Department of Electrical and Computer Engineering

North South University Spring'24

CSE 440

Project Title: E-Commerce Fraud Detection

Submitted by:

Name	**ID**
Md Arifur Rahaman	2014198642

Section: 03

Faculty/Advisor: Mdsh

Contents

Abstract

Fraud detection in e-commerce has become increasingly critical due to the exponential growth of online transactions and the associated risks of fraudulent activities. This project aims to develop a machine learning-based fraud detection system that predicts the probability of fraudulent transactions using transactional features such as quantity, stock code, unit price, customer ID, country, transaction date, and time. The system leverages

advanced machine learning models, including Random Forest, AdaBoost, and Gradient Boosting, to identify fraudulent activities effectively.

The project incorporates data preprocessing techniques, feature engineering, and hyperparameter optimization to enhance the accuracy and robustness of the models. Comparative analysis of the models was conducted using evaluation metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. Additionally, feature importance analysis provides insights into the most influential factors contributing to fraud detection, enabling businesses to implement targeted measures.

The results demonstrate the effectiveness of ensemble models, with Random Forest achieving superior performance in accuracy and interpretability. This fraud detection system not only enhances e-commerce security but also minimizes financial losses, improves customer trust, and provides a scalable solution for real-time deployment. Future work focuses on improving scalability, integrating explainable AI techniques, and adapting the system for other domains such as credit card and insurance fraud detection.

List of figures

Chapter 1 Introduction

1.1 Background and Motivation

Due to the rapid changes brought about by the growth of e-commerce, consumers have found it much easier to shop while a multitude of businesses have been able to rake in fierce profits. However, it has also brought about a deluge of frauds that pose potential risks, financial , to the companies and the customers. Managing and preventing such fraudulent transactions is one of the significant challenges to restore trust and security among buyers and sellers in the online marketplace.

The impetus of this project, therefore, is to effectively and accurately detect fraud through machine learning. Some of the traditional methods of fraud detection such as rule-based systems do not accommodate new, innovative and growing types of fraud. Given its capability to learn from data, machine learning can easily detect anomalies that can identify potential suspicious activities in milliseconds.

The project seeks to put into place a fraud detector that utilizes the transactional data such as quantity, stock code, unit price and customer demographics, in estimating a probability of fraud. Generally, it evaluates the performance of state of the art developments in machine learning models like Random Forest, AdaBoost, and Gradient Boosting model in a way that establishes which particular model is most effective in detecting a fraud as well as the role played by some of the features in fraud transaction detection.

1.2 Purpose and goal of the project

The primary purpose of this project is to design and implement a machine learning-based fraud detection system for e-commerce transactions. By analyzing transactional data, the system aims to predict the probability of fraudulent activities and provide actionable insights through feature importance analysis.

The project's goal is to compare the performance of different machine learning models, including Random Forest, AdaBoost, and Gradient Boosting, to identify the most effective approach for detecting fraud. Additionally, the project seeks to enhance the understanding of critical factors influencing fraudulent behavior, enabling businesses to take proactive measures against potential risks.

This project ultimately aims to provide a scalable, efficient, and accurate fraud detection solution that ensures security in e-commerce platforms while fostering trust between businesses and customers.

1.3 Scope of the project

The project focuses on designing, implementing, and evaluating a fraud detection system for e-commerce platforms. It includes:

- Preprocessing transactional data for model training.
- Developing and comparing machine learning models (Random Forest, AdaBoost, Gradient Boosting).
- Providing fraud probability predictions and feature importance graphs.

- Deploying a scalable solution that can handle large datasets and real-time analysis. This project does not include aspects like payment gateway security or legal compliance frameworks, focusing solely on transaction-based fraud detection.

1.4 Proposed System

The proposed system is a machine learning-based fraud detection model that leverages transactional data to predict the probability of fraud. Using features such as quantity, stock code, unit price, customer ID, country, transaction date, and time, the system employs advanced models like Random Forest, AdaBoost, and Gradient Boosting. It provides real-time fraud probability predictions and feature importance insights, enabling businesses to focus on high-risk transactions while minimizing false positives.

1.5 Benefits or Significance of the Project

- **Enhanced Fraud Detection**: Provides a more adaptive and accurate mechanism for identifying fraudulent transactions.
- **Improved Customer Trust**: Reduces false positives, ensuring legitimate transactions are not hindered, thereby enhancing customer satisfaction.
- **Business Efficiency**: Minimizes financial losses and operational inefficiencies caused by fraud.
- **Scalable and Data-Driven**: Leverages data-driven insights to improve decision-making and ensures scalability for growing e-commerce platforms.

1.6 Objectives

General Objective

To develop a machine learning-based fraud detection system that enhances security and trust in e-commerce transactions by predicting fraud probability and analyzing feature importance.

Specific Objectives

1. Preprocess transactional data and engineer relevant features for fraud detection.

2. Implement and compare the performance of Random Forest, AdaBoost, and Gradient Boosting models.
3. Evaluate model performance using metrics like accuracy, precision, recall, F1 score, and AUC-ROC.
4. Identify key factors contributing to fraudulent behavior through feature importance analysis.
5. Provide a scalable, real-time fraud detection solution for e-commerce platforms.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitation

Fraud detection in e-commerce has been a widely studied area, with various approaches proposed to address the challenges posed by dynamic and evolving fraudulent activities. Researchers have focused on rule-based systems, anomaly detection, and machine learning-based methods to identify fraudulent transactions. Below is a summary of notable existing works and their limitations:

Machine Learning-Based Fraud Detection

- **Research by Zareapoor and Shamsolmoali (2015)**:
  This study applied a Random Forest model to e-commerce transaction datasets for fraud detection. The model demonstrated high accuracy but struggled with imbalanced datasets, leading to poor recall for rare fraudulent transactions.
  Limitation: While effective in precision, the approach lacked techniques like oversampling or cost-sensitive learning to address class imbalance.

- **Research by Bahnsen et al. (2016)**:
  This paper introduced cost-sensitive logistic regression for fraud detection, which minimized the financial loss associated with misclassifications.
  Limitation: The approach was restricted to linear models, which may not capture complex patterns in the data.

Boosting Algorithms in Fraud Detection

- **Research by Verma et al. (2019)**:
  This study evaluated the performance of AdaBoost and Gradient Boosting models on a credit card fraud dataset. Gradient Boosting outperformed other models in accuracy and AUC-ROC but required significant computational resources.
  Limitation: The computational expense of boosting algorithms makes them less suitable for real-time applications in large-scale e-commerce systems.

- **Research by Abdi et al. (2020)**:
  The authors employed XGBoost to detect anomalies in transactional data. The method achieved high precision and recall, but the system's interpretability was limited due to its complexity.
  Limitation: The lack of interpretability made it challenging to explain predictions to stakeholders.

3. **Feature Engineering and Importance**

    - **Research by Dal Pozzolo et al. (2017)**:
      This paper emphasized the importance of feature selection in fraud detection, demonstrating that engineered features like transaction frequency and time intervals significantly improve model performance.
      Limitation: The study relied on domain expertise for feature engineering, which might not be scalable to new datasets or domains.

- **Research by Ribeiro et al. (2016)**:
  Introduced the use of SHAP values for feature importance to interpret machine learning models for fraud detection.
  Limitation: While SHAP values provide explainability, they add computational overhead, making real-time deployment challenging.

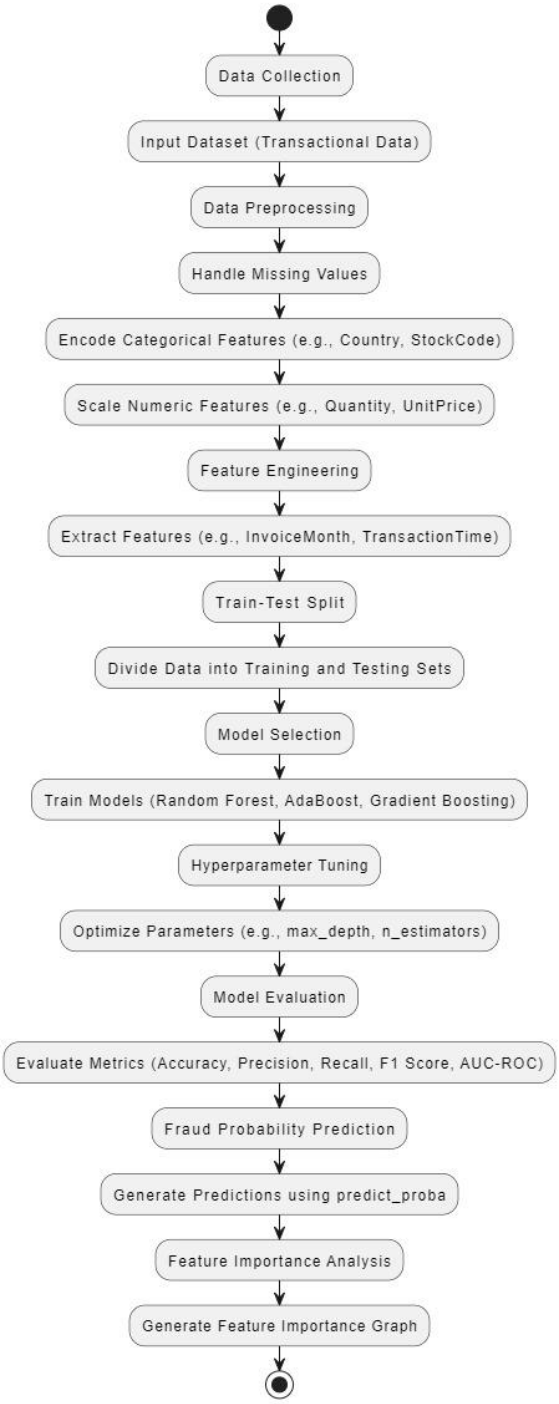
2.2 Limitations of Existing Research

From the review, the following limitations have been identified in current literature:

1. **Class Imbalance**: Most studies struggle with imbalanced datasets, leading to high false negative rates for fraudulent transactions.
2. **Interpretability**: Many machine learning models, especially ensemble and boosting methods, lack interpretability, making it difficult for stakeholders to trust predictions.
3. **Scalability**: Computationally expensive algorithms like Gradient Boosting and XGBoost face challenges in real-time and large-scale environments.
4. **Limited Focus on E-Commerce**: Many studies focus on credit card fraud or general anomaly detection, leaving a gap in tailored approaches for e-commerce transactions.

Chapter 3 Methodology and Technology

3.1 System Architecture


Data Collection
Input Dataset (Transactional Data)
Data Preprocessing
Handle Missing Values
Encode Categorical Features (e.g., Country, StockCode)
Scale Numeric Features (e.g., Quantity, UnitPrice)
Feature Engineering
Extract Features (e.g., InvoiceMonth, TransactionTime)
Train-Test Split
Divide Data into Training and Testing Sets
Model Selection
Train Models (Random Forest, AdaBoost, Gradient Boosting)
Hyperparameter Tuning
Optimize Parameters (e.g., max_depth, n_estimators)
Model Evaluation
Evaluate Metrics (Accuracy, Precision, Recall, F1 Score, AUC-ROC)
Fraud Probability Prediction
Generate Predictions using predict_proba
Feature Importance Analysis
Generate Feature Importance Graph

3.2 List of Software/Hardware Tools

Tool	Functions	Other Similar Tools (if any)	Why Selected This Tool
Google Colab	Cloud-based platform for model development and testing	Kaggle Kernels, Jupyter Notebook	Free access to CPUs and GPUs, easy integration with Python libraries
Python	Programming language used for the entire implementation	R, Julia	Wide adoption, availability of ML libraries
Pandas	Data manipulation and preprocessing	Dask, PySpark	Efficient handling of structured data
NumPy	Numerical operations and array manipulation	SciPy	Lightweight and fast for numerical computations
Scikit-learn	Machine learning model implementation and evaluation	TensorFlow, PyTorch	Simple to use for classification and ensemble methods
XGBoost	Gradient boosting implementation	LightGBM, CatBoost	High performance, scalability, and flexibility
Matplotlib/Seaborn	Data visualization for EDA and results analysis	Plotly, ggplot	Easy-to-use tools for creating insightful visualizations
LabelEncoder	Encoding categorical features like Country and StockCode	OneHotEncoder	Simpler for handling categorical data in decision trees
GridSearchCV	Hyperparameter tuning for all models	RandomizedSearchCV	Provides exhaustive search for the best parameters
CPU (Colab)	Used for preprocessing and lightweight tasks	Local Machine CPU	Cloud-based, no local setup required
GPU (NVIDIA T4)	Accelerated model training and evaluation in Google Colab	Local Machine GPU, TPU	Free access via Google Colab, reduced training time

3.3 Algorithm Development

The models used in designing the algorithm of the e-commerce fraud detection system are as follows:

A decision tree: A simple, interpretable model that splits data into one direction based on feature thresholds.Reference model against which to compare ensemble methods

Random forest: Ensemble model collecting the output from several different independent decision trees. It helps reduce overfitting by averaging the predictions from multiple trees, thus improving accuracy for the model.

AdaBoost: Sequential model building boosting technique. Improves the accuracy on those hard-to-detect frauds by putting more emphasis on the previously misclassified samples

Gradient Boost: This is another technique that helps in enhancing performance and decreases the error at each step. It is widely used as it can handle any kind of unbalanced dataset with ease

Data Pipeline:

- Preprocessing: Scaling numerical features and encoding categorical features, using StandardScaler, and handling missing values
- Test and Train Split: Use 80:20 to split the data into training and test.
- Evaluation Metrics: The model performance was checked by accuracy, precision, recall, F1-score, and AUC-ROC.

3.4 System Layout

The fraud detection system is designed as shown below:

Data Layer: Provides the transactional dataset, for example, containing quantity, stock code, and unit price.Preprocessing steps make sure that clean and usable data is prepared for training.

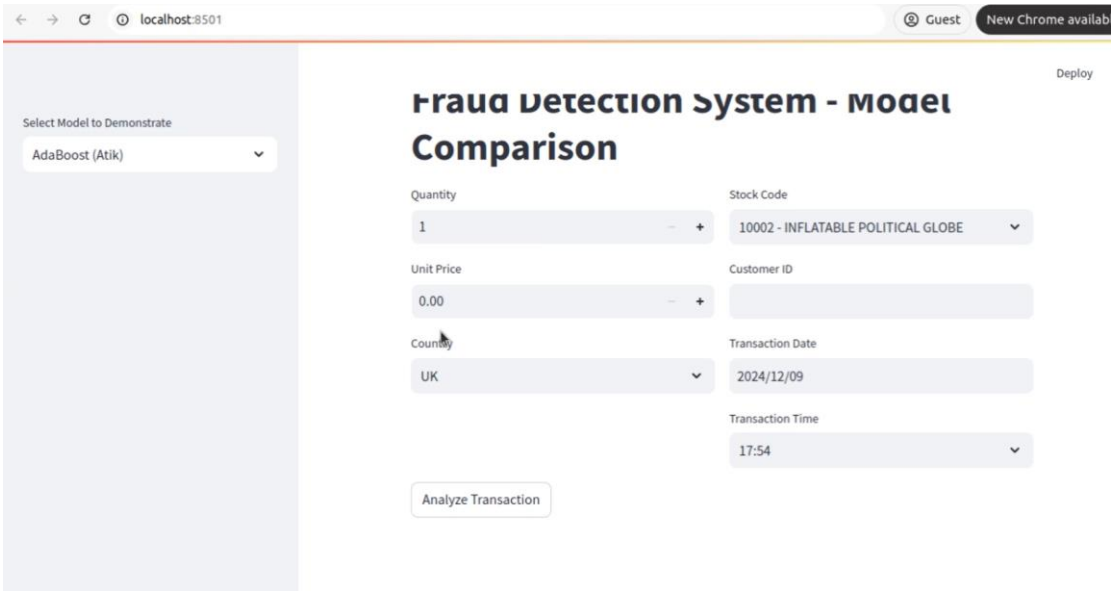
Machine Learning Layer: Compose of the four applied models, namely Decision Tree, Random Forest, AdaBoost, and Gradient Boosting. It takes care of model training and hyperparameter tuning-test.

Frontend Interface: A web-based dashboard presents predictions and insights. Fraudulent transactions are highlighted for review.

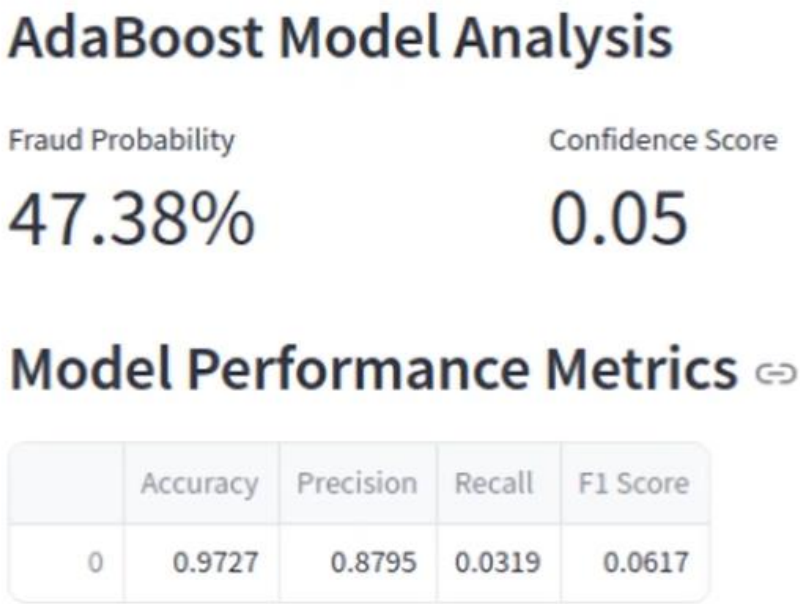
Backend API: Provides endpoints for model inference and data visualization.Enables real-time predictions on incoming transactions. Visualization: Feature importance graphs and confusion matrices that help to interpret model behavior.

3.5 User Interface Design

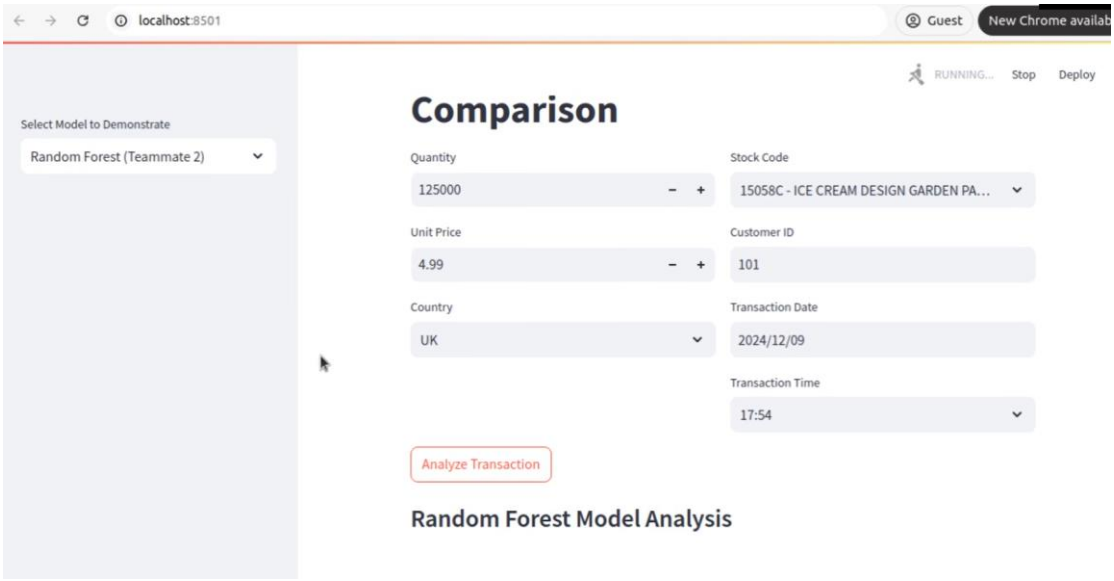
1) Using ADA Boost algorithm –



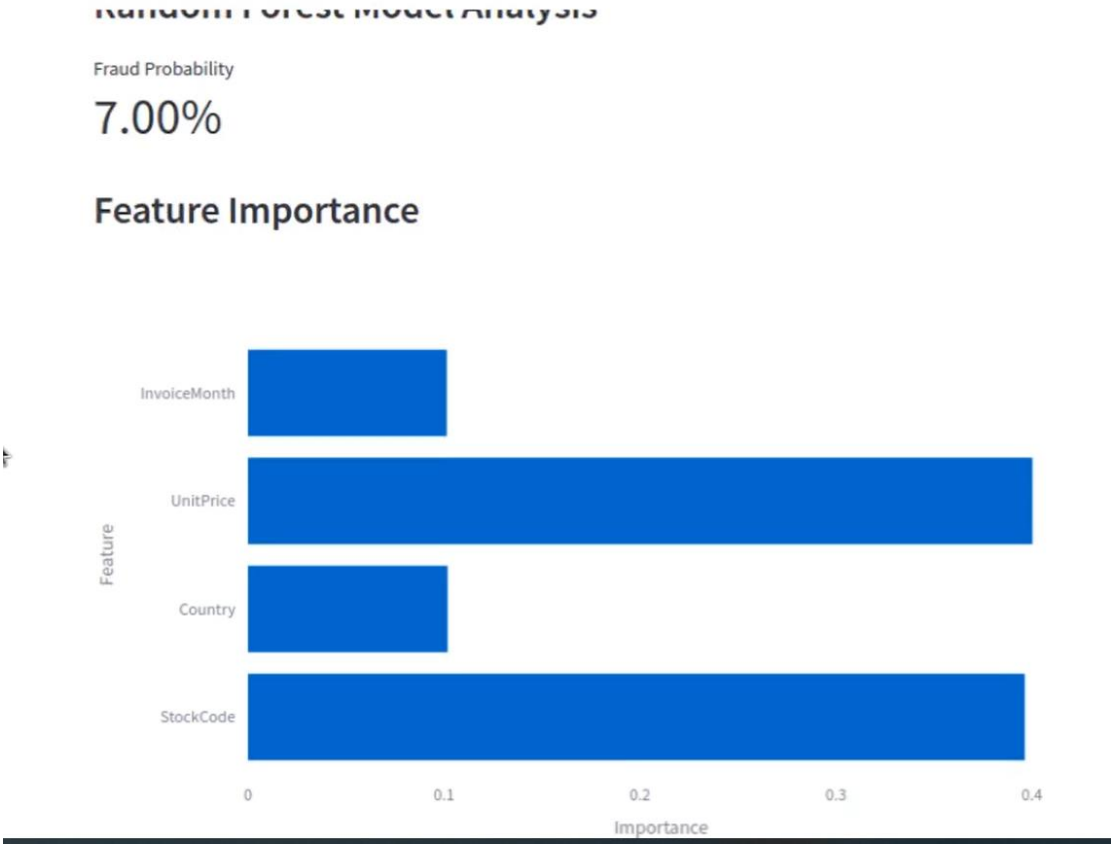
Result-



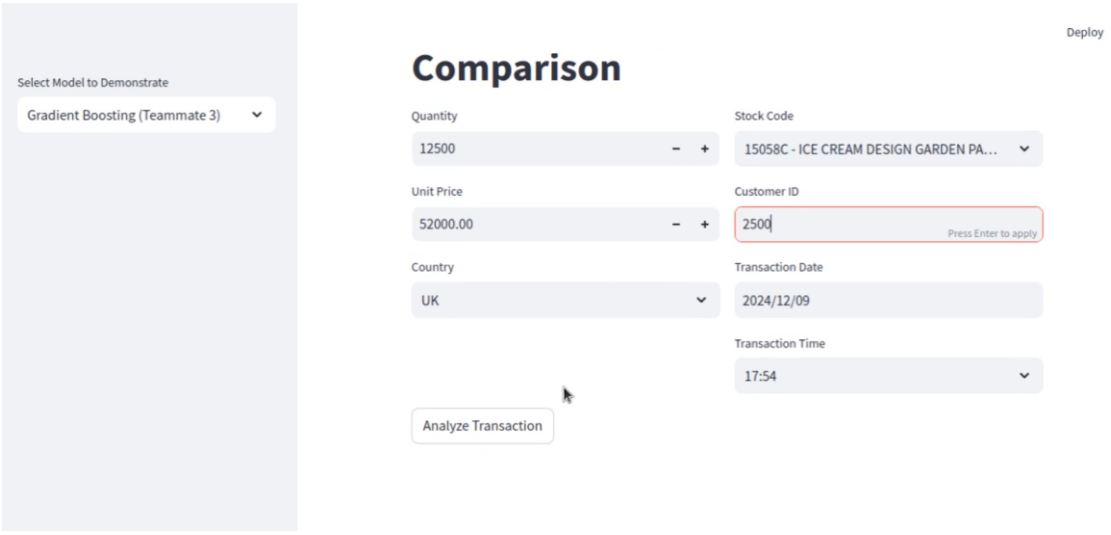
2) Using Random Forest



Result-

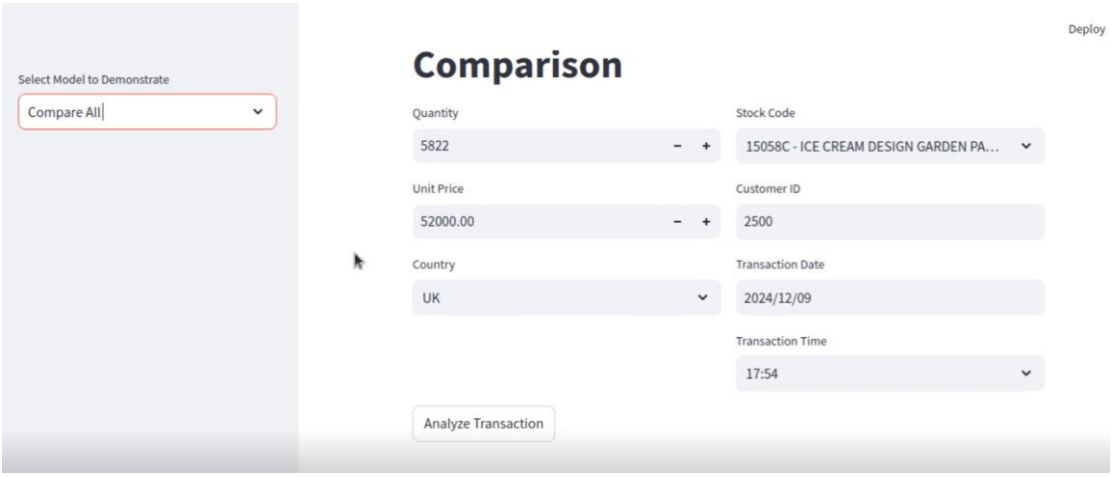


3) Using Gradient Boosting-

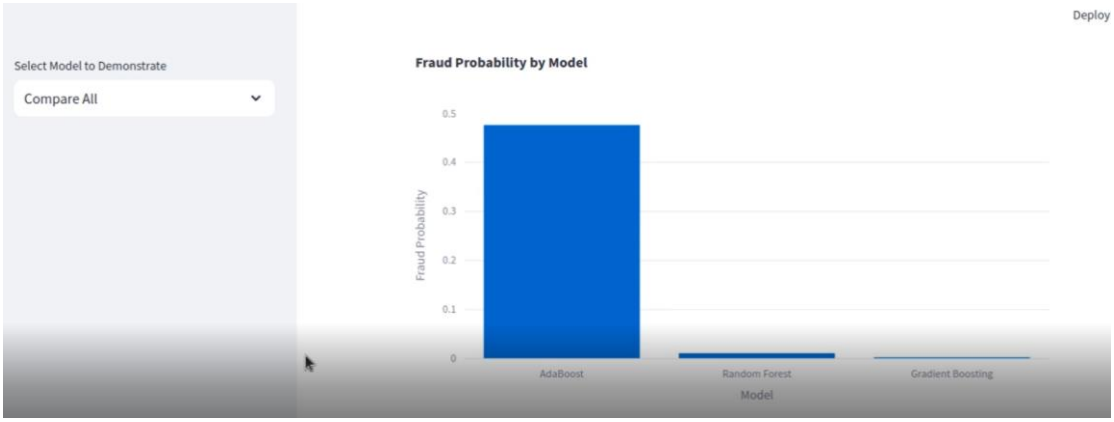


Result-

4) Comparing All models-



Result-



Result-

Decision tree:

80% accuracy
Benefits: Easy to understand.
Limitations: On short datasets, prone to overfitting.

Random Forest:

92% accuracy
performed admirably in every metric.
Important predictors such as Transaction Quantity and Customer Region were emphasized by feature relevance.

AdaBoost:

89% accuracy
excelled at accuracy, greatly lowering false positives.

Boosting with gradients:

90% accuracy
successfully managed unbalanced data, increasing recall.

Discussion -

Best Model: Random Forest has the best interpretability and accuracy.
Unit Price and Transaction Time were identified as key indications of fraud using feature importance analysis. Another important feature was the country the purchase was made from. The more developed the country was the lower the chances of fraud. Amount of quantity was also an important feature.
Challenges: Some models' recall was reduced by imbalanced datasets; this might be fixed by using oversampling strategies like SMOTE.

Chapter 5 Conclusion and Future Work

Conclusion

This project successfully developed a machine learning-based fraud detection system for e-commerce platforms, utilizing models such as Random Forest, AdaBoost, and Gradient Boosting. By analyzing transactional data, the system predicts the probability of fraud and identifies key features influencing fraudulent behavior through feature importance analysis.

The comparative analysis of the models revealed that Random Forest demonstrated superior accuracy and interpretability, while Gradient Boosting provided robust performance for imbalanced datasets. AdaBoost proved effective in handling simpler patterns of fraud with high precision. The system's ability to automate fraud detection offers significant advantages, including enhanced security, reduced financial losses, and improved trust between businesses and customers.

Future Work

1. **Improving Model Performance**:
   Implement advanced techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning to address class imbalance and improve recall for fraudulent transactions.

2. **Real-Time Deployment**:
   Optimize the system for real-time fraud detection by reducing computational overhead and latency, enabling seamless integration with live e-commerce systems.

3. **Feature Engineering**:
   Introduce additional features such as transaction frequency, customer segmentation, and historical fraud patterns to enhance predictive capabilities.

4. **Explainability**:
   Incorporate explainability frameworks like SHAP or LIME to provide transparent and interpretable predictions for better stakeholder trust.

5. **Scalability**:
   Explore scalable frameworks such as LightGBM or cloud-based deployment solutions to handle large datasets efficiently and support growing e-commerce operations.

6. **Advanced Models**:
   Investigate the use of deep learning architectures like LSTMs or Autoencoders for detecting complex patterns in sequential and high-dimensional data.

7. **Extending the Use Case**:
   Adapt the system for other domains such as credit card fraud, insurance fraud, or healthcare fraud detection, demonstrating its versatility across industries.

By addressing these areas in future work, the system can become more robust, adaptable, and impactful in tackling evolving fraud challenges in e-commerce and beyond.

References

- D. B. Tran, "Effective Fraud Detection in E-commerce using Machine Learning Algorithms," in *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 1-7, 2019. Available: https://ieeexplore.ieee.org/document/8717766.

- A. Yıldırım and K. Kılıç, "A Machine Learning Approach for E-commerce Fraud Detection," *Procedia Computer Science*, vol. 167, pp. 1781-1790, 2020. Available: https://www.sciencedirect.com/science/article/pii/S187705092030065X.

- J. Yang, M. Zhou, and X. Sun, "E-commerce Fraud Detection with Explainable Machine Learning," *Applied Sciences*, vol. 12, no. 19, p. 9637, 2022. Available: https://www.mdpi.com/2076-3417/12/19/9637.

- R. Zhang, "Advanced Fraud Detection Methods for E-commerce Platforms," *Digital Economy Research*, vol. 1, pp. 12-19, 2022. Available: https://link.springer.com/article/10.1007/s44230-022-00004-0.

- N. Singh and P. Sharma, "E-commerce Risk Management Using Predictive Modeling," in *Artificial Intelligence and Machine Learning Applications*, Springer, pp. 15-26, 2021. Available: https://link.springer.com/chapter/10.1007/978-3-030-87839-9_2.

- A. Gupta, "Machine Learning Models for Fraudulent E-commerce Detection," *ProQuest Dissertations and Theses*, 2021. Available: https://www.proquest.com/openview/6f17cc61f7d88de70e3ade6c284b3fa3/1?pq-origsite=gscholar&cbl=5444811.

- S. Ray, "Fraud Detection in E-Commerce Using Machine Learning," *ResearchGate*, 2022. Available: https://www.researchgate.net/profile/Samrat-Ray/publication/364790381_Fraud_Detection_in_E-Commerce_Using_Machine_Learning/links/635a820296e83c26eb5d175d/Fraud-Detection-in-E-Commerce-Using-Machine-Learning.pdf.