Analysis for Credit Card Fraud Detection

Your Name Ryan Habis
Student ID D00245309

November 22, 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

The rapid growth of digital transactions has been accompanied by a corresponding increase in credit card fraud, resulting in significant financial losses globally. Financial institutions face the ongoing challenge of developing effective systems to identify fraudulent activities in real-time while minimizing false positives that inconvenience legitimate customers. The fundamental characteristic that makes credit card fraud detection particularly challenging is the severe class imbalance inherent in transactional datasets, where fraudulent cases typically represent less than 0.2% of all transactions.

This research proposal focuses on establishing a comprehensive analytical framework for credit card fraud detection through rigorous exploratory data analysis (EDA) and data preprocessing. Rather than developing predictive models, this study emphasizes the crucial preliminary steps of understanding data characteristics, identifying patterns, and preparing the dataset for future analysis. The Credit Card Fraud dataset from Kaggle provides an ideal case study for this investigation, containing real transaction data with documented fraud cases.

The importance of thorough EDA cannot be overstated in the context of fraud detection. Before any modeling can occur, analysts must develop an intimate understanding of the data's structure, distribution, and peculiarities. This study aims to demonstrate how systematic EDA can reveal critical insights about fraudulent transaction patterns and establish a solid foundation for subsequent analytical work.

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

The literature on fraud detection consistently highlights class imbalance as the primary analytical challenge. **?** demonstrated that standard analytical approaches often fail when the minority class represents such a small proportion of the data, leading models to favor the majority class and ignore fraudulent patterns.

Previous studies have shown that effective fraud detection systems begin with comprehensive exploratory analysis. Phua, Lee, Smith, and Gayler (2010) emphasized that understanding the statistical properties of fraudulent versus legitimate transactions is crucial for developing effective detection strategies. Research by Bolton and Hand (2002) revealed that fraudulent transactions often exhibit distinct patterns in terms of transaction amount, timing, and frequency that can be identified through careful EDA.

The importance of data preprocessing in fraud detection has been well documented. He and Garcia (2009) discussed various sampling techniques for handling class imbalance, noting that creating balanced samples for exploratory purposes can reveal patterns that would otherwise remain hidden in the full dataset. Visualization techniques have proven particularly valuable in fraud analysis, with **?** demonstrating how graphical representations can help identify anomalous patterns and relationships.

While much of the literature focuses on predictive modeling, there is growing recognition of the importance of foundational data analysis. **?** argued that the quality of insights derived from fraud detection systems is directly related to the depth of initial data understanding established through EDA. This research proposal builds upon this perspective by focusing exclusively on the analytical groundwork necessary for effective fraud detection.

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 there analyzing to favor 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 et al. (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.

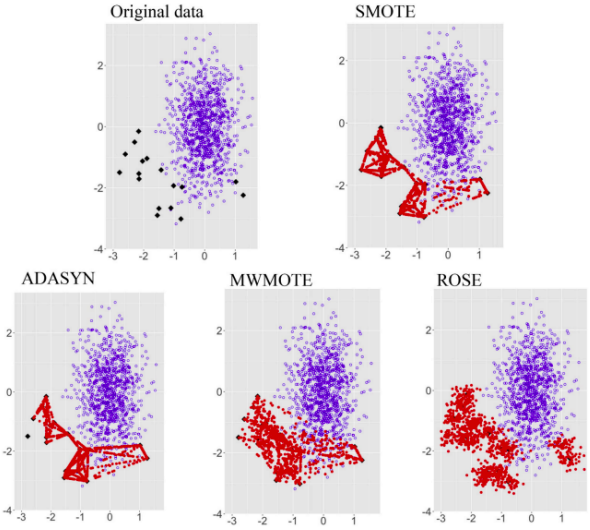

Figure 1: There are 5 charts showing the same data in a different way

- **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.

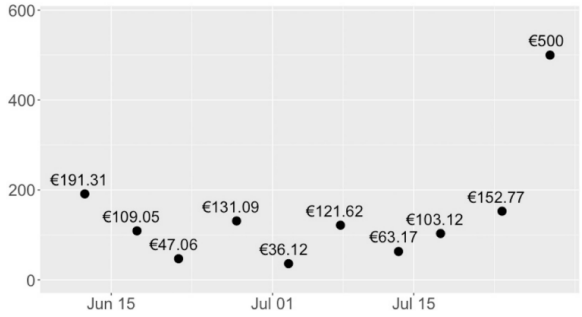

Figure 2: Here is an example of an outlier which is 500 euro which raises suspension

3 Research Objectives

The primary aim of this research is to develop a comprehensive analytical framework for credit card fraud detection through systematic exploratory data analysis and data preprocessing. The specific research objectives are:

1. To perform comprehensive Exploratory Data Analysis (EDA) of the credit card fraud dataset to characterize and contrast the statistical properties of fraudulent and legitimate transactions across all available features.
2. To identify and quantify significant statistical differences and relationships between the target variable ('Class') and key features, particularly 'Time', 'Amount', and selected principal components, using appropriate statistical tests and visualization techniques.
3. To investigate the practical effects of extreme class imbalance on initial data summaries and analytical outcomes, and to evaluate simple data sampling strategies for creating balanced datasets that facilitate clearer examination of fraudulent transaction patterns.
4. To develop a thoroughly documented, preprocessed version of the dataset and provide data-driven recommendations for feature selection and preprocessing steps that should be prioritized in future predictive modeling efforts.

4 Research Questions and Hypotheses

4.1 Research Questions

This study will address the following research questions:

1. What are the key statistical characteristics and visual patterns that distinguish fraudulent transactions from legitimate ones across the principal components and original features in the dataset?

2. How are temporal patterns ('Time') and transaction amounts ('Amount') related to the likelihood of fraudulent activity, and do these relationships differ significantly from legitimate transaction patterns?

3. What is the impact of extreme class imbalance on initial data exploration and summary statistics, and how can strategic sampling techniques provide enhanced visibility into the characteristics of the minority class (fraudulent transactions)?

4.2 Research Hypotheses

Based on preliminary examination of the problem domain, the following hypotheses are proposed:

- **H1:** The distribution of transaction amounts will be statistically different for fraudulent transactions compared to legitimate ones, with fraudulent transactions showing distinct central tendency and dispersion characteristics.

- **H2:** Fraudulent transactions will exhibit non-random temporal patterns, potentially clustering during specific periods that differ from the temporal distribution of legitimate transactions.

- **H3:** Creating balanced samples through strategic undersampling will reveal patterns and relationships in fraudulent transactions that are statistically obscured in the original imbalanced dataset.

5 Methodology

This research will employ a descriptive and diagnostic analytical design, focusing on understanding data characteristics and discovering relationships rather than building predictive models. The methodology is structured around four main phases:

5.1 Data Source and Tools

The study will utilize the publicly available Credit Card Fraud Detection dataset from Kaggle, containing 284,807 transactions from European cardholders recorded over two days in September 2013. The dataset includes 31 features: 28 principal components (V1-V28) obtained from PCA transformation, 'Time' (seconds elapsed between each transaction and the first transaction), 'Amount' (transaction amount), and 'Class' (target variable: 0 for legitimate, 1 for fraudulent). Analysis will be conducted using Python with key libraries including Pandas for data manipulation, NumPy for numerical computations, Matplotlib and Seaborn for visualization, and SciPy for statistical testing.

5.2 Data Quality and Preliminary Analysis

The initial phase will involve comprehensive data quality assessment including checking for missing values, examining data types, and generating summary statistics for the entire dataset and stratified by the target class. This will establish a baseline understanding of data completeness and basic distributional characteristics.

5.3 Exploratory Data Analysis Framework

The EDA will be conducted through multiple complementary approaches:

- **Univariate Analysis:** Distribution analysis using histograms, box plots, and density plots for all features, with particular focus on 'Time' and 'Amount' stratified by transaction class.
- **Bivariate Analysis:** Correlation analysis using heatmaps and scatter plots to identify relationships between features and the target variable. Statistical testing (t-tests or Mann-Whitney U tests) to compare feature distributions between fraudulent and legitimate transactions.
- **Multivariate Analysis:** Pattern analysis through dimensionality reduction visualization and segmented analysis to identify complex interactions between multiple features.

5.4 Handling Class Imbalance for Analysis

To address the analytical challenges posed by class imbalance (fraudulent transactions constitute only 0.172% of the dataset), the study will implement and evaluate sampling strategies including random undersampling of the majority class to create balanced datasets for exploratory purposes. This approach will facilitate clearer visualization and statistical comparison of the minority class characteristics without the overwhelming influence of the majority class.

5.5 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

This research holds both practical and pedagogical significance in the field of data analytics. From a practical perspective, the study addresses a critical real-world problem with substantial financial implications for the banking and e-commerce sectors. By establishing a systematic framework for analyzing fraud detection data, the findings can inform the development of more effective monitoring systems and contribute to reducing financial losses due to fraudulent activities.

From an educational perspective, this research serves as an exemplary case study in applied exploratory data analysis. It demonstrates the critical importance of thorough data understanding and preprocessing before embarking on predictive modeling. The study provides a template for approaching complex, imbalanced datasets that are common in real-world analytics scenarios but often underrepresented in academic curricula.

The expected outcomes of this research include:

- A comprehensive profile of fraudulent versus legitimate transactions, identifying the most discriminative features and patterns that characterize fraudulent activity.
- A curated collection of visualizations and statistical summaries that effectively communicate the story of the data and highlight key differences between transaction classes.
- A demonstrated methodology for handling extreme class imbalance during the exploratory analysis phase, providing practical strategies for gaining insights into minority class characteristics.
- Data-driven recommendations for feature selection, engineering, and preprocessing steps that should be prioritized in subsequent predictive modeling efforts.
- A thoroughly documented analytical process that can serve as an educational resource for students and practitioners approaching similar imbalanced classification problems.

This research will contribute to the broader understanding of how foundational data analysis techniques can extract meaningful insights from challenging datasets, emphasizing that valuable knowledge can be gained before the application of complex machine learning algorithms.

References

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.