Credit Card Fraud Detection

Andromeda - Pranathi Alluri, Sanjana Kandunoori, Anagha Kadoo, Sinuo Song, Chen Li

Introduction

- Ongoing challenges

- Real-time detection importance

- Safeguarding financial systems

Credit Card Fraud Detection

Using the Machine Learning Classification Algorithms to detect Credit Card Fraudulent Activities

Dataset Generation & Preparation

- Stimulated Data
  - Less than 1% Fraud
  - Two types of fraud:
    - Phishing
    - Stolen accounts
- BorderLine SMOTE to deal with class imbalance
- Feature Engineering
  - Fraud rates
  - Total transactions
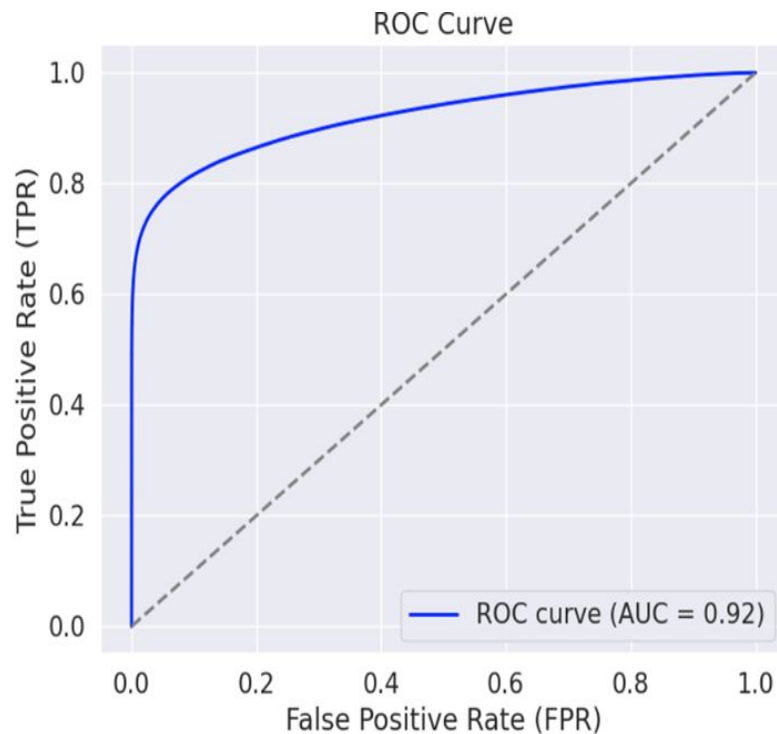 - 30 day windows
- Normalized using min-max scalar
- 80/20 split
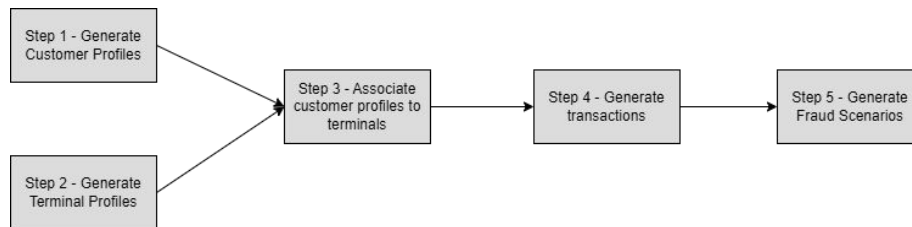
# Models

Logistic Regression

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.91	0.87	383551
1	0.90	0.81	0.85	384292
accuracy			0.86	767843
macro avg	0.86	0.86	0.86	767843
weighted avg	0.86	0.86	0.86	767843

ROC Curve

SVM (using Linear Kernel)

```
Classification Report:
              precision    recall   f1-score   support

          0       0.82       0.92      0.87      383551
          1       0.91       0.80      0.85      384292

   accuracy                            0.86      767843
  macro avg       0.87       0.86      0.86      767843
weighted avg      0.87       0.86      0.86      767843
```


Receiver Operating Characteristic

ROC curve (AUC = 0.92)
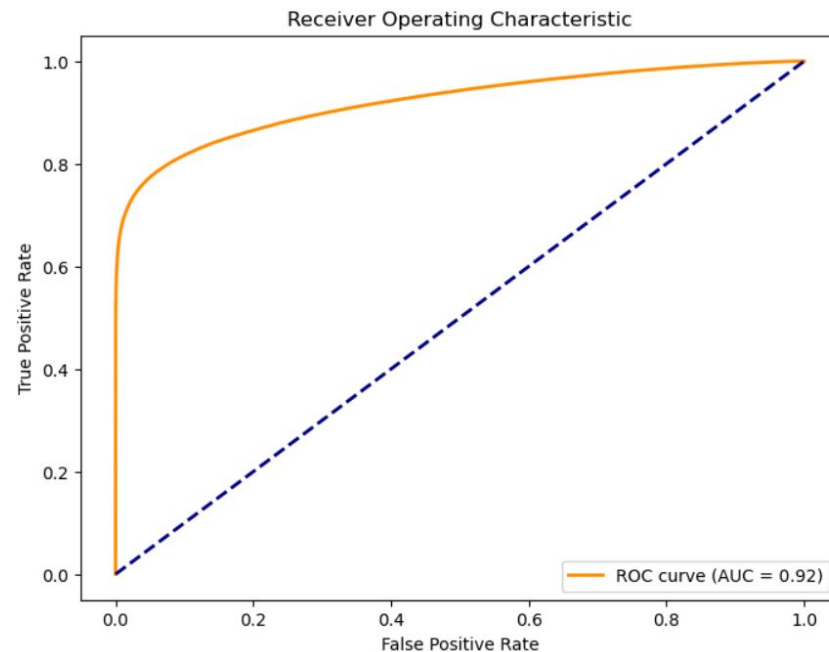
Decision Trees

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97    383551
           1       0.97      0.97      0.97    384292

    accuracy                           0.97    767843
   macro avg       0.97      0.97      0.97    767843
weighted avg       0.97      0.97      0.97    767843
```
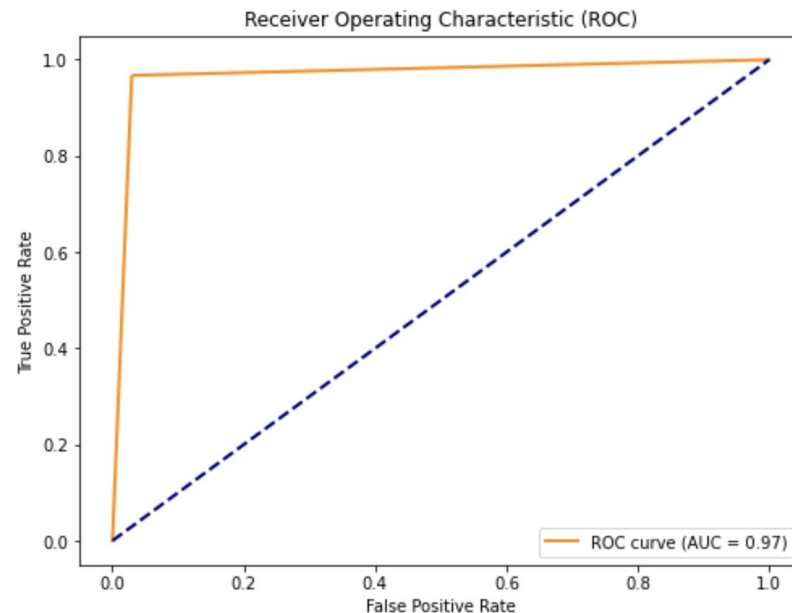
```
Precision: 0.9705350138591078
Recall: 0.9658072507364192
F1-score: 0.9681653606569349
ROC-AUC: 0.96896281828251
```


Receiver Operating Characteristic (ROC) — ROC curve (AUC = 0.97)

Random Forest

```
Classification Report:
              precision    recall   f1-score    support

           0       0.97      0.99       0.98     383551
           1       0.99      0.97       0.98     384292

    accuracy                            0.98     767843
   macro avg       0.98      0.98       0.98     767843
weighted avg       0.98      0.98       0.98     767843

Precision: 0.9927671762863846
Recall: 0.9668663412196975
F1 Score: 0.9796455907129052
ROC-AUC Score: 0.9980771196410041
```
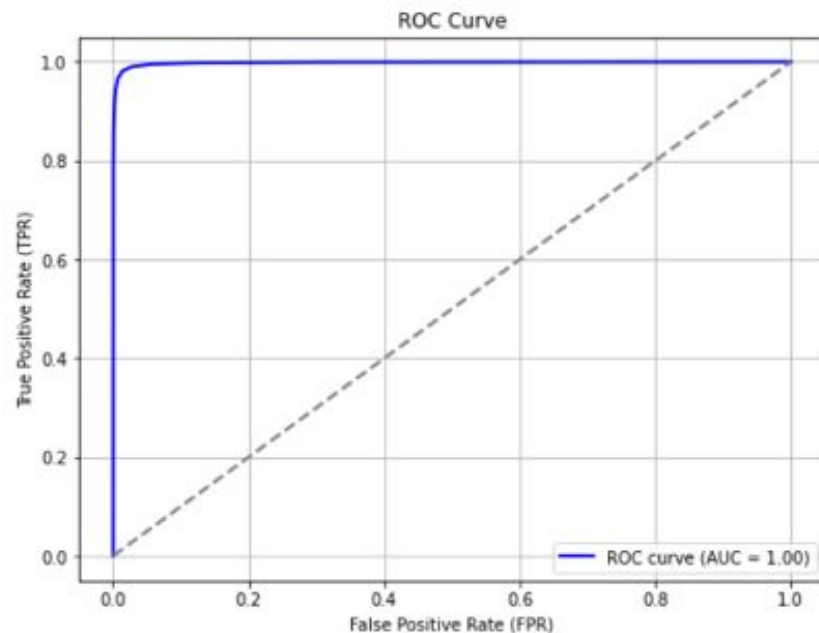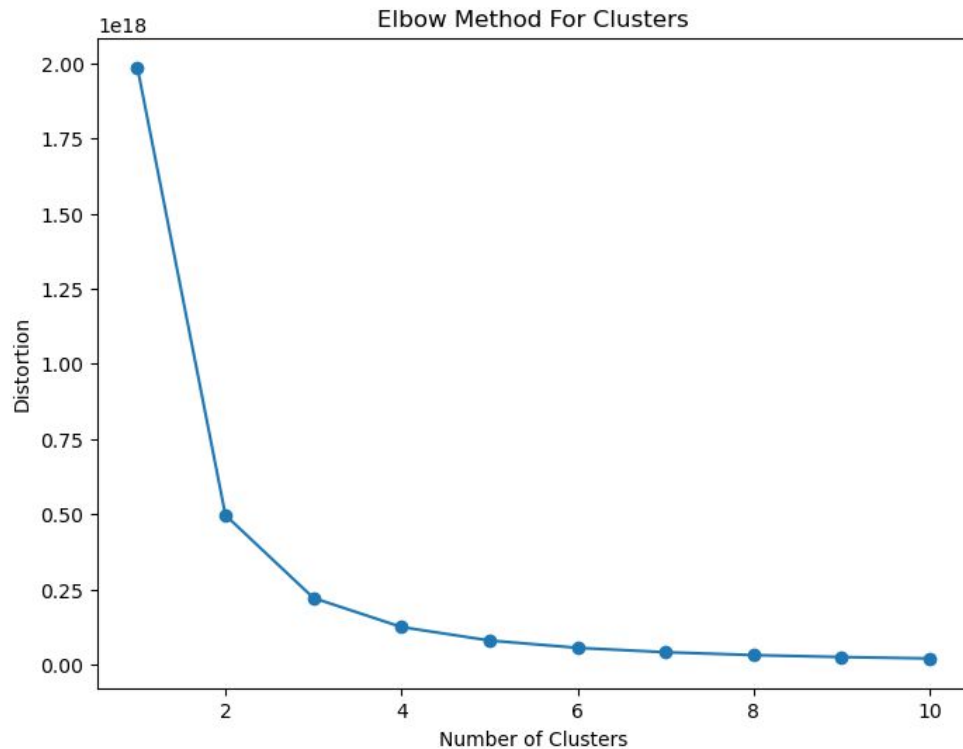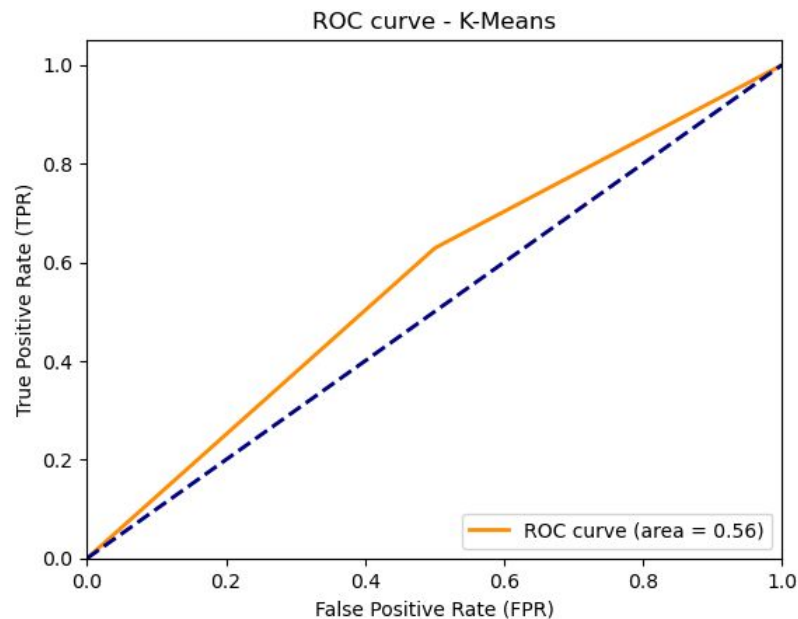
K-Means

Elbow Method For Clusters

Classification Report :

	precision	recall	f1-score	support
0	0.99	0.50	0.67	383551
1	0.01	0.63	0.02	384292
accuracy			0.50	767843
macro avg	0.50	0.56	0.34	767843
weighted avg	0.99	0.50	0.66	767843

Accuracy : 0.5002667259390976
F-1 score : 0.01672395847000271
Recall score : 0.6289473684210526
Davies Bouldin Score : 0.5000290444611917

ROC curve - K-Means

Results - Comparing different models

Model	Accuracy (in percent)
Logistic Regression	86
SVM	86
Decision Tree	97
Random Forest	98
K-Means Clustering	50

- Random Forest performs the best amongst all models

- Decision Tree also has a comparable accuracy to Random Forest

- K-Means Clustering has the least accuracy amongst all models

- Ensemble methods of Decision Tree and Random Forest prove to be the most accurate

- They capture patterns within the dataset most accurately to produce predictions

Conclusion

- Overview of the project
- Random Forest works the best
    - Use multiple decision trees to capture complex relationship
    - Results show high accuracy and ROC scores
- K - Means works the worst
    - Clustering data points into different clusters based on minimum distance between each point and the centroid of the cluster
    - Results show low accuracy and ROC scores

Future Work

- Extending the scope beyond the existing radius.

- Incorporate more sophisticated techniques and encompassing a broader spectrum of fraud types.

- Develop and optimize real-time fraud detection systems.

Thank you!