Project Report: Fraud Detection Using Machine Learning

Objective:

To develop machine learning models that predict fraudulent transactions from a bank transaction dataset. The goal is to classify transactions into **fraudulent** or **non-fraudulent** categories.

Dataset Overview:

- **Source**: Kaggle
- **Features**:
  - **TransactionAmount**: The amount of the transaction.
  - **AccountBalance**: Account balance during the transaction.
  - **TransactionHour**: The time of day the transaction occurred.
  - **DaysSinceLastTransaction**: Number of days since the last transaction.
  - **DeviceID**: The device used for the transaction.
  - **Location**: The location of the transaction.
  - **TransactionCount**: Number of transactions by the account.
  - **Fraudulent Flag**: Whether the transaction is fraudulent (1 = fraud, 0 = non-fraud).

Model Development:

1. Logistic Regression:

- **Accuracy**: 98.94%
- **Precision** (Fraud): 1.00
- **Recall** (Fraud): 0.11
- **Issues**: High precision but low recall for fraud. The model misses most fraud cases.

2. Support Vector Classifier (SVC):

- **Accuracy**: 98.94%
- **Precision** (Fraud): 1.00
- **Recall** (Fraud): 0.11
- **Issues**: Similar performance to Logistic Regression. The model struggled with fraud detection.

3. Random Forest Classifier:

- **Accuracy**: 99.34%
- **Precision** (Fraud): 0.70
- **Recall** (Fraud): 0.78
- **F1-Score** (Fraud): 0.74
- **Improvements**: Better recall for fraud detection, but still some false positives.

4. XGBoost:

- **Accuracy**: 99.60%
- **Precision** (Fraud): 0.75
- **Recall** (Fraud): 1.00
- **F1-Score** (Fraud): 0.86
- **Key Insight**: Perfect recall (1.00) for fraud detection, minimal false positives (3).

Model Performance Comparison:

Model	Accuracy	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)	Precision (Non-Fraud)	Recall (Non-Fraud)
Linear Regression	0.02	N/A	N/A	N/A	N/A	N/A
Logistic Regression	98.94%	1.00	0.11	0.20	0.99	1.00
Support Vector Classifier	98.94%	1.00	0.11	0.20	0.99	1.00
Random Forest	99.34%	0.70	0.78	0.74	1.00	1.00
XGBoost	99.60%	0.75	1.00	0.86	1.00	1.00

Key Insights:

- **XGBoost** is the best-performing model with **perfect recall for fraud detection (1.00)** and minimal false positives.
- **Random Forest** also showed strong performance, with **78% recall** for fraud detection, but more **false positives** compared to XGBoost.

- **Logistic Regression** and **SVC** performed well on **non-fraud transactions** but had **low recall for fraud** (only 11%).
- The dataset is **imbalanced**, with fraud transactions being rare, which challenges the model to detect fraud effectively.

Future Work:

1. **Hyperparameter Tuning**: Optimize the hyperparameters of **XGBoost** and **Random Forest** to improve performance.
2. **Resampling**: Use techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** or **undersampling** to balance the dataset and enhance fraud detection.
3. **Real-time Fraud Detection**: Implement the **XGBoost** model for real-time fraud detection in a live system.

Conclusion:

- The **XGBoost** model provided **outstanding performance**, achieving **perfect recall for fraud detection**, which is crucial for fraud detection tasks.
- The project demonstrates the importance of choosing the right model for imbalanced datasets, and **XGBoost** proved to be the most suitable for this task.