Abstract

With more consumers shifting to alternative forms of payment beyond cash, the adoption of the use of credit cards continues to increase at a rapid pace. In Canada alone, there are “76.2 million Visa and MasterCard cards in circulation” (Canadian Bankers Association, 2023). Though most credit cards come with their advantages ranging from cash back and rewards to easier and faster transactions, they too possess significant risk for consumers. Such risk is the increased exposure to credit card fraud. By definition, credit card fraud occurs when either a person steals or uses without permission a person's credit card or credit card information (Government of Canada, 2023). In fact, the Chartered Professional Accounts of Canada (CPA) found that 3 in 5 individuals between the ages 18-34 years have reported being a victim of at least one form of financial fraud (Chartered Professional Accountants Canada, 2023). Knowing this, especially at a time when data is easily and widely accessible and exchanged, how can financial institutions utilize such information to its optimum to create models that are able to detect and minimize the risk of credit card fraud for their customers? Using a credit card fraud dataset containing 1,296,675 observations and 23 features, published by Shenoy (2019) on Kaggle, this project strives to address and explore the occurrence of fraud while providing a sufficient model to classify whether a transaction is fraudulent or not in Python. Moreover, this project will follow similarly to Afriyie et al. (2023) as the authors utilize the same data set and similar approaches. Some of the key research questions under consideration include where and how credit card frauds happen more frequently, what sampling technique works best for creating a fraud detection model, and what are the best features that can better predict credit card fraud. With respect to the nature of the subject, classification will be the theme of this task, as the key objective is primarily to predict the correct label for fraud as accurately as possible given the

training data. For context, classification is a supervised machine learning approach in which the model attempts to predict the appropriate label for a given set of data and is trained using the training set, where it is then evaluated on the testing set (Keita, 2022). Some of these approaches include logistic regression, decision tree, and k nearest neighbour (KNN) and these will be the three algorithms used and compared against using Python's sklearn package to identify the appropriate model. Moreover, techniques for balancing the dataset and feature selection will be considered.

Introduction

Credit fraud can happen to anyone at any point of time, especially at a time when technology is advancing significantly. To define, credit card fraud occurs when a person uses a stolen credit card or information to make unauthorized purchases under the person's credit card name (Rafter, 2017). In fact, such fraudulent activity can occur when an individual steals someone's actual card or through hacking their account on the internet (White, 2020). Hacking one's credit card information through the internet is considered as a card-not-present fraud, which is a type of credit card scam where the consumer does not present their physical card to the seller during a fraudulent transaction, either on the internet or through the phone (Dollarhide, 2021). In fact, this is more difficult to prevent than having your credit card present since the seller is not able to analyze the card for suspecting fraud (Dollarhide, 2021).

According to Rafter (2017) from LifeLock, there are several approaches for criminals to attain one's credit card information/card. One way is where a fraudster directs an individual to a fake website in which they trick them into giving their credit card information and then proceed to use their credit card to make fraudulent purchases (Rafter, 2017). Similar to this are in the form of phishing emails in which such emails direct a person to clicking either a button or link,

prompting them to enter their account information (Dickey, 2021). Or instead, the emails may ask them to download files containing spyware that fraudsters can utilize to distribute one's credit card information (Dickey, 2021). Another method is when a person's credit card information is caught in a data breach (Rafter, 2017). Large establishments, such as businesses and banks are more prone to targeted data breaches, placing consumers' credit cards and any additional personal information at risk (Dickey, 2021). Through this, fraudsters can utilize such information to accumulate online payments with one's credit card(s) (Rafter, 2017).

Although it is known that the senior population is the prime target for fraud, the Chartered Professional Accounts of Canada found, surprisingly, that in their annual fraud study, "three-in-five 18-34 year olds (63%) report being a victim of at least one type of financial fraud in their lifetime" (Chartered Professional Accountants of Canada, 2023) whereas for individuals between the ages of 35-54 and those over 55 saw a drop in the number of reports, 39% and 31% respectively. Moreover, their study found that credit card fraud remains to be the most common type of financial fraud as more Canadians are relying on online services to conduct their daily lives. For instance, with respect to online banking, approximately four-in-five individuals who participated in the study with debit cards reported that they initiate their banking online whereas 72% of the participants with credit cards state that they manage their finances online, such as viewing their balances and making payments (Chartered Professional Accountants of Canada, 2023). Online shopping, as another example, accounts as a large portion of consumers spending habits in that those between the ages of 18 to 54 are roughly twice as likely to make large purchases online ranging from appliances to vehicles in comparison to their older counterparts, which stands at roughly 18% (Chartered Professional Accountants of Canada, 2023).

Looking at the impact of credit card fraud on a more macro level, it is known that this type of fraud not only causes financial losses among businesses and financial institutions but also destroys the trust with their customers (Falco, 2023). To put in perspective, on a global scale, credit card fraud "cost the global economy $5.127 trillion annually" (Falco, 2023) with account takeover attacks making up 29.8% of total fraud cases on e-commerce websites (Falco, 2023). In this sense, financial institutions and e-commerce websites are consistently at risk of losing revenue and credibility to which it may lead to long-term damage of bankruptcy or closure. In fact, in 2021, 72% of businesses globally reported an increase in fraud within the span of 12 months, more specifically credit card fraud, along with a 28% increase in fraud attempts (Falco, 2023).

Though there have been different methods proposed to reduce the risk of fraud, such as two-factor authentication, data science has paved the way for the creation of detecting fraud. This is all due to the increased availability of structured data within the banking industry (Kosourova, 2022). When detecting fraud, there exists key analytical techniques where on one hand, there are statistical techniques that include regression and probability distributions while on the other hand, artificial intelligence techniques that encompass data mining, machine learning, and deep learning (Kosourova, 2022). However, the main technique used for fraud detection lies in machine learning in which it can be approached in two ways, supervised or unsupervised (Kosourova, 2022). Examples of supervised machine learning algorithms are k-nearest neighbours, random forest, neural networks, logistic regression, and decision trees whereas for unsupervised machine learning algorithms, there are cluster analysis, principal component analysis (PCA), anomaly recognition and several other algorithms (Kosourova, 2022). Knowing that fraud detection is primarily a classification and prediction problem (Afriyie

et al., 2023), for this project, the techniques and algorithms used will be supervised and the following literature examined have as well used such an approach.

Critique of Credit Card Fraud Detection

Although there have been a number of models created and made available for fraud detection, it is worth highlighting some of its critical issues. In most credit card fraud datasets, there exists a high imbalance of the two classes, namely fraud and not fraud, where majority of the observations are classified as not fraudulent. The issue that arises as a result from this is that machine learning models will be biased towards non-fraudulent transactions, meaning that fraudulent transactions can mimic non-fraudulent transactions and won't be able to detect or differentiate between the two frauds accurately.

Moreover, most financial data is very confidential as it contains personal information. Observing a number of credit fraud datasets, their features are in the form of principal components (PC) where all the features are captured and encapsulated in each uncorrelated PC with different weights being assigned. Though this is a good approach for protecting consumers' privacy, the downside is that it is difficult to capture the essence of fraud and what features constitute fraud. Furthermore, when analyzing the models, an interpretation cannot be made since it is unknown (publicly) what variables are contained in each PC. Rather, only the performance of the models containing PCs can be observed.

As already mentioned, several models have been proposed to detect credit card fraud. Despite this, fraudsters are developing new ways to retrieve people's credit card information. As more models become outdated, the more opportunities it provides for fraudsters to understand and adapt to the model's algorithm and eventually create ease in committing fraud. Therefore, models will need to be updated routinely to adapt to the different forms of fraud.

In what follows, this paper will first discuss the existing literature that uses similar machine learning algorithms to the ones used in this project and examine how these authors have approached (or somewhat approached) in addressing the issues mentioned above. Following that, it will proceed into a discussion about some of the related work where authors have utilized other supervised machine algorithms and methods that are not used for this project and highlight their results. Afterwards, the paper will elaborate on this project's contribution to the literature in fraud detection along with providing a description about the data set used as well as a tentative methodology outline.

Similar work

With the intention of analyzing and identifying the best performing machine learning model for fraud detection, Haddab (2022) utilizes three machine learning algorithms, namely logistic regression, random forest, Naive Bayes, and multilayer perceptron to achieve such an objective. In her study, she used a credit card fraud detection dataset from Kaggle that comprised 274,807 transactions from European cardholders within the span of two days during the 2013 period, where 490 of the observations were classified as fraudulent while the remaining were non-fraudulent. As well, the dataset contained 31 numeric features, as principal component analysis (PCA) was conducted due to the issue of confidentiality of financial information. Given the high imbalance of fraud, the author used the Synthetic Minority Oversampling Technique (SMOTE) to overcome this. Splitting the dataset into their train and test set (80% training and 20% testing) and after training the model for each algorithm, the author found that random forest provided the best model for classifying credit card fraud in which it had a precision of 97.38%, recall of 83.63%, and an accuracy of 98.96%.

In a similar fashion, Afriyie et al. (2023) explores similar machine learning algorithms to Haddab (2022) but with the inclusion of decision trees to identify the best performing classifier for fraud. Throughout the authors research, they utilized a dataset, generated by Sparkov data generation, on credit card transactions in the United States between the January 2020 and December 2020 period where it included transactions of 800 businesses using credit cards along with 1000 customers. Moreover, the data used consists of 555,719 observations in which it contains a total of 23 columns, 12 of which were categorical. Rather than using SMOTE to overcome the issue of imbalance, the authors enforced an undersampling technique to decrease the number of instances of the majority class (no fraud) to approximately equal to the minority class. Though the authors did not specify how they split the data for training and testing (train/test split or cross validation), their findings showed that random forest performed better in predicting fraudulent transactions than the decision tree model and the logistic regression model with an area under the curve (AUC) value of 98.8%, satisfying the findings of Haddab (2022). Furthermore in their analysis, the authors also found that most fraud cases happen between 10pm and 5am and cardholders over the age of 60 are prime targets for fraud.

Observing the performance of three machine learning algorithms on highly skewed credit card data is documented by Awoyemi, Adetunmbi, & Oluwadare (2017). Working with the same dataset as Haddab (2022) and using Python, the authors of the study carried out a hybrid technique of undersampling and oversampling to overcome the skewness that was present. Essentially, their objective was to understand the effect of hybrid sampling on the performance of naive bayes, k-nearest neighbours (KNN), and logistic regression for detecting fraud. For evaluating the models, the authors divided the dataset into 70% training and 30% for testing and validation. As a result, the authors found that KNN outperformed naive bayes and logistic

regression techniques in terms of accuracy, precision, recall, Matthews Correlation Coefficient (MCC), and Balanced Classification Rate (BCR) and that hybrid sampling on skewed data significantly improves the performance of binary classification.

Related Work

In the literature of credit card fraud detection, other authors have approached such a task in alternative ways, such as Yee, Sagadevan, & Malim (2018). In their paper, the authors explored supervised classification models through Bayesian network classifiers, mainly K2, Tree Augmented Naive Bayes (TAN), Naive Bayes, logistics and J48 classifiers, using WEKA. As for the data under investigation, they used dummy data which depended on the manipulation of specific features that were considered as significant indicators for fraud. More specifically, the features included in the creation of the dataset comprised credit card number, reference number, terminal id, actual pin, entered pin, transaction amount etc. and was developed manually using spreadsheet and auto data generation script. Along with this, they conducted two experiments in which the first experiment verifies the authentication of the variables used in the credit card data whereas the second experiment evaluates the performance of the classifiers after undergoing the data preprocessing stage, namely data normalization and Principal Component Analysis (PCA). Again, the authors of the study used WEKA as their machine learning tool to measure each of the performances of the classifiers and used 10-fold cross validation techniques. Under their results, they found that all of the Bayesian classifiers in the second experiment attained an accuracy of more than 95% and achieved better processing speed than the results found from the first experiment.

Alternatively, Maes, Tuyls, Vanschoenwinkel & Manderick (2002) applied two different machine learning algorithms that were appropriate for reasoning under uncertainty within the

credit card transaction system, artificial neural networks (ANN) and Bayesian belief networks (BBN). For their analysis, the authors were provided real world data by Serge Waterschoot at Europay International (EPI) and found that although BNN provided better results with respect to fraud detection and a quicker training period, the fraud detection process with ANNs is significantly faster.

Using a big data analytic framework to undertake large volumes of data and implement machine learning algorithms for fraud detection is documented by Patil, Nemade, & Soni (2018). Throughout their research, the authors strive to overcome the issue of processing and modelling incoming data with fewer delays by offering a solution that uses SAS with a Hadoop framework with the addition of constructing an analytical framework for fraud. In doing so, to build such a model, the authors used a German credit card fraud dataset of approximately 1000 transactions containing 20 attributes, where 7 of the features are numerical while the remaining 13 being categorical, and performed logistic regression, decision tree, and random forest algorithms. In their results, the authors found that out of the three models, random forest and decision tree performed better than logistic regression with respect to their accuracy, precision, and recall. However, the authors note an issue in regards to random forest in which overfitting may occur when data increases and that future work should try and resolve such an issue.

Contribution to the Literature and Objectives

Given what has been going on in the field of fraud detection, this project aims to provide an updated model for detecting credit card fraud transactions. Following similarly to Afriyie et al. (2023), the project will apply 3 supervised machine learning algorithms, logistic regression, decision tree, and k-nearest neighbours and will be using the same credit card transactions fraud data set from Kaggle, published by Shenoy (2019). However, what makes this project different

from Afriyie et al. (2023) is the size of the data set. In their study, the authors only used the test data of 555,719 observations and did not include the training data that was provided in the website. Knowing this, this project will be combining both the test and training data to increase the sample size and obtain more accurate results. Furthermore, the project will be examining 3 different sampling techniques (oversampling, undersampling, SMOTE), rather than solely using undersampling, to examine which sampling technique performs best on one of the three algorithms. Therefore, this project will contribute to the literature in the following ways:

- Uses a dataset that is not in the form of principal components to obtain better interpretation, analysis, and understanding of credit card fraud
- Compares 3 sampling techniques on 3 machine learning algorithms and examines the performance of each one to generate the best possible model

Following that, this project aims to address the following research questions:

1. Knowing that there does exist a bias towards non-fraudulent transactions, what sampling technique will perform better for creating a more accurate fraud detection model?
2. What are the essential features that aid in classifying and differentiating what is a legitimate transaction or a fraudulent transaction?

Data and Methodology

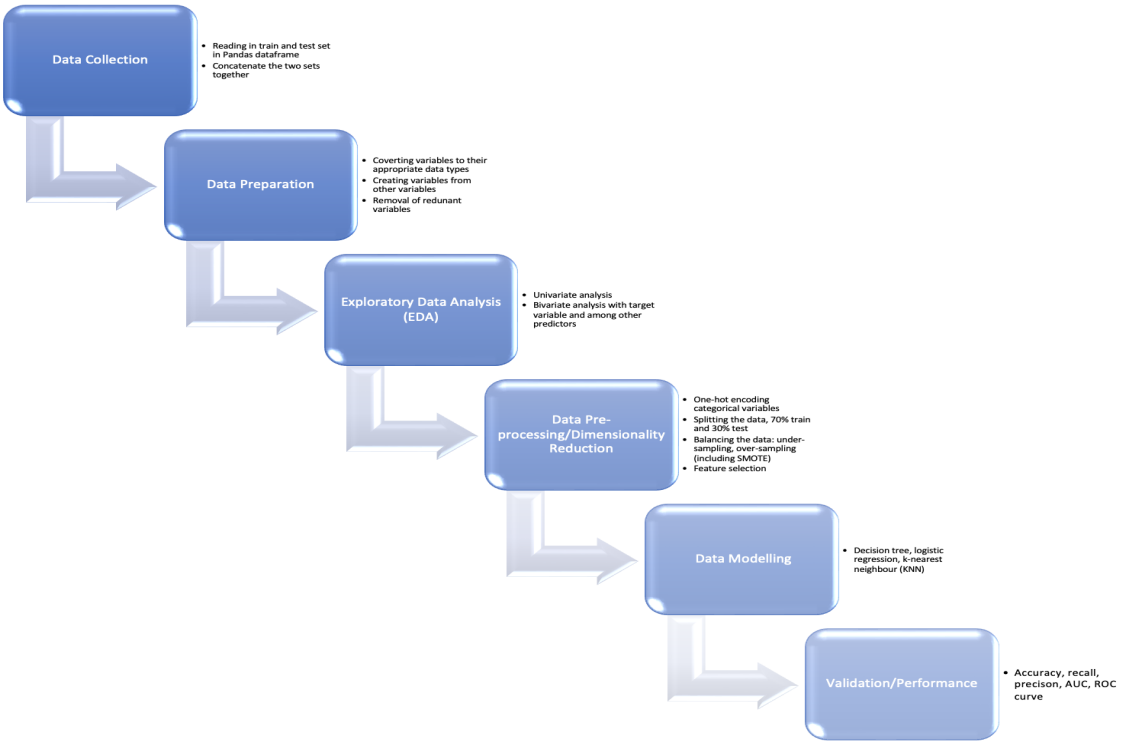
As briefly mentioned, the data set that will be examined is from Kaggle and is published by Shenoy (2019). This data set, according to the publisher, is “a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration of January 1, 2019 - December 31, 2020” (Shenoy, 2019). As well, it includes credit cards of 1000 consumers

in the United States performing transactions, along with a set of 800 merchants. To create the dataset, Shenoy (2019) mentioned that the data was generated using a Sparkov Data Generation tool designed by Brandon Harris. Within the data, and after combining both the train and test sets, there are a total of 1,296,675 observations with 23 features, 11 of them being numeric while the remaining 12 being categorical.

Variable Name	Description
trans_date_trans_time	Transaction date and transaction time
cc_num	Credit card number
merchant	Name of merchant
category	Product category of transaction
amt	Transaction amount
first	First name of credit card holder
last	Last name of credit card holder
gender	Gender of credit card holder
street	Street address of transaction location
city	City of transaction
state	State of transaction
zip	ZIP code of transaction location
lat	Latitude of transaction location
long	Longitude of transaction location
city_pop	City population of transaction location
job	Job of card holder
dob	Date of birth of card holder
trans_num	Transaction number
unix_time	Transaction time in UNIX format
merch_lat	Latitude of merchant
merch_long	Longitude of merchant
is_fraud	Whether transaction is legitimate or not (target variable)

Below displays the general methodology for this project. To begin, I will read both the train and test sets into a Pandas dataframe and concatenate the two together. After examining the dimensions (rows and columns) of the data set, I will proceed with checking the data types of each of the variables to ensure that they are assigned to their appropriate type. If some of the variables are assigned to an incorrect data type, they will be converted accordingly. As well, some variables will be used to create additional variables, such as dob to create an age variable,

to enrich the data set with more information for each observation. Once the original variables are transformed to their newer variable(s), they will be dropped from the data set since they will no longer be used for analysis. Next, I will transition over to exploring the data set by applying univariate analysis for each of the variables and a bivariate analysis between the target variable and predictors and between predictors with the other predictors. After gaining a better understanding of the data set, the data will need to be preprocessed in order to create the models, and some of the steps that will be performed at this stage includes one-hot encoding of the categorical variables, splitting the data into their train (70%) and test set (30%), using an algorithm such as recursive feature elimination (RFE) for feature selection, and applying the three sampling techniques to overcome the issue of imbalance. Normalizing the numeric attributes may also be considered in this stage. Next, the 3 machine learning algorithms will then be applied for each of the sampling techniques. Lastly, each of the 9 models will be tested with their associated test sets and will be evaluating each of their performances by using metrics such as accuracy, recall, precision, and its area under the curve (AUC) value along with plotting the receiver operating characteristic curve (ROC) to analyze the difference in performance between the models.



References

Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., ... & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, *6*, 100163.

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-9). IEEE.

Canadian Bankers Association . (2023, March 31). Focus: Credit Cards: Statistics and Facts. Cba.ca. https://cba.ca/credit-cards

Chartered Professional Accountants of Canada. (2023, February 15). *Unlikely targets: More young Canadians report being a victim of financial fraud than older Canadians: CPA Canada survey reveals*. Cpacanada.ca.

https://www.cpacanada.ca/en/the-cpa-profession/about-cpa-canada/media-centre/2023/february/cpa-canada-fraud-survey-2023

Dickey, C. (2023, March 13). *Ways your credit card info might be stolen and how to prevent it*. Bankrate; Bankrate.com. https://www.bankrate.com/finance/credit-cards/5-ways-theives-steal-credit-card-data/

Dollarhide, M. (2021, July 12). *Who is Liable for Credit Card Fraud?* Investopedia. https://www.investopedia.com/ask/answers/09/stolen-credit-card.asp

Falco, A. (2023, February 21). *Understanding the Threat of Card Transaction Fraud and its Impact on the Financial Ecosystem | Waylay Blog*. Waylay.io. https://www.waylay.io/articles/understanding-the-threat-of-card-transaction-fraud-and-its-impact-on-the-financial-ecosystem#:~:text=Globally%2C%20card%20fraud%20is%20making,less%20than%2024%20hours%20old.

Government of Canada. (2023). Credit card fraud . Canada.ca. https://www.canada.ca/en/financial-consumer-agency/services/credit-fraud.html

Haddab, D. M. (2022). Data Science & Machine Learning Methods for Detecting Credit Card Fraud. *International Journal of Data Science and Advanced Analytics (ISSN 2563-4429)*, *4*(4), 71-75.

Keita, Z. (2022, September 21). Classification in Machine Learning: An Introduction. Datacamp.com; DataCamp. https://www.datacamp.com/blog/classification-machine-learning

Kosourova, E. (2022, March 21). *Data Science in Banking: Fraud Detection*. Datacamp.com; DataCamp. https://www.datacamp.com/blog/data-science-in-banking

Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002, January). Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the 1st international naiso congress on neuro fuzzy technologies* (Vol. 261, p. 270)

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia computer science*, *132*, 385-395.

Rafter, D. (2017, September 14). *What Is Credit Card Fraud?* LifeLock by Norton. https://lifelock.norton.com/learn/fraud/what-is-credit-card-fraud

Shenoy, K. (2019). *Credit Card Transactions Fraud Detection Dataset*. Kaggle.com. https://www.kaggle.com/datasets/kartik2112/fraud-detection

White, A. (2020, September 3). *Here's how credit card fraud happens and tips to protect yourself*. CNBC; CNBC. https://www.cnbc.com/select/credit-card-fraud/

Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit card fraud detection using machine learning as data mining technique. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(1-4), 23-27.