MBA USP ESALQ

# Maximizing Efficiency and Accuracy in Credit Analysis through Machine Learning and Deep Learning Techniques

*Mateus Raimundo da Cruz*
*Fabiano Guasti Lima*

# INTRODUCTION

## Credit Analysis and Its Significance

- Credit analysis is a crucial part of the lending process for businesses.
- Assessing the creditworthiness of borrowers is essential for risk management.
- Effective credit analysis safeguards financial stability and cash flows.

### *Traditional Credit Risk Assessment*

- Historical financial data, credit scores, income, and employment history are traditionally used.
- Challenges arise with the increasing number of loan applicants.

### *The Rise of Machine Learning*

- Machine learning (ML) is transforming credit analysis.
- ML models can process vast data quickly and predict credit-related risks.
- Detect patterns and trends that humans may miss.

# INTRODUCTION

## ML Enhances Credit Analysis

- ML provides accurate systems for credit risk analysis.
- Captures non-linear dynamics in financial data.
- Optimizes both micro and macro supervision of credit risk.
- Customized Approach

## ML Enhances Credit Analysis

ML techniques offer a more personalized approach to financial institution supervision. Supports better decision-making and informed regulatory policies. While ML is powerful, human expertise remains essential.

Combining ML and human judgment optimizes credit analysis. Credit analysis continues to evolve with technology. Combining ML with human expertise for competitive lending decisions.

# LITERATURE REVIEW ABOUT FRAUD DETECTION

### Decision Tree Algorythm

Regression-based decision trees have been explored for fraud detection, promising results but with limitations in relation to classes.

### Random Forest Ensemble

Random Forest has been extensively studied for fraud detection with reported accuracies ranging from 90% to 99%

### We have a lot of experience in various projects

Different machine learning techniques offer unique strengths and may complement each other. The choice of model depends on the specific fraud detection task and dataset characteristics.

### Naïve Bayes and K-NN Algorithm

Naïve Bayes is a probabilistic algorithm and K-NN makes predictions based on the similarity. Studies show Naïve Bayes outperformed K-NN in accuracy

### Logistic Regression and LightGBM

Logistic Regression and LightGBM has also used to credit fraud detection and has shown promise with superior accuracy in comparison to other techniques like Random Forest.

### Dummy Classifier Algorithm

Sophisticated models consistently outperform the Dummy Classifier.

# METHODOLOGY

Such progress is seldom abrupt, but rather a steady, progressive, and intricate process that occurs over a relatively lengthy period of time

## Data and Feature Extract Preparation

Gather and prepare the dataset and identify the most predictive features.

Use statistical methods to select relevant features with high correlation to credit risk.

## Model Developing and Training

To help the client improve their IT infrastructure and ease them with the best Service

Compare model predictions with ground truth credit risk to measure performance.

## Model Testing and Selection

Select a suitable model for credit risk analysis based on training and evaluation results

Choose the final model based on performance metrics specific to the dataset.

# APPLICATION OF MACHINE LEARNING FOR FRAUD DETECTION

### Capturing the Patterns of Fraud

Machine learning is applied in fraud detection to analyze and identify patterns in large datasets, enabling the automatic detection.

### Adapting the Model for New Data

By continuously learning from new data, machine learning models can adapt to evolving fraud tactics and help financial institutions

# DEVELOPING THE MODELS AND COMPARING THE MAINLY METRICS

### 1 Data Imbalance and Preparation

The dataset exhibits significant class imbalance for credit risk. Only 14.8% of observations represent credit risk, while the majority are non-fraud cases.

### 2 Modeling the Data for ML Training

Categorical variables are transformed into numerical ones. LabelEncoder and OneHotEncoder from Scikit-Learn are used for this purpose.

### 3 Model Performance Metrics

The F1-Score, AUC Curve and Recall metrics are used to evaluate the model's performance.

### 4 The Importance on the use of Metrics

To help the client improve their IT infrastructure and ease them with the best Service. They help in assessing the model's accuracy and its performance in fraud detection.

# DEVELOPING THE MODELS AND COMPARING THE MAINLY METRICS

### 5 First Step on Model Development.

Popular classification models in Scikit Learn are instantiated for initial training. Dense Neural Network (DNN) is developed using TensorFlow and Keras for more detailed results.

### 6 Hyper Parametric Optimization

A search for the best parameters for each model was applied in order to find the best final model version.

### 7 Choosing the best model

Scikit-learn offers a versatile set of tools for machine learning model development and evaluation

### 8 Deep Neural Network Approach

A Dense Neural Network model for fraud detection was developed and trained.

# METRICS:

## 1 F1-Score

F1-Score is a metric that combines both precision and recall into a single value to provide a balance between accurate positive predictions and capturing all positive instances.

## 2 Recall

Recall, also known as true positive rate or sensitivity, measures the proportion of actual positive instances that were correctly identified by a classification model.

## 3 AUC Courve

AUC Curve (Area Under the Curve) is a metric used in binary classification to measure the overall performance of a model by calculating the area under the Receiver Operating Characteristic (ROC) curve.

## 4 Confusion Matrix

A confusion matrix is a table used in machine learning and statistics to describe the performance of a classification model, displaying the counts of true positives, true negatives, false positives, and false negatives, allowing for a detailed evaluation of model performance and error types.

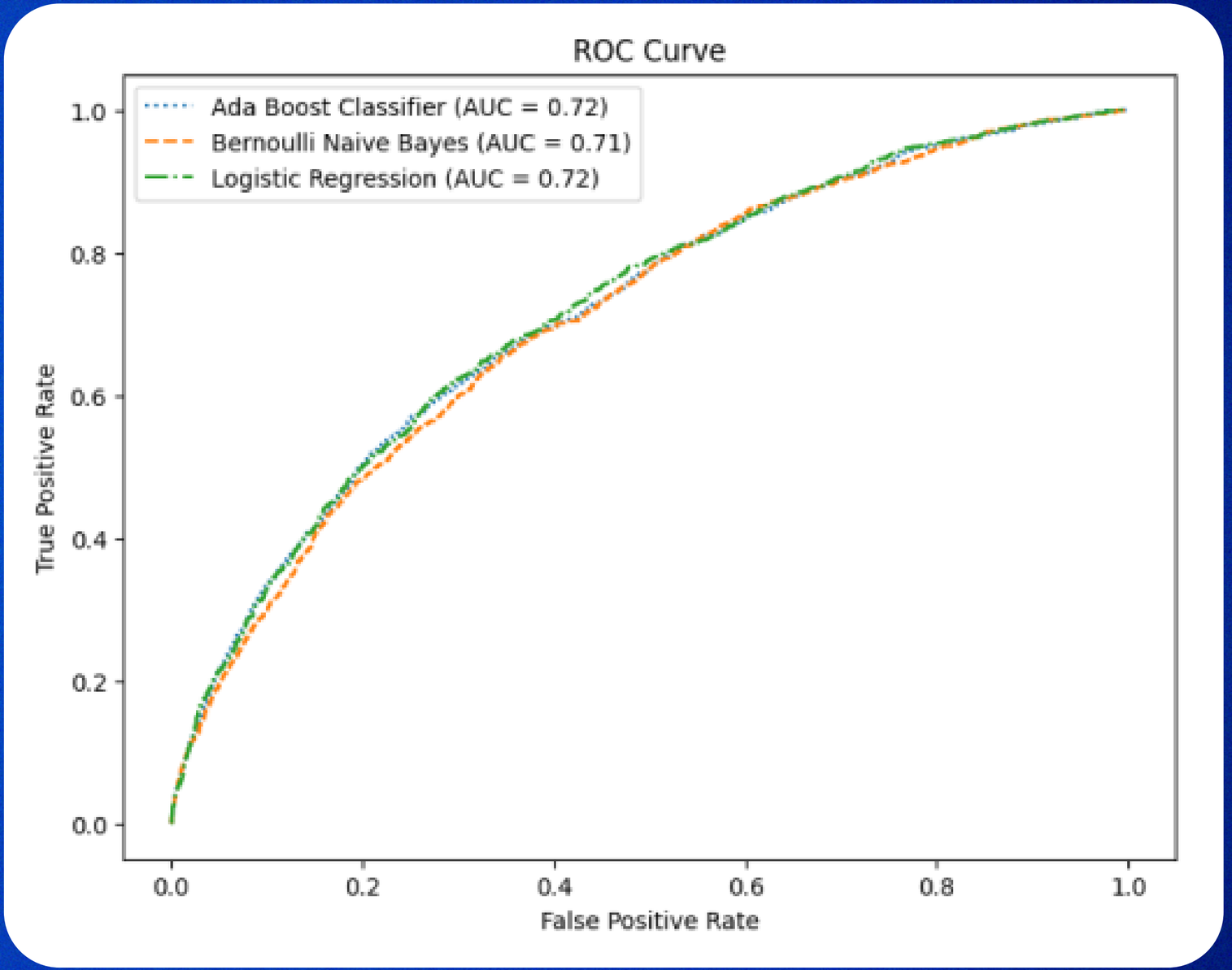
# RESULTS

The results of all the models were captured and the average between them calculated. This average is considered the threshold for rejecting models that perform below average. The three best models were chosen based on the best result for each metric. The red cells show results below the global average, while the green cells show the best results.

| Model | AUC Curve | Recall | F1-Score | Average |
|---|---|---|---|---|
| Ada Boost Classifier | 71.11 | 65.19 | 65.15 | 67.15 |
| Bagging Classifier | 63.36 | 52.69 | 56.95 | 57.67 |
| Gradient Boosting | 70.95 | 65.48 | 64.87 | 67.10 |
| Hist Gradient Boosting | 69.23 | 64.43 | 64.16 | 65.94 |
| Random Forest | 69.38 | 63.66 | 63.81 | 65.62 |
| Bernoulli NB | 69.80 | 66.67 | 65.01 | 67.16 |
| Gaussian NB | 65.02 | 33.78 | 44.01 | 47.60 |
| Calibrated Classifier CV | 70.98 | 66.09 | 65.73 | 67.60 |
| Decision Tree | 56.18 | 56.00 | 56.08 | 56.09 |
| Extra Tree | 54.68 | 55.17 | 54.90 | 54.92 |
| SVC | 69.69 | 64.19 | 64.10 | 65.99 |
| Nus SVC | 64.79 | 60.07 | 60.56 | 61.81 |
| Linear SVC | 71.04 | 66.05 | 65.72 | 67.60 |
| Logistic Regression | 71.04 | 66.14 | 65.66 | 67.61 |
| Logistic Regression CV | 71.27 | 66.16 | 65.81 | 67.75 |
| Passive Aggressive | 56.54 | 56.23 | 54.51 | 55.76 |
| Perceptron | 59.27 | 54.49 | 54.79 | 56.18 |
| Ridge Classifier CV | 71.10 | 66.04 | 65.67 | 67.60 |
| SGD Classifier | 66.64 | 63.35 | 62.55 | 64.18 |
| K Neighbors | 58.18 | 54.65 | 55.43 | 56.09 |
| Linear Discriminant | 71.00 | 66.12 | 65.79 | 67.64 |
| Quadratic Discriminant | 59.63 | 34.08 | 51.27 | 48.33 |
| Gaussian Process | 60.24 | 56.35 | 57.31 | 57.97 |
| Multi-Layer Perceptron | 63.12 | 57.17 | 58.98 | 59.76 |
| **Average (Threshold)** | **65.59** | **62.39** | **60.37** | **61.71** |

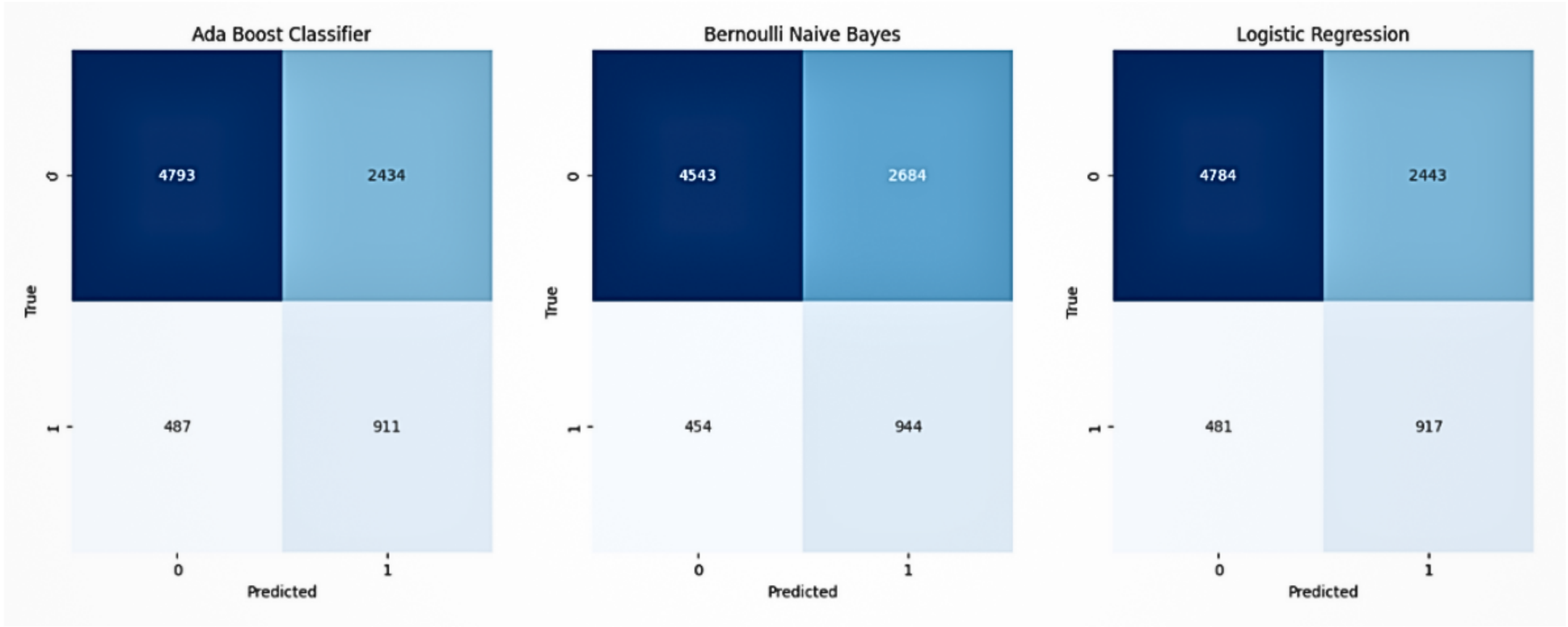
# RESULTS

The ROC curve visually shows the relationship between the rate of true positives and false positives presented by the model. Therefore, the greater the area over the curve, the greater the model's ability to distinguish. This in turn helps to detect fraud in systems.



# CONFUSION MATRIX

Best 3 models developed.



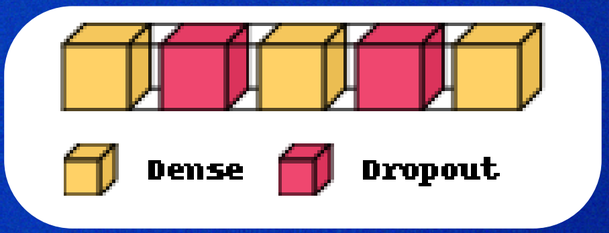
# DEEP LEARNING MODEL:

## PARAMETERS:

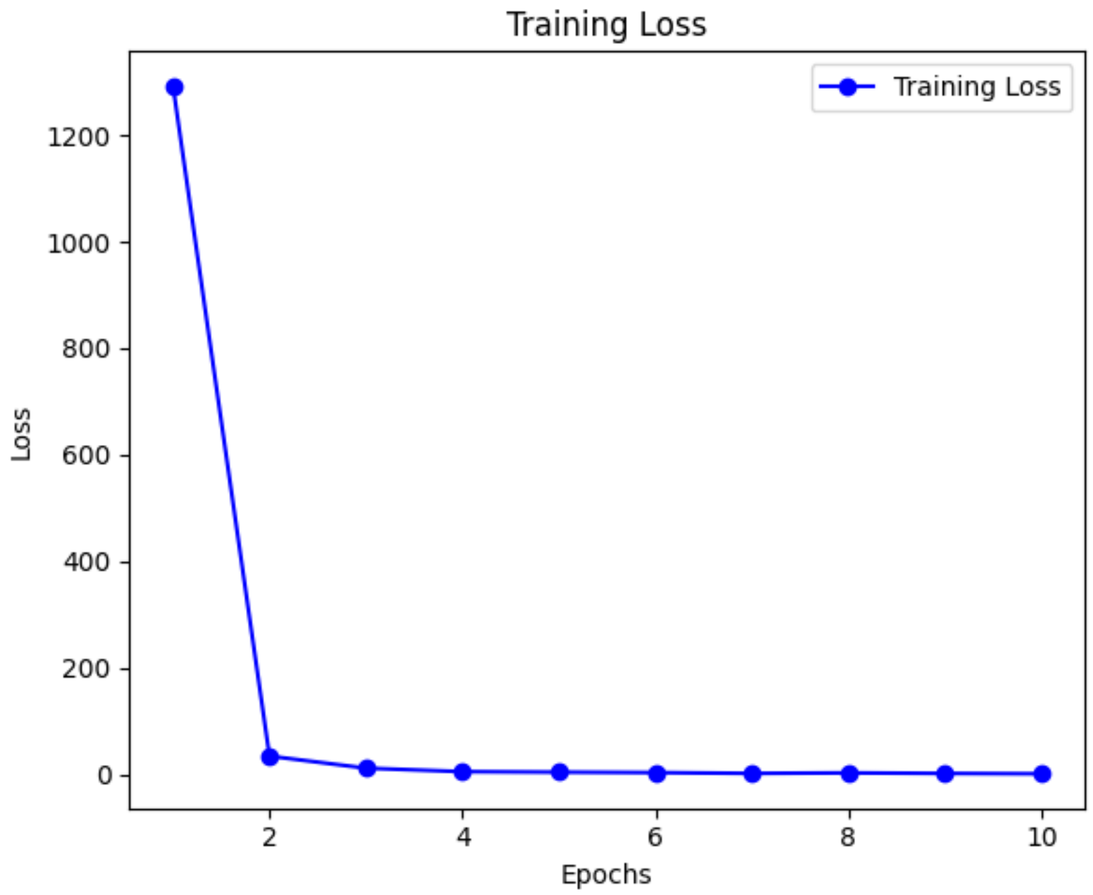
- Learning Rate = 0.001
- Optimizer = Adam
- Loss = Binary Crossentropy
- Metrics = Accuracy
- Epochs = 10

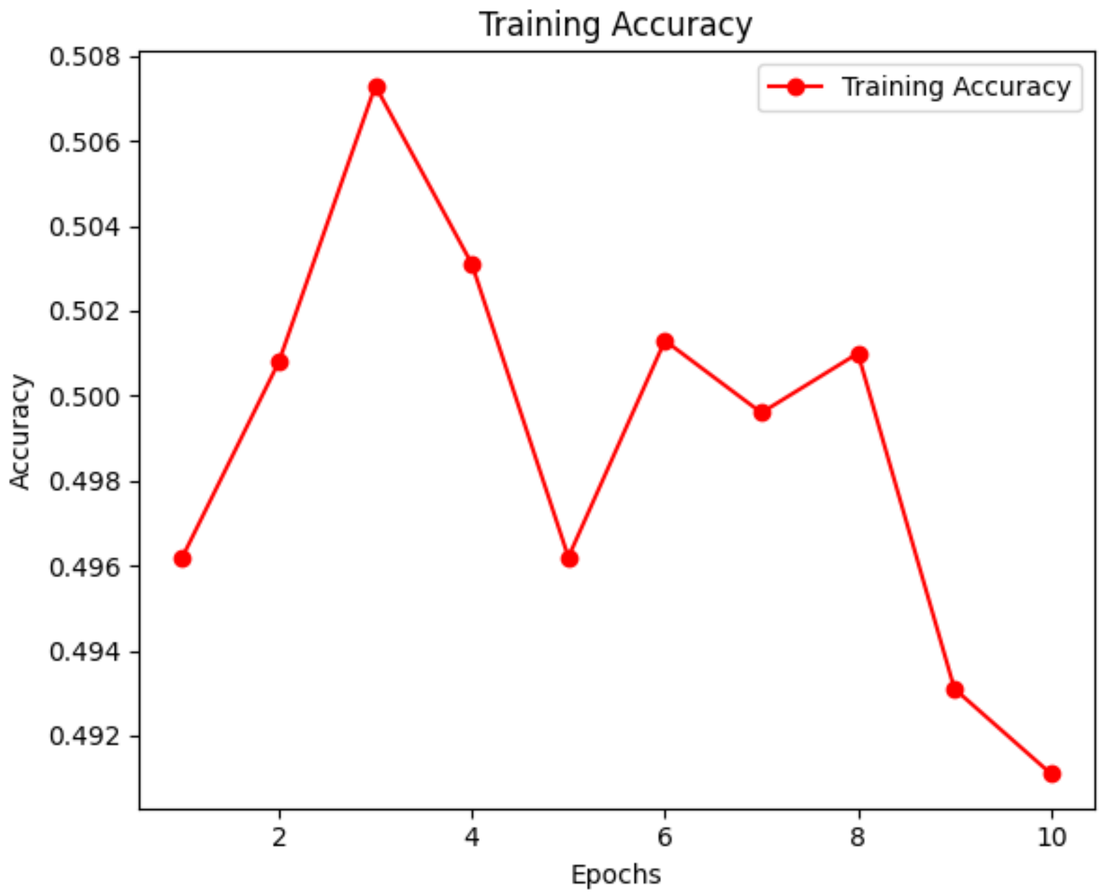
Sequential()

1. Dense(units=64, activation='relu', input_shape=(X_train.shape[1],))
2. Dropout(0.5)
3. Dense(units=32, activation='relu')
4. Dropout(0.6)
5. Dense(units=1, activation='sigmoid')



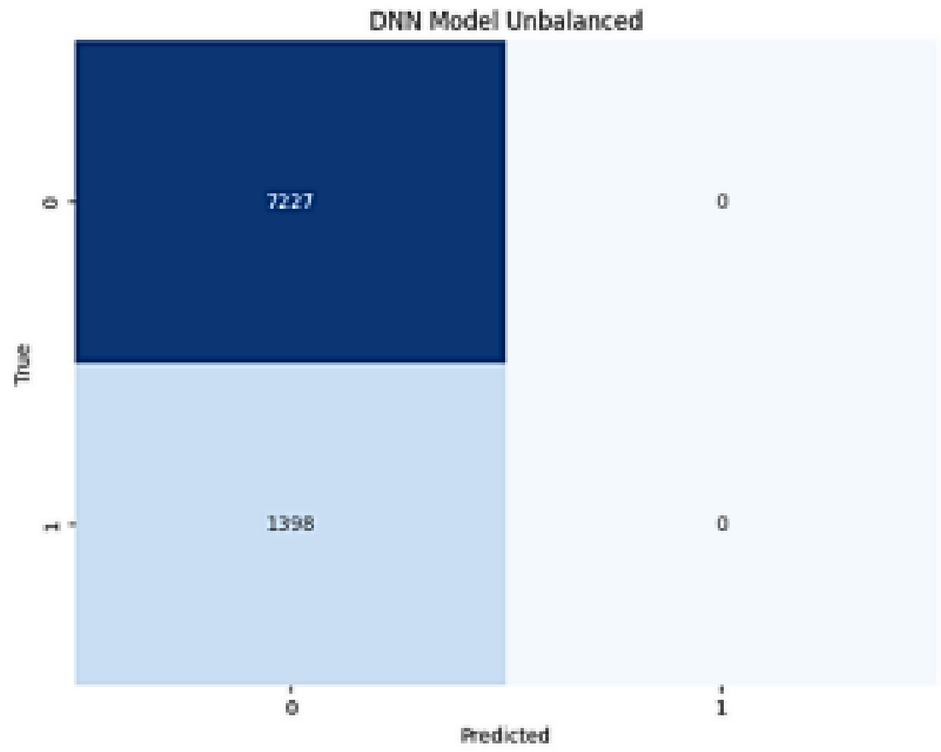
# TRAINING RESULTS:

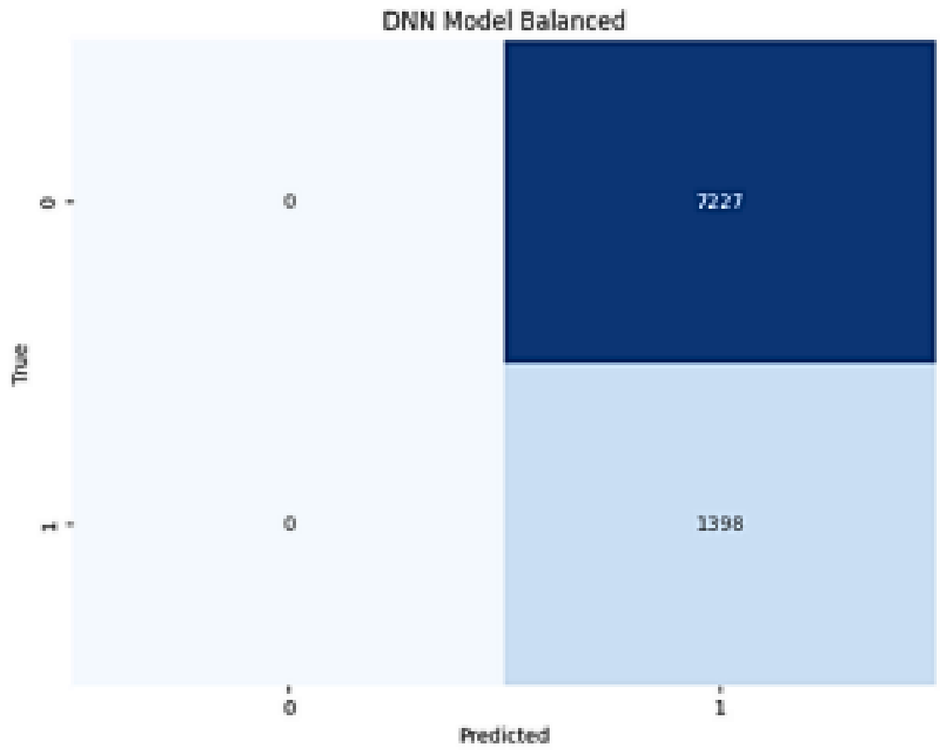


# CONFUSION MATRIX

2 DNN’s developed.



# CONCLUSIONS

1. **Complexity of Credit Risk Analysis and Fraud Detection:**
   - Credit risk analysis for fraud detection is a complex task, even with current AI techniques.
2. **Need for Greater Depth in Data and Models:**
   - The study suggests a need for more extensive data, improved models, and better techniques.
   - The final model should yield better metrics for future research to build upon.
3. **Parameter Experimentation and Data Pre-processing:**
   - Experimentation with new parameter combinations for models and further data pre-processing is essential.
   - Given the sensitivity of fraud detection, fine-tuning is crucial, especially when dealing with customers' financial data.
4. **Possibility of Future Automation:**
   - Despite the challenges, the results indicate the potential for future automation of fraud detection tasks using machine learning models.

# THANK YOU

Learning doesn't stop, it's continuous, it's lifelong.