Analysis for Credit Card Fraud Detection

Your Name Ryan Habis
Student ID D00245309

November 23, 2025

Module: RESA C9009 - Research Process for Data Analytics
Assessment: Research Proposal
Word Count: Approximately 1500 words
Submission Date: Nov,23

1 Introduction and Background

Credit card transaction is the norm now a days since the rise of credit card being used around the world credit card fraud as well happens this results in a major lose of profits and harms the credit card business as a whole it ruins the companies reputation as well there clients are unhappy which they might move banks or sue the bank itself which affects there revenue, on a more severe cases it could bankrupt the companyas a whole.

Companies are trying to find a way to stop fraudulent credit card transaction which will stop the loss of there profit margin and unsatisfied customers that causes them inconvenience. In order to stop credit card fraud companies are working on a way to develop sophisticated systems to prevent theft of people's funds within there bank accounts.

The problem with credit card fraud is that it happens 0.2% Dal Pozzolo, Caelen, Johnson, and Bontempi (2015) of all transactions therefore it is complected to find the people that are committing the act since there is hundreds of thousands of transactions a day and the bank database is absolutely massive its like fining a needle in a haystack.

Within this research proposal it will be focused on analyzing and detecting fraudulent activities by using data analytical technics in order to make cense of the data we acquired. Understanding the data analyzing it will then allow us to identify reoccurring patterns on where fraud may have accrued, once the pattern is found we can investigate it further to be certain and to prevent the illegal activity from happening again. The dataset that will be analyzed is from a site called kaggle, it provided ideal data to be studied because it is a dataset that has fraudulent activities within itself it is a real dataset from a bank it contains "transactions made by credit cards in September 2013 by European cardholders." Dal Pozzolo et al. (2015)

2 Literature Review

Within this literature review the fraud that is detected consistently shows that the class imbalance is the main primary challenge for the annalist.Within the paper Weiss (1995) demonstrates that the standard analytical way often fails when the data we are looking for is in the minority, the dataset records "284,807 transactions"Dal Pozzolo, Caelen, Johnson, and Bontempi (2013) and 0.2% of the transactions are fraudulent having such a big number of translations often leads data analysist down the wrong path since there analyzing the data as a whole it can be easily misinterpreted making the model, therefore analyzing the majority of transactions therefore ignoring the fraudulent patterns.

Past studies show that in order to be effective in finding fraudulent data it all begins in exploring the dataset and understanding what what makes the difference in a fraudulent transaction and a normal transaction once these is understood then we can find out whats the root cause to the problem Phua, Lee, Smith, and Gayler (2010). Studies shown

by Bolton and Hand (2002) indicates that fraudulent transactions often shows distinct patterns in a few fields such as the transaction amount, the timing of the withdrawal of funds and the frequency on how often they they do it.

The importance of analyzing and having the data preprocessed so it can be well documented is a crucial step. Within this paper He and Garcia (2009) it discusses a few different types of sampling technics in order to handle the task of finding fraudulent transaction, the method that is being used is to dissect the dataset into smaller more manageable chunks, if the dataset was looked at as a whole it would be too much data and nothing would be found but by creating a smaller dataset with sample data from the main dataset once this is done patterns start to show but if the dataset was looked at as a whole it would be hidden to the human eye. Once the dataset is in a sample formate tools that help visualize the data in a chart formate so that the data make sense and understand it Baesens, Höppner, and Verdonck (2021) within this paper it demonstrates how different types of ways to visualize datasets.

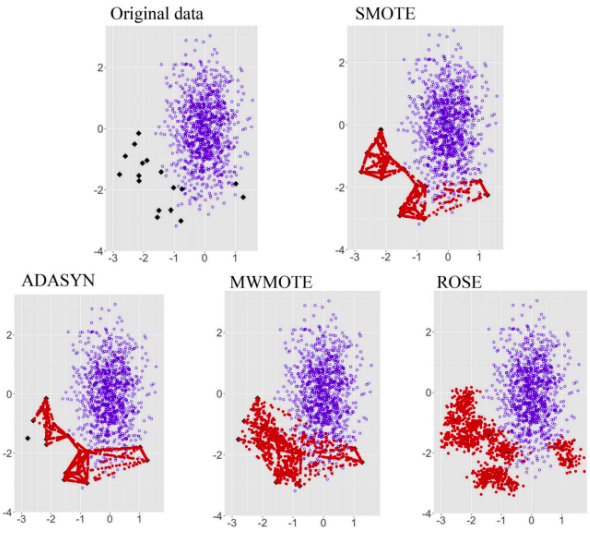

Figure 1: There are 5 charts showing the same data in a different way

- **Original data:** This is the original chart what we see here is black dots are the fraud cases and the blue dots are the legitimate cases.
- **Smoke:** Appear along the straight lines connecting the original black squares.
- **Adasyn:** Red dots are generated, but they are not uniformly distributed.
- **MWMOTE:** The red dots appear to form cleaner, more distinct clusters around the original black squares. The generation seems more focused and less noisy.
- **Rose:** The red dots are not just on lines between existing points. They appear in a small, smoothed cluster or "cloud" around each original black square.

Figure 2: Here is an example of an outlier which is 500 euro which raises suspension

The importance of understanding the data thats being analyzed is crucial. Within a study by Ahmed, Mahmood, and Islam (2016) talk about the quality of the data and how important it is to have it well organized and understood. This research proposal builds on this prespective by focusing exclusivley on the analytical work necessary for effective fraud detection, without having a sophisticated modeling approach and good understood of the dataset finding fraudulent transactions would be futile

3 Research Objectives

The main aim for this research project is to develop a way in order to visualize the data in an analytical way having the dataset that has credit card fraud within it and having a way to detect the fraudulent activities is the main goal, to do this the research objectives are:

1. To perform a thorough exploration of the data in the credit card fraud dataset to be able to see the characteristics and the statistical properties of what makes this data legit or fraudulent transaction.

2. To identify and be able to sum up all the fraud transactions within the dataset finding the difference and the relationship and what patterns the data show by looking at some key features in the data set such as the time, amount, and from whomever the transaction was authorized made by.

3. To investigate the the huge dataset that has imbalanced data and take sample data so it can be studied more thoroughly because the dataset looking at it as a whole will lead to the wrong result, creating a clear dataset so we can visualize will create a better way to examine the fraudulent transaction patterns.

4. The last step involve documentation of the findings of the fraudulent data and steps that should be taken in order to prioritize the and stop fraudulent transaction.

4 Research Questions and Hypotheses

4.1 Research Questions

This study will talk about the following research questions:

1. What are the key characteristics and visual patterns that show fraudulent transactions from real legitimate ones across the dataset?
2. How are patterns like ('Time') and transactions ('Amount') related to the likelihood of fraudulent activity, and do these relationships differ from legitimate transaction patterns?
3. How can sampling techniques show the data in a more visible and sectioned into the characteristics of the minority (fraudulent transactions)?

4.2 Research Hypotheses

Based on the problem, the following hypotheses are proposed:

- **H1:** The distribution of transaction amounts will be statistically different for fraudulent transactions compared to legitimate ones.
- **H2:** Fraudulent transactions will show no random patterns, potentially clustering during specific periods that differ from the temporal distribution of legitimate transactions.
- **H3:** Creating balanced samples through strategic under sampling will reveal patterns and relationships in fraudulent transactions that are statistically obscured in the original imbalanced dataset.

5 Methodology

Within this research proposal it will showcase a descriptive and detailed way to diagnose the data in an analytical way, it will be focusing on understanding the data the characteristics of it and how to discover the relationship between one aspect and another.

The methodology is broken up into four main phases:

5.1 Data and its tools

The dataset acquired uses the public available data from the site Kaggle called Credit Card Fraud Detection dataset, this dataset contains 284,807 transactions from the European cardholders recorded over two days in September 2013 during that time 492 fraudulent cases (0.172%) among all transactions were recorded.

Within this dataset there is 3 main categories to look at:

- **Time:** Second since the transaction was made.
- **Amount:** The amount of money for the transaction.
- **v1 - v28:** Each one of these columns represent a person and there translations the reason it is labeled v1 to v28 is because of confidentiality of the people's transactions, these represents customers and there spending behaviors without giving away there personal details.

Analysis will be conducted using Python with key libraries including:

- **Pandas:** for data manipulation.
- **NumPy:** for numerical computations.
- **Matplotlib and Seaborn:** for visualization

5.2 Data Quality and Analysis

Within this section it will involve the processing of the data making sure there is no missing values in the dataset, as well examining the data types in the dataset. Once that is established a generated summary of the data will be made. This is the baseline to understand the dataset it will allow us to visualize the data and make better sense of whats going on.

5.3 Exploratory the Data with Charts

There will be two different types of methods that will be used in this dataset.

- **Univariate Analysis:** This analysis will use charts such as histograms, box plots these plots will focus on looking at one feature at a time for example 'Time' and 'Amount' look for fraud vs. legitimate transactions.
- **Bivariate Analysis:** Correlation analysis using heatmaps and scatter plots will focus on connecting two different relationships with one another for example fraud transactions are genuinely different from legitimate ones.

5.4 Rationale

This methodological approach is specifically designed to align with first-semester data analytics competencies, emphasizing the foundational skills of data understanding, visualization, and preprocessing that form the basis of all advanced analytical work.

6 Significance and Expected Outcomes

Within this research it shows what is possible from data analysis point of view in a practical way. From a practical perspective, this study talks about real world problems that affects peoples financial well being and hurst the business as a whole. While establishing a system that can analyze fraudulent data transactions and detect it will allow the developing team the necessary knowledge to rectify the problem by creating or fixing there banking system as well this research will contribute to the broader understanding of how foundational data analysis techniques can show meaningful insights from challenging datasets, emphasizing that valuable knowledge can be gained before the application of complex machine learning algorithms.

References

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, *55*, 278–288.

Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems*, *150*, 113492. doi: https://doi.org/10.1016/j.dss.2021.113492

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, *17*(3), 235–255. doi: 10.1214/ss/1042727940

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2013). *Credit card fraud detection dataset.* Kaggle. Retrieved from `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud` (European cardholders, September 2013 transactions)

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 ieee symposium on computational intelligence and data mining (cidm)* (pp. 159–166). IEEE. doi: 10.1109/CIDM.2015.7400677

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. doi: 10.1109/TKDE.2008.239

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*. Retrieved from `https://arxiv.org/abs/1009.6119` (14 pages)

Weiss, G. M. (1995). Learning with rare cases and small disjuncts. In *Proceedings of the 12th international conference on machine learning* (pp. 558–565). Morgan Kaufmann.