Fraud Fiends

Team 07: Anant Sahoo, Deep Patel, Sulaiman Mohyuddin

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Abstract

Credit card fraud represents a substantial financial threat in today's increasingly digital economy [1]. This study introduces a logistic regression (LR) model for real-time detection of potentially fraudulent transactions, employing a publicly available dataset from Kaggle [4]. Preprocessing steps include addressing missing values and scaling categorical and numerical features. For model validation, we employ k-fold cross-validation with 10 folds to split the data into training and testing sets. After applying more robust feature scaling, Lasso (L1) regularization and stratification for our k-fold cross-validation, we improved our accuracy for our initial model from 92.31% to 94.65%. The platform's real-time notifications are designed to help mitigate the impact of fraudulent activities on individuals and organizations alike.

Motivation

- Credit card fraud is a growing issue, impacting millions globally.
- Real-time detection is critical to prevent financial losses.
- Logistic Regression offers a fast and effective solution.
- The model is designed for scalability and low-latency environments.
- Our platform aims to enhance trust in digital transactions.

Fig. 1: General Scenario of Online Fraud
Source: Alenzi et al. [2].

Methodology

Our project began with a Logistic Regression model as the baseline due to its computational efficiency and interpretability. Preprocessing steps included handling missing values, encoding categorical variables, and ensuring data integrity [3]. To address class imbalance in the dataset, we implemented stratified k-fold cross-validation with ten folds, which ensured balanced class distributions across training and testing sets. Advanced feature optimization techniques, such as Principal Component Analysis (PCA) were explored to improve model performance further. Additionally, feature scaling and stratified sampling were introduced to enhance the accuracy and robustness of the model. For evaluation, metrics such as accuracy scores, confidence intervals, confusion matrices, and especially false positive rates (FP) were used.

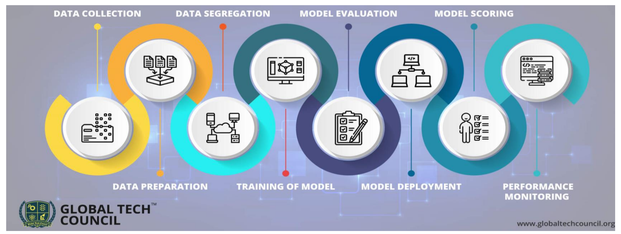

Fig. 2: Machine learning pipeline steps used
Source: Global Tech Council [5].

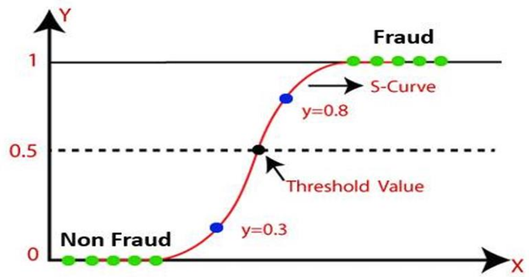

Fig. 3: Shows how a logistic regression model works
Source: Alenzi et al. [2].

Results

To find the total accuracy of the model, we calculate the arithmetic mean of the accuracies of each fold. This turned out to be **94.65%**, a 2.24% increase from our previous model. We also built a confusion matrix for each fold and averaged those results into one final matrix, as shown in Figure 3. Here, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates can be attained. We were able to keep this rate at **0.87%,** around a 0.13% decrease from our previous model. While this is a good baseline for our purposes, the scale of transactions can still produce many false positives. Previous implementations have been able to keep this rate around 0.30%. Finally, we constructed 95% confidence intervals for each fold in Figure 4.

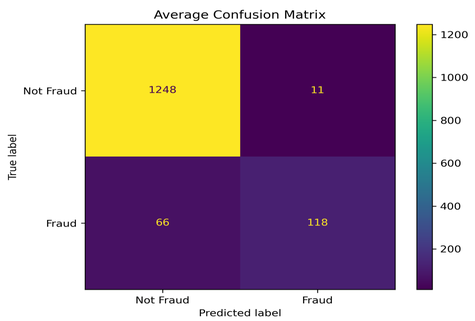

Fig. 4: Average Confusion Matrix across all K-Folds

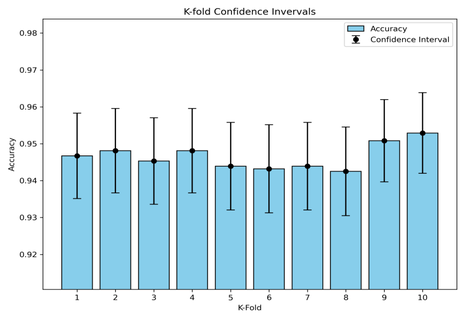

Fig. 5: K-Fold model accuracies and their confidence intervals

Lessons Learned

Fig 6: Lessons Learned
Source: LinkedIn [6].

The Logistic Regression model presented in this report demonstrates strong potential for real-time credit card fraud detection, achieving an accuracy of 94.65%. Through improved feature scaling, selection, and stratification, we were able to adequately refine our baseline model. While there is still room for improvements, we believe we have created a strong foundation for any future work. Along the way, we have learned some important lessons:

1. The challenges of feature engineering with unstructured, high dimensional data
2. Importance of feature scaling
3. Having realistic expectations for model performance

Selected References

[1] Security.org, "Credit Card Fraud Statistics and Reports", 2023. [Online]. Available: https://www.security.org/digital-safety/credit-card-fraud-report/.
[2] H. Z. Alenzi and N. O, "Fraud Detection in Credit Cards using Logistic Regression," International Journal of Advanced Computer Science and Applications, vol. 11, no. 12, 2020, doi:https://doi.org/10.14569/ijacsa.2020.0111265
[3] C. Liu and D. P. Robinson, "A Comparison of Machine Learning Models for Fraud Detection," IEEE Trans. Neural Networks, vol. 29, no. 4, pp. 987–998, 2019
[4] Kaggle, "Credit Card Fraud Detection Dataset," [Online]. Available: https://www.kaggle.com/datasets/neharoychoudhury/credit-card-fraud-data/data.
[5] Global Tech Council, "How to Build a Machine Learning Pipeline," 2023. [Online]. Avaliable: https://www.globaltechcouncil.org/blockchain/how-to-build-a-machine-learning-pipeline/.
[6] LinkedIn, "Article Cover Image: The Power of Lessons Learned: Enhancing Project Mangement Success," 2023. [Online]. Available: media.licdn.com/dms/image/v2/D5612AQG2X7t6jD5ysA/article-cover_image-shrink_720_1280/article-cover_image-shrink_720_1280/0/1685891462829?e=1738800000&v=beta&t=VM9U_4dipPiYFS2fRv4hRAHhU9K9kxsKjIT8Xx2fplk.

Acknowledgments

This poster would not have been possible without the selected references, the lecture notes and class code provided by Dr. Santos.