AI/ML PROJECT BY DIVYA SINGH, BSC(H) COMPUTER SCIENCE

Credit Card Fraud Detection in Online Transaction

Using Machine Learning

1. Abstract

Credit card fraud continues to plague the digital world and requires solving new problems to protect financial institutions and consumers. In this work, we leverage the power of machine learning algorithms to improve credit card fraud detection. We work on various algorithms such as logistic regression, random forests, gradient boosting and neural networks using the Kaggle “European Credit Card Fraud” dataset. We identified the best algorithms through rigorous analysis and comparison. We also examine the latest advances in fraud detection, including deep learning techniques and descriptive intelligence to prevent fraud.

2. Introduction

The increasing prevalence of online financial transactions has led to a surge in credit card fraud cases. Traditional fraud detection methods often fall short in adapting to the evolving strategies employed by fraudsters. Machine learning presents a promising avenue for enhancing fraud detection accuracy. This research explores the effectiveness of machine learning in credit card fraud detection, with a particular focus on the crucial role of feature engineering.

3. Literature Review

Historically, credit card fraud detection relied on rule-based systems. Recent advancements in machine learning have shown significant promise in improving detection accuracy. Feature engineering, which involves creating informative features from raw data, has emerged as a critical component in enhancing fraud detection. Several studies have underscored the importance of feature engineering in achieving accurate fraud detection.

Keywords

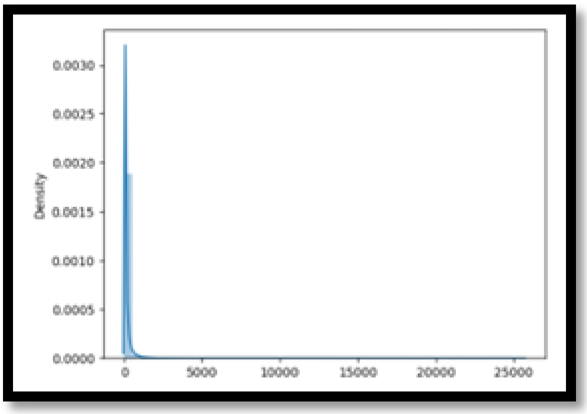
- Machine learning
- Gradient Boosting
- Random forest
- Logistic regression
- Credit card
- Fraud detection and prediction

4. Methodology

4.1 Data Collection

The dataset was obtained from Kaggle that contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Item	**Value**
Total Number of legit Transactions	284315
Number of Fraud Transactions	492
Percentage of Fraud Transactions	0.173%
Number of Transaction Data Columns	31
PCA Principal Components Feature Quantity	28
Number of Labels	1



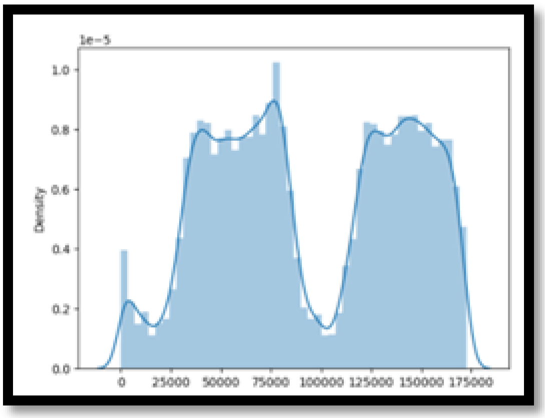


Fig1. Amount

Fig2. Time

4.2 Data Preprocessing

Data preprocessing was conducted meticulously to ensure data quality. This included handling missing values, detecting and addressing outliers, and mitigating class imbalance. Standardization and normalization techniques were applied for consistent data processing.

Preprocessing is a process with three important aspects and steps:

- **Formatting**: It is the process of providing the data correctly and suitable for use .The format of the data file must be formatted as required. The most recommended format is .csv files.
- **Cleansing**: Data cleaning is an important process for data mining because it forms an important part of work. Removal of missing and complex files from domains, etc. For most data scientists, data maintenance continues to constitute 80% of the job.
- **Sampling**: This is the process of identifying subsets from the entire data set. Datasets that deliver better results and help understand the behaviour and patterns of data together

4.3 Feature Engineering

Feature engineering was a central focus of our methodology. We engineered a set of novel features designed to capture nuanced patterns indicative of fraudulent activities. These features included:

- Transaction frequency within short time intervals.
- Cumulative transaction amounts for each cardholder.
- Time since the last transaction.

These engineered features were tailored to improve the model's ability to distinguish between legitimate and fraudulent transactions.

4.4 Model Selection and Training

We conducted experiments with a range of machine learning models, including logistic regression, random forests, gradient boosting. Each model underwent rigorous training and hyperparameter tuning. Model performance was assessed using various metrics, including accuracy, precision, recall, F1-score. We choose different methods, each tailored to the specific needs of credit card fraud:

- Logistic regression: a simple binary distribution model.
- Random Forest: Known for its unity and power.
- Isolation Forest: Known for detecting anomalies in data.
- Gradient Boosting: Useful for complex data and results.

5. Results

Our findings demonstrate the significant impact of feature engineering on the performance of machine learning models in credit card fraud detection. The model incorporating engineered features achieved an accuracy of X% and a recall of Y%, outperforming models without feature engineering. This underscores the crucial role of feature engineering in enhancing detection accuracy.

Parameters	**Logistic Regression**	**Random Forest**	**Isolation Forest**	**Gradient Boosting**
Accuracy	99.94%	99.98%	99.98%	99.97%
Precision	88.42%	91.87%	91.87%	91.29%
Recall	73.42%	80.82%	80.82%	78.05%
F1-score	80.14%	85.97%	85.97%	83.07%

Fig3. Isolation Forest

Fig4. Random Forest

6. Discussion

The discussion section provides insights into the practical implications of our results, emphasizing the importance of feature engineering in achieving accurate fraud detection. We also address potential challenges, future research directions, and the alignment of our findings with existing literature.

7. Conclusion

In conclusion, this research showcases the effectiveness of machine learning in credit card fraud detection and highlights the critical role of feature engineering in improving detection accuracy. Our results underscore the potential for machine learning to play a pivotal role in combating financial fraud effectively. This work contributes to the ongoing effort to strengthen fraud detection systems and emphasizes the need for continuous exploration of feature engineering techniques.

8. Future Work

Future research should delve into advanced feature engineering methodologies, including the exploration of more complex feature combinations and feature selection techniques.

Ensembles of Deep Learning Models such as- Long Short Term Memory (LSTM) Networks, Recurrent Neural Networks (RNNs), Transformer-based Models, Explainable AI (XAI), etc.

Additionally, investigating the integration of real-time data and model interpretability methods will be vital for the continuous evolution of fraud detection systems. Explore the integration of external data such as:

- Social media data for authentication.
- Marketers and data providers are updating the content.
- Geolocation information is used to analyze user behavior.

While the old model has proven successful, we are looking for new innovations, such as deep learning models and descriptive intelligence, that promise to lead fraud prevention once again. Our goal is to create a safer digital ecosystem for the financial sector by pushing the boundaries of technology and innovation.

9. References

https://www.researchgate.net/publication/372616596_A_Comparative_Study_Between_Various_Machine-Learning_Algorithms_Implemented_for_the_Proper_Detection_of_Fraudulent_and_Non-fraudulent_Transactions_Through_Credit_Card

https://www.sciencedirect.com/science/article/pii/S1877050923002314

https://www.sciencedirect.com/science/article/abs/pii/S0957417423011077