MGT 6203 Project Final Report Team 36 - Fraud Detection

Leslie Francis, Alejandro De Jesus, Tania De Pena, Vaisag Radhakrishnan

Abstract

E-commerce fraud is defined as any type of fraudulent transaction that occurs on e-commerce platforms. Many data mining techniques and machine learning algorithms can be used to detect fraudulent transactions based on various features about the transaction. This study is intended to test several machine learning classification techniques for determining if a transaction is fraudulent or not. Fraud Detection dataset from Kaggle is used in the research. The dataset is imbalanced and required balancing to ensure equal class distribution. The classification algorithms like logistic regression, KNN, Random Forest, and SVM are used as part of this study. Logistic regression produced the highest cross-validated mean F1 score, a metric that measures a model's accuracy, making it the best model to classify fraudulent transactions in this study.

Introduction

According to the new "State of Fraud 2023" study from fraud prevention solution provider Signifyd, the total cost of e-commerce fraud in 2023 will reach $206.8 billion. These frauds affect both companies as well as customers. Although most banks will reimburse fraudulent transaction cost, it gets passed to the customer. To identify and stop potential financial loss for both the businesses and their clients, high accuracy fraud detection is essential. Effective fraud detection can protect customer data and maintain trust between companies and their clients. In this project, we are using e-commerce transaction data to detect fraudulent transactions. The aim is to find out whether a transaction is fraud or not. Identifying this will help banks and payment providers alert their customers in case of suspicious activities, so they can block these transactions.

Even though this study is done on e-commerce transactions, the techniques used in this research can be used in a variety of fields like financial institutions for detecting credit-card frauds(Varmedja et al, Awoyemi et al, Chaudhary et al), healthcare for fraudulent claims(Kumaraswamy, N. et al), insurance industry for detecting fraudulent accident claims(Viaene et al) etc. As fraudsters evolve and businesses battle to contain fraud, manual methods to detect and avoid fraud in most cases are slow and ineffective. Machine learning techniques like the ones used in this research will help detect the fraudulent transactions faster with less manual intervention saving time and money for both the business and their customers. Financial institutions may utilize these techniques to contact the customers in case of suspicious activities to confirm the transactions. Healthcare institutions and insurance providers can use these methods to further investigate those transactions that are classified as fraudulent.

The general approach followed is to explore the data, carry out feature engineering by using techniques like one-hot encoding, balance the dataset, train, and test the data using various machine learning techniques. The initial hypotheses are that young people are more likely to commit frauds as compared to others, and that fraud is committed on higher valued items.

Literature Review

The simplicity and ability to explain the model are typically ignored in favor of focusing on predictive performance. Most papers focus on predictive performance as opposed to the interpretation of the model - while predictive performance is important, interpreting a model is equally important to make critical business decisions. As fraud detection occurs in the financial services industry, many datasets will include masked columns for privacy concerns. Both Varmedja et al and Awoyemi et al used methods like Principal Component Analysis to keep the data anonymous. Due to this, they were able to make accurate predictions, but the model's interpretability suffered.

Fraud detection being a key factor in gaining and maintaining customer trust, businesses come up with various machine learning techniques to predict fraudulent transactions. Numerous researchers are trying their best to better the existing methods. Varmedja et al used machine learning techniques like logistic regression, Naïve Bayes, Random Forest and Multilayer perception to obtain high accuracy in credit card fraud detection. Awoyemi et al used Naïve Bayes, logistic regression, and K-nearest neighbor techniques to obtain high predictive performance.

Most fraud detection datasets are imbalanced. An imbalanced dataset is a dataset where the number of records belonging to one class is much higher than the other. This usually occurs in the fraud detection use case as the number of fraudulent transactions are much lesser than that of the number of non-fraudulent ones. It is extremely important to balance the dataset to get good predictive performance. Varmedja et al used SMOTE technique to balance the dataset by a technique called oversampling where the records in the smaller class is increased in a balanced way to make sure that the samples of both classes are equal. While other research papers like Awoyemi et al explored a combination of oversampling and under sampling to achieve balance in the dataset.

Problem Statement and Exploratory Data Analysis

The objective of the assignment is to predict whether an e-commerce transaction is fraudulent or not based on various available features. Accurately predicting this will help increase the customers' trust on the companies and help reduce losses for both the customers as well as the companies.

Source Data

The datasets for this project are from Kaggle. There are two datasets used here. First, a transaction dataset that contains the fraudulent as well as non-fraudulent transactions. Next, ip address to country cross reference table to find out the country associated with transactions in the transaction dataset.

Fraud dataset is shown in *Table 1* below.

Feature Description	Type
Customer ID	Numeric
Amount of the purchase	Numeric
Date and time of the purchase	Date
Used Browser for the transaction.	Categorical
Source (Direct, Ad, etc.)	Categorical

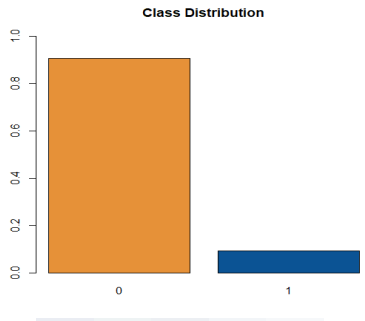
Date and time of the sign up	Date
Age	Numeric
Gender	Categorical
Ip address	Numeric

Table 1: Shows the important features and types in the fraud transactions dataset

The second dataset includes the Ip address boundaries for each country.

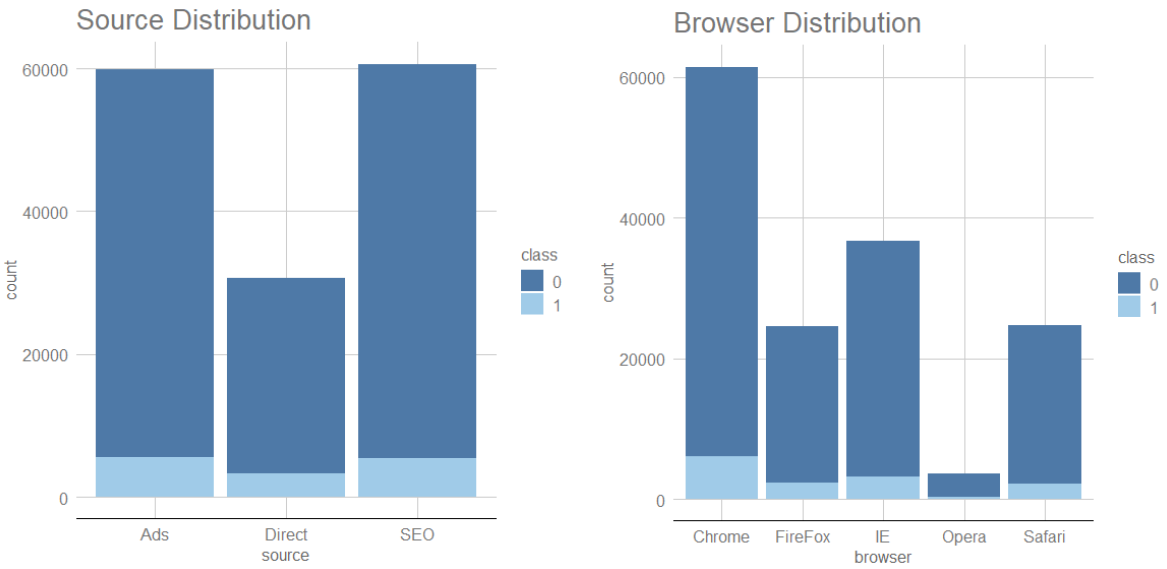
Exploratory Data Analysis and Data Wrangling

We have 151,000 unique transactions made from January 2015 to December 2015 across different locations. The dataset contains 14K fraudulent transactions and 138K non-fraudulent transactions, which means we are working with an imbalanced dataset, since only the 9% of the transactions are marked as class 1 (fraudulent trans.) *Figure 1* shows the distribution of the fraudulent(class = 1) and non-fraudulent(class = 0) transactions.



Post the initial exploration and merging of the datasets, the next step was to check for data cleaning process. The initial search was for null values; however, the dataset contains data in every column. The next step was to search for outliers and create some graphs to show how the data is distributed.

- The least used source of purchase is Direct purchases. This is expected, as companies spend money on advertisements and do search engine optimization to appear on top of the search results. See *Figure 2*.
- The most used Browser is Chrome and the least used one is Opera. See *Figure 3*.
- The variables age and purchase value have a skewed distribution to the right. There are no outliers in both of these numeric variables. *Figure 4* and *Figure 5* show these distributions.



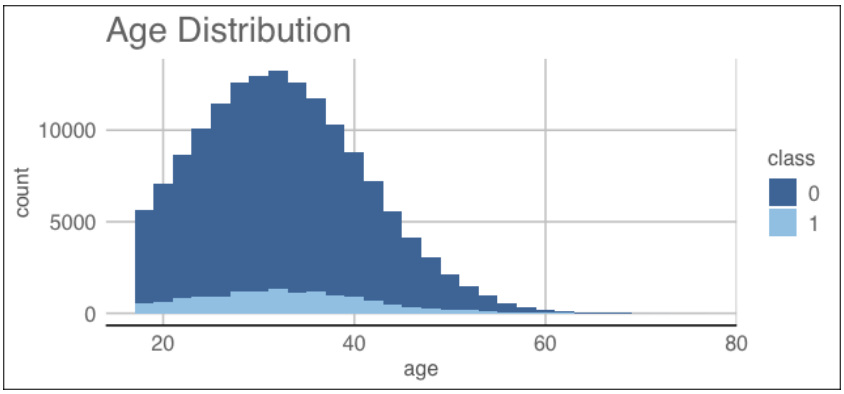


Figure 4: Age Distribution

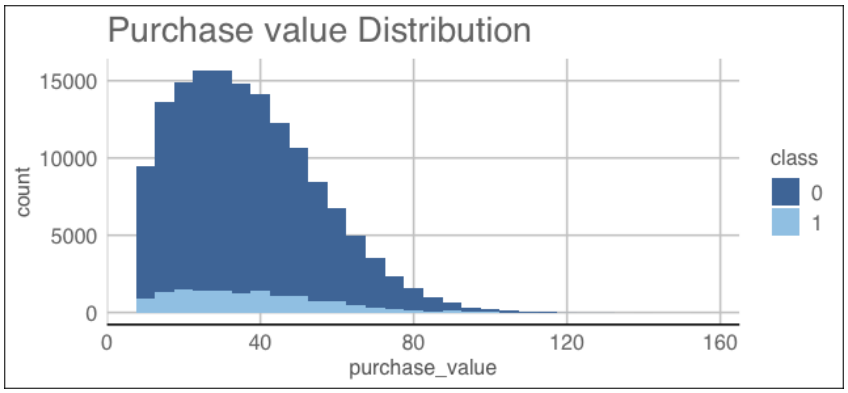


Figure 5:Purchase Value Distribution

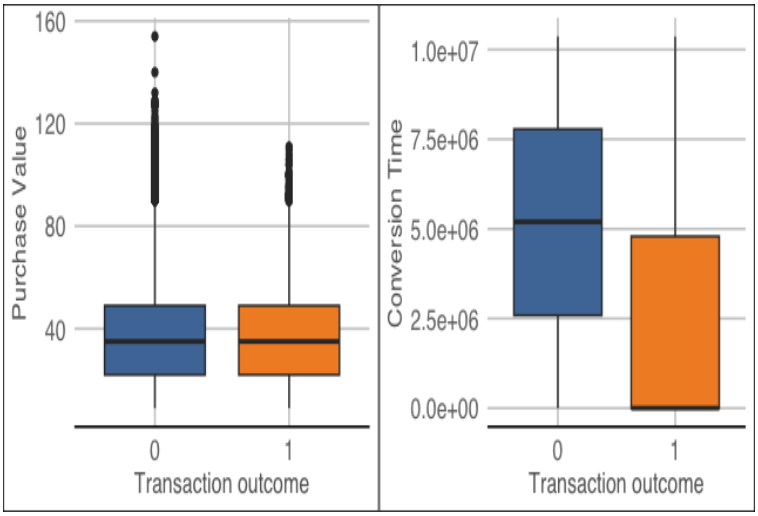


Figure 6: Boxplot of Purchase Value and Conversion Time vs Transaction Outcome

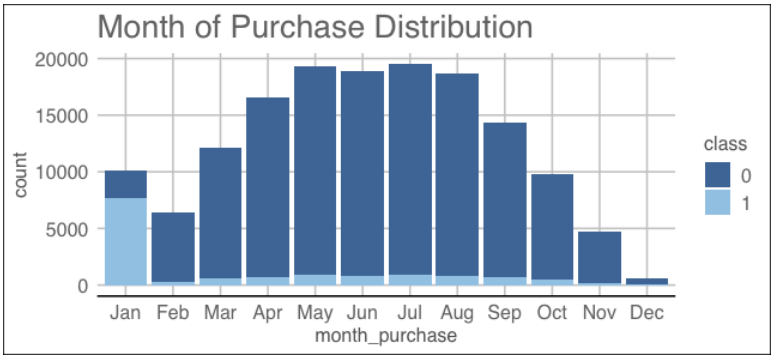
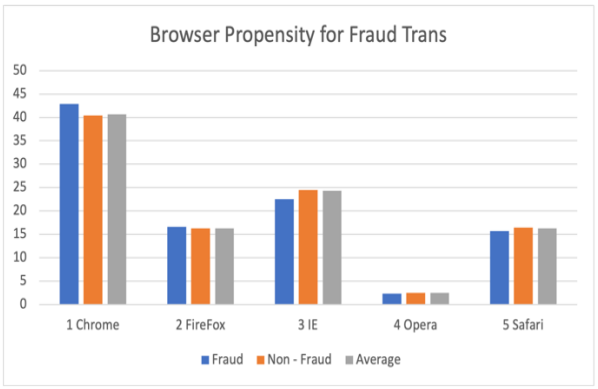


Figure 7: Distribution based on Month of Purchase

Next, we analyzed the difference between the fraudulent and non-fraudulent transaction, by looking at the propensity of both groups and comparing it with the average of the dataset. See Figure 8 for detailed analysis on various fields. Below are some of the insights on this analysis.

- The browser with most fraudulent transaction's propensity is chrome (43%) vs the average (40.7%). See Figure 8.1
- On weekends we have a higher propensity for fraudulent transactions than the rest of the week (46.4% weekend propensity for fraud transactions vs 43.1% the average). See Figure 8.2
- Tuesdays and Wednesdays are the days of the week with lowest propensity for fraudulent transactions. See figure 8.3
- As discussed earlier, purchases made the same day of sign up have higher propensity for fraudulent transactions. See figure 8.4



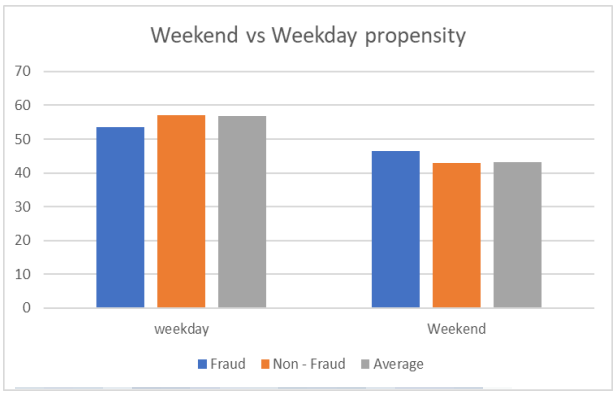
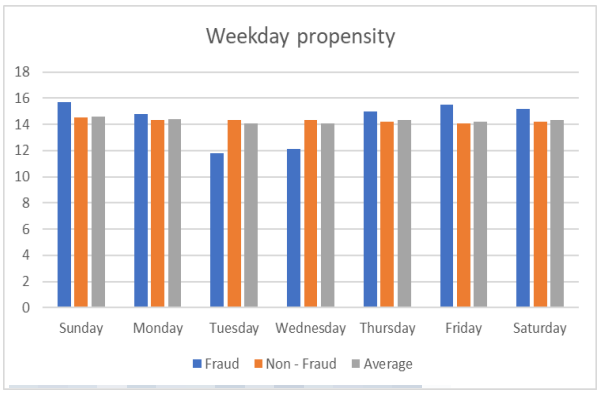


Figure 8.1 Browser Propensity



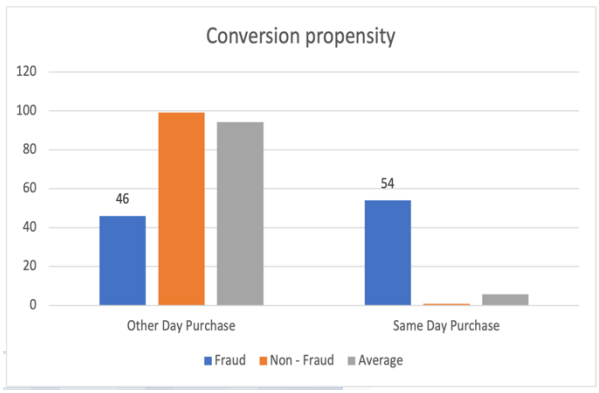


Figure 8: Various Propensity Analysis

Modelling Approach

Post the initial exploratory data analysis, the team carried out checks for outliers, erroneous data or fields that require data cleaning prior to model training. New features were added to both explore the data and to include in the models as explained in the feature engineering section.

Balancing the Data

Fraud detection datasets tend to be imbalanced as there are considerably more legitimate transactions than fraudulent ones. For this reason, the dataset needs to be balanced by under sampling legitimate transactions or by oversampling fraudulent ones. This will serve the purpose of making the class that we are interested in identifying (frauds) more relevant. The team utilized SMOTE(Chawla et al) technique to balance the dataset. SMOTE is a popular technique that is proven effective on imbalanced datasets as the one we are working on.

Feature Engineering

Table 2 shows the new variables derived from existing data.

Variable Name	Description
seo	1 when purchase was made from SEO otherwise 0
ads	1 when purchase was made via ad else 0
Chrome	1 when browser used is Chrome else 0
Safari	1 when browser used is Safari else 0
Opera	1 when browser used is Opera else 0
Ie	1 when browser used is IE else 0
Male	1 when male else 0
Young	1 when age < 30 else 0
Middle_aged	1 when age is between 30 and 50 else 0
Monday – Saturday (6 variables)	Monday = 1 when purchase day is Monday, Tuesday = 1 when purchase day is Tuesday etc
Weekend	1 when purchase is on a weekend else 0
Night	1 when purchase is in the night else 0
Afternoon	1 when purchase is in the afternoon else 0
Evening	1 when purchase is in the evening else 0
Country dummy variable	Country dummy variable for each country

Table 2: Dummy variables created

Build and Test Models

Our approach for model building was to build and compare a few classification models. We used N-fold cross validation to build and a test sample of 40% to compare models. Here are the results from model testing phase. Since we wanted to reduce false positive and false negative, we used F1 score as the determinant to choose the best model. A higher F1 indicates a better balance between precision and recall. This balance is the goal because we want to correctly identify a high percentage of frauds to mitigate losses, but we do not want to saturate clients by blocking their accounts/credit cards when they are making legitimate transactions.

1. Logistic Regression:

Confusion Matrix:

Predicti	Truth		
		0	1
	0	41048	3977
	1	38	268

o n			

2. KNN
   Grid search was performed for KNN for the following K values (1,3,5,7,9,11,13,15). Among these the best K was found to be 15.

   Confusion Matrix:

Prediction		Truth	
		0	1
	0	11768	569
	1	1939	836

3. Random Forest Classification
   Grid search was performed on Random Forest and the number of trees chosen was 500 with 101 variables tried at each split.

   Here's confusion matrix from Random Forest. This was not considered as this was run on the unbalanced dataset. After balancing the number of records grew which caused us problems running the model.

Prediction		Truth	
		0	1
	0	40828	1905
	1	258	2340

Other Metrics of both models.:

Model	F1	Precision	Recall	Sensitivity	Specificity	Accuracy
Logistic regression	0.67	0.87	0.54	0.54	0.99	0.95
KNN (k=15)	0.40	0.30	0.60	0.60	0.86	0.83

Note: Full results for all hyperparameters is in Appendix.

Conclusion

While doing the propensity analysis, we hypothesized that variables with high propensity would be crucial in correctly predicting fraud as they show clear distinctions between legitimate and fraudulent transactions. This hypothesis is confirmed in the results (see Appendix for full results) as the logistic regression model with high propensity predictors outperforms the other trained models. It has the best performance when F1 scores are compared. It also outperforms with respect to Accuracy, Precision and Specificity. This would be the model that we recommend that the business use.

Businesses could predict a fraudulent transaction and reject the transaction thereby avoiding potential fraud using this model. Post this, customers could be notified or contacted whenever there is a potential fraudulent transaction, so they are aware why their transaction was rejected if it was a legitimate transaction. Same day purchases could be restricted by additional two factor authentication or other similar security measures.

Future Work and Lessons Learned

As part of this study, we tried out multiple models including ensemble models. Although ensemble models tend to give the best predictive performance, the model's ability to explain is compromised. Additionally, running ensemble methods on a huge dataset like the one we tried was computationally intensive and need better methods to optimize the run time.

Although there were multiple models tried out as part of this study, there are multiple extensions that this research could take. Other evaluation metrics like ROC could be tried out as part of future work. Also, researching on the importance of variables based on the results from Random Forest will give us an idea of what variables are the most important to predict a fraudulent transaction. Although the results from this study were good, we could expand this by using more variables to make the results more reliable.

Appendix

Complete Model Results:

	model_name	mean_thr	mean_f1	mean_precision	mean_recall	mean_sensitivity	mean_specificity	mean_accuracy
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Logistic regression: high …	0.98	0.667	0.873	0.540	0.540	0.992	0.950
2	KNN (k=15)	0	0.405	0.304	0.603	0.603	0.858	0.834
3	KNN (k=13)	0	0.401	0.299	0.605	0.605	0.854	0.830
4	KNN (k=11)	0	0.396	0.294	0.607	0.607	0.850	0.827
5	KNN (k=9)	0	0.390	0.287	0.611	0.611	0.843	0.821
6	KNN (k=7)	0	0.383	0.278	0.616	0.616	0.835	0.814
7	KNN (k=5)	0	0.372	0.265	0.622	0.622	0.822	0.803
8	KNN (k=3)	0	0.354	0.247	0.629	0.629	0.801	0.785
9	KNN (k=1)	0	0.334	0.225	0.651	0.651	0.768	0.757
10	Logistic regression: all p…	0.99	0.169	0.884	0.0935	0.0935	0.999	0.914

References

1. Varmedja, Dejan, et al. "Credit card fraud detection-machine learning methods." 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). IEEE, 2019.
2. Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." 2017 international conference on computing networking and informatics (ICCNI). IEEE, 2017.
3. “New FTC Data Show Consumers Reported Losing Nearly $8.8 Billion to Scams in 2022.” Federal Trade Commission, 23 Feb. 2023, www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022.
4. “New Data Shows FTC Received 2.8 Million Fraud Reports From Consumers in 2021.” Federal Trade Commission, 22 Feb. 2022, www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0
5. Berthiaume, Dan. “Study: Total E-commerce Fraud in 2023 Will Exceed $200 Billion.” Chain Store Age, 25 Jan. 2023, https://chainstoreage.com/study-total-e-commerce-fraud-2023-will-exceed-200-billion#:~:text=According%20to%20the%20new%20%E2%80%9CState,tangible%20losses%20for%20online%20retailers
6. Brownlee, Jason. “Cost-Sensitive Learning for Imbalanced Classification”, 7 Feb. 2020, https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/
7. Brownlee, Jason. “A Gentle Introduction to Threshold-Moving for Imbalanced Classification”, 10 Feb. 2020, https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/
8. Brownlee, Jason. “SMOTE for Imbalanced Classification with Python”, 17 Jan. 2020, https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
9. Kumaraswamy, Nishamathi, et al. "Healthcare fraud data mining methods: A look back and look ahead." Perspectives in Health Information Management 19.1 (2022).
10. Viaene, Stijn, et al. "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection." Journal of Risk and Insurance 69.3 (2002): 373-421.
11. Chaudhary, Khyati, Jyoti Yadav, and Bhawna Mallick. "A review of fraud detection techniques: Credit card." International Journal of Computer Applications 45.1 (2012): 39-44.

12. Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.