**Report on Anomaly Detection for Credit Card Fraud**

**1. Introduction**

* **Context**:
  + Credit card fraud detection is crucial for protecting customers from unauthorized transactions. The aim is to accurately identify fraudulent transactions without disrupting genuine transactions.
* **Dataset**:
  + The analysis is based on a dataset of credit card transactions, where the goal is to classify transactions as legitimate (0) or fraudulent (1). The dataset is highly imbalanced, with fraudulent transactions being a small minority.

**2. Data Preprocessing**

* **Data Overview**:
  + The dataset contains various features, most of which are scaled values. The notable columns include Time, Amount, and Class (fraud indicator).
  + A summary of the dataset was produced using descriptive statistics and the distribution of the classes was observed (i.e., a highly skewed dataset).
* **Feature Scaling**:
  + Applied RobustScaler to scale the Time and Amount columns to handle outliers effectively. Other columns were already scaled.

**3. Methodology**

* **Data Splitting**:
  + Employed stratified k-fold cross-validation to ensure balanced class distributions between training and testing sets.
* **Handling Imbalance**:
  + Addressed the class imbalance by creating a new dataset with a balanced number of fraudulent and non-fraudulent transactions.
  + Used techniques like t-SNE (t-distributed Stochastic Neighbor Embedding) for dimensionality reduction, facilitating the visualization and clustering of similar data points.

**4. Model Training and Evaluation**

Several models were applied and evaluated for detecting fraudulent transactions:

* **Logistic Regression**:
  + Implemented as a baseline model.
  + Evaluated using accuracy, precision, recall, and F1-score metrics.
* **Random Forest Classifier**:
  + Showed significant improvement over Logistic Regression.
  + Accuracy and other metrics were computed, revealing a better balance between precision and recall for identifying fraudulent transactions.
* **XGBoost**:
  + Another powerful method for handling imbalanced datasets.
  + Results indicated competitive performance with a focus on minimizing false negatives.
* **Support Vector Machine (SVM)**:
  + Applied for further exploration of the decision boundary between fraudulent and legitimate transactions.
  + Achieved good results but was more computationally intensive.

**5. Results and Analysis**

* **Best Model**:
  + The Random Forest model emerged as the most effective based on the evaluation metrics, achieving a good balance between precision and recall.
  + The results indicated that decision trees could benefit from parameter tuning.
* **Grid Search for Hyperparameter Tuning**:
  + Applied GridSearchCV to optimize the Random Forest model parameters, improving its performance in detecting fraud.

**6. Conclusion**

* **Summary**:
  + The analysis demonstrated that with appropriate data preprocessing and model selection, it is possible to significantly improve the detection of fraudulent transactions.
  + Random Forest, supported by hyperparameter tuning, was the best-performing model for this dataset.