, viaducts or on surface level but with physical separation. Metropolitan railways are the optimal public transport mode for a high capacity line or network service. Some systems run on rubber-tires but are based on the same control-command principles as steel-wheel systems. In different parts of the world metro systems are also known as the underground, subway or tube." (ERRAC & UITP, 2014).

In 2017, the metro/rail portion of public transit systems were operating in 56 countries, 178 cities, and carried upwards of 53 billion passengers per year. This segment experienced a 19.5% growth rate between 2012 and 2017 (UITP, 2018).

Complementing these rail systems are bus systems comprised of a variety of buses, vans, trolleys etc. of different capacities and vehicle lengths that use existing roads. In the case of bus rapid transit systems, vehicles are sometimes afforded dedicated lanes and preferential traffic signal arrangements. A 39-country survey in 2015 revealed that the annual modal distribution of trips for participating systems was heavily dominated by bus with 63% of all trips and an estimated 153 billion trips provided (UITP, 2017).

To support such massive transit operations, an enormous amount of public funding is necessary. Total public funding for transit in the U.S. in 2015, was $24 billion with approximately $12.2 billion each from federal and state funds (Bureau of Transportation Statistics, 2015). This funding is augmented by passenger fares, other transit authority revenues, and local assistance. The total funding for U.S. public transportation in 2015 was approximately $68.3 billion. Total revenues generated from passenger fares for the same period was $15.9 billion (Hughes-Cromwick & Dickens, 2018).

Because passenger fares afford the greatest opportunity for transit agencies to directly impact revenues, it is critical that fare capture rates are high. The farebox recovery ratio is the percentage of a trip's direct cost that is recovered via passenger fares (NTD, 2019). The recovery ratio varies by mode of transportation, but according to NTD findings, never reaches 100%. In 2018, the highest recovery rate by mode was vanpools at 73.6% while heavy rail and commuter bus were only 61.1% and 47.9% respectively (NTD, 2019). Given that revenues are substantially lower than operating expenses and that passenger revenues are the most direct route to close the revenue deficit, fare capture is critical to public transit sustainability. Even relatively moderate levels of transit fraud can undermine the financial viability of a transit system.

Impact of fraud

Transit fraud is committed by fare evasion or fare media fraud. Fare evasion is the act of avoiding detection, and consequently payment, while entering or riding the transit system. Fare media fraud is entering or riding the system under seemingly legitimate means while actually using theft related, modified, or counterfeit media. Typical transit operator responses to transit fraud include increased or enhanced physical controls (i.e., updated gates/turnstiles), upgraded ticketing systems, use of ticket inspectors, increased fine levels and/or fine enforcement, and attempts to optimize fare levels, service levels, and customer satisfaction (Delbosc & Currie, 2019).

While the scale of transit fraud varies widely between systems, an international survey of 31 systems in 18 countries found an average fare evasion rate of greater than 4% (Bonfanti &Wagenknecht, 2010). Some systems report fraud levels as high as 25% or more (Troncoso & de Grange, 2017). The economic impact of transit fraud is extensive, and the scale can be shocking.

A 2011 study of New York City Transit, excluding certain customer types and using an adjusted fare average method, estimated losses at greater than $23 million per year (Reddy, Kuhls, & Lu, 2011). A more recent report on New York based MTA (Metropolitan Transit Authority), estimates that fare evasion in 2018 resulted in a $240 million loss (WABC, 2019).

Table 11 shows examples of reported transit fraud (fare evasion and fare media fraud) as published by a variety of news outlets and transit industry publications. These values are often submitted as an educated guess when a transit official or law enforcement representative is questioned about the estimated impact of fraud involved in recent or ongoing investigations. While all forms of fraud effect the governing authority, some forms also directly impact other customers (e.g., using stolen credit cards to buy fare media). Some research even asserts that fare evasion increases the customer perception that transit is unsafe (Reddy, Kuhls, & Lu, 2011). The result can be the catalyst for a general distrust of public transit and brand erosion.

Fare evasion is inherently difficult to track and estimate because of the lack of recorded transaction details (Reddy, Kuhls, & Lu, 2011). Alternatively, fare media fraud involving theft, counterfeiting, data manipulation etc. potentially leaves the necessary digital clues for detection and countermeasure development. Despite the superior data trails produced by fare media fraud, it is far less commonly studied than fare evasion. Locating a transit authority willing to provide access to the necessary data is just one of several difficulties common to fraud research.

Transit System	Year	Estimated Fraud Losses
New York	2019	$300,000,000
New York	2018	$240,000,000
London	2019	$200,000,000
Paris	2015	$97,000,000
Toronto	2020	$75,000,000
Paris	2019	$68,000,000
Washington D.C.	2020	$40,000,000
Melbourne	2005-2011	$24,000,000

New York	2011	$23,000,000
London	2019	$21,000,000
San Francisco	2018	$19,000,000
Barcelona	2018	$10,000,000
Helsinki	2020	$9,700,000
Bern	2018	$7,600,000
Toronto	2014	$5,000,000
New Jersey	2012	$3,000,000
Dallas	2002	$2,200,000
Gold Coast	2018	$1,400,000
Santiago	2016	$1,000,000
Guangzhou	2019	$564,000
San Francisco	2011-2016	$500,000
Beijing	2019	$61,000

Table 11: Transit Fraud Examples

Challenges in Fraud Research

Fraud research typically encounters one or more fraud-specific data obstacles. Three of the most common issues are gaining access to the data, dealing with evolving fraud techniques, and coping with the highly imbalanced nature of fraud data. Gaining access to detailed transaction data can be difficult as companies seek to safeguard customer transaction details and suppress news involving operational fraud. Major data breaches are often widely reported and closely followed by the public. Backlash against companies who compromise customer data can be significant. Recent breaches include Yahoo (2016), Marriott (2018), LinkedIn (2012), Equifax (2017), and eBay (2014) and totaled 3.8 billion compromised records (Privacy Rights Clearinghouse, 2019). In most instances, the standard practice is to not discuss or release to the public any details related to ongoing criminal investigations. The concern is that any information divulged may inadvertently damage the case, generate additional negative attention from customers/stakeholders, or provide future fraud perpetrators information that assists their efforts to defraud the company and/or avoid detection.

In addition to scarcity of data, fraud methods are constantly refined in an attempt to frustrate detection techniques. Van Vlasselaer et al. (2016), note that fraud is time-evolving, well-considered, and organized. Jenson (1997), calls criminals "intelligent adversaries" and warns that they are highly adaptive and can quickly change tactics to avoid detection systems. Effective solutions for detecting fraud in complex and shifting environments require a level of adaptability that exceeds traditional rule-based methods which can be difficult to implement and maintain (Kou et al. 2004).

The third challenge is shared with other research areas (e.g., medical diagnosis, intrusion detection, text classification, and risk management), and deals with imbalanced data (Chawla, Japkowicz, & Kotcz, 2004). Imbalanced data refers to a set of records where two or more classes of observation occur at different rates. Typically, the minority class is the focus. This is true for essentially all types of fraud research. In a normal system, fraud occurs at a fraction of the rate of legitimate transactions. The resulting classes are therefore highly skewed. As the balance of data begins to reach more extreme levels, some detection methods begin to suffer.

Criminology Considerations

A review of criminology research can help explain some of the factors that contribute to fraud. By understanding how variables rooted in psychology, sociology, and criminology interact and intersect with conditions that cause or accelerate fraud, researchers can work to build mitigating factors. Because of the variety of transportation modes and the assortment of infrastructure elements represented within those modes, it is difficult to create universal transit fraud countermeasures. The Fraud Triangle Theory proposed by Donald Cressey is commonly referenced in discussions around fraud detection methodologies. The Fraud Triangle Theory states that pressure, rationalization, and opportunity combine to influence the rate of fraud. Another significant contributor from the criminology discipline, is the General Deterrence Theory.

The General Deterrence Theory (GDT) is based on the principle that rational actors attempt to maximize their individual satisfaction while simultaneously avoiding risk and/or negative consequences. The GDT suggests that if the punishment for a crime is applied quickly, severely, and with enough certainty, a rational person will weigh the benefits vs. rewards and opt not to commit the crime. Deterrence theory is rooted in the collective works of Thomas Hobbs (1588-1678), Cesare Beccaria (1738-1794), and Jeremy Bentham (1748-1832) (Hobbes, 2010). Based on these principles, developing techniques that produce higher rates of severity, likelihood of detection, or detection speeds will increase the compliance rate. Creating fraud countermeasures that satisfy as many of these criteria as possible increases the effectiveness of the solution. One promising solution is classification via machine learning.

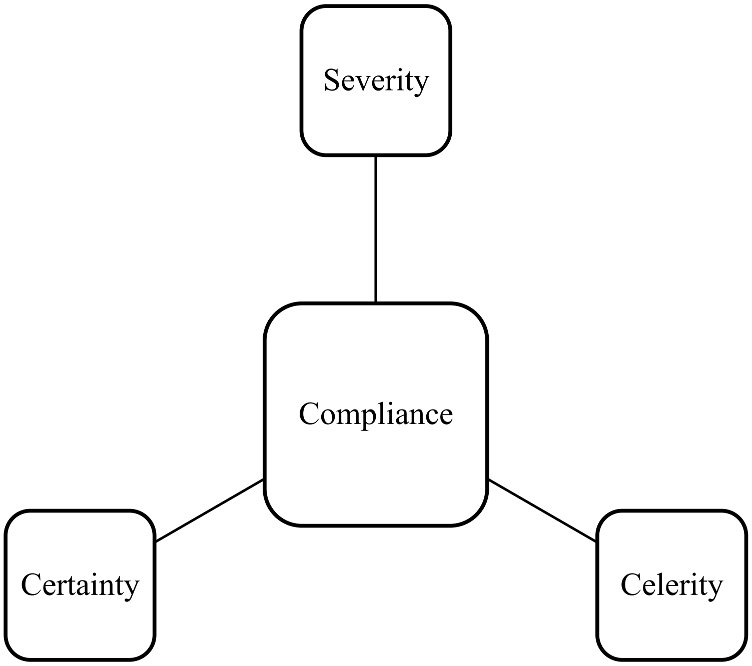


Figure 10: General Deterence Theory

Classification / Machine Learning

In its simplest sense, a classification model attempts to draw on labeled observations to determine which of the identified categories or classes new data points are most likely to be. In the case of fraud, labeled observations (fraud vs. legitimate) are loaded into the model along with a group of independent variables. Depending on the model selected, various statistical/

mathematical techniques are applied to produce the most likely classification of “fraud” or “not fraud”.

Machine learning can be described as the search among eligible solutions, with input from training experience, to optimize the given performance metric (Jordan & Mitchell, 2015). There is a rich body of research focused on the application of machine learning methods to detect credit card and electronic fraud (Bolton & Hand, 2001), (Brause, Langsdorf, & Hepp, 1999).

Advances in transit related technology have resulted in intelligent transportation systems utilizing smart cards and detailed transaction logging. These systems produce and store large repositories of data highly suited to the application of machine learning classification techniques. Fare media known or suspected to be associated with fraud or theft can be labeled in the system using a method called Hotlisting. When a fare media product is flagged as Hotlisted in the system, the next interaction with a system component (e.g., station gate, vending machine, or bus payment terminal) will result in the card being deactivated. Thus, Hotlisting is functionally comparable to a bank deactivating a reported stolen or lost credit card. Because of the cataloged transaction details, the Hotlisted media has been simultaneously deactivated and labeled, thus becoming an ideal candidate for supervised learning classification methods.

While the fluid nature of fraud makes it difficult to detect, it also makes it a highly suitable candidate for machine learning methods. Unlike rule-based detection methods, which rely on a catalog of static trigger scores or events, machine learning techniques offer a means to utilize new training data to frequently update classifiers. Many modern transit authorities utilize fully networked systems that log highly detailed transactions records. As trends in fraud shift over time, these records would allow machine learning methods to continuously recalibrate model parameters and training data sets to maintain the model's efficacy. Examples of appropriate classification methods for transit fraud includes logistic regression, k-nearest neighbor, Naïve Bayes, and random forest classifiers.

Research Goal

A wide variety of machine learning models have been studied from a fraud detection/classification perspective. This study seeks to extend the existing body of fraud and machine learning research by specifically exploring how machine learning classification models perform from a public transit perspective. Machine learning options address some of the principal components of well-established criminology theories. Two of these components are the certainty of being caught and the speed of discovery. Prior research shows that machine learning models can achieve accuracy scores of greater than 90% when predicting some types of fraud. The learning aspect also shortens the effectiveness timespan of new fraud techniques. Considering the well documented financial impacts of fraud, and because lost fares are unrecoverable, it is essential to detect fraud as early and accurately as possible. This research focuses specifically on the under-represented area of transit media fraud. Emphasis is placed on solutions that meet the joint requirements of effectiveness and timeliness. This work seeks to identify which models yield the strongest results when classifying fraud in a transit setting.

Research question 1. *Can machine learning models effectively classify transit fare media fraud?*

2.2 Literature Review

The following sections discuss the relevant research on the definition and impact of fraud, imbalanced data, and machine learning models.

Fraud

A review of the literature reveals a number of definitions for fraud. Hill describes fraud as "the intentional use of deceit, a trick or some dishonest means to deprive another of his/her/its money, property or a legal right" (Hill, 2005). Black's Law Dictionary defines it as "a knowing misrepresentation of the truth or concealment of a material fact to induce another to act to his or her detriment" (Garner, 2004). In 2002, the Auditing Standards Board of the American Institute of Certified Public Accountants issued a statement of auditing standards for fraud. Commonly referred to as SAS 99, the statement emphasizes the difference between error and fraud as the presence of intent (AICPA, 2002). The Federal Bureau of Investigation defines fraud as *"the intentional perversion of the truth for the purpose of inducing another person or other entity in reliance upon it to part with something of value or to surrender a legal right. Fraudulent conversion and obtaining of money or property by false pretenses. Confidence games and bad checks, except forgeries and counterfeiting, are included"*. The common elements in these and other popular definitions are the inclusion of intent and deceit. These characteristics are evident in transit fraud.

Private and public interests in fraud detection/prevention create pressure that ensures fraud research will be produced frequently and in high volumes. In a review of fraud research as a subset of academic publishing, it was found that the number of published fraud related articles increased from 140 in 1995 to 910 in 2016 (Gantman & Zinoviev, 2017). Several literature reviews and research surveys have focused on the associated detection methodologies and techniques. As machine learning techniques have grown in popularity and applicability, their emphasis in academic literature has also grown.

At the individual level, the scale of fraud can be surprisingly large. The Federal Trade Commission reports that in 2019 there were approximately 1.7 million consumer reports of fraud totaling $1.9 billion in total fraud losses. This total indicated a $300 million increase from the 2018 totals. Identity theft, imposter scams, and telephone/mobile services were the top 3 reported categories (FTC, 2020). At the corporate level, the impact of fraud can be shocking.

PwC's Global Economic Crime and Fraud Survey is an annual corporate survey of 5,000+ participants, the majority of participants are C-suite members in companies with greater than $10M in global revenues. Their 2020 survey found that U.S. losses were approximated at $42B and were most commonly either customer fraud, cybercrime, asset misappropriation, or bribery/corruption (PwC, 2020).

Bolton & Hand (2002), reviewed several subsets of fraud where statistical and data analytic tools were used effectively. A summary of the research and techniques was described for credit card fraud, money laundering, telecom fraud, computer intrusion and medical & scientific fraud (Bolton & Hand, 2002). Due in part to a recent increase in large scale financial fraud, extensive overviews of financial fraud were conducted by Nqai et al. (2011), Hogan et al. (2008), and Trompeter et al. (2013). Phua et al. (2010), produced an exhaustive review of fraud literature including analysis of sample sizes, performance measures, and methods/techniques by fraud category. They critiqued their findings for a shortage of examples of implemented research, studies utilizing temporal or spatial information, and research utilizing faster but simpler algorithms.

Machine learning has been incorporated into many instances of fraud detection and deterrence research. Sharma & Panigrahi (2013), reviewed the literature on financial accounting fraud. They reported that neural networks, decision trees, Bayesian networks, k-nearest neighbor, support vector machines, logistical regression, and other techniques were represented in the fraud literature between 1995 and 2011. Likewise, a review of over 50 fraud studies between 2004-2014 revealed techniques including Bayesian belief networks, genetic algorithms, text mining, response surface methodology, neural networks, logistic regression, group method of data handling, support vector machines, decision trees, hybrid methods, self-organizing maps, fuzzy logic, and artificial immune systems to detect credit card and financial statement fraud (West & Bhattacharya, 2016). Despite the widespread and well documented use of machine learning in other domains, insufficient research has been conducted around its use in public transit fraud studies.

Fraud research involves numerous challenges for researchers including the use of private/sensitive information, the changing nature of fraud, and the use of imbalanced data. Bolton & Hand (2002), note that advances in fraud detection are inhibited by the reluctance of firms to share data for fear that criminals will exploit the information to further enhance their capabilities to commit fraud. Reluctance to publicize sensitive information, the need to maintain consumer confidence, and a desire to safeguard market value for stakeholders, combine to create a scarcity of fraud test data for researchers. This was reiterated in later research that noted that a chief criticism of fraud detection research is the scarcity of publicly available data (Phua, Lee, Smith, & Gayler, 2010). While even general transit fraud research is scarce, those based on publicly available transit fraud data sets suitable for machine learning studies are especially rare.

Imbalanced Data

Imbalanced data is understood to be data that exhibits high levels of skewness or unequal distribution among classes. Fraud detection is focused on identifying relatively few instances of deviant behavior among thousands or even millions of legitimate transactions. This imbalance of classes can create misleading results from machine learning techniques, as choosing the majority class as the default prediction will often produce highly accurate if not precise results. To achieve this goal, a variety of statistical techniques can be employed.

Several classification methods, including decision trees, backpropagation neural networks, Bayesian classification, support vector machines, association classification, and k-nearest neighbor, have been shown to exhibit suboptimal performance when applied to imbalanced datasets (Sun, Wong, & Kamel, 2009). He and Garcia (2009), detail a number of techniques for dealing with imbalanced data, including sampling methods, cost-sensitive methods, kernel-based methods, and active learning methods (He & Garcia, 2009).

Over-sampling the minority class while simultaneously under-sampling the majority class was explored by Ling & Li (1998). Under-sampling the majority class is common approach to dealing with imbalanced data. While there may be redundant data points in the majority class, there can be disadvantages to removing the data. In the case of support vector machines, which rely on support points near the decision boundary, under-sampling risks omitting key data points (He & Garcia, 2009).

SMOTE (Synthetic Minority Over-sampling Technique) is one popular resampling technique. A 2002 study of imbalanced data introduced SMOTE, which over-samples the minority class by creating synthetic examples (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In SMOTE, the minority class is over-sampled by taking each sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen (Chawla et al., 2002). The majority class is simultaneously undersampled so that a balanced dataset is produced.

Machine Learning Models

This section reviews the various models utilized and examples of associated academic work. The following table provides a brief description of the models selected and examples of associated fraud research. It is important to note the absence of transit related fraud research being conducted with these models.

Model	Definition	Fraud Research Examples
Logistic Regression	Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several X's to a dichotomous dependent variable (Kleinbaum, Dietz, Gail, Klein, & Klein, 2002).	Bell & Carcello (2000) – Fraudulent financial reporting Spathis (2002) – Fraudulent financial statements Spathis et al., (2002) – Factors of fraudulent financial statements Owusu-Ansah et al. (2002) – Corporate fraud detection in New Zealand Guoxin et al. (2007) – Accounting fraud detection Yuan et al. (2008) – Impacts of compensation and competition on fraud Perols (2011) – Financial statement fraud detection
K Nearest Neighbor	Unclassified data points are assigned to the class represented by a majority of its k nearest neighbors in the training set (Fix & Hodges, 1951).	Kotsiantis et al., (2006) – Fraudulent financial statements Yeh (2009) – Credit card default prediction Senator et al., (1995) – Money laundering He, Graco, & Yao (1999) – Medical fraud
Naïve Bayes	Most likely class is assigned to a data point described by its feature vector. To simplify the calculation an assumption of feature independence is made (though the reality is often very different). Highly successful in application even when compared to more complicated techniques. (Rish, 2001).	Viane et al., (2004) – Insurance claim fraud Balaniuk (2012) – Government audit Phua (2004) – Fraud detection Yeh (2009) – Credit card defaults Panigrahi (2009) – Credit card fraud detection
Random Forest	Random Forest is a meta estimator that fits a given number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (scikit-learn documentation).	Liu, Chan, Kazmi, & Fu (2015) – Financial fraud detection Bhattacharyya et al., (2011) – Credit card fraud Whiting et al., (2012) – Management fraud Patel et al., (2019) – Financial statement manipulation Carneiro et al., (2017) – Credit card fraud in e-tail

Table 12: Machine Learning Fraud Research

2.3 Methodology

Model

The research model for this study is based on the Transit Fraud Framework and the primary emphasis is on variable selection, testing length impacts, and model performance.

Transit fraud framework

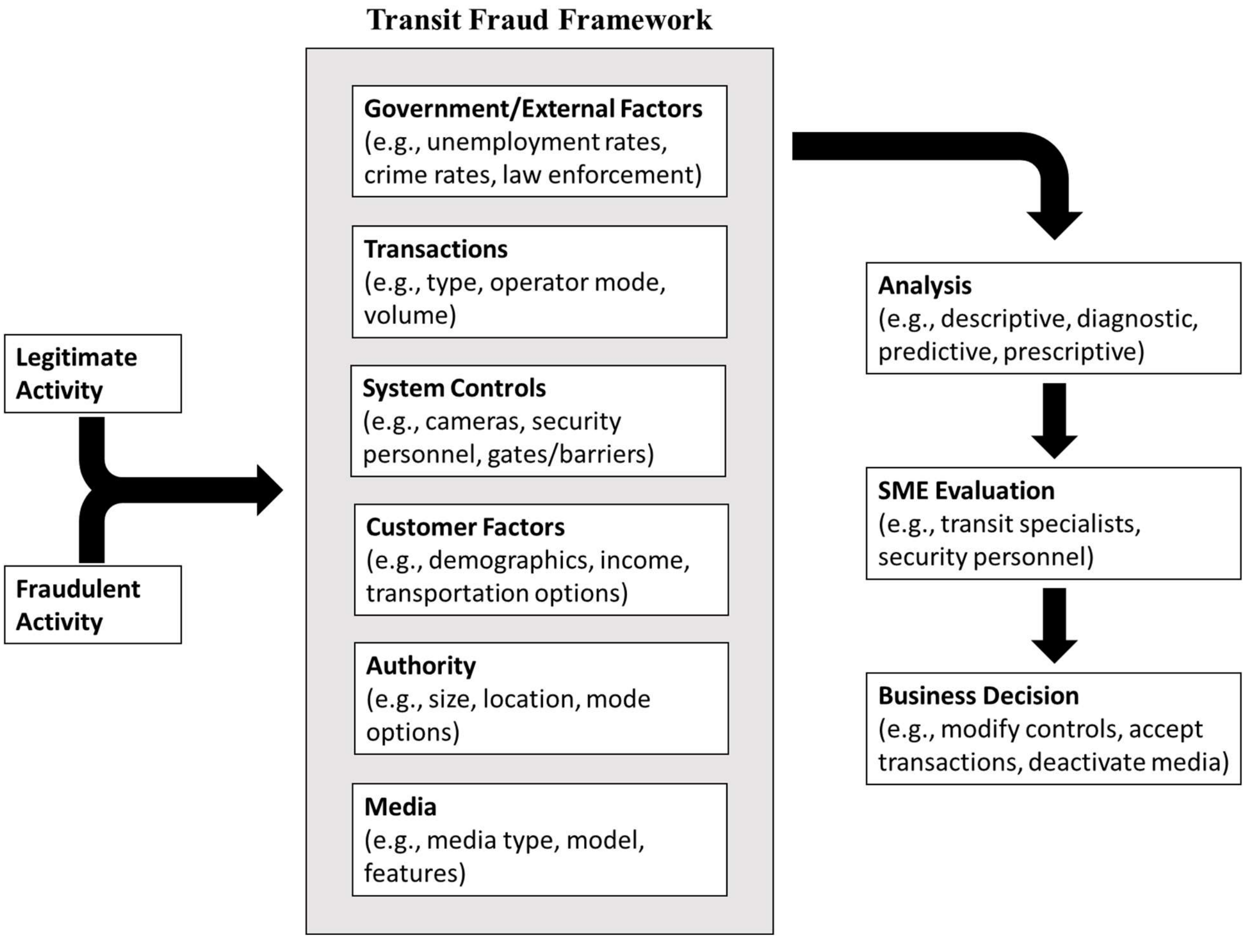

Transit Fraud Framework
Legitimate Activity
Fraudulent Activity
Government/External Factors (e.g., unemployment rates, crime rates, law enforcement)
Transactions (e.g., type, operator mode, volume)
System Controls (e.g., cameras, security personnel, gates/barriers)
Customer Factors (e.g., demographics, income, transportation options)
Authority (e.g., size, location, mode options)
Media (e.g., media type, model, features)
Analysis (e.g., descriptive, diagnostic, predictive, prescriptive)
SME Evaluation (e.g., transit specialists, security personnel)
Business Decision (e.g., modify controls, accept transactions, deactivate media)

Figure 111: Transit Fraud Framework

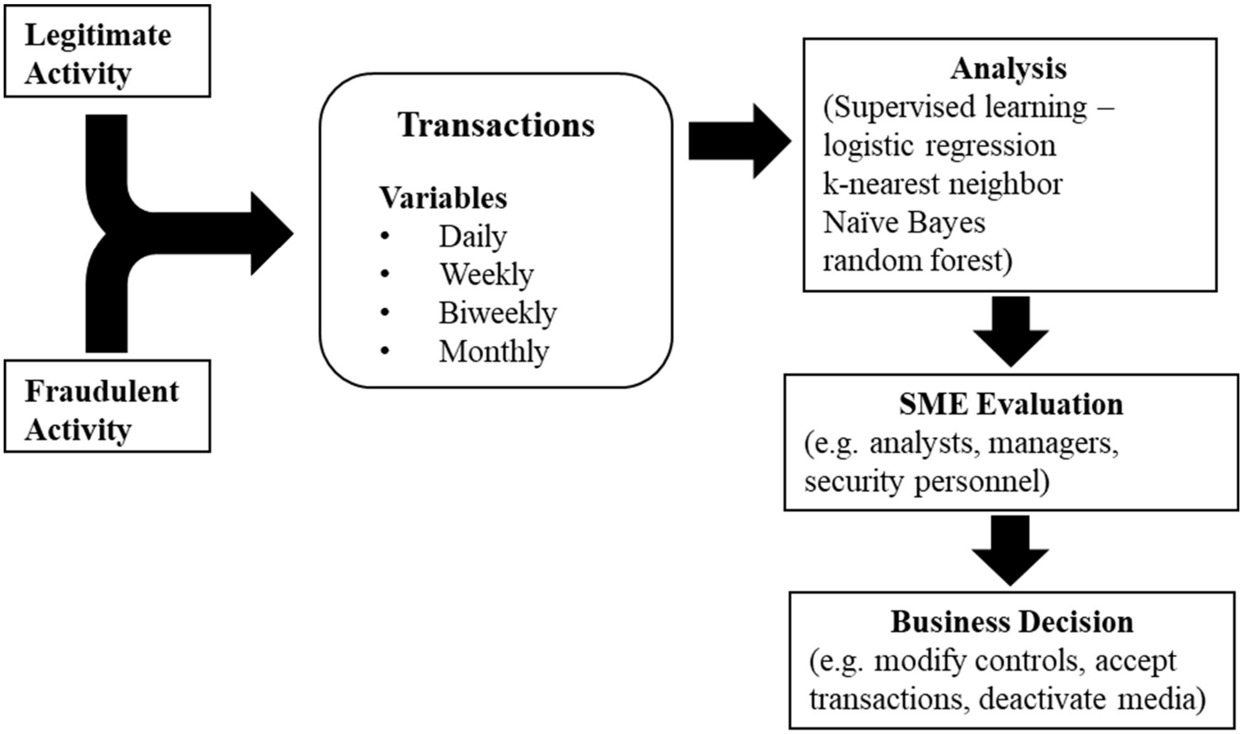


Figure 122: Research Model

Research Methodology

Testing was conducted by training a series of machine learning models to classify records as predicted "fraud" or "not fraud". Predictions were compared to labeled data to generate accuracy values for each model. The research strategy was to develop a binary classifier ("fraud", "not fraud") using a supervised learning approach. Accuracy, precision, recall, F1-score, and ROC AU are calculated for each modeling technique and time span and presented for comparison.

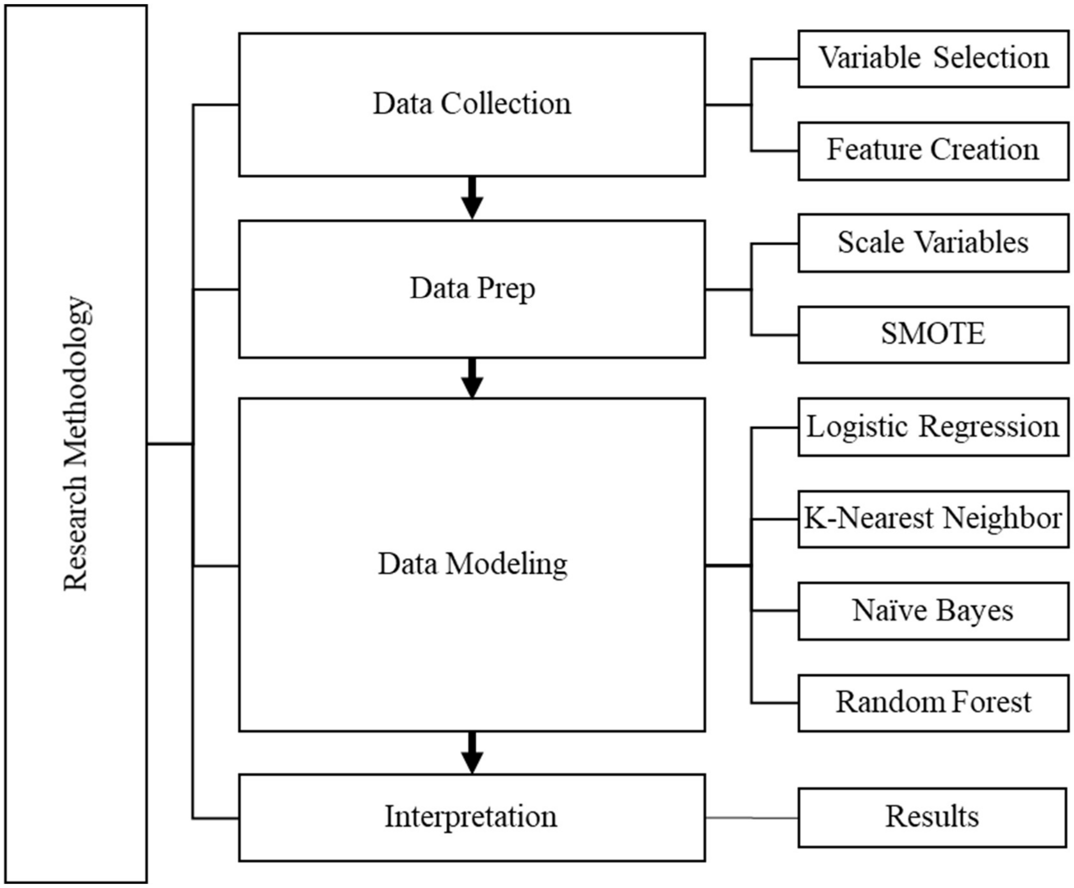


Figure 133: Research Methodology

2.4 Data

Source

Data for this study was gathered from transaction logs of the Metropolitan Atlanta Rapid Transit Authority (MARTA). MARTA is a top 10 U.S. transit authority and meets the Transit Fraud Framework assumption of a modern operation utilizing smart cards as the primary form of fare media. Using a set of 20 numeric variables, testing lengths of daily, weekly, biweekly, and monthly data sets were generated.

Variable Selection / Creation

Variables were selected based on previous findings (paper 1), that indicate they offer predictive potential regarding fraudulent vs. legitimate classification. Native variables include

ENTRY_TAG_ON, EXIT_TAG_OFF, MARTA_BUS, MARTA_RAIL, DEVICES, FACILITIES, FI_CATEGORIES, MODES, and USES. All other variables were derived. Time spans were in increments of 1, 7, 14, and 30 days for the daily, weekly, biweekly, and monthly data sets respectively.

Feature	Description
ENT_EXT_RATIO	Entries minus exits
ENTRIES_PER_DAY	Total entries divided by time span
ENTRY_TAG_ON	Entries (gates or buses)
EXIT_TAG_OFF	Exits (gates or buses)
EXITS_PER_DAY	Total exits divided by time span
BUS_RATIO	Percentage of transactions that were bus related
MARTA_BUS	Bus related transactions
MARTA_RAIL	Rail related transactions
RAIL_RATIO	Percentage of transactions that were rail related
CATEGORIES_PER_DAY	Total categories divided by time span
DEVICES	Total devices utilized
DEVICES_PER_DAY	Total devices divided by time span
FACILITIES	Total facilities visited
FACILITIES_PER_DAY	Total facilities divided by time span
FI_CATEGORIES	Total fare categories utilized
MODES	Total modes utilized
MODES_PER_DAY	Total modes divided by time span
TRANSIT_DAYS	Total days with transactions
TRANSIT_DAYS_RATIO	Total days with transactions divided by time span
USES	Total uses

Table 133: Data Variables

Scaling

To avoid issues of scale, where features using larger variable values might be assigned disproportionate impact, variables were preprocessed using the sklearn StandardScaler. Features were standardized by subtracting the mean and dividing by the standard deviation. The standard score of a sample x is calculated as:

$$z = (x - u) / s$$

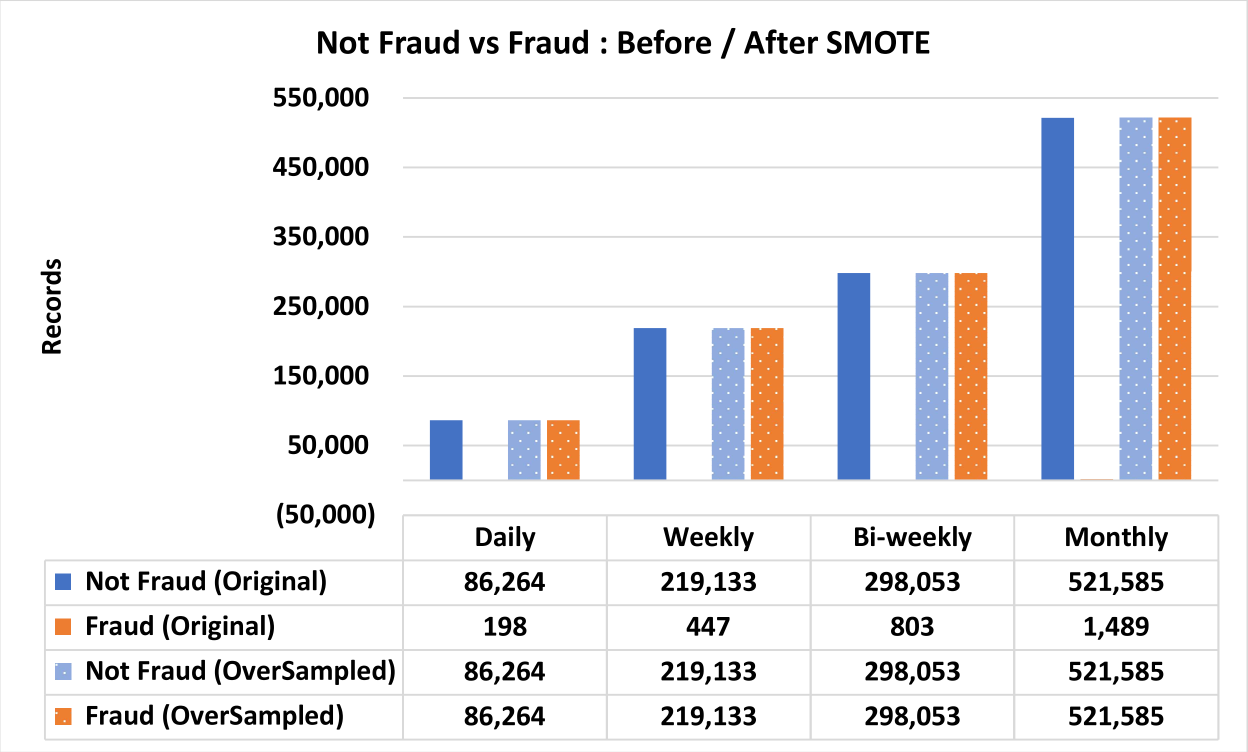
Where u equals the mean of the training samples and s is the standard deviation of the training samples (sklearn ref).

Imbalanced Data

The dataset beginning totals and ratios are shown in table 14. The average ratio of non-fraudulent to fraudulent records across all time spans was approximately 415 to 1. The dataset was split into training (80%) and testing (20%) portions, and the imbalance was addressed utilizing SMOTE.

	Records	Fraud	Fraud %	Ratio
Daily	108,078	244	0.23%	443:1
Weekly	274,475	566	0.21%	485:1
Bi-weekly	373,571	983	0.26%	380:1
Monthly	653,843	1,865	0.29%	351:1

Table 144: Fraud Ratio



	Daily	Weekly	Bi-weekly	Monthly
🟦 Not Fraud (Original)	86,264	219,133	298,053	521,585
🟧 Fraud (Original)	198	447	803	1,489
▨ Not Fraud (OverSampled)	86,264	219,133	298,053	521,585
▨ Fraud (OverSampled)	86,264	219,133	298,053	521,585

Figure 144: SMOTE Balanced Data

2.5 Results

The models were first trained and tested using the original unbalanced data. Each model was tested for accuracy and ROC AUC scores using each of the testing time frames (i.e., daily, weekly, biweekly, & monthly). These two criteria were selected to help demonstrate the misleading impact of unbalanced data on the models. The traditional accuracy calculation is calculated by summing the True Positives and True Negatives and dividing by the total number of predictions. The ROC AUC score plots the True Positive Rate against the False Positive Rate. The benefits of ROC AUC have been thoroughly explored (Huang & Ling, 2005). For this paper, the primary applicable advantage of ROC AUC is that it considers both true and false predictions to provide a more meaningful model score.

Unbalanced Data Accuracy Scores

	LR	KNN	NB	RF
Daily	0.9978	0.9978	0.8735	0.9980
Weekly	0.9977	0.9979	0.9225	0.9979
Biweekly	0.9975	0.9978	0.9278	0.9976
Monthly	0.9970	0.9976	0.9463	0.9977
Average Score:	0.9975	0.9978	0.9175	0.9978

Table 155: Imbalanced Accuracy Scores

Unbalanced Data ROC AUC Scores

	LR	KNN	NB	RF
Daily	0.5000	0.5000	0.6980	0.5434
Weekly	0.5000	0.5420	0.6509	0.5587
Biweekly	0.5027	0.6110	0.7116	0.5970
Monthly	0.5000	0.6580	0.6455	0.6899
Average Score:	0.5007	0.5778	0.6765	0.5973

Table 166: Imbalanced ROC AUC Scores

In the second phase, each model was retested using the SMOTE based balanced data. Again, each model was tested for accuracy and ROC AUC for each of the timeframes.

Balanced Data Accuracy Scores

	LR	KNN	NB	RF
Daily	0.8039	0.9133	0.8317	0.9209
Weekly	0.8379	0.9577	0.8748	0.9602
Biweekly	0.8690	0.9517	0.8896	0.9628
Monthly	0.8962	0.9641	0.9071	0.9719
Average Score:	0.8518	0.9467	0.8758	0.9540

Table 177: Balanced Accuracy Scores

Balanced Data ROC AUC Scores

	LR	KNN	NB	RF
Daily	0.7716	0.7722	0.7855	0.7543
Weekly	0.7594	0.7818	0.7528	0.7494
Biweekly	0.8650	0.8705	0.8366	0.8262
Monthly	0.8458	0.8680	0.7877	0.8226
Average Score:	0.8105	0.8231	0.7907	0.7881

Table 188: Balanced ROC AUC Scores

2.6 Discussion

In this section the impact of balancing the data, feature importance ranking, impact of the various timeframes, model interpretation via partial dependence plots (PDP), and managerial implications will be examined.

Balancing impact

Testing with unbalanced data resulted in the expected model bias toward the majority class. Logistic regression and KNN were most severely impacted by unbalanced datasets. Logistic regression initial testing resulted in an average 99.75% majority class prediction rate across all timeframes. While all of the tested models showed a marked decrease in accuracy rates calculated on unbalanced data, they also saw substantial improvement in their ROC AUC scores. Naïve Bayes appeared least impacted by the unbalanced data. While Naïve Bayes scores for accuracy and ROC AUC weren't the highest of the models, they did show the least amount of change when SMOTE was utilized to balance the datasets.

High impact variables

Feature importance testing based on coefficient values revealed that UT_EXIT_TAG_OFF, EXITS_PER_DAY, UT_ENTRY_TAG_ON, ENTRIES_PER_DAY, & OP_MARTA_BUS scored highest respectively. Interestingly, the identified variables are all based on use counts of various kinds. None of the fare category, fare type, device based, or ratios based on these features scored high in terms of feature importance.

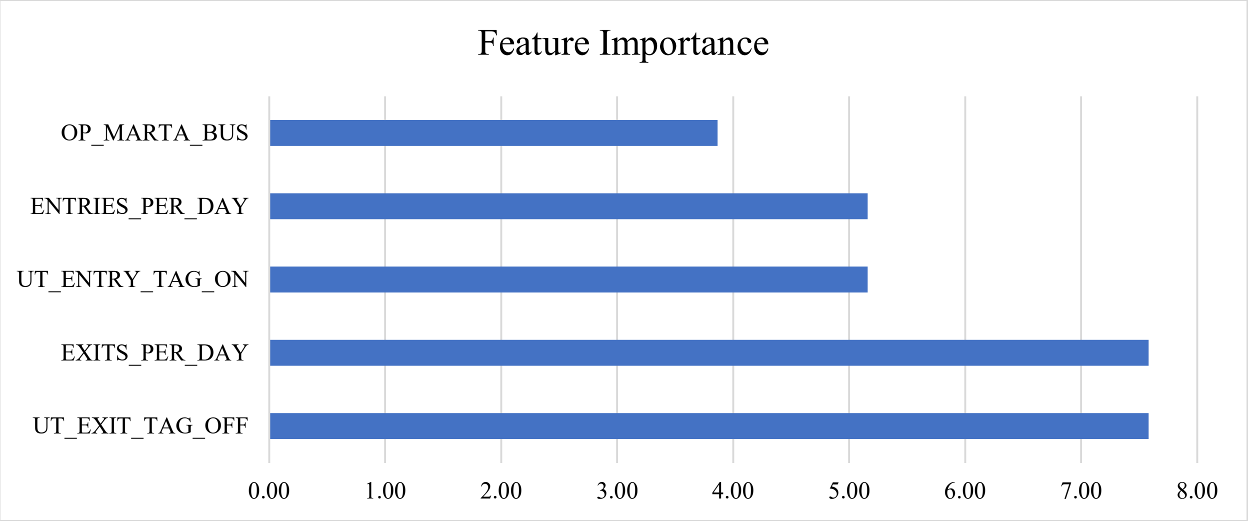


Figure 15: Feature Importance

Timeframes

Monthly and Biweekly timeframes were found to be consistently superior across all models in terms of ROC AUC scores. This was also true for accuracy scores with the one exception of KNN where the weekly score (.09577) was marginally higher than the biweekly score (0.9517). Accuracy was consistently highest for all models using the monthly timeframe. Conversely, the ROC AUC scores were highest for the biweekly testing period.

Managerial Implications

To gauge the impact of these findings from a management perspective, subject matter experts from the participating transit authority were asked to give an assessment of the findings. With the exception of logistic regression, they were uniformly unfamiliar with the machine learning models being used. They agreed that transactional variables, especially those associated with use type, should be the focus of future analysis based on the feature importance findings. There was debate about how the model scores should be used to determine an optimum timeframe. The two predominant opinions were largely based on focusing either on saving money or minimizing impact on customers. The first view suggested using shorter time frames, despite their lower model scores, to deactivate suspicious fare media earlier and mitigate the fraud based financial losses. The second view was that the highest scoring combination of model and timeframe should be used, despite allowing longer periods of potentially fraudulent activity, to minimize the likelihood of false positives that would negatively impact the customer experience. The discussion was essentially reduced to a decision on which party should bear the burden. The transit authority continues to lose money by Hotlisting media later based on the monthly model, while the customer is potentially impacted by a greater number of false positives by Hotlisting media based on the biweekly timeframe.

2.7 Limitations and Future Research

This research is subject to several challenges and constraints. Limitations include the study being conducted with data from a single transit authority, samples limited to identifiable instances of fraud, initial data considered only 20 features, and a wide range of classification techniques still to be explored.

Future research should explore the minimum required transaction history to increase the practical application of the research. From a loss prevention perspective, emphasis on early detection and Hotlisting is key. Once the card has been fraudulently used to enter the system there is no mechanism for the transit authority to recoup the lost fare value. Regarding Hotlisting, a certain amount of data noise is inherent because the technique is not strictly used for fraud. There are some instances where Hotlisting is related to card defects, media reported as lost, etc. By applying a fraud specific marker to deactivated media most of this could be eliminated.

It would also be beneficial to explore specific types of fraud associated with each instance of labeled fraud. Different types of media fraud (e.g., media purchased with a stolen credit card, counterfeit media, media tampering etc.) are likely suited to a variety of models vs. an attempt to create a one-size-fits-all approach to detection. Better data labeling would assist transit professionals to see where security gaps may exist and researchers to determine optimal approaches to those fraud categories.

2.8 Conclusion

There are several research implications of this study. One outcome of this work is that it is possibly the first of its kind. As discussed earlier, it is difficult to gain access to fraud records for research purposes. This research benefited by being able to directly access system transactions to include labeled fraud data. Secondly, this work explores a new application of the general deterrence theory as it applies to transit fraud. Lastly, it supports the growing body of research around the use of machine learning techniques as fraud classifiers. The fast, flexible, and updatable nature of machine learning models makes them especially adept at accurately classifying the type of fraud seen in public transit.

The primary practice implication of this work relates to the scale of financial losses being considered. The transit authority used in this study comprises approximately 1% of the ridership of the top 20 North American transit authorities. Recalling that fraud related chargebacks are accumulating at approximately $200k per year, the extrapolated total for the top 20 could be approaching $20M per year. While this figure is a crude approximation, it does help to demonstrate the potential scale of the issue. To further illustrate this point, consider how $200k per year at one agency scales when considering that there are over 920 transit authorities of various sizes just in the U.S.

It has been demonstrated that machine learning in the form of logistic regression, KNN, Naïve Bayes, and random forests can be effective classification techniques for transit fraud transactions. All of the selected models performed well, with an approximate 91% accuracy rate and an ROC AUC score of 0.80. The challenge of unbalanced data was effectively overcome by the application of SMOTE to balance the datasets.

It is evident that machine learning can classify transit-based fraud with a high degree of efficacy. Based on supporting sociology/criminology theories (e.g., the fraud triangle theory and the general deterrence theory) which emphasize the correlation between positive detection rates and negative crime rates, machine learning based detection systems should be considered a mitigating response to transit fraud.

2.9 References

American Institute of Certified Public Accountants. Auditing Standards Board. (2002). Consideration of Fraud in a Financial Statement Audit:(supersedes Statement on Auditing Standards No. 82, AICPA, Professional Standards, Vol. 1, AU Sec. 316; and Amends SAS No. 1, Codification of Auditing Standards and Procedures, AICPA, Professional Standards, Vol. 1, AU Sec. 230," Due Professional Care in the Performance of Work," and SAS No. 85, Management Representations, AICPA, Professional Standard, Vol. 1, AU Sec 333). American Institute of Certified Public Accountants.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control VII, 235-255.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical science, 235-249.

Bonfanti, G., & Wagenknecht, T. (2010). Human factors reduce aggression and fare evasion. Public Transport International, 59(1).

Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural data mining for credit card fraud detection. In Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on (pp. 103-106). IEEE.

Bureau of Transportation Statistics. (2015). Federal and State Funding of Public Transit, 2015. Retrieved January 8, 2020, from https://www.bts.gov/content/federal-and-state-funding-public-transit-2015

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), 1-6.

Delbosc, A., & Currie, G. (2019). Why do people fare evade? A global shift in fare evasion research. Transport Reviews, 39(3), 376-391.

ERRAC & UITP. (2014, June 6). Metro, light rail and tram systems in Europe. Retrieved from http://www.uitp.org/metro-light-rail-and-tram-systems-europe

Federal Trade Commission. (2020, April 15). Consumer Sentinel Infographic. Retrieved April 20, 2020, from https://public.tableau.com/profile/federal.trade.commission#!/vizhome/ConsumerSentinel/Infographic

Fix, E., Hodges, J.L. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Gantman, S., & Zinoviev, D. (2017). Conceptual Structure of Fraud Research and Its Dynamics.

Garner, B. A. (2004). Black's law dictionary.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

Hill, G. N. (2005). The people's law dictionary.

Hobbes, T. Early Classical Philosophers of Deterrence Theory. Criminology, 41(1), 99-130.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310.

Hughes-Cromwick, M., & Dickens, M. (2018). APTA 2017 Public Transportation Fact Book.

Jensen, D. (1997, July). Prospective assessment of ai technologies for fraud detection: A case study. In AAAI Workshop on AI Approaches to Fraud Detection and Risk Management (pp. 34-38).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression. New York: Springer-Verlag.

Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004, March). Survey of fraud detection techniques. In IEEE International Conference on Networking, Sensing and Control, 2004 (Vol. 2, pp. 749-754). IEEE.

Lee, J. (2011). Uncovering San Francisco, California, Muni's proof-of-payment patterns to help reduce fare evasion. Transportation research record, 2216(1), 75-84.

NTD. (2019, October 17). 2018 National Transit Summaries and Trends (NTST). Retrieved August 15, 2020, from https://www.transit.dot.gov/node/134401

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

Privacy Rights Clearinghouse. (2019). Retrieved August 1, 2020, from https://privacyrights.org/data-breaches.

PwC. (2020). Fighting fraud: A never-ending battle. Retrieved April 20, 2020, from https://www.pwc.com/gx/en/forensics/gecs-2020/pdf/global-economic-crime-and-fraud-survey-2020.pdf

Reddy, A. V., Kuhls, J., & Lu, A. (2011). Measuring and controlling subway fare evasion: improving safety and security at New York City transit authority. Transportation Research Record, 2216(1), 85-99.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Sharma, A., & Panigrahi, P. K. (2013). A review of financial accounting fraud detection based on data mining techniques. arXiv preprint arXiv:1309.3944.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. International journal of pattern recognition and artificial intelligence, 23(04), 687-719.

Troncoso, R., & de Grange, L. (2017). Fare evasion in public transport: A time series approach. Transportation Research Part A: Policy and Practice, 100, 311-318.

UITP. (2017, October 31). Urban Public Transport in the 21st Century. Retrieved from https://www.uitp.org/urban-public-transport-21st-century

UITP. (2018, October 23). World Metro Figures 2018. Retrieved from https://www.uitp.org/world-metro-figures-2018

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2016). Gotcha! Network-based fraud detection for social security fraud. Management Science.

WABC-TV. (2019, June 17). MTA getting 500 additional officers to fight fare evasion, worker assaults. ABC7 New York. https://abc7ny.com/traffic-transit-mta-subway-riders/5349988/.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers & security, 57, 47-66.

Paper 3: Employing Deep Learning to Detect Transit Fraud

3.1 Introduction

Abstract – The scale of both public transit and transit related financial losses are enormous. Billions of passenger trips are provided annually by public transit systems that are heavily dependent on revenue collections via passenger revenues. Transit media fraud costs authorities millions of dollars per year and has thus far been largely unexplored in academic research. This research describes the difficulties associated with transit research fraud and then addresses them via a demonstration of data techniques (SMOTE & ADASYN) and machine learning methods (deep learning). A series of 10 deep learning model variations, pretreated with SMOTE, are tested with the highest performing model achieving approximately 93% accuracy. These results represent compelling findings for both transit fraud researchers and public transit authorities.

Keywords – Public transit, deep learning, neural networks

Public Transit

The American Public Transportation Association defines public transportation *(also referred to as transit, public transit, or mass transit*) – as "transportation by a conveyance that provides regular and continuing general or special transportation to the public" (APTA, 2020). In 2019 there were approximately 6,800 public transportation organizations operating in the U.S. Located in every state, and in both urban and rural areas, they provided 9.97 billion passenger trips. These trips were conducted via a variety of modes including bus systems, paratransit service (for passengers with disabilities), bus-rapid transit, light rail, commuter rail, heavy rail, and water-based systems (APTA, 2021). Figure 16 is a map produced by the Bureau of Transportation Statistics that shows transit agency headquarters (for participating agencies) in the U.S. This visual helps demonstrate the volume and geographical dispersion of domestic transit systems.

Scale /Growth

Globally, approximately 168 million people utilize mass transit each day. Metropolitan transit systems currently operate in 56 countries and 178 cities with systems being added and/or expanded each year (UITP, 2018).   Ridership numbers are predicted to continue to grow for the foreseeable future.

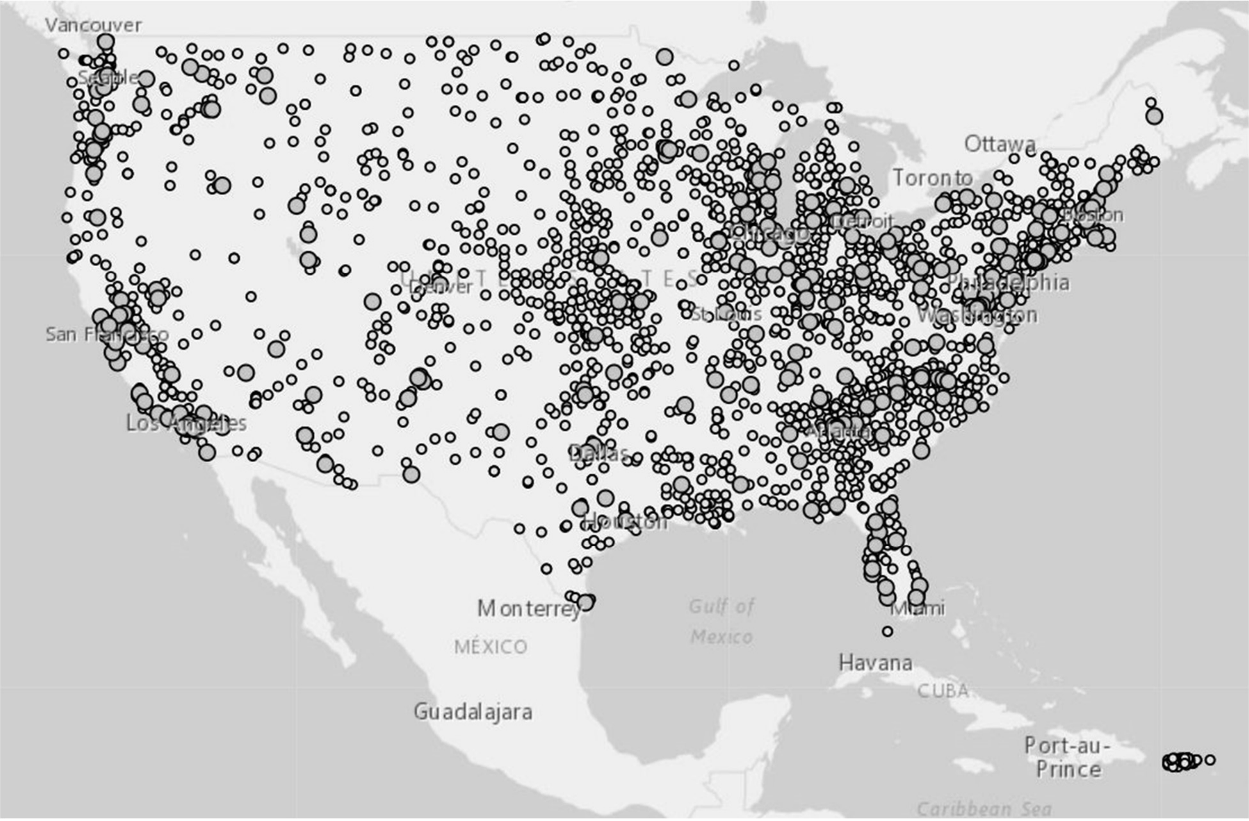


Figure 16: Transit Agency Headquarters

(Bureau of Transportation Statistics, 2021)

Unique features of transit business model

Transit authorities operate on a unique business model. They enjoy a geographically captive market to whom they can largely dictate both the price of service and the choice of payment methods. This allows them to control payment technology adoption rates (Quibria, 2008). Transit authorities use an array of fare media including paper tickets, tokens, magnetic stripe cards, and smart cards. One of the most significant ways they influence customer behavior is by controlling electronic fare media. Electronic fare media is “any portable media that contains the ability to store and retrieve data in a non-volatile manner by a method of electronically reading, writing, or both”. (Trends in Electronic Fare Media Technology, 2004).

The smart card utilizing radio frequency identification (RFID) has emerged as the fare media of choice for modern transit authorities. APTA reports that smart card adoption rates rose from 12% in 2009 to 47% in 2020 (APTA, 2021). RFID can generally be defined as any method of identifying unique items using radio waves. Typically, a reader (also called an interrogator) communicates with a transponder, which holds digital information in a microchip (RFID Journal, 2017). In the case of transit, the chip is embedded into a smart card. Smart cards are typically RFI enabled, credit card shaped, plastic cards with an imbedded computer chip. The chip maintains data regarding trips or stored value which are adjusted during transit use by interacting with card reader/writers at gates and on buses. Additional trips or value can be added using transit media vending machines or web applications.

Smart card systems offer several advantages over traditional fare media, including simplified customer transactions, cash replacement options, lower operating costs, enhanced revenue management capabilities, and remote access to fare media (Smart Card Alliance, 2003). Smart card implementation has been linked to reduced maintenance costs while simultaneously enhancing data collection and reporting capabilities. Smart cards offer transit authorities the ability to store and analyze customer transaction data. This allows for a better understanding of both system complexities and user travel behaviors (Gokasar et al. 2015).

Smart cards also offer a means for transit authorities to remotely edit the stored value, add or remove products, or completely disable an individual card. The process of disabling a card is called “Hotlisting” and is utilized when a card is reported lost, stolen, or is associated with suspicious/fraudulent transactions. Hotlisting is a common control mechanism used throughout the banking and financial services industries. When the system encounters a ticket/card that has been Hotlisted, the fare media is automatically deactivated, and the card holder is forced to purchase a new ticket/card or contact customer service to continue utilizing the transit system.

Transit fraud

Like most industries, public transit is impacted by fraud. As it relates to transit, fraud can be considered an umbrella term that includes the extremely simple (e.g., token slugs and turnstile hopping) to the much more complex (e.g., counterfeit media or credit card fraud). One description of fraud that highlights the complexities of preventing and identifying fraud, is "Fraud is an uncommon, well-considered, time-evolving, carefully organized and imperceptibly concealed crime which appears in many different types and forms." (Van Vlasselaer et al. 2016). As the transit customer base increases and familiarity grows with both systems controls and the various forms of transit media, the frequency and complexity of fraud continues to grow.

Characteristics

There are two primary forms of fraud in public transit, fare evasion and fare media fraud. Fare evasion is primarily physical in nature and is focused on avoiding detection while not paying the fare. It includes climbing over turnstiles, forcing open gates, gate surfing (entering illegally behind a paying customer), and sneaking onto unattended buses. This type of fraud is prevalent because it requires no equipment, little skill, and no specialized system information. Alternatively, fare media fraud attempts to fool the system into accepting a transaction based on counterfeit, stolen, illegally purchased, or illegally altered fare media. While fare media fraud is less frequent it includes the use of a wide range of equipment, skills, and system security information.

Characteristics	Fare Evasion	Fare Media Fraud
Difficulty	Easily executed, low skill activity	Varies, high-end fraud requires significant technical skill
Frequency	High, present in most transit systems	Low, but with strong indications of growth
Fiscal Impact	Individual instances are negligible, cumulative impact is substantial	Individual instances vary widely from hundreds to millions of dollars each

Table 19: Transit Fraud Characteristics

The complexity of counterfeiting fare media is primarily dictated by the individual transit system. Systems utilizing low-tech paper tickets can be circumvented using paper copies of authentic tickets (primarily in systems that rely on transit employees to visually inspect a customer's ticket). In more modern systems, fare media counterfeits may be produced using stolen data or illegally produced tickets/cards.

Scale

Because of a general lack of access to transit operational data, and specifically data related to fraud, little academic research has been conducted in this area. By default, news media becomes the best source for examples of transit fraud scale. The Toronto Transit Commission estimated in 2008 that it was losing as much as $400,000 per month to counterfeit ticketing operations (CTV News, 2008). In the U.S., a 2011 counterfeit ticket scheme conducted by a transit industry contractor was uncovered with an estimated $5M loss to the MBTA (Massachusetts Bay Transportation Authority) (Moskowitz, 2011). New Jersey transit police made over 200 arrests in

2010-2011 related to counterfeit ticketing (Medina, 2012). In 2012, Italian authorities seized 2 million counterfeit train tickets worth $35M (Natanson, 2012). In a reported 10-year review by a public advocate in New York, 3,300 arrests were made for “swiping”. “Swiping” is the illegal act of selling individual train station entries from a monthly card. The perpetrator buys a monthly unlimited card and sells entries for $2 each. Interviewed participants claimed that they made hundreds of dollars per day and a New York Transit Police spokesman claimed that stolen swipes are part of the millions of dollars lost to illegal “fare-beating” crimes (Stewart, 2018). In 2017, the Transit for London (Tfl) reported that each year fare evasion (includes fraudulent tickets) costs Londoners approximately £70million per year (roughly $96.4 million) (Hall, 2019). Additional examples exist in news archives, but these samples illustrate the scale of transit media fraud.

Transit fraud challenges

Effectively researching transit fraud requires solutions for several complexities. Three issues of particular concern are the availability/complexity of the data, the imbalanced nature of the data, and the evolution rate of fraud. Operational data is rarely distributed in the transit community. In modern transit authorities, customer and transaction data is recorded in minute detail. The result is a mix of static and time-stamped data with a range of several dozen to hundreds of data fields and variables. Data that includes Hotlisted media suspended for suspected fraud is even less accessible. Understandably, companies prefer not to share details of how their system controls failed to stop or detect fraud. There is also a common concern that any attention given to information involving fraud will expose system weaknesses and/or encourage more fraudulent activity. The complexity of the dataset requires complex models to attempt to classify unlabeled transactions as either fraudulent or legitimate. Related to the complexity issue, but significant enough to warrant a separate discussion, is the fact that fraud data is inherently imbalanced.

Imbalanced data is simply data that contains highly skewed data. In the context of transit fraud, the vast majority of transactions are entirely legitimate. Only a fraction of the overall transaction population will ultimately be associated with fraud. This heavy skewing of the data causes issues when researchers attempt to build classification models to help categorize transactions as fraud or not fraud. To overcome the issue of imbalanced data, methods of oversampling minority classes, undersampling majority classes, and creating synthetic samples have been developed.

A third obstacle when dealing with fraud is the elusive and evolving nature of fraud. By design, fraud is an intentional attempt to deceive the victim. As victims discover the fraud and begin adapting and establishing new controls, offenders continue to strengthen existing techniques and develop new methods to commit fraud. Any detection or deterrent method developed by victims must account for shifting methodologies of attack and increasingly sophisticated attempts to commit fraud. An example of evolving fraud is the slug.

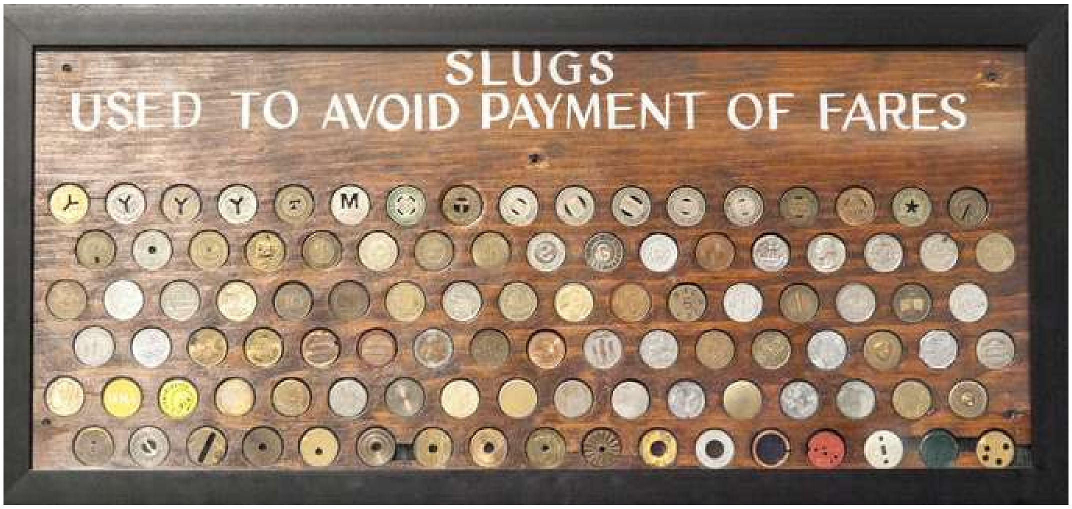


Figure 17: Subway Slugs from 1953 Until 2003 at the New York Transit Museum

Many public transit systems use or have used tokens to represent converted monetary value into transit value. Simply put, each token purchased can be redeemed for a trip. A slug is a counterfeit coin (or token in this example). Slugs attempt to approximate the shape, size, and weight of the actual coin or token using low-cost material. The photo in Figure 17 shows an array of slugs collected by the New York Transit Authority. As the authority shifted designs to defeat known slugs, fare evaders simply updated their slugs to continue bypassing the need to pay for tokens. The variety of materials as well as alterations to change the weight and shape demonstrate the resourcefulness and tenaciousness of committed fare evaders.

Psychology of Fraud

When exploring fraud and fraud detection related material, it is common to reference psychology, sociology, or criminology theories to help explain the underlying motivations and/or causes of offender behavior. Many criminology theories approach the issue of crime from an offender perspective. They focus on social factors (e.g., biological tendencies, social learning, labeling theories etc.) to explain why an offender commits crime. An alternative approach emphasizes the concept of opportunity. These theories, collectively labeled rational choice theories, emphasize the intersection of capability and opportunity to explain crime.

An example of the latter approach is routine activity theory. Routine activity theory adopts the perspective of crimes as individual events that rely on the convergence of a motivated offender, a suitable target, and the absence of a capable guardian (Felson & Cohen, 1980). A motivated offender is a person or group with both the willingness and the capability to commit an illegal act. When the offender encounters a situation (constant or temporary) wherein an attractive opportunity is presented, a rational determination will determine if a crime is attempted. If the offender determines that there is a weak or non-existent crime deterrent, there is some degree of likelihood that a crime will be committed. If the offender lacks the capability to commit the offense without being detected or punished, there is a smaller chance that the offense will occur.

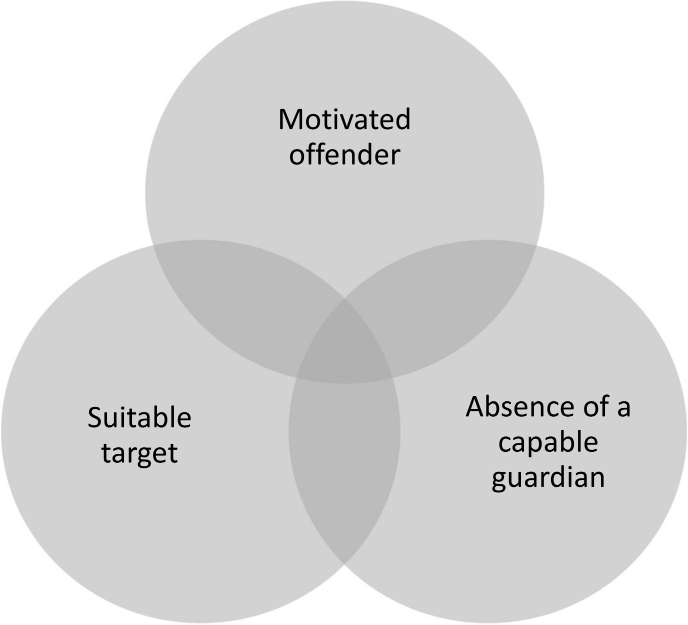


Figure 18: Routine Activity Theory

While there is very little that a transit authority can do to mitigate an offender’s willingness to commit a crime, they can institute system controls to make it more difficult to avoid fraud detection and/or punishment. Based on the three-pillar premise of the routine activity theory, by demonstrating a more robust detection system the occurrence of crime (i.e., fraud) will decrease. There is evidence that higher detection and enforcement rates do have an inverse relationship with crime rates for theft and fraud (Bandyopadhyay, 2011), (Harbaugh, Mocan, & Visser, 2013).

Research

With the advent of intelligent transit systems, public transit authorities generate, record, and analyze massive amounts of customer transaction data. Facing growing concerns around fraud and counterfeit ticketing, public transportation companies have realized that conventional purchase security standards (e.g., spending limits and velocity checking) are insufficient to adequately address the level of associated risk. Transit systems offer a semi-controlled environment where research variables can be more readily tracked and controlled, and findings can be tested.

This study seeks to extend the fraud detection literature by focusing on transit related fraud classification. Recalling the scope of public transit and the scale of transit fraud, practical application of these findings to real-world settings could have significant financial and operational impacts. By overcoming the challenges presented earlier (e.g., availability of data, imbalanced datasets, and complex/evolving methods of fraud), and while considering fraud motivations and business concerns, this research will demonstrate a replicable and practical research methodology.

Research question:

RQ: *Can a deep learning approach to transit fraud detection accurately categorize transactions?*

Layout

The following sections of this paper will highlight relevant research and explain the choice of research model and classification method. A review of the data will follow, as well as the results of the modeling exercise. A detailed discussion of the findings and the practical implications of the work will be delivered. Next, a list of limitations and potential future studies will be examined. Finally, conclusions of the research will be presented along with references and an appendix.

3.2 Literature Review

Fraud

Fraud is an umbrella term that can be difficult to adequately define. In Weiss v. United States in 1941, court proceedings note that "*the law does not define fraud; it needs no definition; it is as old as falsehood and as versable as human ingenuity*" (Weiss v. United States, 1941). This reinforces the view that fraud is complex and evolving. Fraud is often opportunistic. As goods, services, or systems change and adapt to new demands, restraints, and technologies, associated methods of fraud often develop in parallel. Viewing fraud from a historical perspective helps to highlight the speed of adaptation. Ancient examples include false weights for measuring agricultural goods and insurance fraud. Modern instances include 15th century forged art, 17th century coin counterfeiting, 18th century share price manipulation, and 19th century patent medicines and real estate fraud. In the 20th and 21st centuries an explosion of fraud variations is well documented, including con artists, phone scams, data breaches, Ponzi schemes, credit card fraud, cryptocurrency fraud, falsified accounting, healthcare fraud, bank and wire fraud, identity theft, invoice fraud, and disaster fraud (Trulioo, 2020).

A comprehensive review of fraud detection research conducted in 2010 listed the general types of fraud reviewed in over 50 published papers. The categories of fraud were loosely grouped into management, employee, home insurance, crop insurance, automobile insurance, medical insurance, credit application, credit transactional, telecom subscription, and telecom superimposed (Phua, Lee, Smith, & Gayler, 2010). It is particularly noteworthy that the current body of fraud research is largely devoid of transit fraud studies.

In terms of fraud, transit media fraud can be most directly associated with credit card fraud. Credit card fraud is the illegal use of credit card information either physically or virtually

(Zarepoor et al. 2012). Credit card fraud costs to all U.S. retailers was approximately $23 billion and $32 billion for 2013 and 2014 respectively (Insider, 2015). Credit card fraud is committed in transit by using stolen credit cards to purchase fare media online or at point-of-sale devices. When victims discover the charges, they typically contact the card issuer (Visa, MasterCard, Discover, etc.) and initiate a chargeback. A chargeback is a reversal of transferred funds back to the customer without the return of the merchandise to the merchant. This is typically accompanied by a chargeback fee to the merchant who can contest the process (Zilenovski, 2017). Once a chargeback has been initiated against a transit related charge there is little recourse for the transit authority. When dealing with credit/debit transactions, the semi-anonymous, intangible, and consumable nature of public transit service requires that providers collect payments and provide service based on the assumption that payment is legitimate.

The Metropolitan Atlanta Rapid Transit Authority (MARTA) lost an estimated $1 million during the period of 2014-2018 due to credit card chargebacks, which are primarily due to instances of stolen credit cards used to buy fare media for unauthorized resale.

Transit Fraud

Public transit systems lose millions of dollars per year to fraud. Of the limited number of available transit fraud studies, most focus on fare evasion because of both the scale of the issue and the ease of collecting relevant data. A 2009 study conducted at the San Francisco Municipal Transit Authority reported an estimated fraud loss of approximately $19 million annually (Lee, 2011). A similar 6-year study conducted at a Melbourne Australia transit authority reported an average annual loss of over $24 million (converted from Australian dollars at today's rates) between 2005-2011 (Currie & Delbosc, 2016).

Narrowing the focus to fare media fraud significantly reduces the already small pool of transit fraud studies and relies more heavily on news sources rather than academic studies. Examples of transit fare media fraud show the diversity of both techniques and the size/maturity of the impacted transit authorities. For instance, a study in San Fracisco based on passenger surveys estimated that approximately 10% of passengers could not provide valid proof-of-payment. Of those passengers approximately 36% were using media that was expired, invalid, were a misuse of age/disability-based passes, or were counterfeit media (Lee, 2010). In a case of ticket counterfeiting, New Jersey Transit police arrested more than 200 transit card counterfeiters between 2000 and 2002. Officials used an ultraviolet scanner to detect fraudulent tickets and voiced concerns that ticket fraud was a growing issue (Mass Transit, 2012). While fraud might be expected in larger transit systems it can also be found in smaller operations. The Tri Delta Transit authority is a very modest operation of approximately 100 buses and vans in the Bay Area of California. Tri Delta approved a $1.1 million upgrade of ticketing systems to offset the $500,000 they estimate they lost to ticket fraud in the preceding 5 years. A year over year comparison showed that fare box revenues rose by approximately $200,000 in the first 8 months (Bay Area News Group, 2016). Evidence that transit operations are the target of fraud regardless of size or location highlights the need for effective fraud countermeasures.

Classification

One well documented method for detecting fraud is binary classification modeling. Binary classification assigns datapoints to one of two classes. In the case of transit fraud, datapoints are evaluated and classified as either “fraud” or “not fraud”. Subsequent datapoints are then compared to the members of these groups to evaluate similarity and a determination is made regarding to which group the datapoint most likely belongs. There are several methods available for binary classification. They include support vector machines, Naïve Bayes, nearest neighbor, decision trees, logistic regression, and artificial neural networks.

Artificial neural networks (ANN) are particularly interesting based on their capacity to extract meaning from complex and oftentimes incomplete data. Advantages of ANNs include a high level of adaptive learning, the ability to self-organize, appropriateness for real-time operations, and their potential for high fault tolerance (Maind & Wankar, 2014). ANNs have been used extensively in fraud detection. Examples showcasing the variety of applications for ANNs include credit card fraud (Aleskerov, Freisleben & Rao, 1997), management fraud (Fanning & Cogger, 1998), financial reporting fraud (Lin, Hwang, & Becker, 2003), online transaction fraud (Zhang, Zhou, Zhang, Wang & Wang, 2018), and insurance fraud (Yan, Li, Liu & Qi, 2020).

Deep learning is a more complex extension of the traditional artificial neural network. By expanding the ANN from a single input, output, and hidden layer to multiple layers, deep learning reaches higher levels of functionality and flexibility. Nested hierarchies of concepts and representations are "*defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones*" (Goodfellow, Bengio, & Courville, 2016). Examples of deep learning applied to fraud research include money laundering (Paula, Ladeira, Carvalho & Marzagao, 2016), credit card fraud (Roy, Sun, Mahoney, Alonzi, Adams, & Beling, 2018), insurance fraud (Wang & Xu, 2018), electricity fraud (Hu, Guo, Shen, Sun, Wu, & Xi, 2019), and financial statement fraud (Craja, Kim & Lessmann, 2020).

Imbalanced Data

Deep learning has also been proven to be effective in imbalanced dataset situations. A review of 15 studies published between 2015 and 2018 explored the issue of imbalanced data and demonstrated a variety of effective techniques available to deep learning neural networks (Johnson

& Khoshgoftaar, 2019). Common approaches to dealing with unbalanced data include over-sampling, under-sampling, and synthetic over-sampling. Over-sampling uses duplicate samples from the minority class to balance the training dataset, under-sampling randomly drops samples from the majority group, and synthetic sampling is the creation of new datapoints (typically between two existing minority points).

Synthetic Minority Over-Sampling Technique (SMOTE) is a well-documented technique that creates synthetic instances in the minority class at a random distance between two existing minority instances. This is done via a k-nearest neighbor approach for selecting the beginning minority points (Chawla et al., 2002). Examples of SMOTE used in fraud detection include social security fraud (Van Vlasselaer et al., 2013), chargeback fraud (Seo & Choi, 2016), credit card fraud (Sisodia et al., 2017), and auction fraud (Anowar & Sadaoui, 2020).

Another popular and well-documented technique for dealing with imbalanced data is the adaptive synthetic sampling (ADASYN) approach. Influenced in part by the success of SMOTE, ADASYN is also an over-sampling technique. A primary component of ADASYN is the use of a density distribution to determine the quantity of synthetic samples based on their level of learning difficulty. This approach helps reduce the bias in imbalanced data sets by creating more synthetic points for difficult examples (He et al., 2008). Examples of ADASYN used in fraud research include insurance fraud (Subudhi & Panigrahi, 2018), medicare fraud (Bauder et al., 2018), credit card fraud (Ba, 2019), and telecom fraud (Lu et al., 2020).

In this study, the identified challenges of data access, imbalanced datasets, and complex/evolving fraud methods will be addressed via a replicable and practical process. Transactional data from a major U.S. transit authority will be sampled directly from the source. The data imbalance issue will be overcome by use of SMOTE and ADASYN to determine which approach offers the best results. And finally, a deep learning model will be used to overcome data complexity concerns and to create a model capable of adapting to new fraud threats.

3.3 Methodology/Model

Research model

To assist with the continuity of future studies and industry applications, a broad transit fraud framework is pictured below. A targeted research methodology is also included to demonstrate how the specific research plan fits into the landscape of the transit fraud framework.

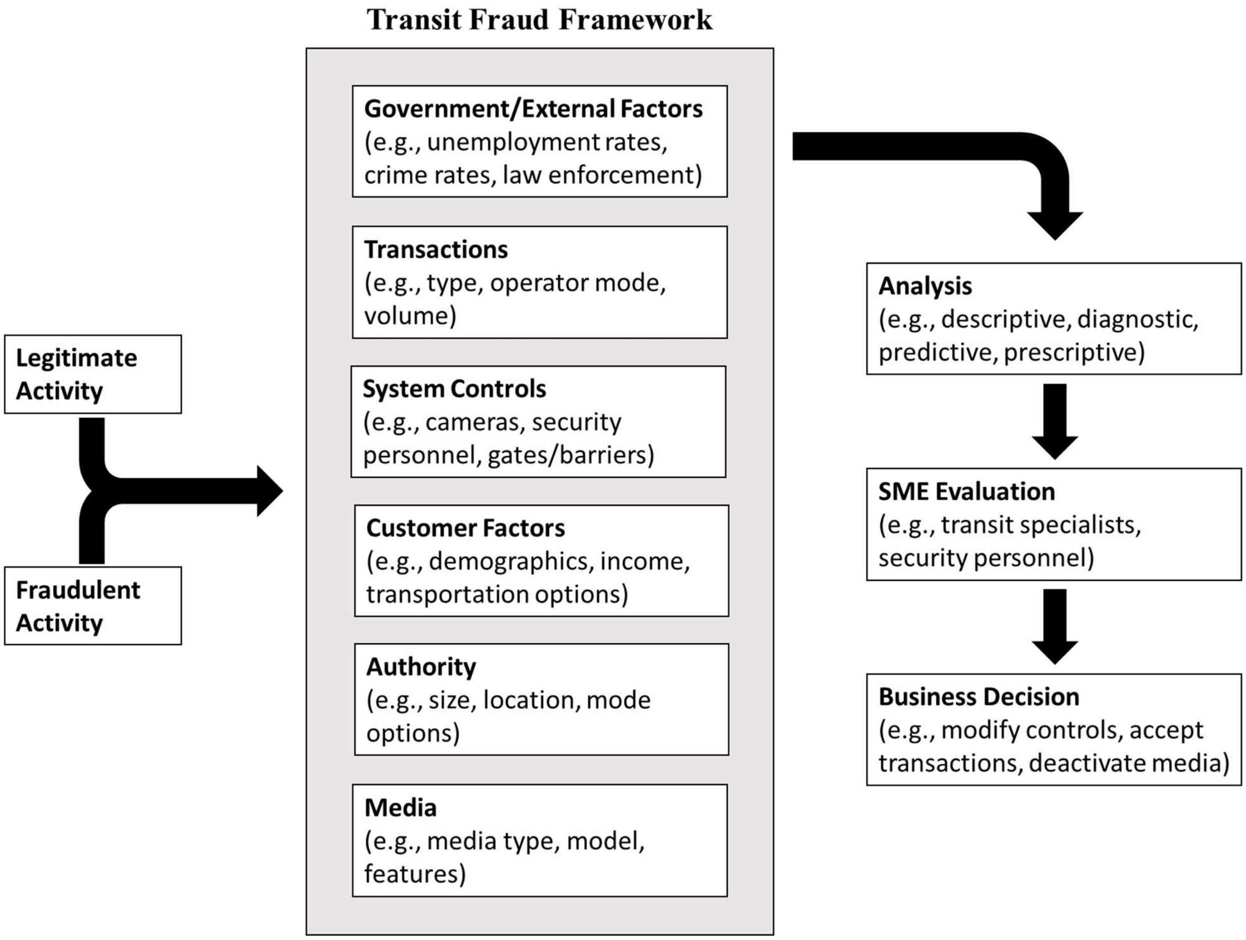

Transit Fraud Framework
Government/External Factors (e.g., unemployment rates, crime rates, law enforcement)
Transactions (e.g., type, operator mode, volume)
System Controls (e.g., cameras, security personnel, gates/barriers)
Customer Factors (e.g., demographics, income, transportation options)
Authority (e.g., size, location, mode options)
Media (e.g., media type, model, features)
Legitimate Activity
Fraudulent Activity
Analysis (e.g., descriptive, diagnostic, predictive, prescriptive)
SME Evaluation (e.g., transit specialists, security personnel)
Business Decision (e.g., modify controls, accept transactions, deactivate media)

Figure 19: Transit Fraud Framework

The transit fraud framework is a generalized approach for transit fraud studies. Based on sample data consisting of both fraudulent and legitimate transit data, a research perspective is selected. Perspectives include governmental/external, transactions, system controls, customer, authority, and media-based approaches. The data, along with any perspective-based variables, are selected and analyzed. Analysis methods vary based on the type of study and data selected (e.g., time-studies, surveys, supervised learning, unsupervised learning etc.). The findings are then presented to transit professionals/experts to check for appropriateness and relevance. The transit authority is then in a position to determine appropriate business steps to address the findings (e.g., increase employee presence, modify fare media, deactivate media involved in suspected fraud, etc.).

The research methodology is a subset of the framework and includes the data, perspective, and analytical approaches that will be used.

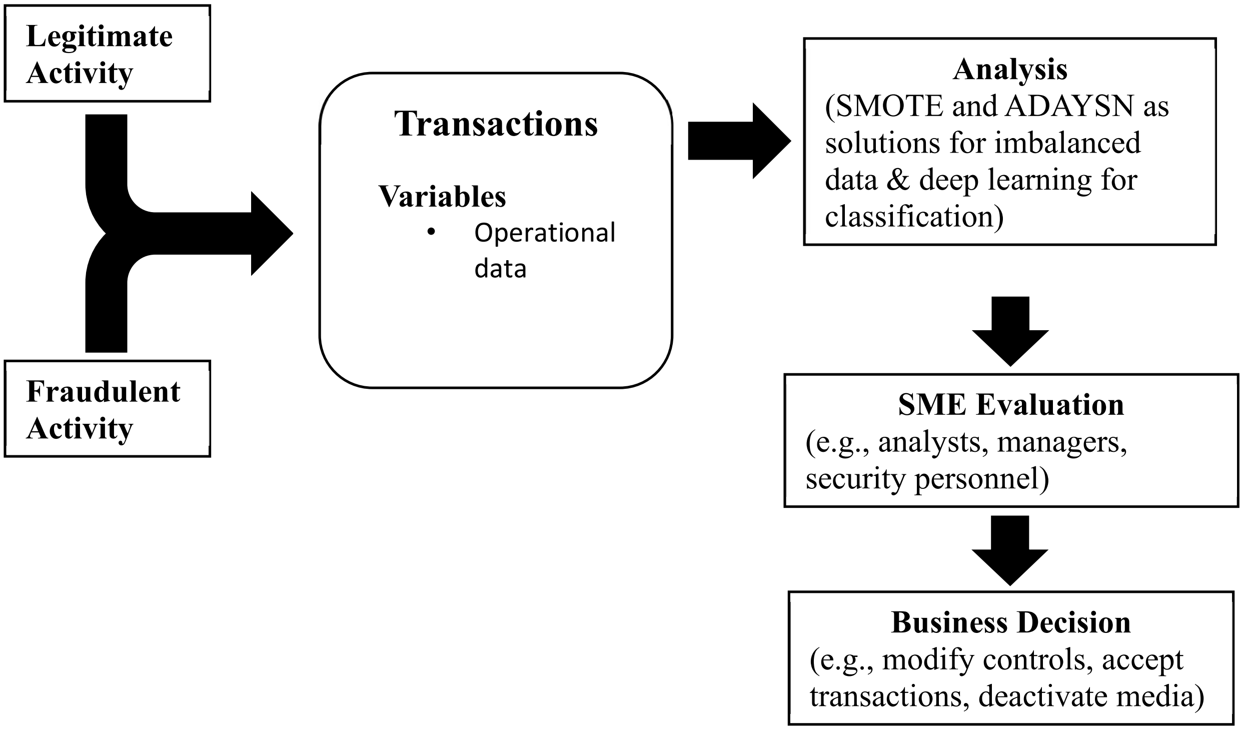

Legitimate Activity
Fraudulent Activity
Transactions
Variables
• Operational data
Analysis
(SMOTE and ADAYSN as solutions for imbalanced data & deep learning for classification)
SME Evaluation
(e.g., analysts, managers, security personnel)
Business Decision
(e.g., modify controls, accept transactions, deactivate media)

Figure 20: Research Methodology

3.4 Data

The Metropolitan Atlanta Rapid Transit Authority (MARTA) is a top 10 U.S. transit agency operating in Atlanta, GA. MARTA operates a system that includes 338 rail cars and 38 rail stations located on 5 major rail lines centered roughly in downtown Atlanta. It also operates 723 buses and vans to service approximately 100 fixed routes (MARTA at a Glance, 2021). First quarter ridership totals for FY21 had an average forecasted total of 8.8 million combined bus and rail boardings per month (MARTA, 2021). The data for this study comes directly from the transaction records of MARTA.

The data consists of a 14-day window of fare media transaction data. It includes a labeled field for Hotlisted media as well as 20 additional fields with potential value for determining fraud classification. Each record is a specific piece of fare media and the system interactions associated with it over the 14-day period. The data fields are a mix of counts, calculated differences between 2 fields, or 14-day averages. The data includes 373,571 unique fare cards/tickets and their associated 5.4 million uses during the sample period. The 2-week sample size was selected as a compromise between gathering sufficient data to accurately classify transactions and mitigating the financial damage of allowing fraudulent media to continue to operate in the system.

The data set includes 372,588 legitimate and 983 fraud samples respectively. This yields the expected unbalanced dataset with a ratio 379:1 or a fraud rate of approximately .26%. The data was split into 80% training and 20% test sets. To rectify the imbalance SMOTE and ADASYN were applied.

Feature	Description
ENT_EXT_RATIO	Entries minus exits
ENTRIES_PER_DAY	Total entries divided by time span
ENTRY_TAG_ON	Entries (gates or buses)
EXIT_TAG_OFF	Exits (gates or buses)
EXITS_PER_DAY	Total exits divided by time span
BUS_RATIO	Percentage of transactions that were bus related
MARTA_BUS	Bus related transactions
MARTA_RAIL	Rail related transactions
RAIL_RATIO	Percentage of transactions that were rail related
CATEGORIES_PER_DAY	Total categories divided by time span
DEVICES	Total devices utilized
DEVICES_PER_DAY	Total devices divided by time span
FACILITIES	Total facilities visited
FACILITIES_PER_DAY	Total facilities divided by time span
FI_CATEGORIES	Total fare categories utilized
MODES	Total modes utilized
MODES_PER_DAY	Total modes divided by time span
TRANSIT_DAYS	Total days with transactions
TRANSIT_DAYS_RATIO	Total days with transactions divided by time span
USES	Total uses

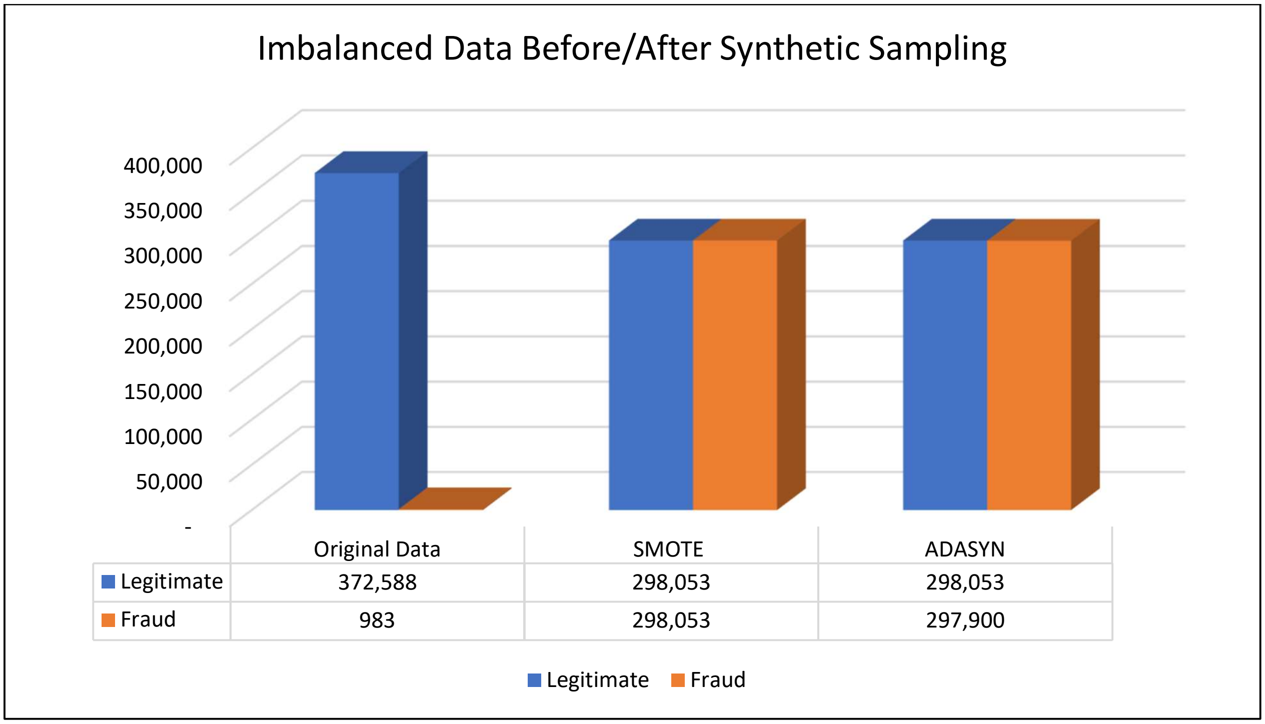
Table 20: Data Fields

3.5 Results

Both SMOTE and ADASN were tested as solutions to the issue of imbalanced data. Without addressing the data imbalance, the models exhibited the anticipated behavior of selecting the majority group (i.e., "not fraud") for each prediction, thereby creating a useless predictor. Models utilizing SMOTE to address data imbalance marginally outperformed ADASYN in testing, so the results shown are for models employing the SMOTE method.

The deep learning models were tuned to test a variety of parameter values for hidden layer counts, optimizers, and loss functions. Static parameter settings included batch size (100), epochs

(5), learning rate (.001), and activation function (Relu). In total, 10 model variations were tested with the metrics summarized in Table 21. The highest scores for accuracy and AUC were highlighted in bold.



	Original Data	SMOTE	ADASYN
🟦 Legitimate	372,588	298,053	298,053
🟧 Fraud	983	298,053	297,900

Figure 21: Treatment of Imbalanced Data

Model	Optimizer	Activation Function	Loss Function	Hidden Layers	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	AUC
1	Adam	Relu	BCE	1	0.9067	0.9064	0.9067	0.9067	0.9064	0.9064	0.9067
2	Adam	Relu	BCE	3	0.9259	0.9247	0.9259	0.9259	0.9246	0.9247	0.9259
3	Adam	Relu	BCE	5	0.9322	0.9313	0.9322	0.9322	0.9312	0.9313	0.9322
4	Adagrad	Relu	BCE	10	0.9229	0.9228	0.9229	0.9229	0.9228	0.9228	0.9229
5	Adamax	Relu	BCE	10	0.9336	0.9329	0.9336	0.9336	0.9329	0.9329	0.9336
6	Adam	Relu	Hinge	10	0.5000	0.5000	0.5000	0.5000	0.6667	0.5000	-
7	Adam	Relu	Square Hinge	10	0.5000	0.5000	0.5000	0.5000	0.6667	0.5000	-
8	Adam	Relu	MSE	10	**0.9362**	**0.9353**	**0.9362**	**0.9362**	**0.9353**	**0.9353**	**0.9362**
9	Adam	Relu	BCE	10	0.9345	0.9335	0.9345	0.9345	0.9334	0.9335	0.9345
10	SGD	Relu	MSE	10	0.9354	0.9344	0.9354	0.9354	0.9343	0.8952	0.8955

Table 21: Deep Learning Model Metrics

*BCE – Binary cross entropy

3.6 Discussion

Results

Regarding the treatment for imbalanced data, no hard conclusion was formed. Both SMOTE and ADASYN outperformed non-treated models. This is consistent with findings from Brandt & Lanzén (2021) where no consistent advantage was identified between SMOTE and ADASYN in testing.

The deep learning models proved to be highly effective at accurately categorizing transactions into fraud or not-fraud classifications. The highest performing model tested used an optimization function of Adam, an activation function of Relu, a loss function of MSE (mean squared error), and 10 hidden layers. It is worth noting that while Adam, which is an extension of and therefore often compared to SGD (stochastic gradient descent), did perform better in testing, there is some evidence that SGD generalizes more effectively in some instances (Zhang, 2018). A direct comparison varying only the optimizer, yielded accuracy scores of 93.53% and 89.52% for the models using Adam and SGD respectively. Additional research will be necessary to determine which optimizer generalizes better from a transit fraud classification perspective.

SME Evaluation & Business Decision

A round table discussion was held with public transit staff members to discuss the models and outcomes. Participants agreed that the models were substantially more complex than analytical methods currently be utilized. While they agreed that the complexity of the techniques was somewhat off-putting, they were pleased with the high degree of accuracy demonstrated across the various iterations. When polled about the possibility of utilizing the models as the basis for an automated response to fraud (i.e., automatically disabling media flagged as fraud by the model), there was a mixed response. The minority group voted that the accuracy rate was sufficient to support an automatic response without supervision. However, the majority of participants felt that media flagged by the model(s) should be reviewed by transit personnel prior to media deactivation. The most common reason cited was a desire to avoid false positives that would inconvenience the customer. There was consensus that with adequate time to conduct tests and sustained positive results, an automated response might be instituted.

3.7 Limitations and Future Research

The primary limitation of this research is the use of a single transit system. To adequately test the generalizability of the findings it is recommended that multiple systems be tested. Because transit systems utilize different system controls, media types, fare products, etc., a wider sample of systems is suggested.

The feature selection would also be improved by adding variables for details associated with initial purchase transactions. Traditional credit/debit card transaction controls include purchase limits, velocity checking, and user verification requests (e.g., zip codes, card security numbers, etc.). This is a clear indication that banks believe collecting and/or monitoring these details provides value. Features of interest include the time of day the purchase was made, the frequency of use for individual credit/debit card numbers, and purchase history details to determine if a particular purchase was consistent with historical activity. By following the lead of mature participants in credit/debit security screening, transit authorities may reap some of the associated benefits without having to independently develop and test new fraud detection methods.

Finally, future transit fraud research should include studies from the additional elements identified in the transit fraud framework (government/external factors, system controls, customer factors, authority, and media). In many instances the transit industry produces rich data sets that may be used to solve transit related issues.

3.8 Conclusion

The goal of this research was three-fold. The first goal was to illustrate the scale of public transit and transit related fraud. This was accomplished by documenting transit metrics and examples of transit fraud to allow the reader to fully appreciate the magnitude of the issue.

Secondly, the objective was to highlight the major obstacles facing transit fraud researchers and to demonstrate an effective solution to those challenges. The 3 major issues of access to scarce data, imbalanced data sets, and a complex and evolving fraud environment were discussed. The research presented used actual transit data, addressed the data imbalance issue using both SMOTE and ADASYN, and utilized deep learning neural networks as a method robust and flexible enough to deal with the evolving and complex nature of fraud.

Finally, a range of deep learning model parameters were tested to determine their effectiveness as a transit fraud classification tool. Several model variations had accuracy scores greater than 93%. Using deep learning models, early detection of transit fraud may have a substantial financial impact for public transit operations. Transit officials now have a viable tool to reliably detect fraud and mitigate financial losses.

3.9 References

122 f.2d 675 (5th cir. 1941), 9735, Weiss v. United States. vLex. (n.d.). Retrieved July 5, 2021, from https://case-law.vlex.com/vid/122-f-2d-675-595329890.

Aleskerov, E., Freisleben, B., & Rao, B. (1997, March). Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr) (pp. 220-226). IEEE.

American Public Transportation Association. (2021, May). 2021 Public Transportation Fact Book. www.apta.com. Retrieved July 24, 2021, from https://www.apta.com/wp-content/uploads/APTA-2021-Fact-Book.pdf.

Anowar, F., & Sadaoui, S. (2020). Detection of auction fraud in commercial sites. Journal of theoretical and applied electronic commerce research, 15(1), 81-98.

Apta.com. (2017). Fact Book Glossary. [online] Available at: http://www.apta.com/resources/statistics/Pages/glossary.aspx#7 [Accessed 3 Nov. 2017].

APTA. (2020, September 14). Ridership Report. Retrieved September 23, 2020, from https://www.apta.com/research-technical-resources/transit-statistics/ridership-report/

Ba, H. (2019). Improving detection of credit card fraudulent transactions using generative adversarial networks. arXiv preprint arXiv:1907.03355.

Bandyopadhyay, S. (2011). An analysis of crime and crime policy. CIVITAS: Institute for the Study of Civil Society.

Bauder, R. A., Khoshgoftaar, T. M., & Hasanin, T. (2018, November). Data sampling approaches with severely imbalanced big data for medicare fraud detection. In 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI) (pp. 137-142). IEEE.

Bay Area News Group. (2016, July 19). Magnetic tickets help repel fraud on Tri Delta Transit bus routes. Retrieved January 7, 2020, from https://www.eastbaytimes.com/2013/08/06/magnetic-tickets-help-repel-fraud-on-tri-delta-transit-bus-routes/

Brandt, J., & Lanzén, E. (2021). A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification.

Bureau of Transportation Statistics. (2021, April 22). National Transit Map (data, maps and apps). National Transit Map (Data, Maps and Apps) | Bureau of Transportation Statistics. Retrieved July 24, 2021, from https://www.bts.gov/national-transit-map/national-transit-map-data-maps-and-apps.

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.

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. Decision Support Systems, 139, 113421.

CTV News (2008, June 12). TTC adult tickets to be discontinued by fall. Toronto. Retrieved October 26, 2021, from https://toronto.ctvnews.ca/ttc-adult-tickets-to-be-discontinued-by-fall-1.301864.

Delbosc, A., & Currie, G. (2016). Four types of fare evasion: A qualitative study from Melbourne, Australia. Transportation Research Part F: Traffic Psychology and Behaviour, 43, 254-264.

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. Intelligent Systems in Accounting, Finance & Management, 7(1), 21-41.

Felson, M., & Cohen, L. E. (1980). Human ecology and crime: A routine activity approach. Human Ecology, 8(4), 389-406.

Gokasar, I., Simsek, K., & Ozbay, K. (2015). Using Big Data of Automated Fare Collection System for Analysis and Improvement of BRT-Bus Rapid Transit Line in Istanbul. In 94th Annual Meeting of the Transportation Research Board, Washington, DC.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Hall, D. (2019, January 1). Fake train tickets sold on Dark Web for 50% less than real deal as rail fares set to increase by 3% This Week. The Sun. Retrieved July 26, 2021, from https://www.thesun.co.uk/news/8024404/fraudsters-selling-fake-cheap-train-tickets-dark-web/.

Harbaugh, W. T., Mocan, N., & Visser, M. S. (2013). Theft and deterrence. Journal of Labor Research, 34(4), 389-407.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.

Hu, T., Guo, Q., Shen, X., Sun, H., Wu, R., & Xi, H. (2019). Utilizing unlabeled data to detect electricity fraud in AMI: A semisupervised deep learning approach. IEEE transactions on neural networks and learning systems, 30(11), 3287-3299.

Insider B (2015) Payments companies are trying to fix the massive credit-card fraud problem with these 5 new security protocols. http://www.businessinsider.com/how-payment-companies-are-trying-to-close-the-massive-hole-in-credit-card-security-2015-3. Accessed 01 Dec 2015

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54.

Lee, J. (2010). Uncovering San Francisco Muni's Proof-of-payment Patterns to Help Reduce Fare Evasion. San Francisco Municipal Transportation Agency.
Lee, J. (2011). Uncovering San Francisco, California, Muni's proof-of-payment patterns to help reduce fare evasion. Transportation research record, 2216(1), 75-84.

Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A fuzzy neural network for assessing the risk of fraudulent financial reporting. Managerial Auditing Journal.

Ling, C. X., & Li, C. (1998, August). Data mining for direct marketing: Problems and solutions. In Kdd (Vol. 98, pp. 73-79).

Lu, C., Lin, S., Liu, X., & Shi, H. (2020, May). Telecom fraud identification based on ADASYN and random forest. In 2020 5th International Conference on Computer and Communication Systems (ICCCS) (pp. 447-452). IEEE.

Maind, S. B., & Wankar, P. (2014). Research paper on basic of artificial neural network. International Journal on Recent and Innovation Trends in Computing and Communication, 2(1), 96-100.
MARTA at a Glance. MARTA. (2021). Retrieved July 7, 2021, from https://www.itsmarta.com/MARTA-at-a-Glance.aspx.

MARTA. (2021, April). Retrieved July 7, 2021, from https://www.itsmarta.com/KPIRidership.aspx.

Mass Transit. (2012, July 17). NJ Transit, NJTPD Step Up Ticked Fraud Crackdown Arrests. Retrieved from https://www.masstransitmag.com/technology/fare-collection/news/10743066/nj-nj-transit-njtpd-step-up-ticked-fraud-crackdown-arrests

Medina, T. (2012, July 16). NJ Transit, NJTPD step up ticket fraud crackdown. NJ TRANSIT. Retrieved July 26, 2021, from https://www.njtransit.com/press-releases/nj-transit-njtpd-step-ticket-fraud-crackdown.

Moskowitz, E. (2011, May 20). Alleged ‘ghost pass’ scheme cost MBTA millions, officials say. Retrieved November 1, 2017, from http://archive.boston.com/news/local/massachusetts/articles/2011/05/20/alleged_ghost_pass_scheme_cost_mbta_millions_officials_say/?page=1

Natanson, P. (2012, June 20). Chinese Counterfeit $35 Million Worth of Italian Train Tickets. Retrieved November 1, 2017, from http://abcnews.go.com/International/chinese-counterfeit-35-million-worth-italian-train-tickets/story?id=16613230

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016, December). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In 2016 15th ieee international conference on machine learning and applications (icmla) (pp. 954-960). IEEE.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

Quibria, N. (2008). The contactless wave: A case study in transit payments. Federal Reserve Bank of Boston.

RFID Journal. (2017). Glossary of RFID Terms. Retrieved November 30, 2017, from http://www.rfidjournal.com/glossary/?R

Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018, April). Deep learning detecting fraud in credit card transactions. In 2018 Systems and Information Engineering Design Symposium (SIEDS) (pp. 129-134). IEEE.

Seo, J. H., & Choi, D. (2016). Feature selection for chargeback fraud detection based on machine learning algorithms. International Journal of Applied Engineering Research, 11(22), 10960-10966.

Sisodia, D. S., Reddy, N. K., & Bhandari, S. (2017, September). Performance evaluation of class balancing techniques for credit card fraud detection. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 2747-2752). IEEE.

Smart Card Alliance (2003). Transit and retail payment: Opportunities for collaboration and convergence. Smart Card Alliance. Princeton Junction, New Jersey, 30.

Stewart, N. (2018, June 17). Selling metrocard swipes remains illegal, and a way of life. The New York Times. Retrieved July 26, 2021, from https://www.nytimes.com/2018/06/17/nyregion/metrocard-swipes-arrests.html.

Subudhi, S., & Panigrahi, S. (2018, September). Effect of class imbalanceness in detecting automobile insurance fraud. In 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA) (pp. 528-531). IEEE.

Trends in Electronic Fare Media Technology (TR-UTFS-FMWG-001-04). (2004). Retrieved from American Public Transportation Association website: (http://www.apta.com/resources/standards/Documents/UTFS_Trends_Electronic_Fare_Media_1-50.pdf)

Trulioo. (2020, June 4). A history of fraud: From ancient egypt to the modern pandemic (part 2). https://www.trulioo.com. Retrieved July 5, 2021, from https://www.trulioo.com/blog/history-fraud-2.

UITP. (2018, March 16). Statistics Brief - World metro figures 2018V4. Retrieved from https://www.uitp.org/data-statistics

Van Vlasselaer, V., Meskens, J., Van Dromme, D., & Baesens, B. (2013, August). Using social network knowledge for detecting spider constructions in social security fraud. In 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013) (pp. 813-820). IEEE.

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2016). Gotcha! Network-based fraud detection for social security fraud. Management Science.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems, 105, 87-95.

Yan, C., Li, M., Liu, W., & Qi, M. (2020). Improved adaptive genetic algorithm for the vehicle Insurance Fraud Identification Model based on a BP Neural Network. Theoretical Computer Science, 817, 12-23.

Zareapoor M, Seeja KR, Alam MA (2012) Analysis of credit card fraud detection techniques: based on certain design criteria. Int J Comput Appl 52:35–42

Zhang, Z. (2018, June). Improved adam optimizer for deep neural networks. In 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS) (pp. 1-2). IEEE.

Zhang, Z., Zhou, X., Zhang, X., Wang, L., & Wang, P. (2018). A model based on convolutional neural network for online transaction fraud detection. Security and Communication Networks, 2018.

Zilenovski, S. E. (2017, February 20). PROTECTING YOUR ONLINE BUSINESS: CHARGEBACK FEE STATISTICS BY INDUSTRY. Retrieved November 30, 2017, from https://blog.clear.sale/protecting-your-online-business-chargeback-fee-statistics-by-industry

Appendix

Appendix A. SQL to generate dataset

```
SELECT
   afc_daily_use_detail_fact_v.SERIAL_NBR,
   afc_daily_use_detail_fact_v.HOTLISTED_FLAG,
   afc_fare_instrument_dim_v.MEDIA_TYPE_DESC,
   afc_fare_instrument_dim_v.RC_DESC,
   Count(Distinct afc_daily_use_detail_fact_v.TRANSIT_MODE_ID)AS MODES,
   Count(Distinct afc_daily_use_detail_fact_v.SK_USE_TYPE_KEY)AS USE_TYPES,
   Count(Distinct afc_daily_use_detail_fact_v.DEVICE_ID)AS DEVICES,
   Count(Distinct afc_daily_use_detail_fact_v.FACID)AS FACILITIES,
   Count(Distinct afc_fare_instrument_dim_v.INSTRUMENT_TYPE_DESC) AS
FARE_INTRUMENTS,
   Count(Distinct afc_fare_instrument_dim_v.FARE_INST_CATEGORY_DESC) AS
FARE_CATEGORIES,
   SUM(To_NUMBER(Case When afc_daily_use_detail_fact_v.SK_USE_TYPE_KEY=9 THEN 1
ELSE 0 END))
   AS ENTRIES,
   SUM(To_NUMBER(Case When afc_daily_use_detail_fact_v.SK_USE_TYPE_KEY=10 THEN 1
ELSE 0   END)) AS EXITS,
   SUM(TO_NUMBER((Case When afc_daily_use_detail_fact_v.SK_USE_TYPE_KEY=9 THEN 1
ELSE 0   END)-(Case When afc_daily_use_detail_fact_v.SK_USE_TYPE_KEY=10 THEN 1 ELSE 0
END))) AS     ENT_EXT_RATIO

FROM
   marta_dw.afc_daily_use_detail_fact_v LEFT JOIN marta_dw.afc_fare_instrument_dim_v ON
marta_dw.afc_daily_use_detail_fact_v.SK_FARE_INSTRUMENT_KEY
=marta_dw.afc_fare_instrument_dim_v.SK_FARE_INSTRUMENT_KEY
   LEFT JOIN marta_dw.afc_device_dim_v ON marta_dw.afc_daily_use_detail_fact_v.SK_DEVICE_KEY
=marta_dw.afc_device_dim_v.SK_DEVICE_KEY  LEFT JOIN marta_dw.afc_transaction_status_dim_v ON
marta_dw.afc_daily_use_detail_fact_v.SK_TS_KEY =marta_dw.afc_transaction_status_dim_v.SK_TS_KEY
   LEFT JOIN marta_dw.afc_use_type_dim_v ON marta_dw.afc_daily_use_detail_fact_v.SK_USE_TYPE_KEY
=marta_dw.afc_use_type_dim_v.SK_USE_TYPE_KEY

WHERE
   CALD_ID BETWEEN 32295 AND 32308 --BETWEEN 32143 AND 32508 FOR 2016
   AND RIDER_CLASS <>131 --Employee Card
   AND afc_daily_use_detail_fact_v.HOTLISTED_FLAG=1

GROUP BY
   afc_daily_use_detail_fact_v.SERIAL_NBR,
   afc_fare_instrument_dim_v.MEDIA_TYPE_DESC,
   afc_fare_instrument_dim_v.RC_DESC,
   afc_daily_use_detail_fact_v.HOTLISTED_FLAG


ORDER BY
   afc_daily_use_detail_fact_v.SERIAL_NBR
```