# Credit Card Fraud Detection

—

By Max Kim

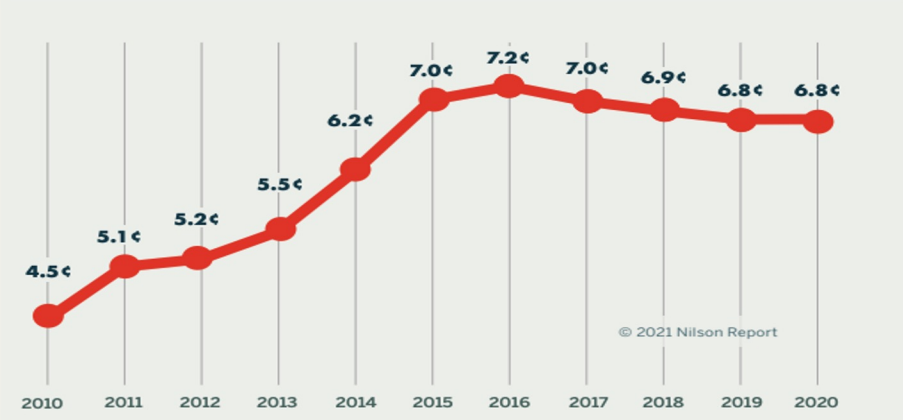
# Agenda

1. Introduction
- Business problem and data questions
1. Data Processes
- EDA, Feature Engineering
1. Value Delivery
- Draw value out of model, accuracy/recall
1. Conclusion
- Summary, next steps and improvement

# 1. Introduction

Card Fraud Worldwide

- Lost $28.58 billion to card fraud in 2020, equal to 6.8cents per $100 in purchase volume




## Card Fraud Projected through 2030

| YEAR | Total Volume (TRIL.) | Fraud (BIL.) | Cents per $100 VOLUME |
|---|---|---|---|
| 2020 | $41.962 | $28.58 | 6.81 |
| 2021 | $47.229 | $32.20 | 6.82 |
| 2022 | $50.868 | $34.36 | 6.75 |
| 2023 | $54.061 | $36.13 | 6.68 |
| 2024 | $57.323 | $38.07 | 6.64 |
| 2025 | $60.583 | $39.89 | 6.58 |
| 2026 | $64.038 | $41.73 | 6.52 |
| 2027 | $67.570 | $43.76 | 6.48 |
| 2028 | $71.221 | $45.54 | 6.39 |
| 2029 | $75.111 | $47.50 | 6.32 |
| 2030 | $79.140 | $49.32 | 6.23 |

# Questions

**Business Question:**

“How to **detect** as many **fraud transactions** as possible to reduce the time and cost invested in the transactions?”

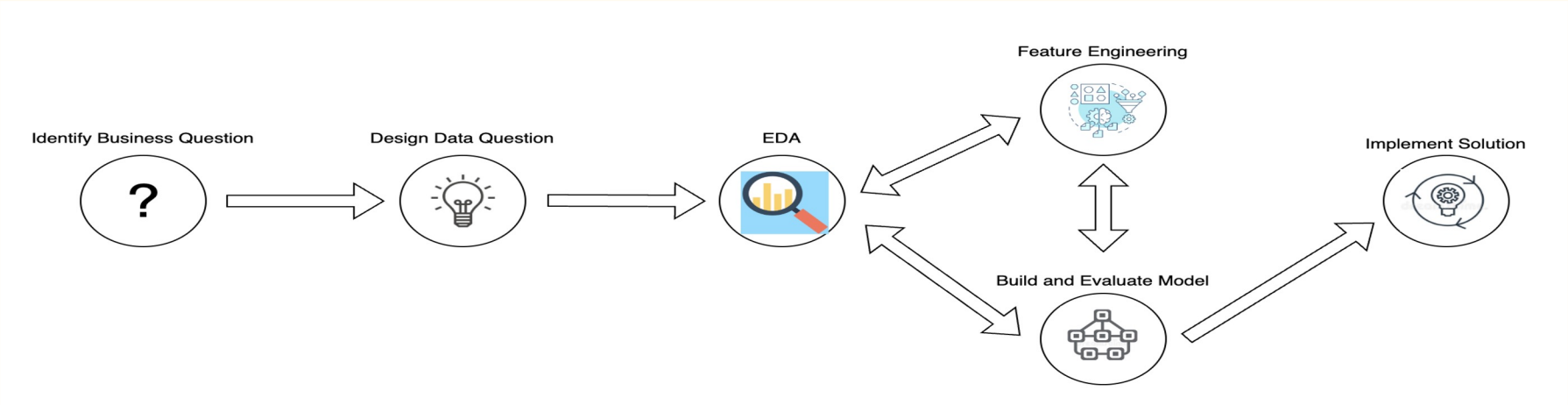
**Data Question:**

“What model can be used to **predict fraud transaction** and what is the accuracy rate of it?

# 2. Data Process

- Domain: Credit card company
- Stakeholder: Director of credit card company & Head of Fraud team

Process Workflow:

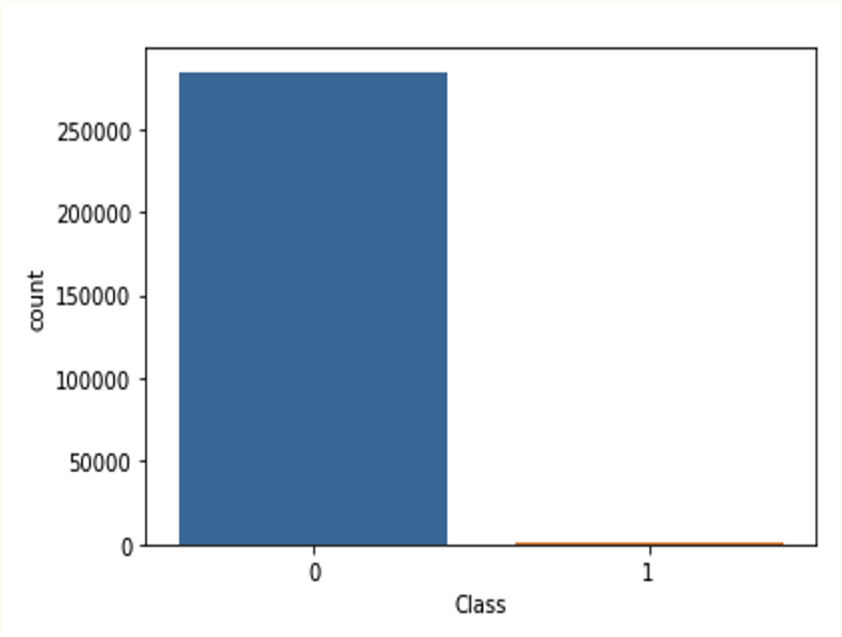


# Target Variable - Class

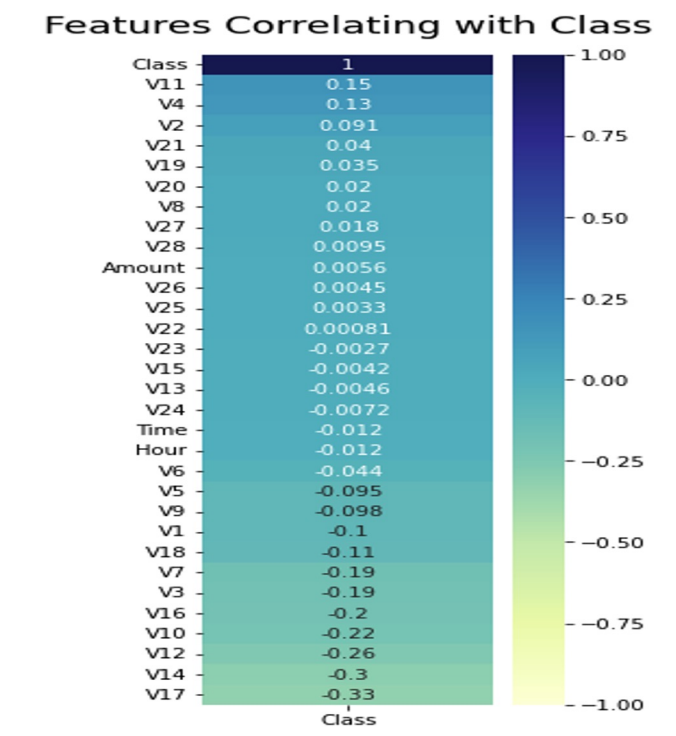
Summary:

- Total 284,807 transactions
- **Class 1**: 492 fraud transactions (0.17%)
- **Class 0:** 284,315 non-fraud transactions (99.83%)

Number of card transactions by Class 1 & 0



# Exploration




- Dataset has features V1 - V28
- This features are numerical input variables from PCA transformation
- Due to confidentiality issues, the original features and background information are not revealed

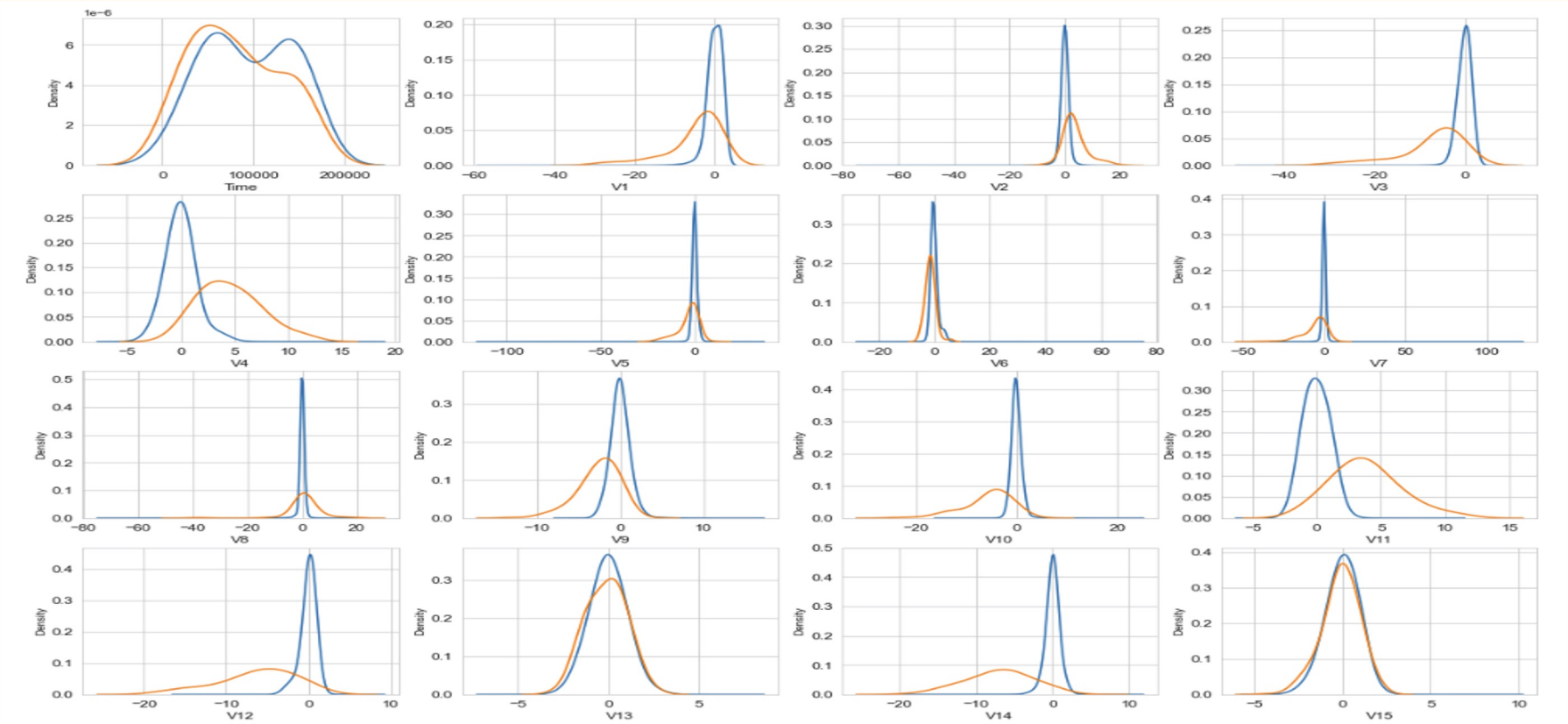
# Exploration

- Amount

| | non-fraud | fraud |
|---|---|---|
| **count** | 284315.00 | 492.00 |
| **mean** | 88.29 | 122.21 |
| **std** | 250.11 | 256.68 |
| **min** | 0.00 | 0.00 |
| **25%** | 5.65 | 1.00 |
| **50%** | 22.00 | 9.25 |
| **75%** | 77.05 | 105.89 |
| **max** | 25691.16 | 2125.87 |

- For each transaction, average amount transacted for non-fraud is $88.29 vs $122.21 for fraud transactions
- There is reasonable difference in transaction amount between non-fraud and fraud transactions

# Feature Density Plot



# 3. Value Delivery

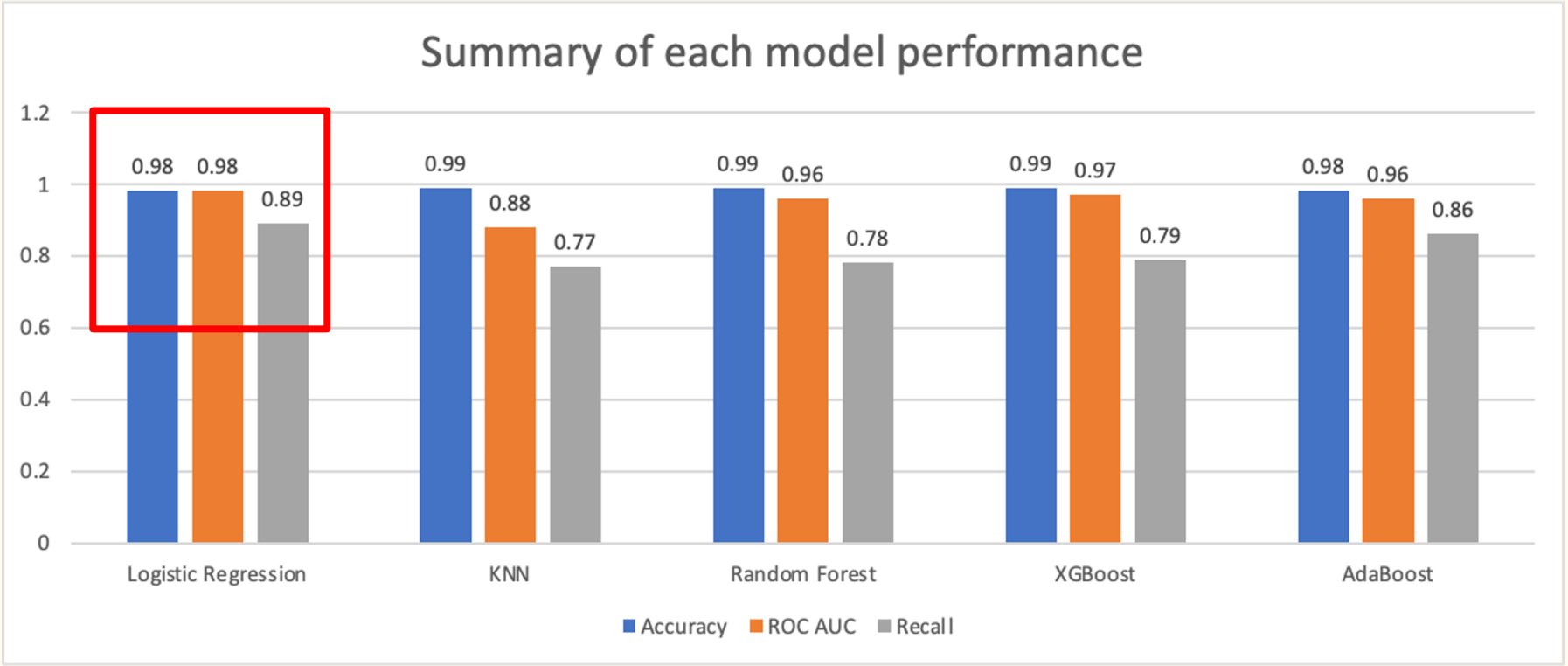
- Which metrics to be considered?
  - This dataset is imbalanced datasets. Thus, just high accuracy will not make a good model. There can be room for improvement by considering other metrics to improve the model performance.
  - It is critical not to predict actual fraud transaction as non-fraudulent transaction.
  - Reducing false negative as much as possible is critical:
    - For our **main metric**, we choose **Recall rate**

# 3. Value Delivery

| Model | Accuracy | Recall |
|---|---|---|
| **Logistic Regression** | **0.98** | **0.89** |
| KNN | 0.99 | 0.77 |
| Random Forest | 0.99 | 0.78 |
| XGBoost | 0.99 | 0.79 |
| AdaBoost | 0.98 | 0.86 |

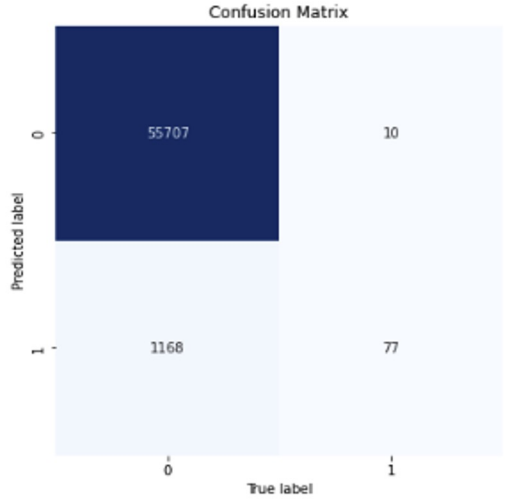
Binary Classification models that define 2 classes of segments using 31 columns

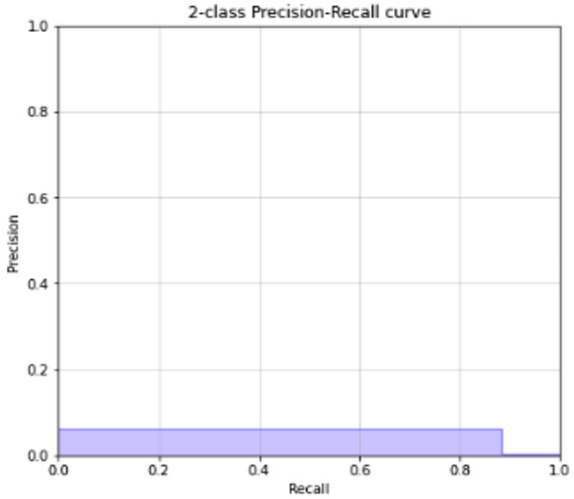
# Model Selection - Logistic Regression

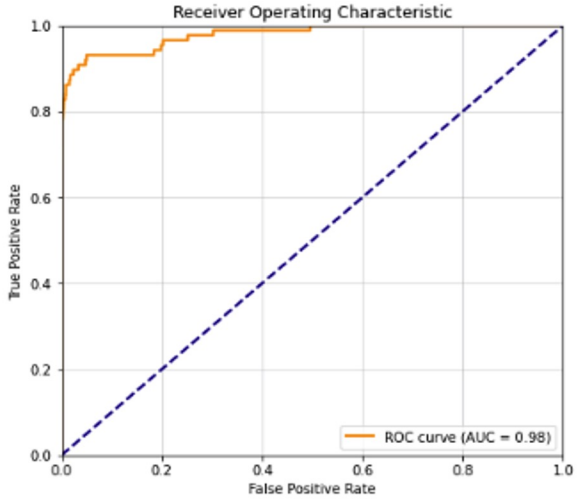


# Logistic Regression Performance

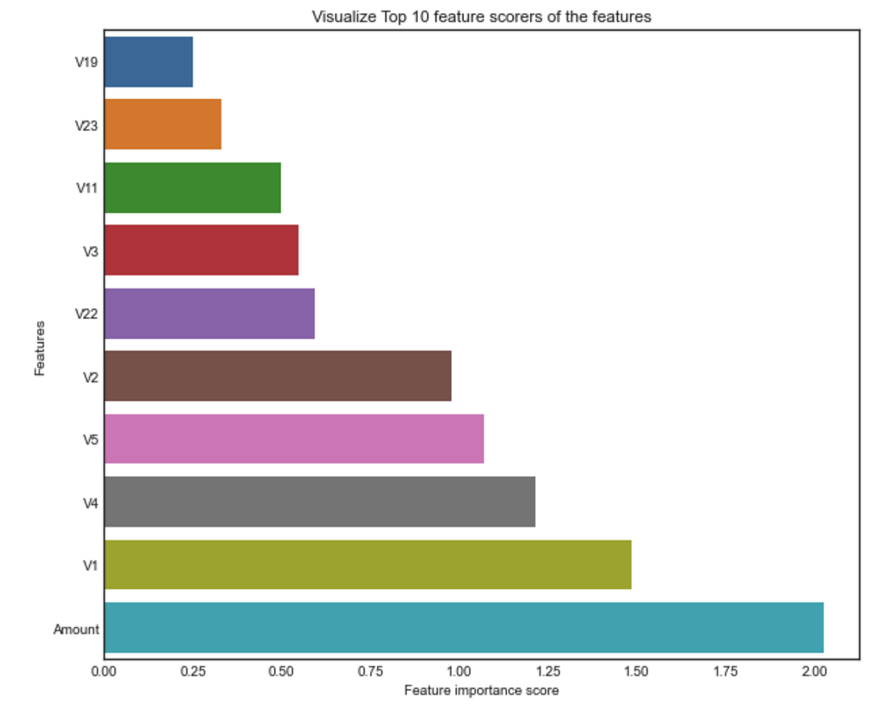
```
Accuracy : 0.9793 [TP / N] Proportion of predicted labels that match the true labels. Best: 1, Worst: 0
Precision: 0.0618 [TP / (TP + FP)] Not to label a negative sample as positive.       Best: 1, Worst: 0
Recall   : 0.8851 [TP / (TP + FN)] Find all the positive samples.                    Best: 1, Worst: 0
ROC AUC  : 0.9785                                                                    Best: 1, Worst: < 0.5
----------------------------------------------------------------------------------------------------
TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, N: Number of samples
```



# Feature Importance




Top 5 Important Features

1. Amount
2. V1
3. V4
4. V5
5. V2

# 4. Conclusion

1. Credit card company can focus on the amount of each transactions among with other key important features to detect fraud transaction more effectively
2. With large number of transaction dataset, company can deploy a machine learning model that has high focus on high recall will distinguish between non-fraud and fraud transactions
3. This model can be further improved from various aspects in the process workflow
    - Data Process:
        - Feature Engineering
    - Apply Deep Learning:
        - Conventional layer
        - Artificial Neural Network
    - Machine Learning:
        - Cat-Boost model

# Reference


- https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?datasetId=310&sortBy=voteCount
- https://nilsonreport.com/content_promo.php?id_promo=16
- https://dataspire.org/blog/what-is-exploratory-data-analysis-anyway
- https://www.istockphoto.com/photos/light-bulbs
- https://www.dreamstime.com/illustration/modelling-icon.html
- https://www.vectorstock.com/royalty-free-vector/feature-engineering-turquoise-concept-icon-vector-42324394