Fraud Detection mini project:

Abstract

Credit card fraud is a pervasive and costly issue in the financial industry, requiring effective and reliable detection systems. This project aimed to assess and compare the performance of Logistic Regression and Random Forest classifiers for credit card fraud detection. The analysis is based on a publicly available, highly imbalanced dataset from Kaggle. To address the challenges of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented. Rigorous methodology was employed to avoid data leakage during cross-validation.

Logistic Regression and Random Forest demonstrated significantly different performance levels across key metrics. Logistic Regression showed a high recall rate, an important metric in fraud detection where failing to identify fraudulent transactions can have severe consequences. However, it suffered from a very large number of false positives, resulting in low precision.

On the other hand, Random Forest demonstrated an excellent balance between precision and recall, establishing its robustness and suitability for complex, high-dimensional, and imbalanced data. The high recall rate signifies the model's ability to capture the majority of fraudulent transactions, which is often the primary objective in fraud detection. The high precision of the Random Forest model could help minimize the operational challenges associated with false positives.

Looking ahead, while the Random Forest model outperformed Logistic Regression in this study, there is room for further improvement. Hyperparameter tuning can optimize the performance of both models, and more advanced machine learning methods such as ensemble learning or neural networks could be explored.

This study highlights the importance of selecting the right machine learning model and preprocessing methods for fraud detection tasks, and the need for methodological rigor to avoid data leakage. The strong performance of Random Forest suggests it is a good candidate for deployment in real-world fraud detection systems, especially where a high degree of reliability is required.

Limitations:

- **Problem**: The dataset is highly imbalanced, with fraudulent transactions making up a tiny fraction of the total transactions. While resampling techniques like random oversampling were used to balance the classes, they have their own limitations.
- **Impact**: Resampling can introduce bias and may not perfectly represent the underlying distribution of the original data. This can affect the model's ability to generalize to new, unseen data.

-S.Saini