Credit Card Fraud Detection Using Machine Learning

Introduction

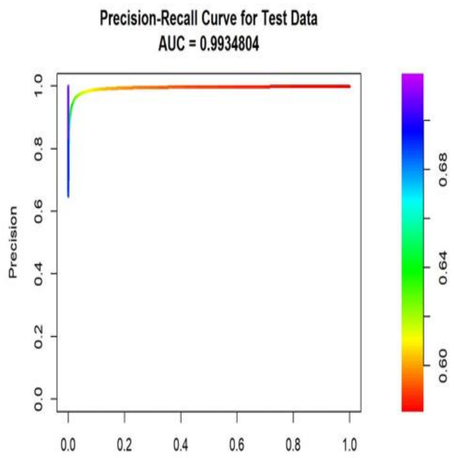
Credit card fraud poses a significant challenge for both financial institutions and customers globally. Developing a fraud detection system holds practical relevance, especially in addressing obstacles like imbalanced datasets and evolving fraud tactics. Our project seeks to overcome these challenges by employing machine learning techniques to bolster security and mitigate financial losses. Through the utilization of advanced algorithms and PCA-based features, our system aims to precisely identify fraudulent transactions, benefiting all parties involved. Success will be gauged through reduced fraud complaints and chargebacks, along with metrics such as AUPRC. While model inaccuracies pose risks, the potential gains in financial savings and security enhancements are substantial. Costs are minimal, utilizing readily available tools and datasets, with a flexible development timeline. Validation against industry benchmarks will ensure effectiveness.

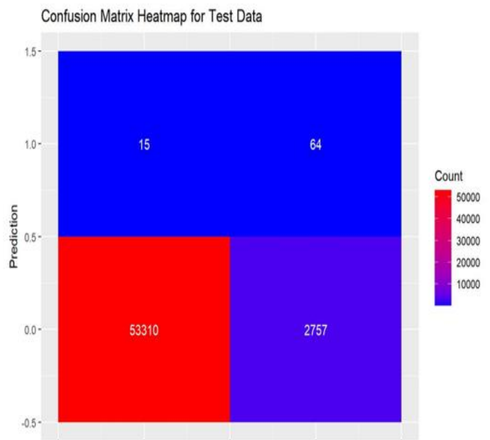
Problem Statement

Our goal is to enhance fraud detection by integrating advanced machine learning and anomaly detection techniques, leveraging PCA-based features and ensemble learning methods. We aim to surpass current methods in identifying various credit card fraud types, including account takeover, card-not-present, counterfeit card, identity theft, and application fraud. Our system addresses challenges like imbalanced datasets and evolving fraud tactics effectively.

Methodologies Used

We integrated advanced machine learning and anomaly detection to enhance fraud detection accuracy. Using PCA-based features and ensemble learning, it outperforms current techniques, addressing challenges like imbalanced data and evolving fraud tactics effectively. We follow a rigorous process including train-test split, model training, anomaly prediction, model evaluation, and visualization using ROC curves and a heatmap.





- **Isolation Random Forest:** An unsupervised algorithm for anomaly detection, efficiently isolating anomalies using random binary trees.
- **Decision Tree Model:** A non-parametric supervised learning method for classification and regression tasks, employing simple rules inferred from data to make decisions.
- **Logistic Regression Model:** A statistical method for binary classification tasks, predicting probabilities of outcomes and evaluating model performance using confusion matrices and visualization techniques.
- **XGBoost Model:** A highly efficient and scalable implementation of the gradient boosting framework, sequentially building an ensemble of decision trees to improve overall model accuracy, with built-in regularization to prevent overfitting.

Results

These are the fraudulent transactions identified by the model, accompanied by a comparative analysis:

```
## Fraudulent transactions predicted by the model:
```

```
head(fraudulent_transactions)
```

```
##          Time        V1        V2        V5         V6         V9
## 221 -1.993313 -1.306267  1.303301 -0.2571324  1.89916501 0.614042788
## 222 -1.993313 -1.306267  1.303301 -0.2571324  1.89916501 0.614042788
## 224 -1.993313 -1.305774  1.304315 -0.2597587  1.90091651 0.613876435
## 226 -1.993271 -1.448397  2.999577  1.0106762 -1.26802022 3.948326704
## 418 -1.990008 -0.544400  1.154011 -0.6655057 -0.05184682 0.005856417
## 469 -1.989124 -1.877608 -2.889081  0.8656660 -0.94494998 1.529419992
##             V10         V11         V12         V14         V15         V16
## 221  0.13744134 -1.19998956  1.3145462 -0.2986622 -2.2044032 -0.8863493
## 222  0.13744134 -1.19998956  1.3145462 -0.2986622 -2.2044032 -0.8863493
## 224  0.13709520 -1.19987504  1.3142392 -0.2988880 -2.2043625 -0.8860504
## 226  5.79713772  2.41778766 -0.7287479 -6.8586143  1.9326695 -0.7473123
## 418 -0.01176964  0.07079537  0.9467465  0.3822799  0.8726434  0.2748692
## 469 -1.80702723 -0.72658489  0.4668575 -0.4622043  0.1340417 -1.2339759
```

Model comparison

	Model	Accuracy	AUC_ROC	AUC_PR
1	Isolation Forest	0.9506287	0.9313276	0.9934804
2	Logistic Regression	0.9991629	0.9803463	0.9915924
3	Decision Tree	0.9994301	0.8606703	0.9926188
4	XGBoost	0.9995013	0.9714112	0.9914756

Conclusion

Our fraud detection system, employing advanced machine learning techniques, enhances security and reduces financial losses caused by credit card fraud. Leveraging various models including Isolation Forest, Logistic Regression, Decision Tree, and XGBoost, we find XGBoost excels in accuracy, Logistic Regression in AUC-ROC, and Isolation Forest in AUC-PR. Despite risks, the potential benefits outweigh concerns, positioning our project to make a significant impact in combatting credit card fraud.

MS in Engineering Science Data Science
School of Engineering and Applied Sciences,
University at Buffalo, NY

Presented by Team – 8
Bhavana Chinnamgari
Ushaswini Punna
Sai Snehitha Ramisetty
Uday Rohith