Optimizing Credit Card Fraud Detection Through Imbalanced Dataset Management and Feature Analysis

Ankitha Dongerkerry Pai
School of Computing and Augmented Intelligence
Arizona State University
apai14@asu.edu

Karthik Mahalingam
School of Computing and Augmented Intelligence
Arizona State University
kmahali2@asu.edu

Abstract

Our objective is to delve into strategies for enhancing credit card fraud detection systems, particularly addressing the challenges posed by imbalanced transaction data. In light of the significant threat that credit card fraud poses to businesses and consumers alike, our motivation stems from the imperative need to develop robust detection mechanisms that safeguard financial assets and uphold trust in digital transactions. By studying this topic, we aim to offer practical solutions to businesses, enabling them to better navigate the complexities of fraud detection in the modern landscape. Our analysis of transaction features such as time and amount seeks to uncover patterns associated with fraudulent activities, refining detection algorithms for heightened accuracy. Ultimately, our contributions aim to empower businesses to fortify their defenses against fraud, minimize disruptions from false alarms, and foster a safer digital economy for all stakeholders.

1 Execution Plan

1.1 Steps to follow

The execution plan entails gathering credit card transaction data and preprocessing it to handle missing values and standardize features. Subsequently, an exploratory data analysis is conducted to discern transaction patterns, particularly focusing on time and amount. Techniques like oversampling and undersampling are then employed to address data imbalance, followed by feature engineering to enhance model performance. Multiple algorithms are experimented with for model selection and training, with hyperparameter tuning conducted to optimize performance. Validation of the model's efficacy is performed using holdout data or cross-validation. Comprehensive documentation and reporting are prepared to summarize findings and recommendations for further enhancements to the fraud detection system.

1.2 Workload Distribution

The workload will be distributed among team members equally for better skill development. Data collection and preprocessing will be handled by one team member, while another will focus on EDA.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

Figure 1: Credit Fraud Dataset

Imbalanced data handling, feature engineering, model selection and training, hyperparameter tuning, model evaluation and validation, deployment and monitoring, and documentation and reporting will be assigned to different team members.

1.3 Time Table

A tentative timeline has been established, allocating specific weeks for each phase of the project, including data collection, preprocessing, EDA, model training, and deployment. Regular progress reviews and adjustments to the timetable will be made as necessary to ensure project milestones are met.

1.4 Expected Challenges and How to Handle Them

Several challenges are anticipated, including imbalanced data, model overfitting, hyperparameter tuning, deployment complexity, and time constraints. Strategies to address these challenges include utilizing techniques like oversampling and undersampling for imbalanced data, employing regularization techniques to prevent overfitting, using automated tuning methods for hyperparameter optimization, collaborating with DevOps for streamlined deployment processes, and prioritizing tasks to meet deadlines effectively. Regular communication and coordination among team members will be essential to address challenges as they arise and ensure project success.

Time	0	0	1	1	2
V1	-1.35981	1.191857	-1.35835	-0.96627	-1.15823
V2	-0.07278	0.266151	-1.34016	-0.18523	0.877737
V3	2.536347	0.16648	1.773209	1.792993	1.548718
V4	1.378155	0.448154	0.37978	-0.86329	0.403034
V5	-0.33832	0.060018	-0.5032	-0.01031	-0.40719
V6	0.462388	-0.08236	1.800499	1.247203	0.095921
V7	0.239599	-0.0788	0.791461	0.237609	0.592941
V8	0.098698	0.085102	0.247676	0.377436	-0.27053
V9	0.363787	-0.25543	-1.51465	-1.38702	0.817739
V10	0.090794	-0.16697	0.207643	-0.05495	0.753074
V11	-0.5516	1.612727	0.624501	-0.22649	-0.82284
V12	-0.6178	1.065235	0.066084	0.178228	0.538196
V13	-0.99139	0.489095	0.717293	0.507757	1.345852
V14	-0.31117	-0.14377	-0.16595	-0.28792	-1.11967
V15	1.468177	0.635558	2.345865	-0.63142	0.175121
V16	-0.4704	0.463917	-2.89008	-1.05965	-0.45145
V17	0.207971	-0.1148	1.109969	-0.68409	-0.23703
V18	0.025791	-0.18336	-0.12136	1.965775	-0.03819
V19	0.403993	-0.14578	-2.26186	-1.23262	0.803487
V20	0.251412	-0.06908	0.52498	-0.20804	0.408542
V21	-0.01831	-0.22578	0.247998	-0.1083	-0.00943
V22	0.277838	-0.63867	0.771679	0.005274	0.798278
V23	-0.11047	0.101288	0.909412	-0.19032	-0.13746
V24	0.066928	-0.33985	-0.68928	-1.17558	0.141267
V25	0.128539	0.16717	-0.32764	0.647376	-0.20601
V26	-0.18911	0.125895	-0.1391	-0.22193	0.502292
V27	0.133558	-0.00898	-0.05535	0.062723	0.219422
V28	-0.02105	0.014724	-0.05975	0.061458	0.215153
Amount	149.62	2.69	378.66	123.5	69.99
Class	0	0	0	0	0

Figure 2: First 5 values of the dataset

2 Evaluation plan

2.1 Outcome Evaluation

The project's success will be measured by the credit card fraud detection system's ability to accurately identify fraudulent transactions while minimizing false alarms. Key metrics such as precision, recall, F1 score, and false positive rate will be used to assess performance. Real-world effectiveness in detecting new fraud patterns will also be considered.

2.2 Performance Evaluation and Peer Review

Internally, the team's performance will be evaluated based on meeting milestones, adhering to timelines, and problem-solving. Peer review:

- Karthik's - He has been very involved and suportive in the discussions. His promptness and curiosity has helped in finalizing the project topic and with the proposal.
- Ankitha's - Ankitha has been forward with her ideas and open to idea's. She has equally distributed the responsibilities and helped us reach our goals.

As a team we plan to divide different ML models and perform their implementations. The aim is to implement all the models and verify the results ensuring industry standards are met and findings are meaningful.

References

[1] Kaggle. (2024). Credit Card Fraud Detection. Kaggle. Retrieved from https://drive.google.com/file/d/1JXS4vbOXuBbYEB-cUiY-hrDMGGgfbzRG/view?usp=sharing

[2] R. Wang and G. Liu, "Ensemble Method for Credit Card Fraud Detection," 2021 4th International Conference on Intelligent Autonomous Systems (ICoIAS), Wuhan, China, 2021, pp. 246-252, doi: 10.1109/ICoIAS53694.2021.00051.

[3] B. A. Smadi, A. A. S. AlQahtani and H. Alamleh, "Secure and Fraud Proof Online Payment System for Credit Cards," in “2021 IEEE 12th Annual Ubiquitous Computing, Electronics Communication Conference (UEMCON),” New York, NY, USA, 2021, pp. 0264-0268, doi: 10.1109/UEMCON53757.2021.9666549.

[4] Marazqah Btoush EAL, Zhou X, Gururajan R, Chan KC, Genrich R, Sankaran P. “A systematic review of literature on credit card cyber fraud detection using machine and deep learning.” PeerJ Comput Sci. 2023 Apr 17;9:e1278. doi: 10.7717/peerj-cs.1278. PMID: 37346569; PMCID: PMC10280638.

[5] A. Maurya and A. Kumar, "Credit card fraud detection system using machine learning technique," in “2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom),” Malang, Indonesia, 2022, pp. 500-504, doi: 10.1109/CyberneticsCom55287.2022.9865466.

[6] I. Vejalla, S. P. Battula, K. Kalluri and H. K. Kalluri, "Credit Card Fraud Detection Using Machine Learning Techniques," in “2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS),” Nagpur, India, 2023, pp. 1-4, doi: 10.1109/PCEMS58491.2023.10136040.