Credit Card Fraud Detection Using Machine Learning and Deploying Model in Streamlit App

Al Mahmud Siam
1902062

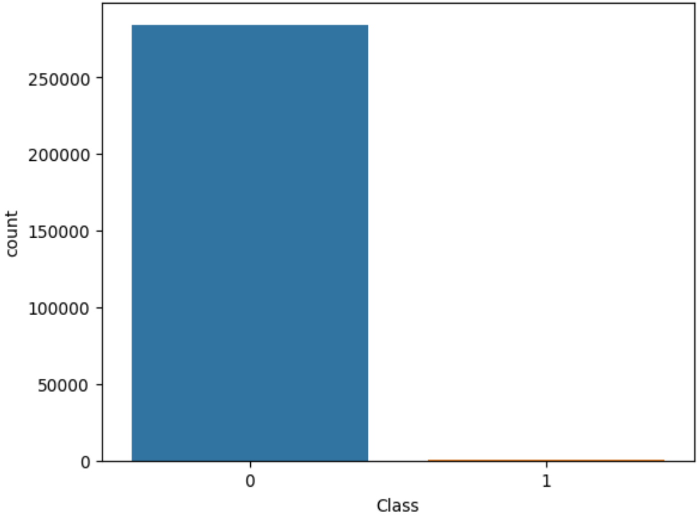
MD. Mostafizur Rahman
1902073

Outline

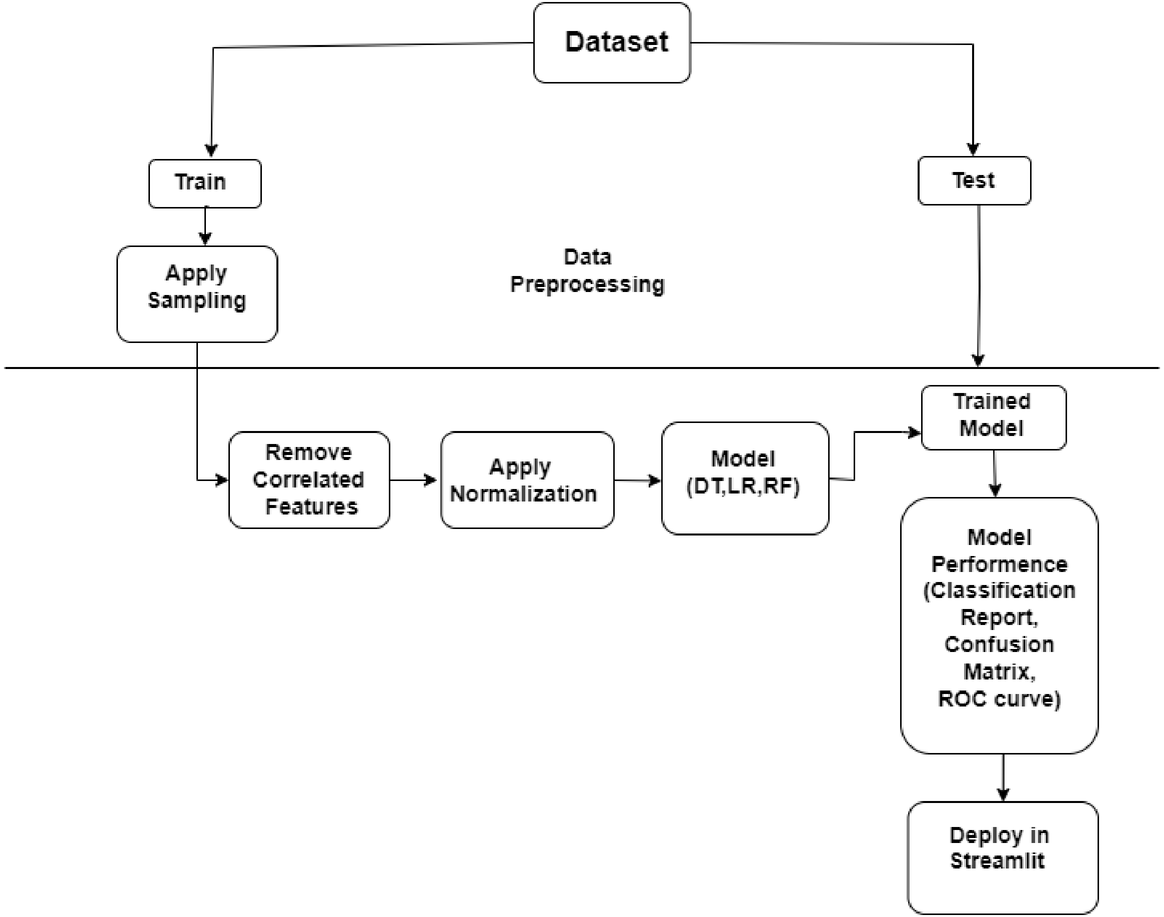
1. Introduction
2. Objective
3. Data description
4. Framework of Project
5. Balancing Data
6. Streamlit App Interface
7. Input Section For each Model
8. Problems and Future Work

Dataset Description and Visualization

- European cardholder’s two-day transactions in September 2013
- 31 features (30 are independent, and the remaining one is a dependent feature).
- Highly imbalanced dataset
- Total Transactions: 284,807
- Fraudulent Transactions: 492 (0.17%)

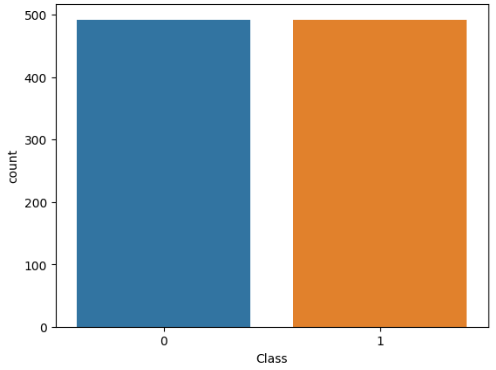


Framework of this Project



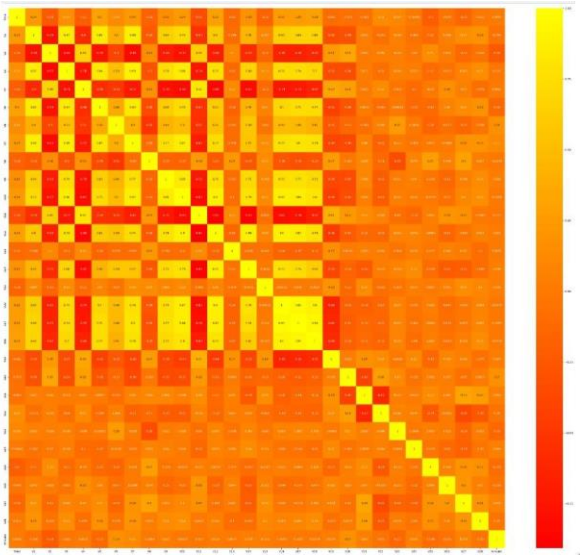
Balancing dataset

- With only 0.17% of transactions being fraudulent, models struggle to effectively learn patterns from the minority class.
- **Undersampling Approach**:
- Balanced the dataset by reducing it to 492 non-fraudulent and 492 fraudulent transactions to prevent bias.
- Reduces computational cost by working with a smaller dataset.
- Provides balanced classes to improve model sensitivity to fraud cases.



Remove correlated features

- Applied Pearson correlation to detect multicollinearity.
- Removed features with a correlation above a certain threshold (0.95) to reduce highly correlated features.
- Helps focus on the most informative features, improving model performance.
- Simplifies the model.



Normalization

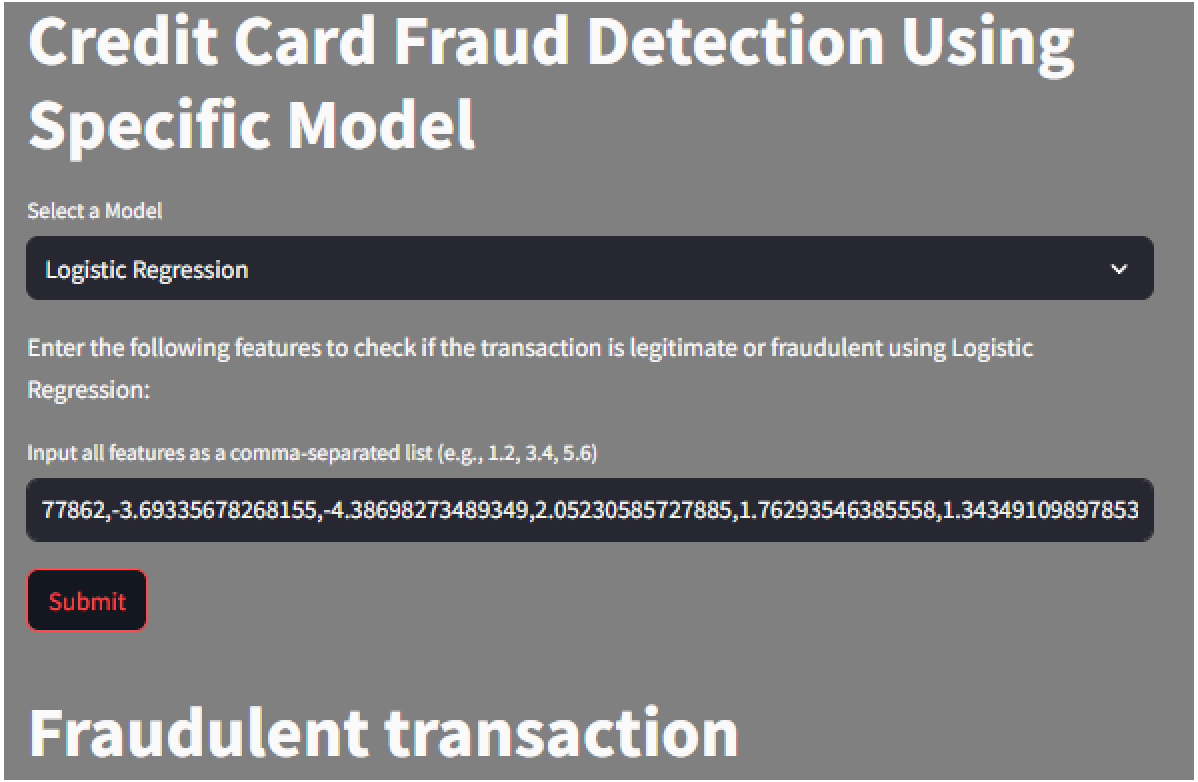
- Scaled features to a range of [0, 1] to ensure uniformity.
- X is the original value,
- Xscaled is the normalized value,
- Xmin is the minimum value in the feature,
- Xmax is the maximum value in the feature.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Benefits:

- Prevents features with larger ranges from dominating the model training process.
- Improves model performance by ensuring all features contribute equally.

Input Section for Each Model(1)



Classification Report and Confusion Martix for Logistic Regression

	precision	recall	f1-score	support
0	0.925234	1.000000	0.961165	99.000000
1	1.000000	0.918367	0.957447	98.000000
accuracy	0.959391	0.959391	0.959391	0.959391
macro avg	0.962617	0.959184	0.959306	197.000000
weighted avg	0.962427	0.959391	0.959315	197.000000

Fig. 4.3 Classification Report for Logistic Regression

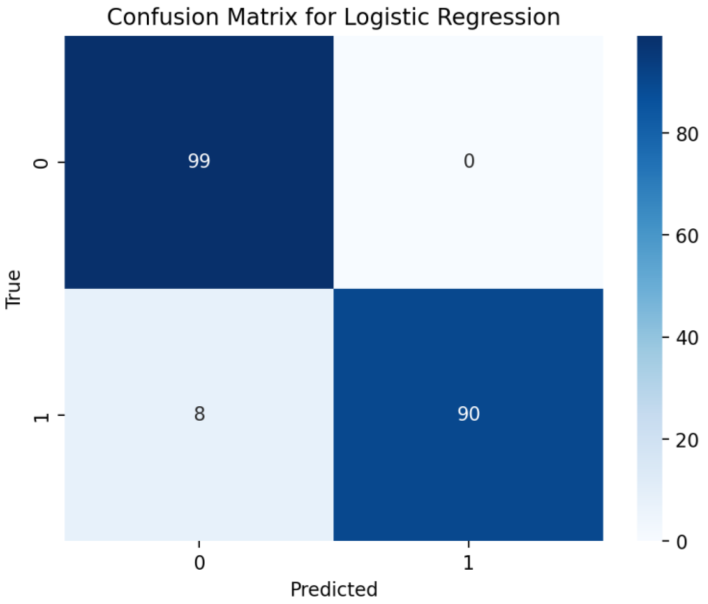


Fig. 4.4 Confusion Matrix for Logistic Regression

ROC curve for Logistic Regression

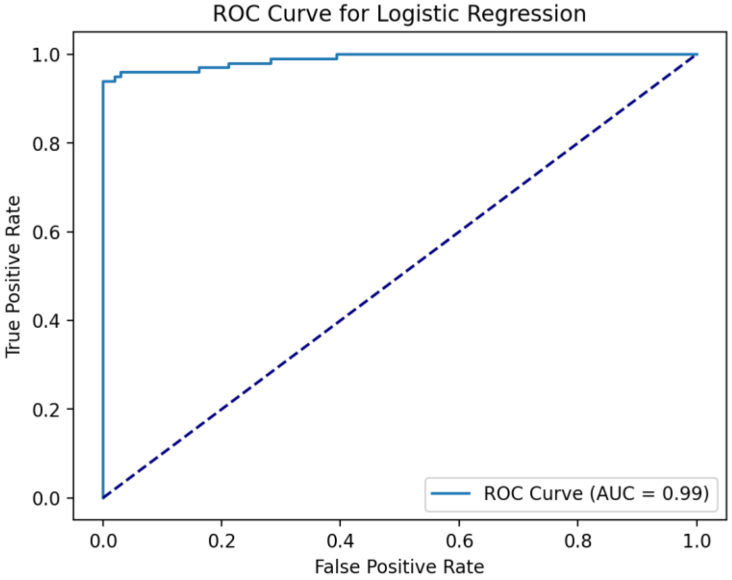
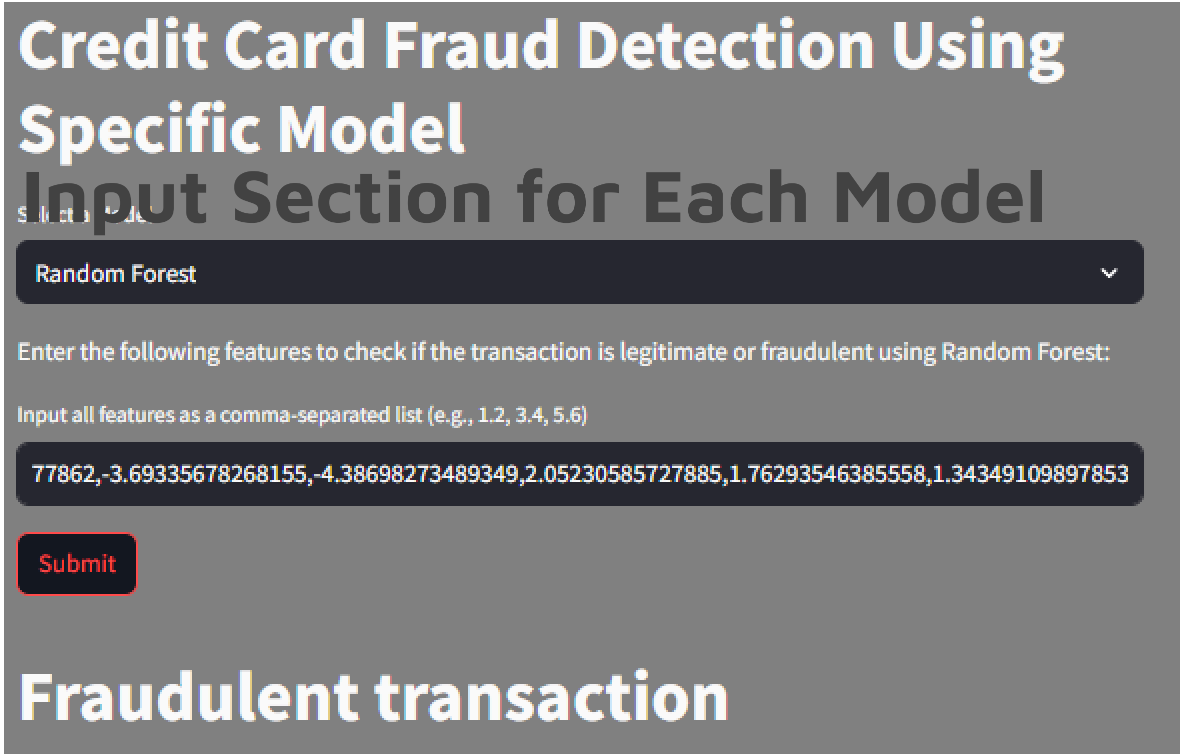


Fig. 4.9 ROC Curve for Logistic Regression

Input Section for Each Model(2)



Classification Report and Confusion Martix for Random Forest

	precision	recall	f1-score	support
0	0.950495	0.969697	0.960000	99.000000
1	0.968750	0.948980	0.958763	98.000000
accuracy	0.959391	0.959391	0.959391	0.959391
macro avg	0.959623	0.959338	0.959381	197.000000
weighted avg	0.959576	0.959391	0.959385	197.000000

Fig. 4.4 Classification Report for Random Forest

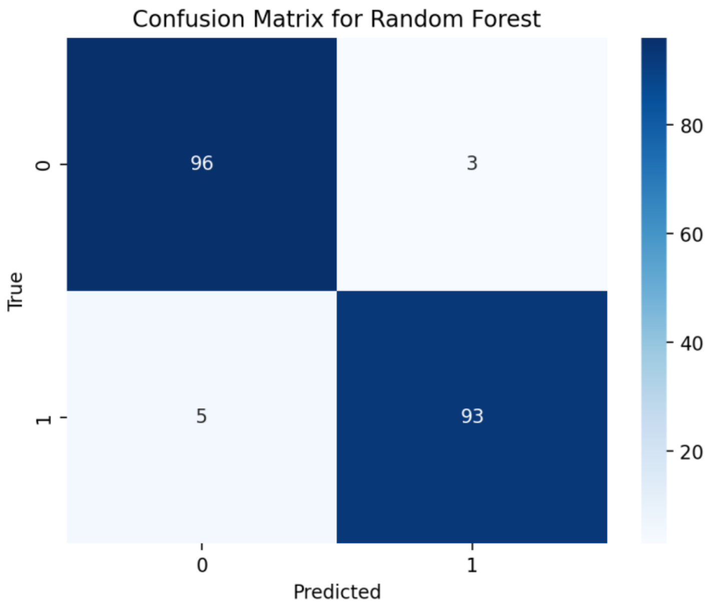


Fig. 4.7 Confusion Matrix for Random Forest

ROC curve for Random Forest

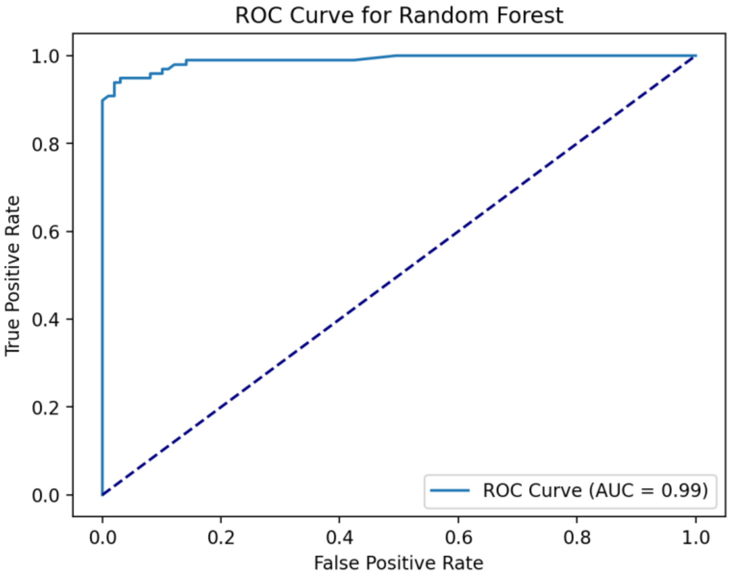
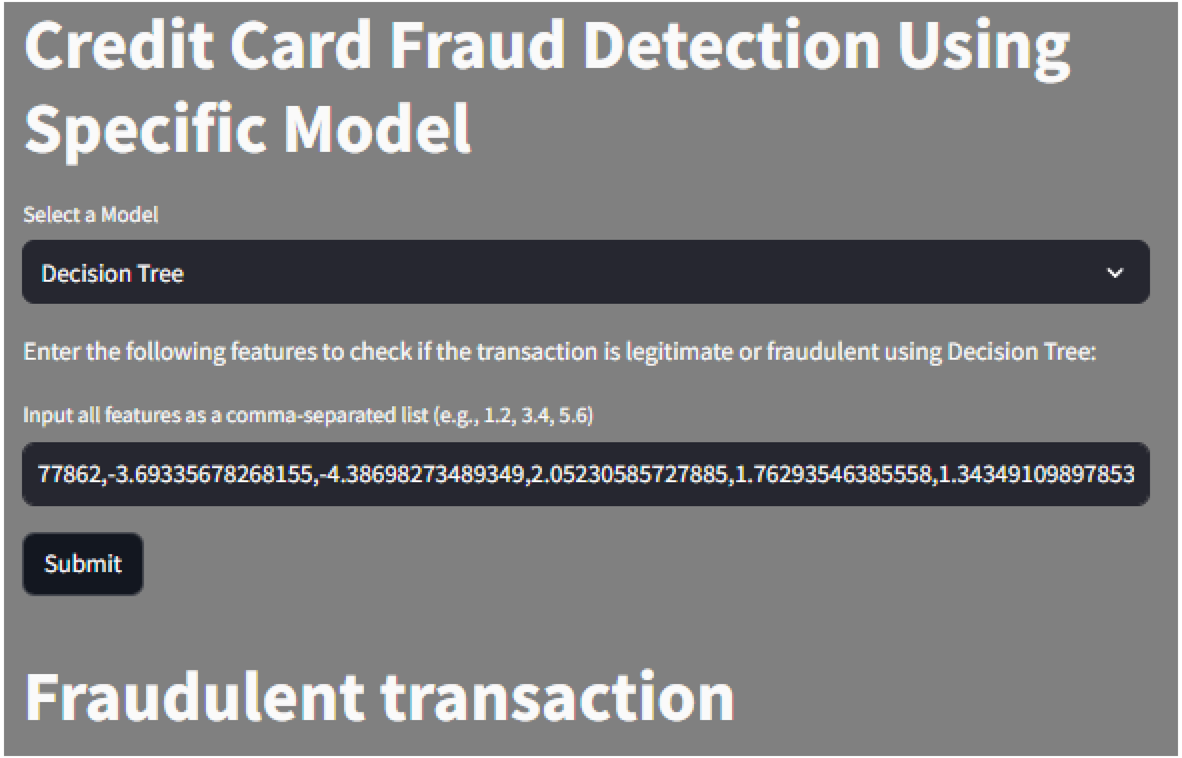


Fig. 4.10 ROC Curve for random Forest

Input Section for Each Model(3)



Classification Report and Confusion Martix for Decision Tree

	precision	recall	f1-score	support
0	0.945652	0.878788	0.910995	99.000000
1	0.885714	0.948980	0.916256	98.000000
accuracy	0.913706	0.913706	0.913706	0.913706
macro avg	0.915683	0.913884	0.913625	197.000000
weighted avg	0.915835	0.913706	0.913612	197.000000

Fig. 4.3 Classification Report for Decision Tree

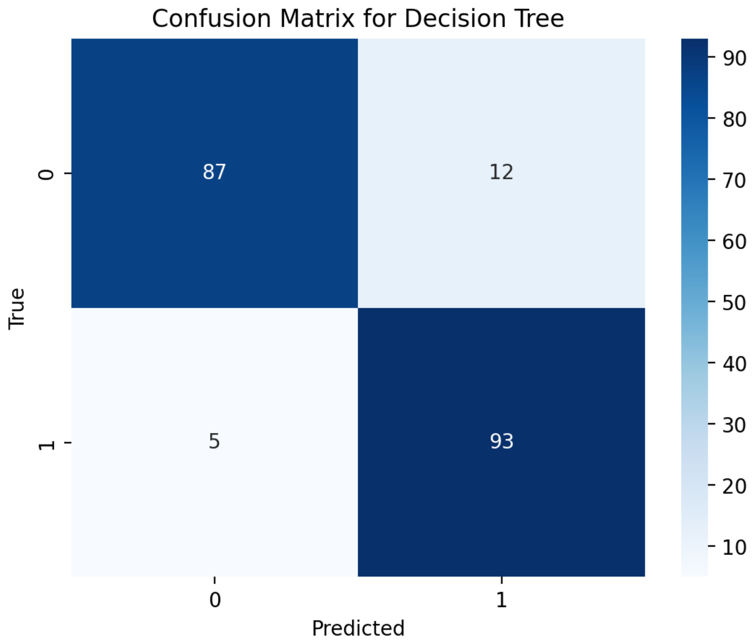


Fig. 4.8 Confusion Matrix for Decision Tree

ROC curve for Decision Tree

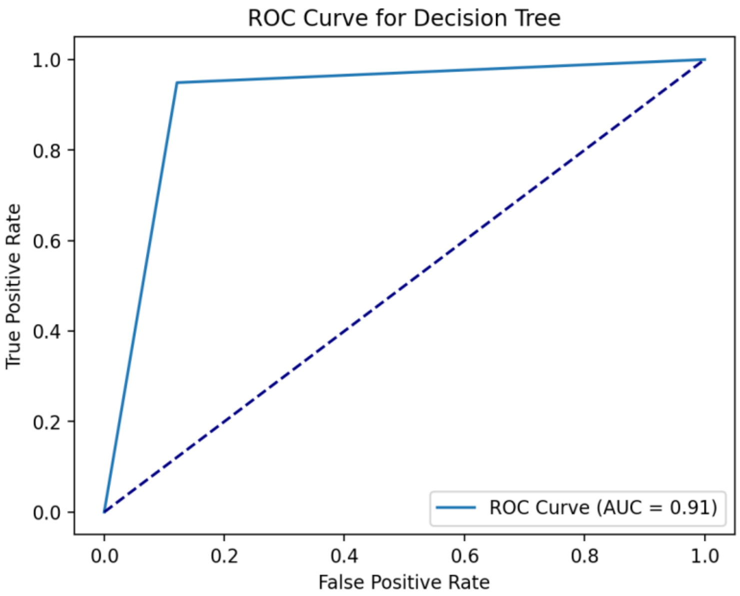


Fig. 4.11 ROC Curve for Decision Tree

Comparison of our Algorithm Based on Accuracy

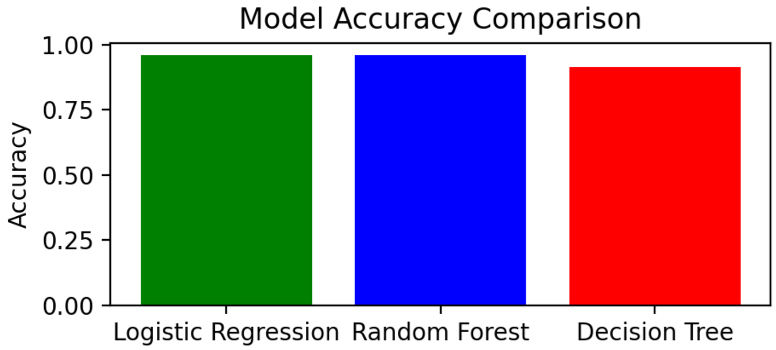
Credit Card Fraud Detection Different Algorithms

Model Comparison

Model Accuracy Scores:


```
{
  "Logistic Regression" : 0.9593908629441624
  "Random Forest" : 0.9593908629441624
  "Decision Tree" : 0.9137055837563451
}
```


Logistic Regressionmodel has the best accuracy among these models.



Problems and Future Work

Problems we Faced

- Time Consuming
- Handling Dataset
- Missing Values
- Model Overfitting and Underfitting
- Visualization Limitations:

Future Work

- Add more Algorithm
- Detect Frauds with Real Time
- Model Interpretability
- Expanding Visualization
- Applying project to Real life

Thank YOU