Project Report: Fraud Detection in Bank Transactions

1. Introduction

The goal of this project is to implement a machine learning-based fraud detection system for bank transactions.

The system aims to classify transactions as either fraudulent or non-fraudulent based on several features,

including transaction details, customer profiles, account activity, and historical fraud patterns.

The project involves data pre-processing, model training using multiple classifiers,

and model evaluation using various performance metrics.

2. Data Pre-processing

The data pre-processing involved several steps to prepare the dataset for model training:

- Merging multiple datasets, including customer profiles, transaction details, merchant data, and fraud indicators.
- Handling missing values by filling numerical columns with mean values and categorical columns with 'Unknown'.
- Checking the distribution of numerical features using Shapiro-Wilk normality tests.
- Encoding categorical variables using Label Encoding.
- Dropping highly collinear features with a correlation coefficient above 0.9.

3. Modeling

For the classification task, four different models were trained and evaluated:

- Logistic Regression
- Decision Tree Classifier

- Random Forest Classifier

- Support Vector Machine (SVM)

These models were chosen due to their popularity in classification tasks and their ability to handle complex datasets.

Each algorithm was evaluated based on accuracy, ROC AUC, precision, recall, and F1-score.

4. Evaluation

The evaluation of models was based on the following metrics:

- Accuracy: The proportion of correct predictions.

- Confusion Matrix: To understand the distribution of predicted and actual labels.

- Classification Report: Provides precision, recall, F1-score for each class.

- ROC AUC: The area under the receiver operating characteristic curve, indicating model's ability to distinguish between

fraudulent and non-fraudulent transactions.

Results:

- Logistic Regression: Accuracy = 0.5314, ROC AUC = 0.5863

- Decision Tree: Accuracy = 0.9162, ROC AUC = 0.9162

- Random Forest: Accuracy = 0.9712, ROC AUC = 0.9974

- SVM: Accuracy = 0.9267, ROC AUC = 0.9812

5. Model Tuning

Hyperparameter tuning was performed on the Random Forest model using GridSearchCV with 5-fold cross-validation.

The following hyperparameters were tuned:

- n_estimators: The number of trees in the forest.
- max_depth: The maximum depth of each tree.
- min_samples_split: The minimum number of samples required to split an internal node.
- min_samples_leaf: The minimum number of samples required to be at a leaf node.

The best parameters found were:
- n_estimators: 200
- max_depth: None
- min_samples_split: 2
- min_samples_leaf: 1

The tuned Random Forest achieved an improved performance with a ROC AUC of 0.9973 and an accuracy of 0.9738.

6. Model Selection

Based on the evaluation metrics, the Random Forest model was selected as the final model due to its superior performance
across all metrics, particularly its high ROC AUC score of 0.9974 before tuning and 0.9973 after tuning.
The model showed a good balance between precision and recall, making it suitable for fraud detection in bank transactions.

7. Conclusion

In this project, a fraud detection system was developed and evaluated using multiple machine learning models.
The Random Forest model was selected as the best performer after tuning its hyperparameters.

Future improvements could involve incorporating additional features such as time series data or exploring deep learning techniques.

Additionally, improving the data collection process and handling class imbalances further could boost model performance.