# Optimising Credit Card Transaction EMI

Monojit Layek
Robotics & Automation, IIT Kanpur

# Contents

# Introduction

- Credit card EMIs allow customers to break down large purchases into smaller, more manageable monthly repayments, avoiding paying revolve interest.
- Banks offer EMIs at various stages - point of sale, post-transaction, and balance on EMI - and earn revenue from processing fees, interest, and foreclosure penalties.
- Despite digital campaigns, Bank A experienced low conversion rates for post-transaction EMIs and set up a call center to improve rates.
- However, the high cost of calling all customers prompted Bank A to prioritize calls to those more likely to convert.
- Our objective is to help Bank A prioritize its Credit Card transactions for EMI calling by analyzing provided datasets.

# Objective

- Help Bank A increase conversion rates of post transaction EMI on their Credit Cards.
- Reduce the cost of EMI calling by prioritizing customers who are more likely to convert.
- Use data analysis to identify key factors that influence EMI conversion.
- Develop a predictive model to identify customers who are most likely to convert to EMI and prioritize them for EMI calling.

# Exploratory Data Analysis

1. Overview of the dataset:
    1. The dataset includes 50,000 randomly sampled Credit Card transactions in the training data directory and 30,000 Credit Card transactions in the validation data directory.
    2. The target variable is binary with a value of 1 indicating whether the transaction was converted to EMI or not.
    3. The dependent variables comprise of 3 categorical features and 341 numerical features.
2. Data quality check:
    - 7.8% missing cells in 51 features, totalling 1,344,546 cells. 33 features have >20% null values, requiring imputation or dropping based on importance and impact on the model
    - 17 features have data inconsistency with only one value, possibly requiring review and removal.
    - Dataset is imbalanced with 94.8% of the target variable containing zeros.

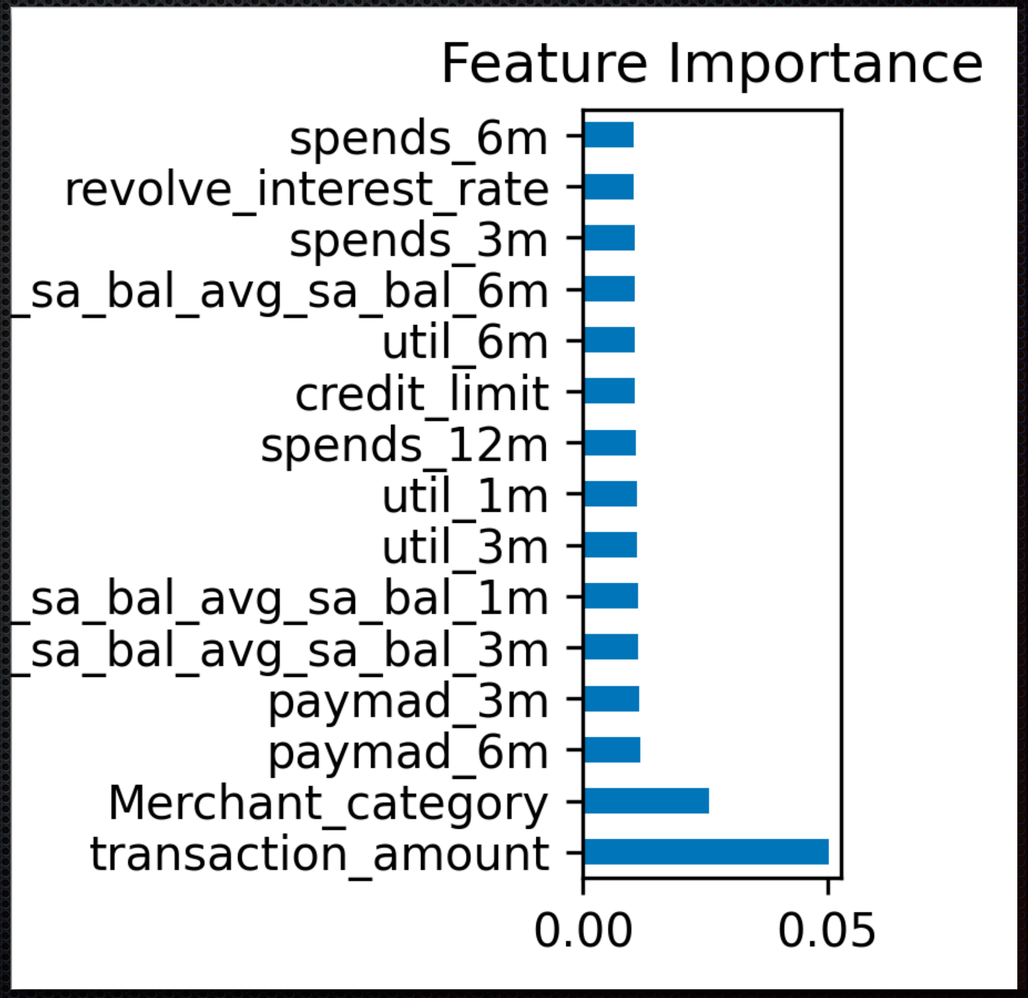
# EDA continued..

## 3. Distribution analysis:

Majority of numerical features are log normally distributed and highly right skewed according to the distribution analysis.

## 4. Correlation analysis

Correlation analysis revealed transaction history and merchant category as highly correlated features. The figure displays 15 features with the highest correlation to the target variable."



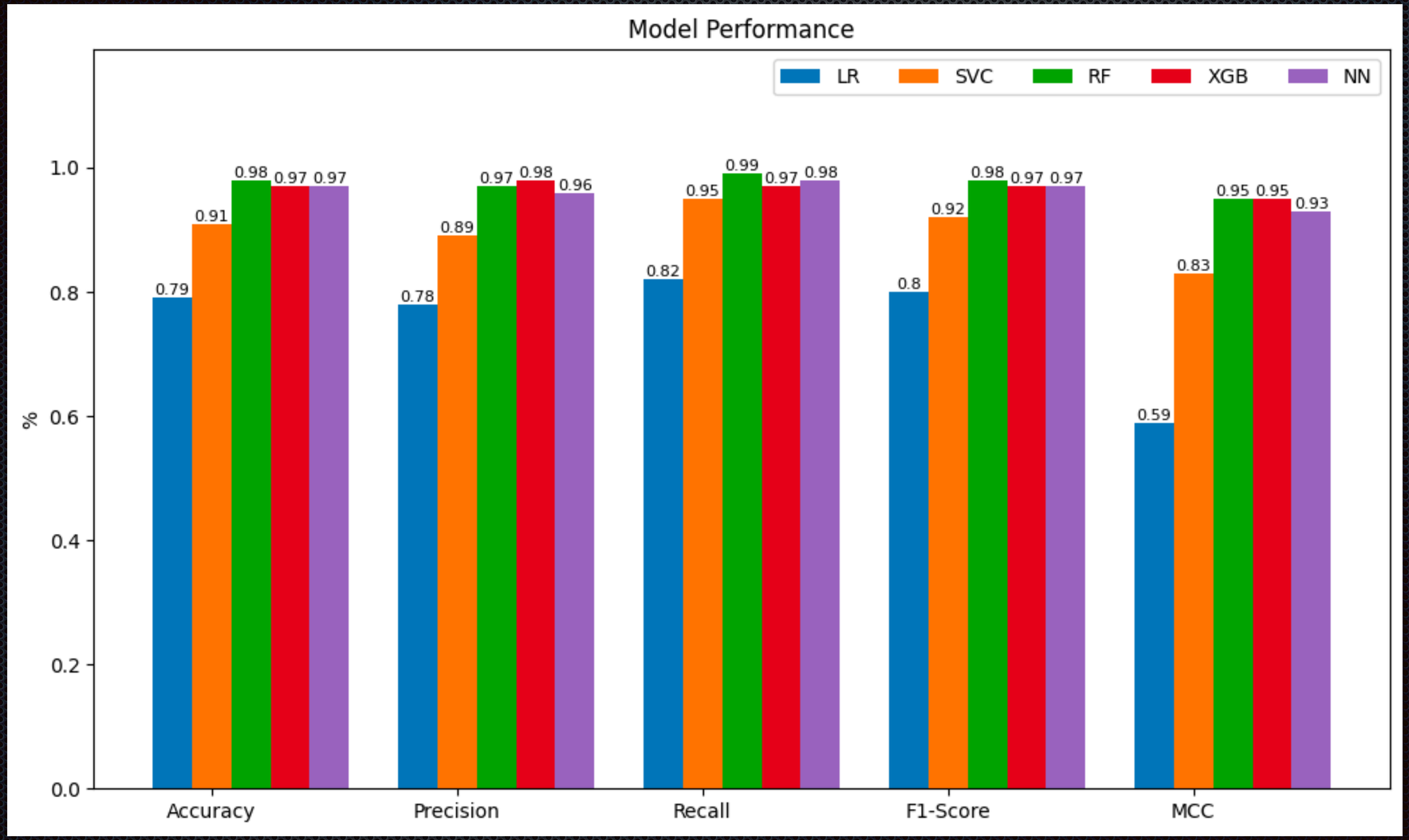
# Feature engineering

1. Potentially drop features with inconsistency of a single value and/or more than 20% null values.
2. Impute missing values using K-Nearest Neighbors (KNN) imputation.
3. Transform categorical features using label encoding.
4. Use correlation analysis and importance ranking techniques to identify the most relevant features for modelling.
5. Map data into normal distribution using techniques such as log transformation to reduce skewness in the data.
6. Scale data to a 0 to 1 range to normalize numerical features.
7. Implement Oversampling to handle class imbalance.

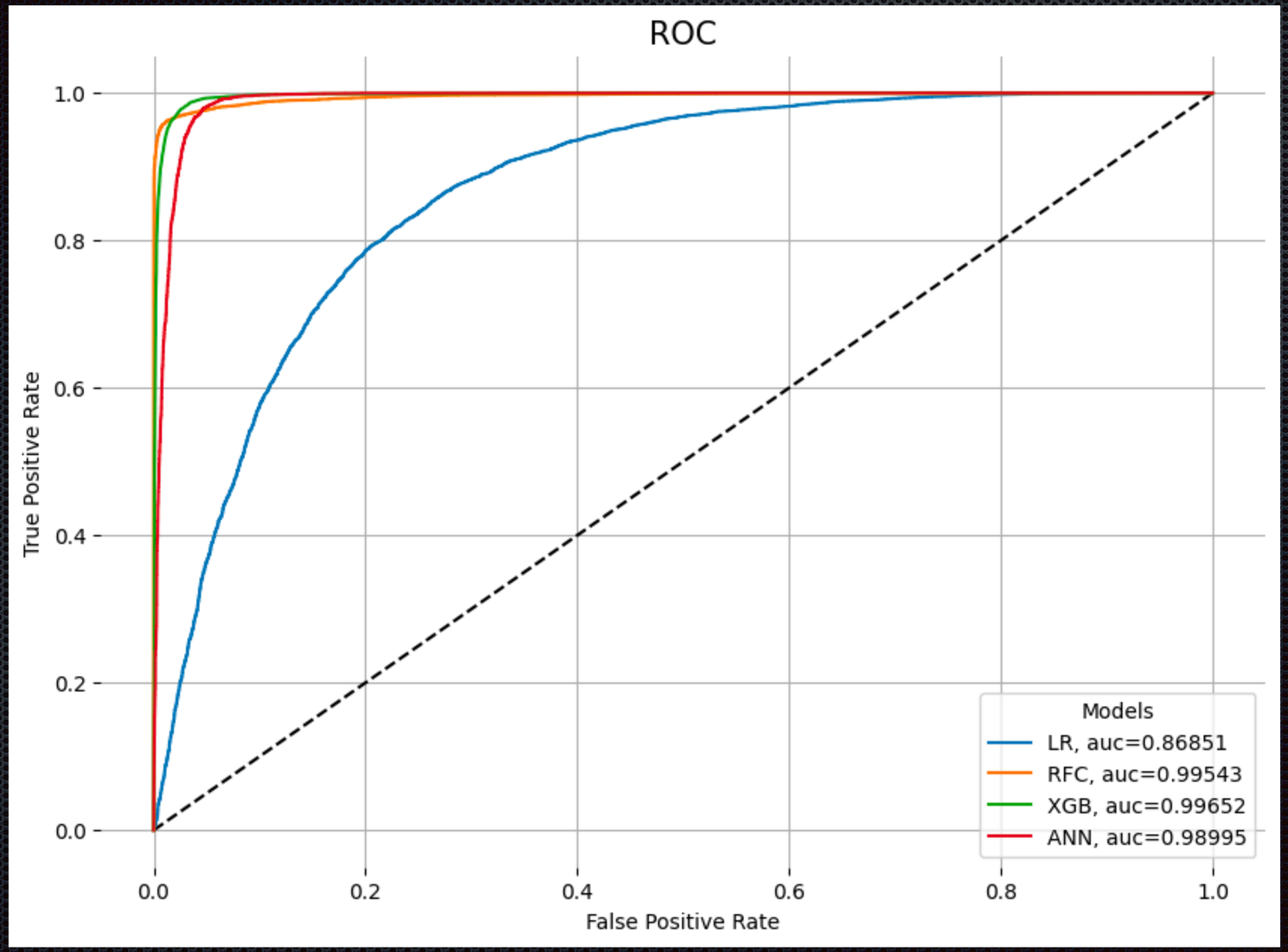
# Modelling approach

1. Due to high data complexity, ensemble-based techniques like Random Forest and XGBoost were recommended for modelling, over ANN, LR, and SVM.
2. Model performance can be evaluated through metrics like accuracy, precision, recall, F1-score, and ROC-AUC to choose the top two models.
3. Hyperparameter tuning is performed using Random Search or Bayesian Optimization will be conducted to optimize the performance of the selected models, particularly XGBoost and ANN.
4. Interpretability techniques, such as SHAP values, can be used to interpret the model's predictions and identify the most important features contributing to the model's output.
5. A pipeline can be constructed to automate the feature engineering and modelling processes and make it easier to reproduce and scale the model.

# Models performance

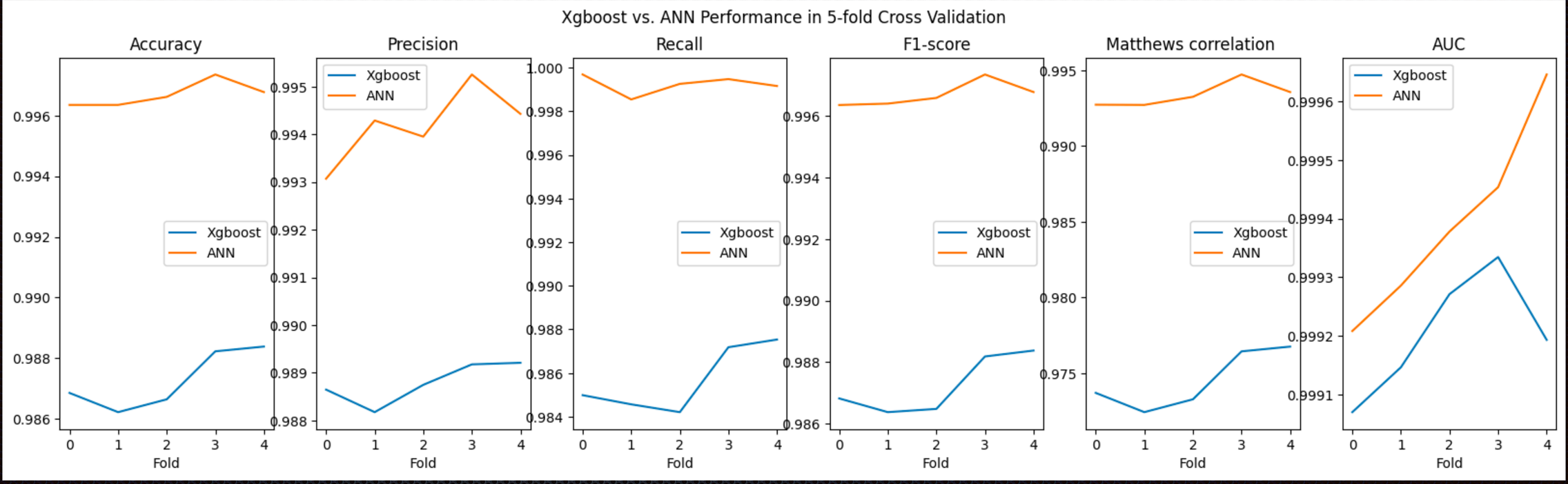


# Models performance



# Validation

For model validation, a 5-fold cross-validation strategy will be used to evaluate the performance of the models. The dataset will be split into training and testing sets in a 80:20 ratio. Stratified sampling will be used to ensure that the class distribution is maintained in both the training and testing datasets. The best model will be selected based on its average performance across the 5 folds.

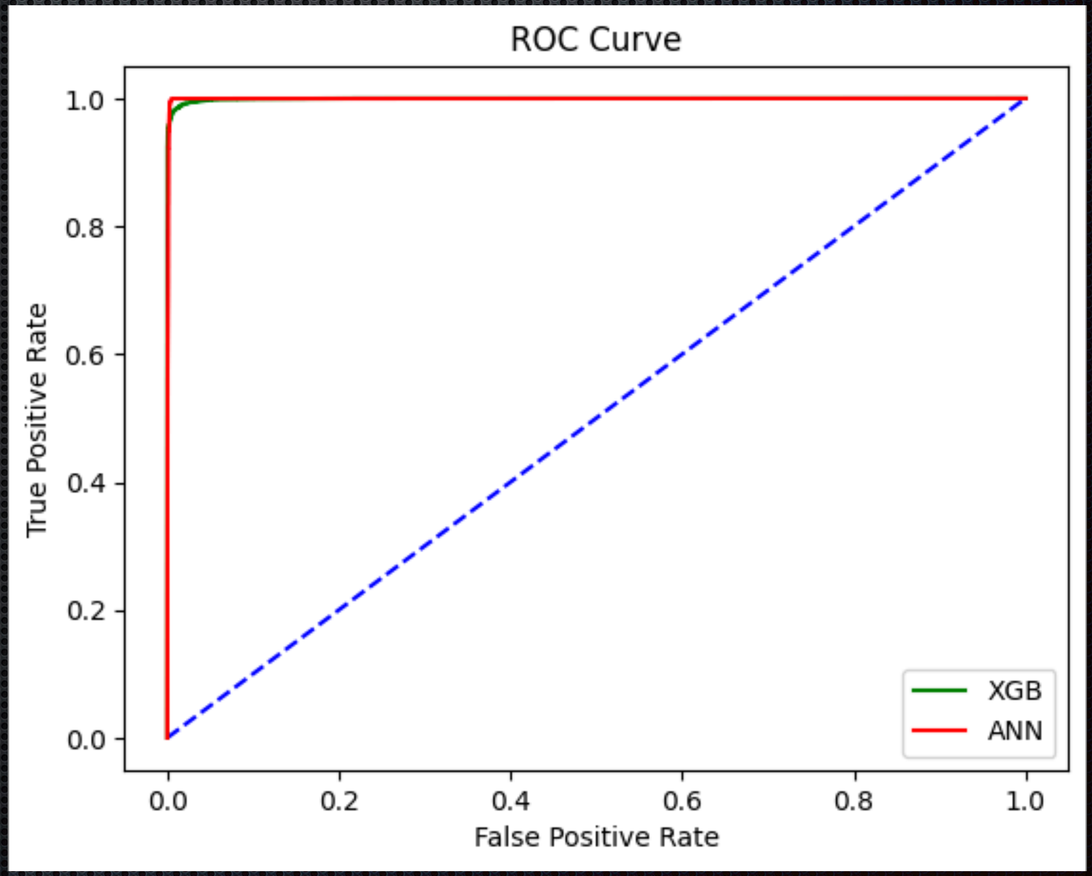


# Conclusion

1. Neural Network Model performance is slightly better than XGBoost model in terms of accuracy, precision, recall, F1-score and Matthews correlation coefficient.

2. Neural Network Model has better AUC score than XGBoost model in most of the tests.

3. So finally we will use Neural Network Model for predictions on validation data



# Thank You

“You can have data without information,
but you cannot have information without data.”

*–Daniel Keys Moran*