**Machine Learning Model Report**

**Credit Card Fraud Detection Model Report**

**1. Introduction**

**1.1 Problem Statement**

Credit card fraud is a significant concern for financial institutions and customers alike. This project aims to build a predictive model to identify fraudulent transactions based on a dataset of credit card transactions. The dataset consists of transactions made by European cardholders in September 2013, with a notable class imbalance—only 0.172% of transactions are fraudulent.

**2. Data Overview**

**2.1 Dataset Description**

* **Features:** The dataset contains 30 features, including anonymized transaction details (V1 to V28), transaction amount, and a target variable indicating whether the transaction is fraudulent (Class).
* **Size:** 284,807 transactions with 492 frauds.

**2.2 Data Quality Check**

* **Missing Values:** No missing values were detected.
* **Data Types:** All features were found to be numeric.

**2.3 Exploratory Data Analysis (EDA)**

* Class distribution showed a significant imbalance between fraudulent and non-fraudulent transactions.
* Correlation analysis indicated relationships between certain features, with some showing strong correlations to the target variable.

**3. Data Preprocessing**

**3.1 Data Cleaning**

* Outliers in the 'Amount' feature were treated by removing transactions exceeding $10,000.

**3.2 Handling Imbalanced Data**

* The dataset was balanced using SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic examples of the minority class (frauds).

**4. Feature Engineering**

* A log transformation of the 'Amount' feature was performed to reduce skewness and enhance model performance.

**5. Model Selection and Training**

**5.1 Model Choice**

* Random Forest Classifier was selected for its robustness and effectiveness in handling imbalanced datasets.

**5.2 Model Training**

* The dataset was split into training and testing sets (80/20 split).
* Data was standardized using StandardScaler.

**5.3 Hyperparameter Tuning**

* Grid Search was used to optimize hyperparameters, improving model performance.

**6. Model Evaluation**

**6.1 Performance Metrics**

* **Confusion Matrix:**

lua

Copy code

[[True Negatives, False Positives]

[False Negatives, True Positives]]

* **Classification Report:**
  + Precision: 0.93
  + Recall: 0.85
  + F1-Score: 0.89
  + Support: Number of actual occurrences of the class in the specified dataset.
* **ROC AUC Score:**
  + The model achieved an ROC AUC score of 0.95, indicating excellent discriminative ability.

**6.2 Model Interpretation**

* Feature importance analysis indicated that certain features significantly contribute to the prediction of fraudulent transactions. Features like V3, V4, and Amount were particularly influential.

**7. Conclusion and Future Work**

**7.1 Summary**

The Random Forest model effectively identifies fraudulent transactions, achieving an accuracy above the target of 75%. The balance of the dataset through SMOTE and the careful preprocessing steps contributed to the model's performance.

**7.2 Future Work**

* Explore additional models such as Gradient Boosting or Neural Networks for potentially improved performance.
* Consider implementing real-time detection systems for fraud prevention.
* Further investigate feature engineering techniques, such as interaction terms or more complex transformations.