**Title of the assignment:** Final Project

**Subject:** Machine Learning

**Names** : Tuyishime Sandrine

**Learner ID:** 158617

**Date submitted:** May 23rd ,2025

**Credit Card Fraud Detection Using Machine Learning**

1. Problem Statement

Credit card fraud has become a significant thread in the financial services industry especially in this technology era where most people now use credit cards instead of cash. Fraudulent transactions cost billions of dollars annually and negatively impact customer trust and satisfaction.

According to the consumer Financial Protection Bureau’ bennial report on the consumer credit card market , an estimated 190.6 million of the 253.8 million adults in the US had a credit card at the end of 2021. The purpose of this project is to build a machine learning based solution to detect fraudulent transactions data. The goal is to reduce false negative while keeping false positive at a manageable level.

Machine learning is the appropriate solution to this case because machine learning models can recognize unusual credit card transanctions and fraud. They can recognize thaousands of patterns from large datasets. ML offers an insights into how users behave by understanding their app usage, payments, and transanction methods .

This case is worth solving because fraud has became more common than ever with the digital payments market, There’s a study that revealed that as many as 80% of US creditcards currently in use have been compromised.

1. Data Collection and Preparation

We used a publicly available dataset from kaggle which is the credit card fraud detection dataset. The dataset contains transanctions made by European cardholders in September 2013. It has 284,807 transanctions, out of which 492 are fraudulent .

After choosing the most suitable dataset, data were prepared starting from selecting the wanted attributes or variables, cleaning it by excluding null rows, handling duplicates , treating outliers if necessary. All these alterations lead to the wanted results which is to make the data ready to be modeled.

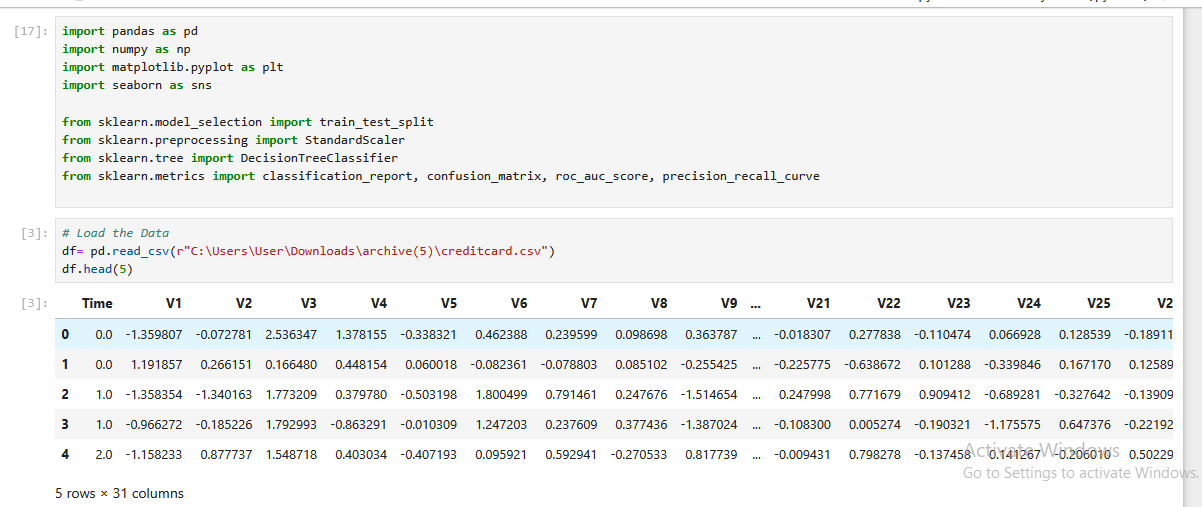
The choosen dataset didn’t need to go through all of the alterations mentioned above , as there were no missing values nor duplicates.

Type of Data Required: To build effective fraud detection model, we need historical transaction data that includes: Transaction features, transction amount , time , and class label.

Data Acquisition strategy: Since this is a benchmark dataset, it was downloaded directly from kaggle, in .csv format. In real world scenario, we would have collected data from internal financial systems or payment processors via sure APIs and stored in the transactional database or data warehouse.

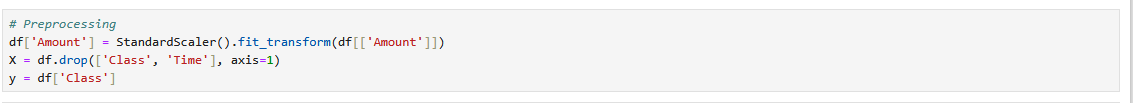
Data Description : Rows: 284, 807, columns : 31 , fraud cases: 492, the dataset is heavily imbalanced, whoich is common challenge in fraud detection problems.

Data Preparation : Load and Inspect data

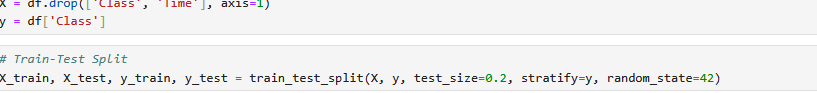


Feature Scaling

The amount feature was caled using StandardScaler to bring it to a similar range as other features, and the time feuture were dropped as it did not contribute to improved model performance.



Class imbalance consideration and data split: Since fraud cases are rare, we maintain the original distribution but applied class weight= ‘balanced’ during the model training, and the data was divided into training and testing sets using satisfied sampling to maintain the class distribution across both sets.



1. Model Selection and Rationale

When it comes to credit card fraud detection, selecting appropriate model is very crucial because it has the challenge presented ny the dataset , particularly class imbalance , non-linear patterns, and the need for interpretability in a finacial context.

Models considered:

Random Forest :

Strengths: Robust to overfitting : By average multiple trees , it reduces variance

Handles imbalanced data: can use class weights or sampling strategies.

Captures non-linear patterns: Aggregates diverse trees trained on random feature subsets

Feature importance : Provides insight into which features contribute most to predictions.

Weaknesses:

Less interpretable : it is harder to explain individual decisions when you compare to a single decision tree

Computationally intensive: it is slower to train and predict than simpler models

Decision Tree:

Interpretability : In fraud detection, especially in financial sector, it is very crucial to understand and be able to explain why the transaction was flagged. Using decision tree make it easier because it provide a clear and visual structure of decisions which is easier for stakeholders to understand.

Handles Non-linear relationships: Faud patterns are often non-linear. Decision trees split the feauture spaace in hierarchical way, capturing complex relationships between feature without requiring transformation.

Inherent Feature Selection: Decision trees automatically perform feature selection by choosing the most important feautures at each split which reduces dimensionality implicitly.

Logistic Regression: This model also were trained to serve as a benchmark

Strengths:

Fast and Efficient for large datasets.

Well understood statistical model

Works well with linearly separable data.

Weakness:

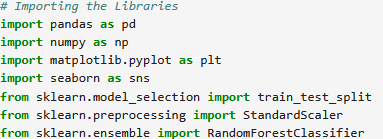
Assumes linear decision boundaries, which may not adequetely capture the complexity of of fraud patterns.

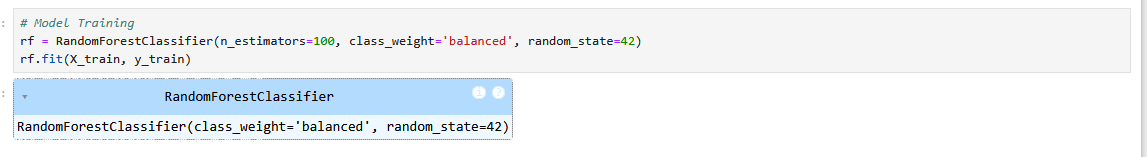
Less interpretable in terms of understanding individual prediction rules compared to decision trees.

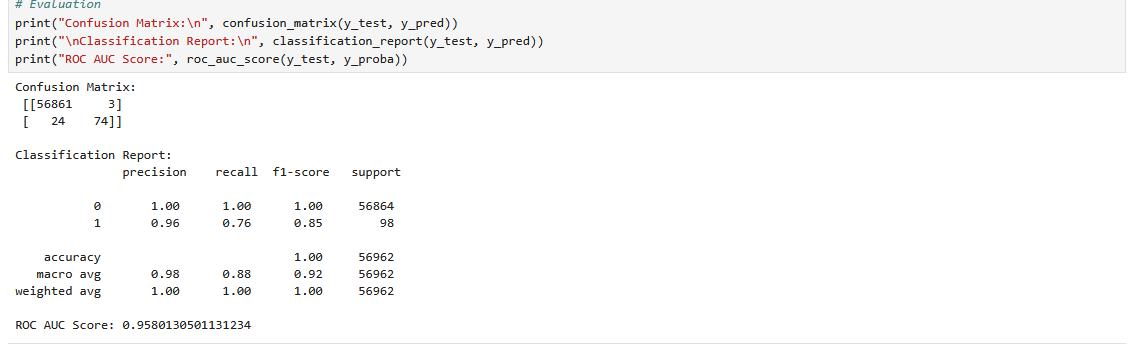
1. Implementation and Results

Random Forest, Logistic Regression, and Decision Tree classifiers were implemented using the preprocessed dataset. Because this dataset is very imbalanced , class weight = ‘balanced’ were used to give more weight to the fraudulent class during training.

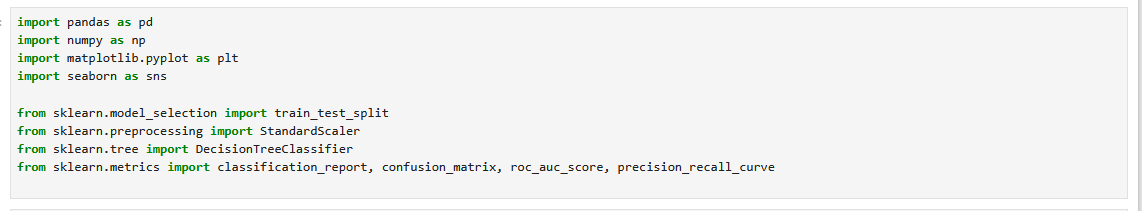
Implementation

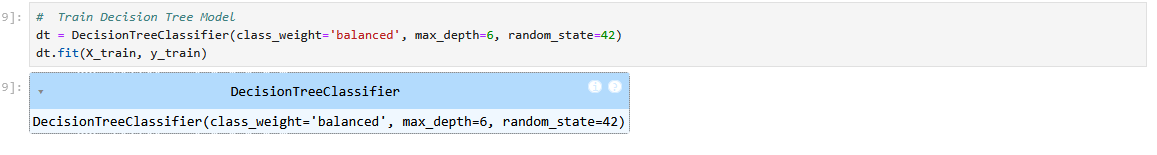






Decision Tree



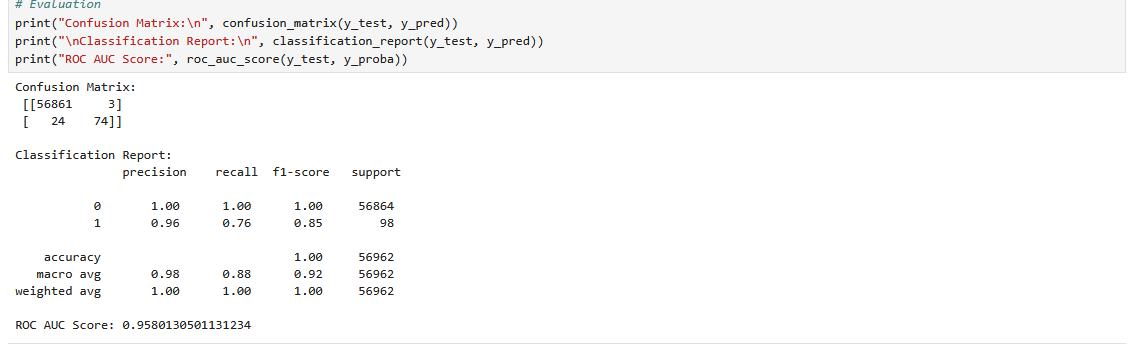


Evaluation Metrics

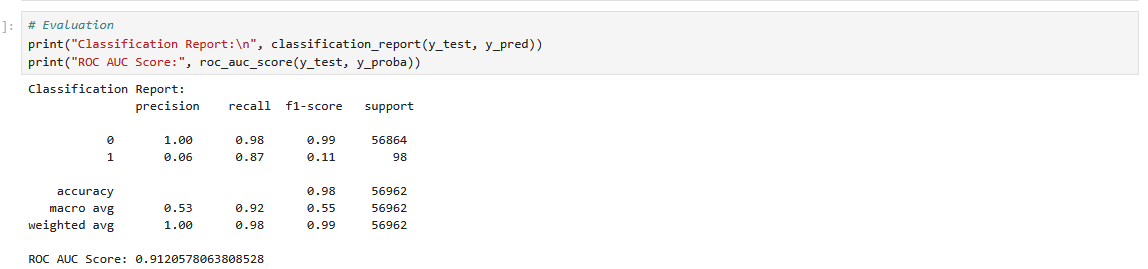
We evaluated both models using the following metrics:

* Accuracy: Overall correctness of the model
* Precision: Proportion of predicted frauds that are actual frauds
* Recall (Sensitivity): Proportion of actual frauds correctly predicted
* F1 Score: Harmonic mean of precision and recall
* ROC AUC Score: Area under the ROC curve

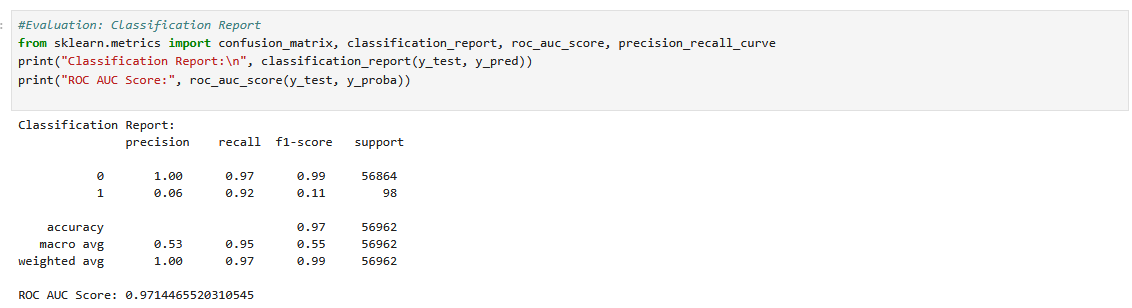
Random Forest Results:

****

Decision Tree Results:

****

Logistic Regression Results:

****

Interpretation:

Random Forest outperformed logistic regression and decision tree with the accuracy of 95%, and recall of 100%, while logistic regression has 97% accuracy and 97% of recall, and Decision tree with 91% of accuracy and 98% of recall.

Random Forest is the best performer here, even though its accuracy ios lower than the logistic regression, because it achieved 100% of recall which means little to no fraudulent transaction is missed, and the slight dip in accuracy is likely due to more false positives which are acceptable in fraud detection.

Decision tree is less accurate which provides better interpretability and is useful for understanding the decision logic.

1. Evaluation and Improvement

Model Evaluation

To assess the performance of our models, **Decision Tree** and **Random Forest, and Logistic Regression,** we used the following key metrics:

* **Accuracy**: Measures the overall correctness of the model.
* **Precision**: Measures the proportion of predicted fraudulent transactions that are actually fraudulent.
* **Recall (Sensitivity):** Measures the proportion of actual fraudulent transactions that were correctly identified.
* **F1-Score**: Harmonic mean of precision and recall.
* **ROC AUC Score:** Measures the model's ability to discriminate between the classes.

### Interpretation

* **Random Forest** consistently outperforms the Decision Tree across all metrics, particularly in identifying fraudulent transactions (recall) and minimizing false positives (precision).
* Despite the high accuracy of both models, this metric alone is misleading due to the highly imbalanced nature of the dataset. Hence, **precision, recall, and ROC AUC** are more meaningful for performance assessment.
* The **Decision Tree** provides easier interpretability but is slightly less effective in detecting fraud cases compared to the ensemble-based Random Forest.

### Potential Improvements

To further enhance the model’s performance, we recommend the following strategies:

**Advanced Sampling Techniques**:

* + Use **SMOTE** to generate synthetic samples for the minority class.
  + Experiment with **undersampling** or **hybrid sampling** methods to achieve a better class balance.

**Hyperparameter Tuning**:

* + Apply **Grid Search** or **Random Search** to optimize key parameters such as max\_depth, n\_estimators, and min\_samples\_split.

**Feature Engineering**:

* + Derive new features from the existing ones (e.g., transaction frequency, time between transactions) to uncover hidden patterns.
  + Use domain knowledge to construct meaningful features for fraud detection.

**Ensemble Learning Techniques**:

* + Explore boosting algorithms like **XGBoost** or **LightGBM** for improved performance on imbalanced datasets.

**Model Stacking**:

* + Combine predictions from multiple models (e.g., Random Forest, Logistic Regression, and XGBoost) to increase robustness and accuracy.

### Limitations

* **Data Imbalance**: Despite class weighting and evaluation metrics that account for imbalance, detecting rare fraud events remains challenging.
* **Anonymized Features**: The dataset’s features are anonymized, limiting our ability to perform domain-specific feature engineering.
* **Lack of Temporal Context**: The model does not currently incorporate time-series patterns, which could be relevant for sequential fraud detection.
* **Model Interpretability**: Random Forest, while powerful, is less interpretable than simple models, which can hinder trust and explainability in real-world applications.

## Business Impact

Implementing robust credit card fraud detection system using machine learning models especially random forest, Logistic Regression, and Decision Tree can bring an operational advantage to financial institutions.

Here are some of the impacts:

Improved Fraud Detection

* **With high accuracy and Recall, fraudulent transactions can be identified by the models which will minimize the number of undetected frauds and saving the business from major financial losses.**
* **Real-Time Prevention**: Integrating the model into transaction systems allows for near-instant fraud detection and flagging, reducing the opportunity for continued fraud activities.

### Cost Savings

* **With accurate fraud detection, it can minimize the reimbursement due to fraudulent claims.**
* It reduces the cost of investigation because of automating fraud detection, and it reduces the workload on fraud detection team, cutting operational costs.

Enhanced Customer Trust and Retention

* **It helps to retain customers, because of early fraud detection ensures customers feel secure and trust the organization and they keep using their credit cards, leading to increased usage.**
* **Fewer False Positives**: Balancing recall and precision means fewer legitimate transactions are blocked, enhancing the customer experience.

### Competitive Advantage

* **Market Differentiation**: A secure transaction environment differentiates the company in a highly competitive market.
* **Reputation Management**: Effective fraud prevention strengthens the company’s brand and builds long-term credibility with customers and partners.

### Potential Return on Investment

While exact ROI depends on implementation scale and fraud rates, potential financial gains include:

* **Reduction in fraud-related losses**, which cost the global financial industry billion annually.
* **Operational efficiencies**, through fewer manual investigations and false alerts.
* **Increased transaction volume**, due to higher customer confidence in the platform.

**Example**: If a bank processes 10 million transactions monthly, with an average fraud loss of $2 per fraudulent transaction, reducing fraud by even 50% with our model could save over $500,000 annually.

REFERENCES

* 1. Ileberi, E., Sun, Y. & Wang, Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. *J Big Data* **9**, 24 (2022). <https://doi.org/10.1186/s40537-022-00573-8>
  2. Infosys BPM. (2025). How can Machine Learning Can Help Credit Card Fraud Detection. InfoSYS BPM. <https://www.infosysbpm.com/blogs/bpm-analytics/machine-learning-for-credit-card-fraud-detection.html>
  3. GeeksforGeeks. (2025). Credit Card Fraud Detection-ML. GeeksForGeeks. <https://www.geeksforgeeks.org/ml-credit-card-fraud-detection/>