Overview

This project aims to predict credit card fraud using machine learning techniques. The dataset used for this analysis is the Kaggle Credit Card Fraud Detection Dataset. The goal is to develop a model that can accurately identify fraudulent transactions from legitimate ones.

Data Understanding

The dataset consists of 32 features including 28 anonymized features labeled V1 to V28, along with Time, Amount, and Class. The Class feature is the target variable where 0 indicates a legitimate transaction and 1 indicates a fraudulent transaction. The dataset is highly imbalanced with a majority of legitimate transactions .

Data Preparation

- No missing values in the dataset.
- The dataset contains numerical features, and only the target variable Class is categorical.
- Features need to be standardized for comparison after balancing the dataset.
- Mean transaction amount: 88.34
- Standard deviation of transaction amount: 250.12
- Time feature is equitably distributed and serves as an independent feature.

Methodology

1. **Data Balancing**: Synthetic Minority Oversampling Technique (SMOTE) was used to balance the dataset by oversampling the minority class.
2. **Feature Standardization**: Numerical features were standardized to ensure comparability.
3. **Model Selection**: Various machine learning models were evaluated, including:
    - Logistic Regression
    - Decision Trees
    - Random Forest
    - K-Nearest Neighbors (KNN)
    - Support Vector Machine (SVM)

4. **Hyperparameter Tuning**: Grid Search was employed to tune the hyperparameters of the models.

Results

- **K-Nearest Neighbors (KNN)** classifier tuned with Grid Search using the Euclidean Distance (p=2) parameter achieved the best results.
- **Evaluation Metrics**:
    - Accuracy: 99.8%
    - Precision: 1.00 for legitimate transactions, 0.50 for fraud transactions
    - Recall: 1.00 for legitimate transactions, 0.86 for fraud transactions
    - F1-Score: 0.63
    - AUC-ROC: 0.93

Conclusion

- The K-Nearest Neighbors Classifier with Euclidean Distance outperformed other models with a test accuracy of nearly 99.8% and minimal overfitting .
- SMOTE effectively mitigated overfitting by synthetically oversampling the minority class labels .

Insights

- Fraudulent transactions typically occur for amounts below 2500, suggesting fraud committers attempt to evade suspicion with smaller amounts.
- No clear temporal pattern was found in the occurrence of fraudulent transactions.
- The dataset's imbalance requires careful handling to prevent overfitting and ensure model deployability .

Installation

To run this project, ensure you have the following Python packages installed:

- pandas
- numpy
- scikit-learn
- imbalanced-learn
- matplotlib

You can install these packages using pip:

bash
Copy code
pip install pandas numpy scikit-learn imbalanced-learn matplotlib

Usage

5. Clone the repository:

bash
Copy code
git clone <repository_link>
cd <repository_directory>

6. Run the Jupyter notebook to execute the code and see the results:

bash
Copy code
jupyter notebook Credit_Card_Fraud_Detection.ipynb

References

- Kaggle Credit Card Fraud Detection Dataset: Link
- https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data