Final Paper: GA-Based Feature Selection for Credit Card Fraud Detection

Aditya Krishna

December 12, 2024

1 Introduction

Credit card fraud represents a significant challenge for financial institutions worldwide. With billions of daily transactions, even a small fraction of fraudulent activities can lead to substantial financial losses for both consumers and credit card providers. Moreover, the increasing reliance on e-commerce magnifies the difficulty of swiftly and accurately identifying fraudulent patterns. High-profile security breaches also raise public anxiety regarding data misuse.

This paper examines a GA-based feature selection approach proposed by Ileberi, Sun, and Wang (2022)1. Their method aims to improve the accuracy and reliability of machine learning (ML) models tasked with detecting fraudulent transactions. By using a genetic algorithm (GA) to select informative subsets of features, the approach may reduce overfitting and enhance predictive metrics like AUC.

However, improved model performance should not overshadow ethical considerations. The collection and analysis of transaction data raise concerns about privacy, fairness, and the trustworthiness of financial institutions. Ensuring that improved fraud detection does not come at the cost of violating consumer rights or introducing biases is a pressing normative issue.

2 Analysis of Methods

Data Loading and Preliminary Analysis We use the widely studied European credit card fraud dataset from Kaggle. It contains 284,807 transactions and a binary target (Class) indicating fraud (1) or non-fraud (0). Due to runtime concerns, we will sample a fraction of the data.

Table 1: Class Distribution (Full Data)

Var1	Freq

0	284315
1	492

```
##     1
## 0.173
```

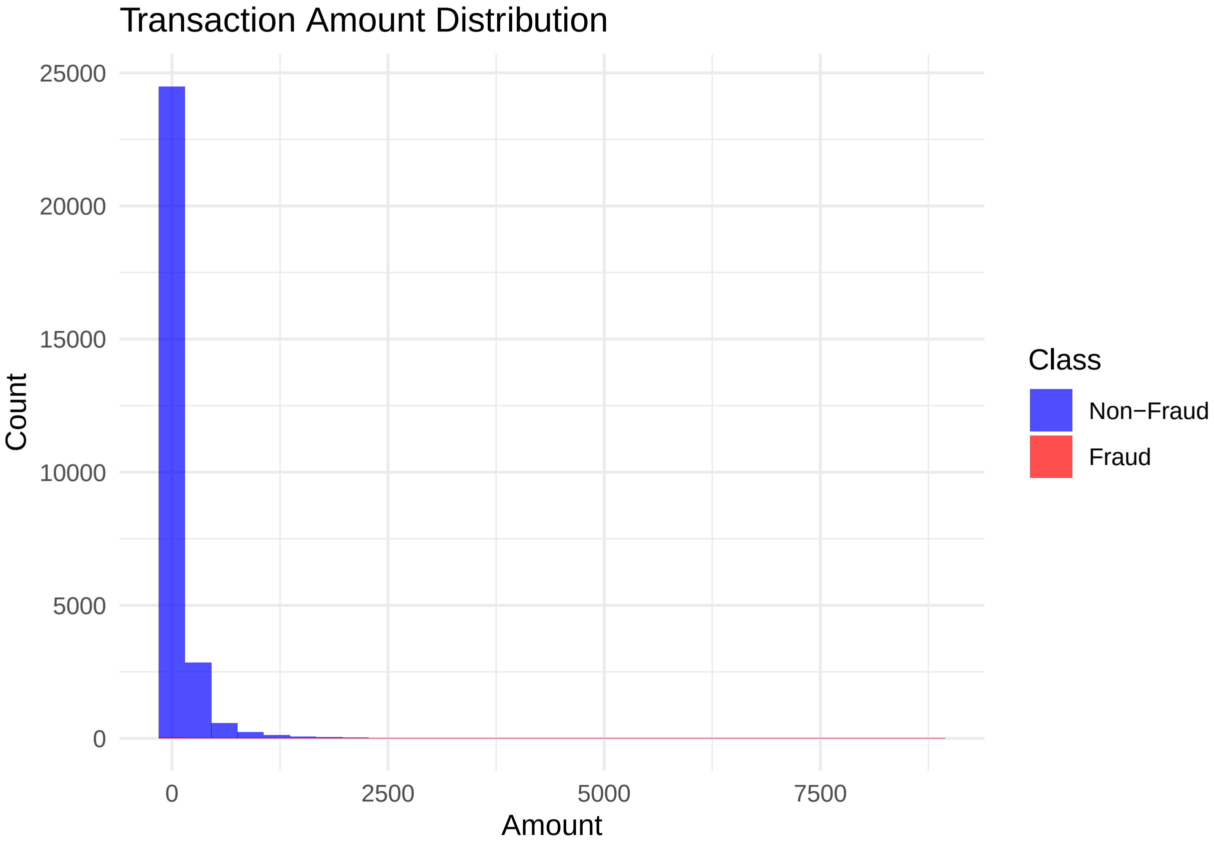
As known, the dataset is highly imbalanced, with fraud cases making up a very small fraction of the transactions.

To speed up computations, we sample 10% of the data from each class. This reduces runtime at the cost of some precision.

Table 2: Class Distribution (Sampled Data)

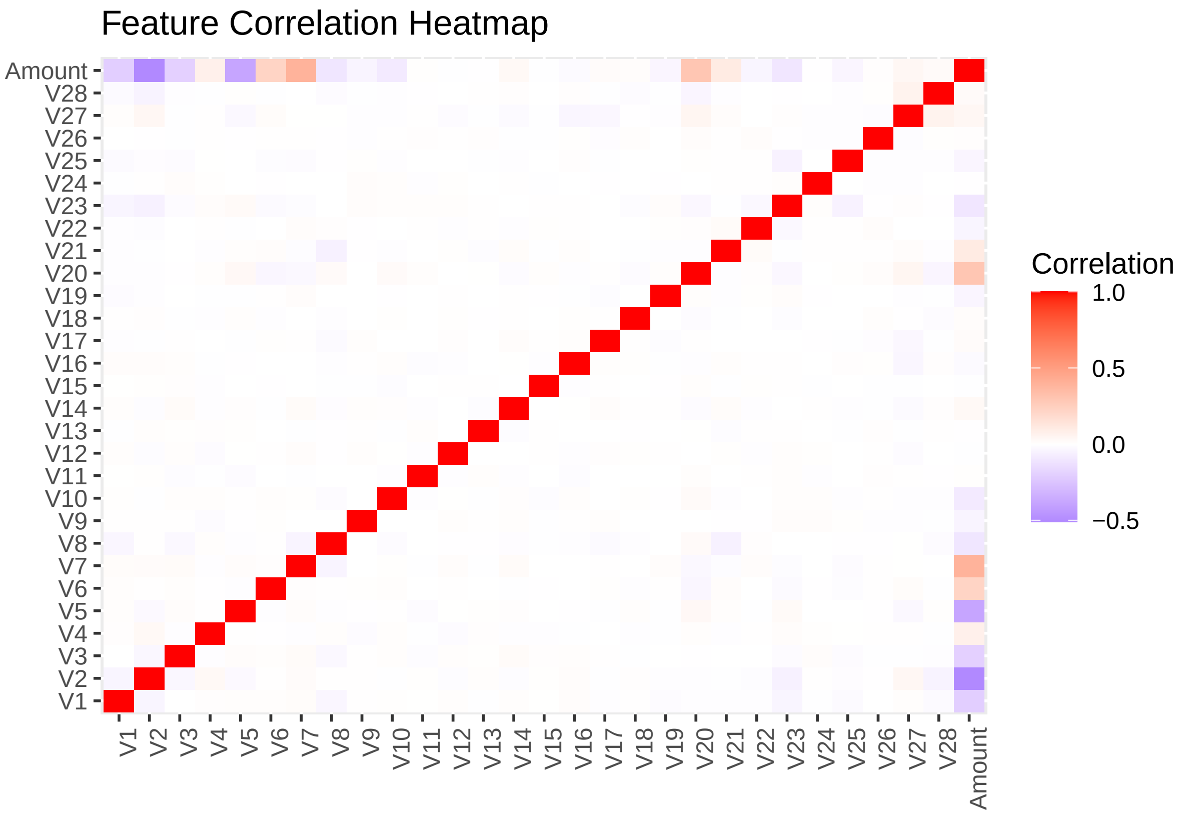
Var1	Freq
0	28432
1	49

##Exploratory Visualizations ##Transaction Amount Distribution



Fraudulent transactions often differ in typical amount distributions, though not always clearly.

Correlation Heatmap (Sampled Data) The dataset's features (V1 to V28) are already transformed by PCA. Still, we can visualize correlations:



GA-Based Feature Selection Explained Genetic Algorithms (GAs) simulate natural selection to find optimal solutions. For feature selection, a GA represents a subset of features as a binary vector (1 = include, 0 = exclude), evaluates its "fitness" (e.g., model AUC), and evolves the population over generations. The paper by Ileberi et al. reports improved accuracy using GA-based selection compared to non-optimized subsets.

To demonstrate the concept, we implement a GA-based feature selection using a Random Forest model. The fitness function takes a binary vector indicating which features to use, trains a RF on the training set, and evaluates AUC on the test set.

```
## Running GA with popSize = 20 , mutationProb = 0.1 , maxGenerations = 10
```

Table 3: GA Parameter Sensitivity Results

popSize	mutationProb	maxGenerations	bestAUC
20	0.1	10	0.9997481

3 Analysis of Normative Considerations

While GA-based feature selection can enhance a model's technical performance, implementing such methods in real-world financial systems raises serious normative questions. Privacy, fairness, and trust are at the core of these concerns and are not peripheral considerations, but integral components of ethical and sustainable ML deployment.

Credit card transactions contain sensitive information. Using them to train sophisticated models—and sharing such data with third-party researchers or vendors—can jeopardize consumer privacy. Even anonymized or aggregated data carries a re-identification risk. Institutions must enforce stringent data governance protocols, such as secure data storage, encryption, and strict access controls. They should also consider implementing differential privacy techniques to protect individual identities, ensuring that improvements in fraud detection do not lead to invasive data collection or breaches.

Algorithms may inadvertently reflect or exacerbate existing social biases. Certain demographic groups could be disproportionately flagged as suspicious if the training data or feature selection process encodes historical prejudices. Fairness auditing tools, bias mitigation techniques, and ongoing performance monitoring can help ensure that no subgroup is systematically disadvantaged. Introducing fairness constraints into the GA's fitness function may balance predictive performance with the moral imperative to avoid discriminatory outcomes.

Financial institutions have a fiduciary and moral responsibility to maintain the trust of their customers. High-performing fraud detection models that raise fewer false alarms and accurately catch fraudulent transactions can bolster customer confidence. However, transparency about how these models operate and the safeguards protecting consumer interests is crucial. Clear communication about why certain transactions are flagged and adherence to regulatory frameworks that protect consumer rights will help sustain trust. Trust, in this context, is both a practical and an ethical requirement—without it, even accurate models risk damaging the long-term relationship between institutions and their clients.

Ultimately, normative considerations are not constraints that oppose technological advancement. Instead, they ensure that improvements in ML-based fraud detection align with societal values, enhance social welfare, and maintain a fair and respectful marketplace.

4 Conclusion

This paper reviewed a GA-based feature selection method for credit card fraud detection. The approach can yield better model performance by evolving subsets of features. We demonstrated a simplified, faster version of this methodology, showing that GA parameters can influence outcomes and that careful tuning is essential.

However, these technical gains must be pursued ethically. Privacy protection, fairness, and trust are central moral considerations. Future work could incorporate fairness metrics into the GA's fitness function, apply privacy-preserving techniques, and conduct statistical tests to ensure that improvements are both robust and ethically justifiable. Ultimately, successful credit card fraud detection is not only a technical achievement but a moral imperative to serve and protect consumers.

References

Ileberi, E., Sun, Y. & Wang, Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. J Big Data 9, 24 (2022). https://doi.org/10.1186/s40537-022-00573-8