# Fraud Detection: Credit Card Data

Allyson Busch

Winter 2019

**Abstract**

Credit card usage as well as mobile payment options, such as ApplePay or SamsungPay, are growing in popularity among consumers. Fraudulent charges, however, are also on the rise. This study utilizes a dataset of over 280,000 credit transactions from European consumers over the span of 48 hours, with over 400 charges being identified as fraudulent. The goal of this project is to be able to define what a typical fraudulent charge looks like in the dataset as well as identify any patterns present in the dataset. The final and overarching goal is to be able to create a machine learning model that can predict whether a transaction is fraudulent or normal, and be able to apply this to future transactions in the dataset. The study was able to determine K-Nearest Neighbors as having the best results when tested against two other models, but there is still opportunity for growth in recall and precision as well as applications to more specific credit usage.

**Research Questions**

The study intends to apply machine learning techniques on a dataset of real credit card transactions in order to create a system that can identify fraudulent charges. The research questions that the study will aim to answer begin with an evaluation of the data in Exploratory Data Analysis. The questions are as followed:

- How many entries in the dataset are labeled fraudulent?
- What do the transaction amounts look like?
- What are the summary statistics of the transaction amounts for fraudulent and normal transactions?
- Do fraudulent transactions occur more often at certain times of day?

After the exploratory questions are answered, the project will move into the construction of the machine learning models, which leads to the final research question:

- How well does the model identify fraudulent transactions?

Specifically, the study intends to investigate the dataset using Isolation Forest and Local Outlier Factor. After these two unsupervised learning algorithms, the study intends to look into density based models, specifically k-nearest neighbor. These models will be compared to see what is the best model for prediction based on the data present..

**Literature Review**

Almost all of historical science worked from the premise of fitting models to data; Galileo, Newton, and Mendel all designed experiments and collected data, with the hope of extracting knowledge by devising theories, or rather by building models to explain the data they had observed (Alpaydin, 2016). The data collection of the past has grown with the adoption of technology in consumers' everyday life, leading to Big Data. While the term Big Data is debated, it is widely accepted that it varies from traditional data in that it has a high volume of data, has a variety of data structures, and has a high velocity (Inge & Leif, 2017). While most of the modern day data analysis can no longer be done manually as it would in the past, the process of data collection and model building continues as technology evolves to handle larger and larger amounts of data (Alpaydin, 2016). There is a growing interest in computer programs due to the enormous amount of data being collected and processed; these computers are expected to analyze and extract information automatically from datasets, or in other words, learn from the datasets (Alpaydin, 2016). The theory that underlies this machine learning is based on statistics but is partnered with computer science as part of artificial intelligence (Alpaydin, 2016).

The realm of machine learning has several applications, even within the world of credit cards and financial accounts. The high volume and accuracy of historical data paired with the quantitative nature of the financial field has made the industry a prime candidate for machine learning and artificial intelligence techniques (Jacobs, 2018). Risk modeling in both financial (credit, market, business, and model) as well as non-financial (operational, compliance, fraud, and cyber) parts of the industry have been a natural domain (Jacobs, 2015).

In 2014, the global credit card business totaled $28.84 trillion, with lost $16.31 billion in fraudulent charges (Kultur & Caglayan, 2017). This fraudulent portion came to a loss of 5.65 cents for every $100, and is expected to grow to exceed $35.54 billion total in 2020 (Kultur & Caglayan, 2017). Fraudulent online transactions have increased and are expected to continue at 51% with the industries growth in 2020 (Manlangit, Azam, Shanmugam, & Karim, 2019). This is being further fueled by the introduction of mobile payment systems, such as ApplePay, which are making credit transactions easier and more convenient (Somasundaram & Reddy, 2018). The growth of consumers' dependency on e-commerce and online payments has significance as organizations and people are suffering substantial financial loss from fraudulent charges (Minastireanu & Mesnita, 2019). Online bank fraud is continuously evolving and can be difficult to analyze and detect as the behavior is dynamic and dispersed (Minastireanu & Mesnita, 2019).

This study intends to apply the use of machine learning algorithms in predicting cases of fraudulent charges on credit card accounts. A fraudulent transaction is defined as utilizing a credit card without the knowledge of the owner or authorization from the owner (Manlangit, Azam, Shanmugam, & Karim, 2019). Fraud can be both detrimental to the victim as well as the merchant, as they then have a compromised product payment (Manlangit, Azam, Shanmugam, & Karim, 2019). Currently, there are two levels of fraud protection: fraud prevention (actions done to stop fraud from happening before it occurs) and fraud detection (identifying fraudulent transactions as they happen) (Manlangit, Azam, Shanmugam, & Karim, 2019). In the realm of fraud detection, complex decision-making systems based on algorithms and analytical technologies that can learn from previous experiences and create patterns have been developed (Minastireanu & Mesnita, 2019). Common techniques applied are Decision Tree, Artificial Neural Network, Artificial Immune Systems, K Nearest Neighbor, Support Vector Machine, Genetic Algorithm, and Hidden Markov Model (Singh & Jain, 2019). The goal of fraudulent detection systems is to maximize the true positive and minimize the false positive predictions of legitimate transactions (Singh & Jain, 2019).

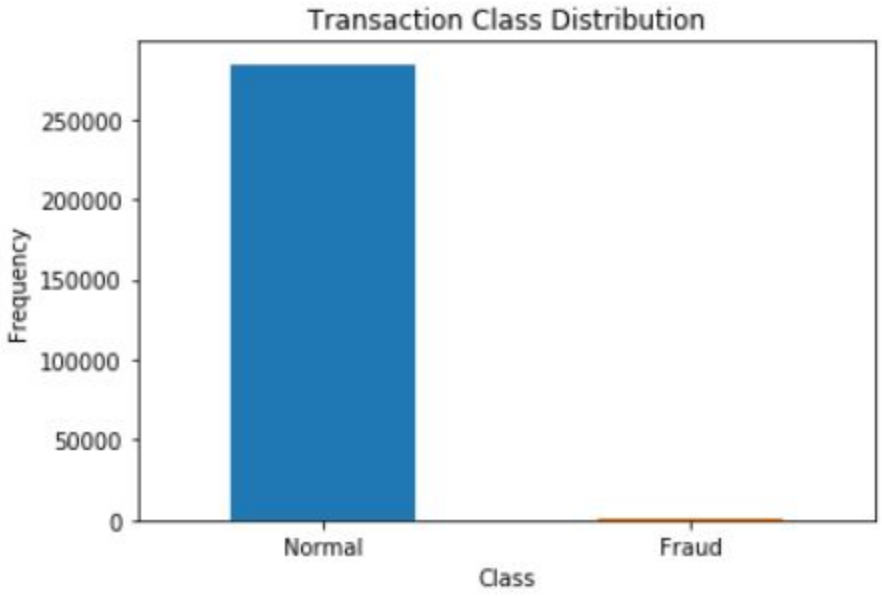
The dataset the study is working with consists of transactions made by credit cards in September 2013 by European cardholders over the span of two days in which 492 transactions were found to be fraudulent of 284,807 transactions (Machine Learning Group, 2018). This dataset was imbalanced, meaning that one of the classes (in this case, normal) exhibited dominance over the existing class (fraudulent) (Somasundaram & Reddy, 2018). The dominating class is referred to as the majority class, with the other class considered the minority class (Somasundaram & Reddy, 2018).

**Methodology**

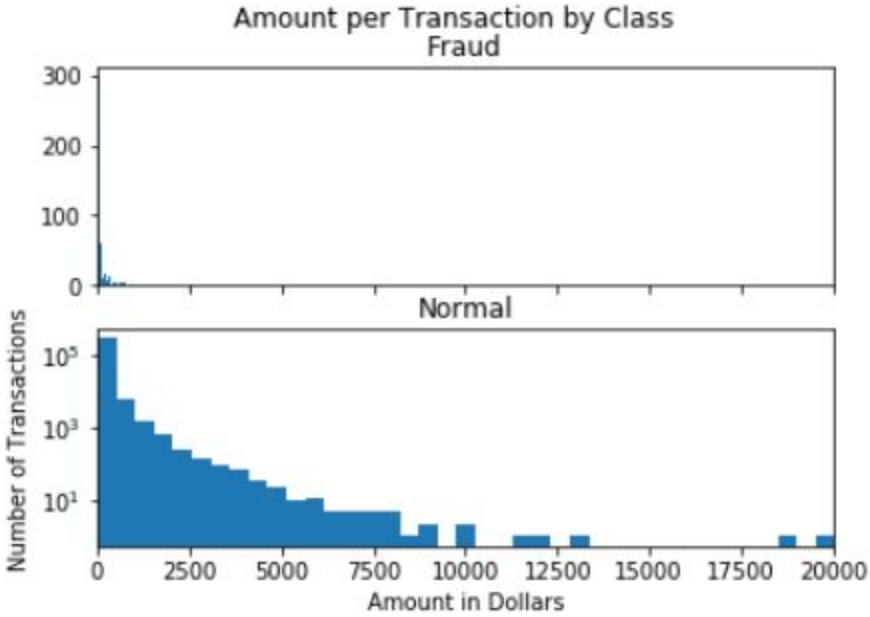
The study started with the importing of libraries, including numpy, pandas, scipy, matplotlib.pyplot, seaborn, plotly, and sklearn. After all the libraries were uploaded, the dataset was imported from the file "creditcard.csv" and the column names were printed out as well as head to ensure that everything imported correctly. The dataset contained the following variables:

- Time: This variable is the number of seconds elapsed between the transaction and the first transaction in the dataset.
- V1 through V28: This variable is the principal components obtained with PCA
- Amount: This variable is the transaction amount in dollars.
- Class: This is a binary variable with 1 as fraudulent transactions and 0 for normal transactions.

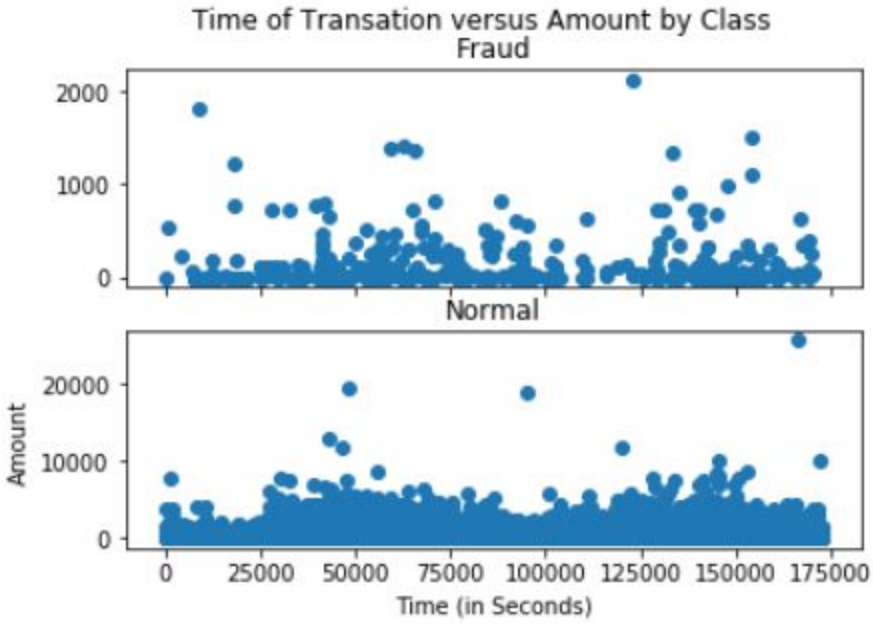
The dataset contained 284,807 transactions with 31 variables, as listed above. There were no null values present in the dataset, meaning we could move forward into Exploratory Data Analysis. The first initiative was looking into the breakdown of the transaction classes of normal and fraudulent. I created a pandas graph of the count of class categories, which provided a graphical representation of the inequality of transaction classes. I then split the dataset into two datasets, separating them by the Class variable; fraudulent data was sorted into the Fraud created set with the requirement of Class equal to 1 while Class equal to 0 was sorted into the newly created Normal set. I then evaluated the shape of the datasets. Fraud has a shape of 492 by 31, meaning all the variables are remaining, but there are only 492 fraudulent transactions in the dataset of 284,807 transactions. The shape of the normal transaction is 284,315 by 31, making up the bulk of the data.




After this we can look into the statistics of the fraudulent and normal datasets separately. Starting with the 492 fraudulent transactions, the mean of the group is $122.21 with a standard deviation of $256.68. 50% of the data fell below $9.25, with a maximum charge of $2,125.87. The normal transactions have a mean of $88.29 and a standard deviation of $250.11. 50% of the data fell below $22.00 with a



maximum charge of $25,691. I then created graphical representation of the amount per transaction by class of fraudulent and normal transactions. The revealed shape of the data shows there are a higher number of small transactions and the count of transactions decreases as the amount in dollars increases at a roughly exponential rate.




After this evaluation, I looked into the time of transaction for each class to see if there was any pattern in the dataset. The graph was created with the Fraudulent and Normal transaction separated. The time of the transaction was put on the x axis while the amount was put on the y axis. There was not a significant difference in the time of transaction versus the amount of transactions in each category and no pattern could be identified based on the graph. This was verified by the correlation analysis that failed to find a significant correlation between time and amount, amount and class, or any of the other variables present in the dataset.

This led into the model creation section of the study. The first model I explored was an Isolation Forest Algorithm, which is based on the decision tree model. This outlier detection method is different than the other popular models in that it explicitly identifies anomalies instead of profiling normal data points (Lewinson, 2018). This model was compared to the Local Outlier Factor Algorithm, which is another unsupervised anomaly detection method (Scikit-Learn, 2019). This method computes the local density deviation of a given data point with respect to its neighbors, and considers outliers as samples that have a substantially lower density than their neighbors (Scikit-Learn, 2019). Isolation Forest had a precision rating of 31% for the fraudulent, meaning it is able to not label an instance positive that is actually negative 31% of the time. Recall was 31% as well, meaning it was able to identify positives 31% of the time. The overall accuracy score was 99.76% for Isolation Forest while Local Outlier Factor has an accuracy score of 99.67%. The precision for Local Outlier Factor dropped to 5%, meaning it was able to not label an instance as positive for fraud when it is actually negative only 5% of the time. The recall score was also 5%, meaning it was only able to identify the fraudulent cases 5% of the time. After running

the K-Nearest Neighbors model, it came to a precision score of 99%, much higher than the other models. The recall score was 74%, meaning it was able to identify fraudulent charges 74% of the time.

**Discussion & Conclusion**

The first research question in the study was to identify how many of the transactions in the dataset were fraudulent. The dataset contained 284,807 transactions, 492 of which were fraudulent charges. Fraudulent charges come up to only 0.17% of the dataset. This makes an incredibly uneven dataset, with the vast majority of data points being considered normal.

The second and third research questions are to look into what the transaction amounts were. I first looked into the fraudulent charges and found the average to be $122.21 with a standard deviation of $256.68. The maximum charge was $2,125.87, but 75% of the data was below $105.89. In comparison, the normal transactions had a mean of $88.29, which was lower than the fraudulent charges. The standard deviation was $250.11, which was very similar to the fraudulent transactions. The maximum charge was $25,691.16 which is much higher than the fraudulent category. However, 75% of the data is below $77.05, which is lower than the fraudulent charges. From this we can determine that overall the fraudulent charges were higher on average than the normal charges, however the range in the fraudulent charges is smaller.

The fourth research question is whether fraudulent transactions have a pattern in regards to the time of day they occur. After graphing the data, there does not seem to be a significant pattern in fraudulent charges or normal charges, and a correlation analysis verified this.

The final research question concerns the overarching model and how accurately the study could predict fraudulent charges. The first model we ran was Isolation Forest which had a precision and recall rate of 31%. In comparison, the second model which is Local Outlier Factor had a precision and recall rate of 5%. In comparison, K-Nearest Neighbors was able to run with a 99% precision rate and a 74% recall rate, much higher than the other two models.

Moving forward with the study, there are opportunities in processing the uneven dataset in more precise ways with more finely tuned algorithms to predict. In addition, looking specifically into mobile transactions or more specified transactions could yield more patterns that would help identify how fraudulent charges are made. For example, looking into transactions at a grocery store to identify if there is a pattern of fraudulent charges across time and if there are identifiers in what people are buying the trigger a fraudulent charge. The limitations in this study was finding alternative ways to deal with uneven

datasets, which showed in the accuracy rates of the models performed. Additionally, the dataset was limited in location, but provided a large amount of data points to be able to train into the model.

**References**


1. Alpaydin, E. (2016). *Machine Learning: The New AI.* Cambridge, MA: MIT Press.
2. Inge, R. & Leif, J. (2017). *Machine Learning: Advances in Research and Applications.* New York, NY: Nova Science Publishers, Inc.
3. Jacobs, M. (2018). The validation of machine-learning models for the stress testing of credit risk. *Journal of Risk Management in Financial Institutions, 11*(3), p. 218-243. Retrieved from https://eds-b-ebscohost-com.ezproxy.bellevue.edu/eds/results?vid=22&sid=a1616536-5c7d-4402-b5bd-9e0a77f5945c%40pdc-v-sessmgr05&bquery=machine+learning+credit&bdata=JmNsaTA9RlQxJmNsdjA9WSZ0eXBlPTAmc2VhcmNoTW9kZT1BbmQmc2l0ZT1lZHMtbGl2ZQ%3d%3d
4. Kultur, Y. & Caglayan, M. (2017). Hybrid approaches for detecting credit card fraud. *Expert Systems, 34*(2), np. Retrieved from https://eds-b-ebscohost-com.ezproxy.bellevue.edu/eds/pdfviewer/pdfviewer?vid=12&sid=a1616536-5c7d-4402-b5bd-9e0a77f5945c%40pdc-v-sessmgr05
5. Lewinson, E. (2018, Jul 2). Outlier Detection with Isolation Forest. Retrieved from https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e
6. Machine Learning Group. (2018). Credit Card Fraud Detection. Retrieved from https://www.kaggle.com/mlg-ulb/creditcardfraud
7. Manlangit, S., Azam, S., Shanmugam, B., & Karim, A. (2019). Novel Machine Learning Approach for Analyzing Anonymous Credit Card Fraud Patterns. *International Journal of Electronic Commerce Studies, 10*(2), p. 175-202. Retrieved from https://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=14&sid=9c5f44d2-c8b6-46f9-9be8-3f9a313b2127%40sdc-v-sessmgr01
8. Minastireanu, E. & Mesnita, G. (2019). An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection. *Informatica Economica, 23*(1), p. 5-16. Retrieved from https://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=6&sid=9c5f44d2-c8b6-46f9-9be8-3f9a313b2127%40sdc-v-sessmgr01
9. N.A. (2019). Outlier Detection with Local Outlier Factor (LOF). *Scikit - Learn.* Retrieved from https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html

10. Singh, A. & Jain, A. (2019, Dec). An Empirical Study of AML Approach for Credit Card Fraud Detection - Financial Transactions. *International Journal of Computers, Communications, & Control, 14*(6), p. 670-690. Retrieved from https://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=140201221&S=R&D=iih&EbscoContent=dGJyMNHX8kSeprY4xNvgOLCmsEieqLBSs6y4TbKWxWXS&ContentCustomer=dGJyMODb6oT06%2BNT69fnhrnb4osA
11. Somasundaram, A. & Reddy, S. (2019). Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Computing and Applications, 31*, p.S3-S14. Retrieved from https://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=9&sid=9c5f44d2-c8b6-46f9-9be8-3f9a313b2127%40sdc-v-sessmgr01

**Appendix: Correlation Heatmap**