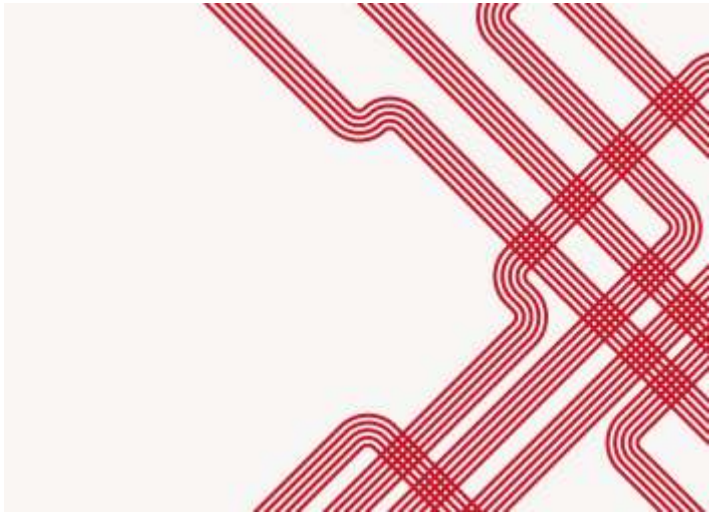
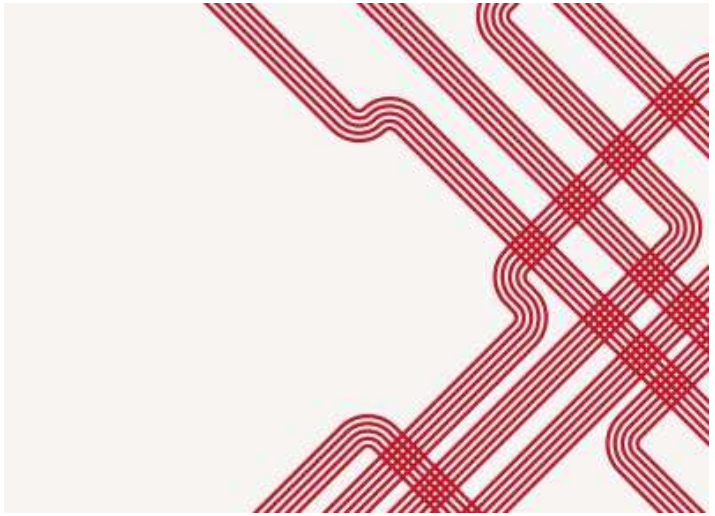
Predicting Customer Churn in a Bank





Presentation Flow

1.Introduction

2.Dataset

3.Machine Learning Models

- Logistic Regression Model
- KNN Classifier
- QDA (Quadratic Discriminant Analysis)
- Random Forest

4.Conclusion

Introduction

- Customer churn, where clients discontinue their relationship with a bank, poses a significant challenge, leading to lost revenue and increased acquisition costs. This project aims to develop a robust predictive model to identify customers at risk of churning, allowing the bank to implement effective retention strategies.

- Our approach involves thorough data preprocessing, exploratory data analysis, and the application of various machine learning algorithms. We will evaluate the model's performance using key metrics and ensure its interpretability for actionable insights. The outcome of this project will enable the bank to deploy targeted interventions, thereby reducing churn rates and fostering long-term customer loyalty.

- This project underscores the importance of data-driven decision-making in enhancing customer satisfaction and achieving sustainable growth in the banking sector.

About the Dataset

Variables	Description
Customer Id	Contains random values and has no effect on customer leaving the bank
Surname	The surname of a customer
Credit Score	Can have effect on customer churn, since a customer with a higher credit score is less likely to leave the bank
Geography	a customer’s location
Gender	male or female
Age	Age, has effect on the output
Tenure	Refers to the number of years that the customer has been a client of the bank
Balance	People with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
NumofProducts	Refers to the number of products that a customer has purchased through the bank.
HasCrCard	Denotes whether a customer has a credit card or not.
IsActiveMember	Active customers are less likely to leave the bank
EstimatedSalary	People with lower salaries are more likely to leave the bank compared to those with higher salaries.
Exited	whether or not the customer left the bank

Data Cleaning

Missing Data & Data Type Correction: Checked for missing data using isnull() function, found none. Also used dtypes method to find the type of variables and changed the data type of exited(churn) from ‘int’ to ‘category’.

Log transformation of Continuous Variables: Histograms of ‘Age’ is right skewed and Histogram of ‘Credit Score’ is left skewed. Log transformation of these two variables didn’t really help, so used the original data.

Exploratory Data Analysis

Descriptive Statistics

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000

Machine Learning Models

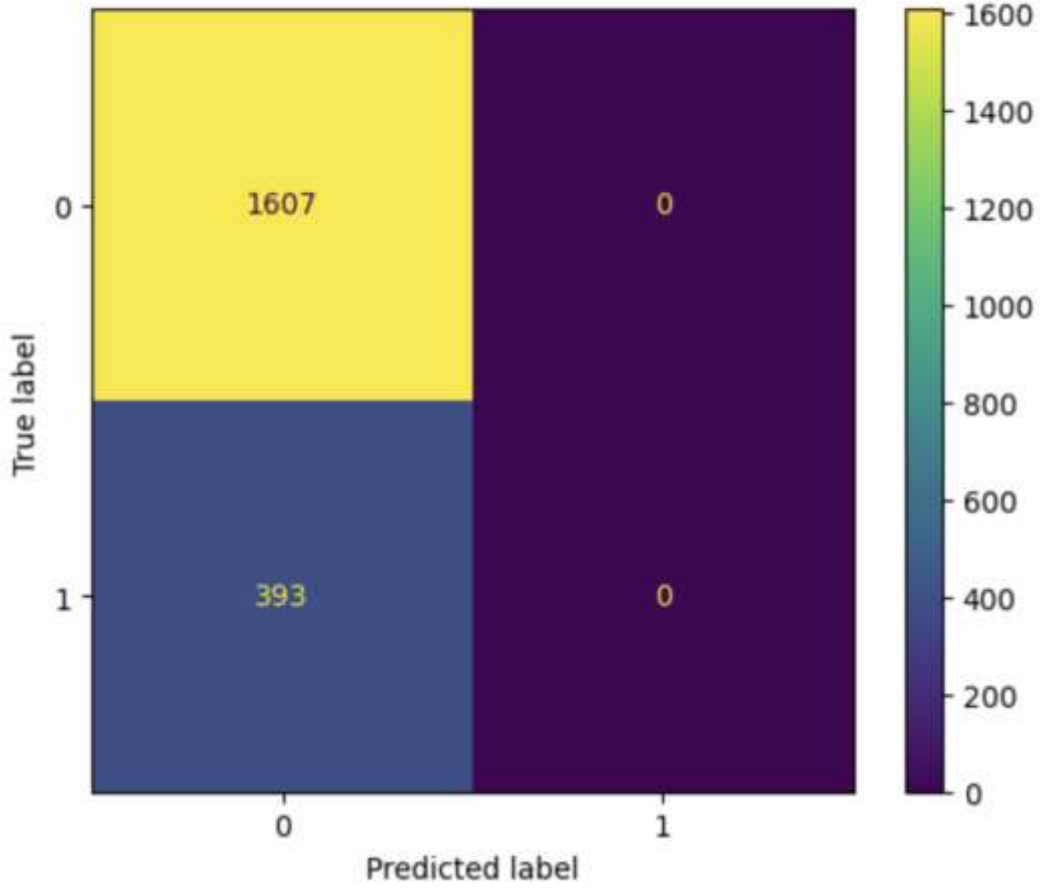
Model Training:

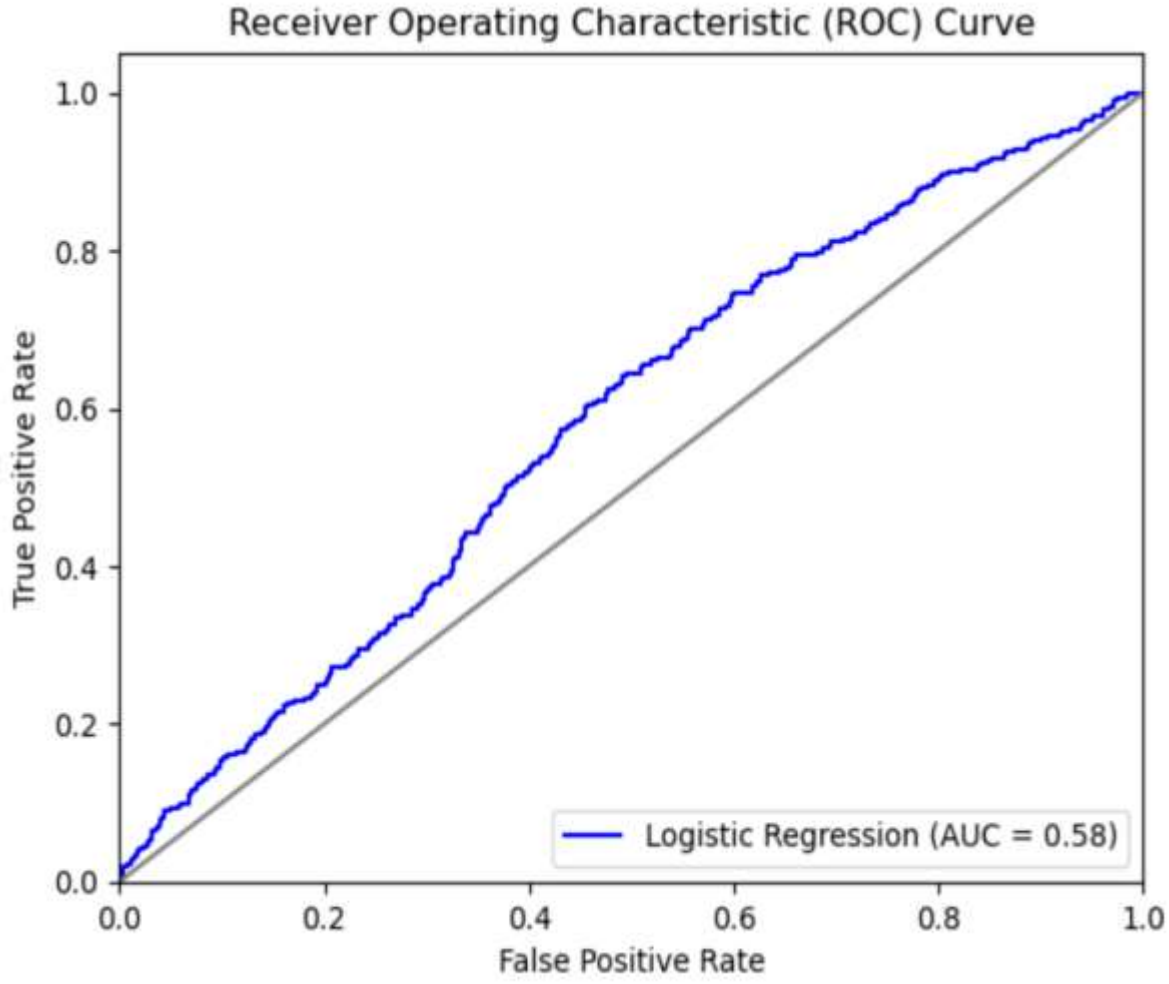
Metric Selection: I believe bank would want to identify as many as customers who might want to churn as possible. True rate positive measures the proportion of actual churn cases correctly identified by the model. Therefore, considering true positive rate will be a suitable metric for model selection.

Model Selection: I am going to build four models, Logistic regression, KNN Classifier, QDA(Quadratic Discriminant Analysis) & Random Forest. Will chose the one with high positive rate.

Model 1 – Logistic Regression

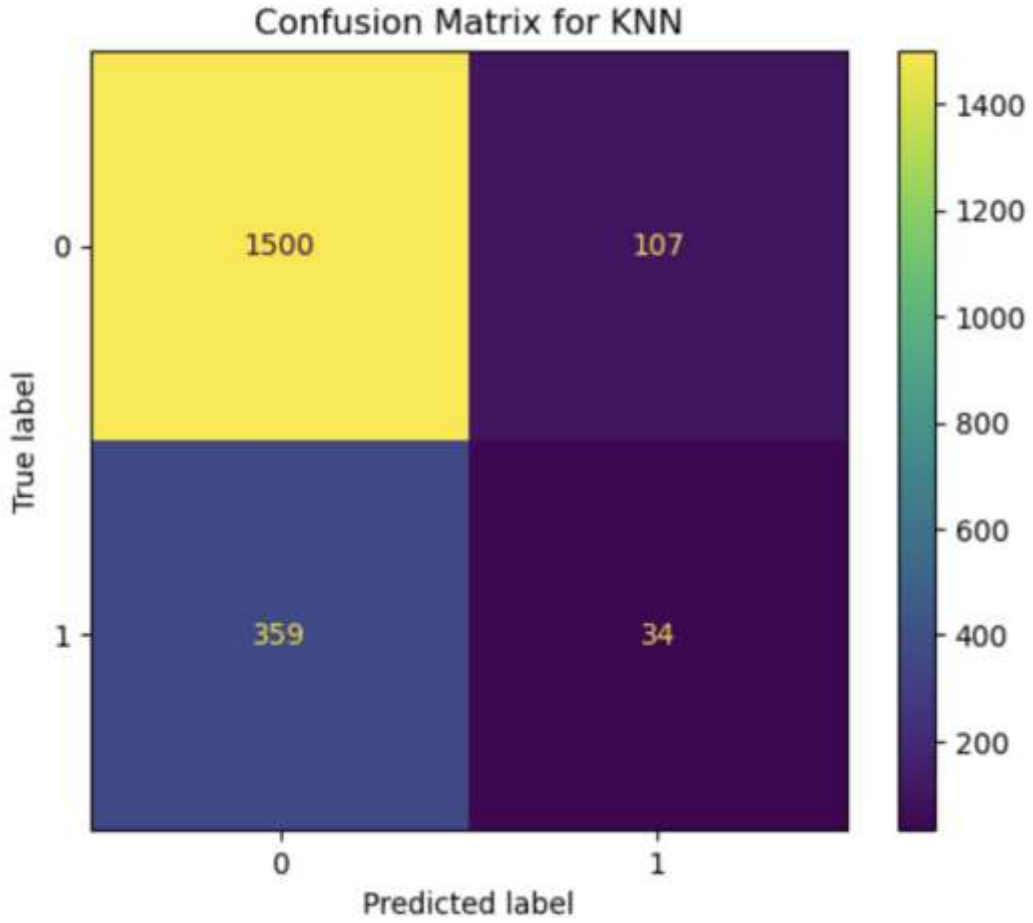
```
Confusion Matrix for Logistic Regression:
 [[1607    0]
 [ 393    0]]
```

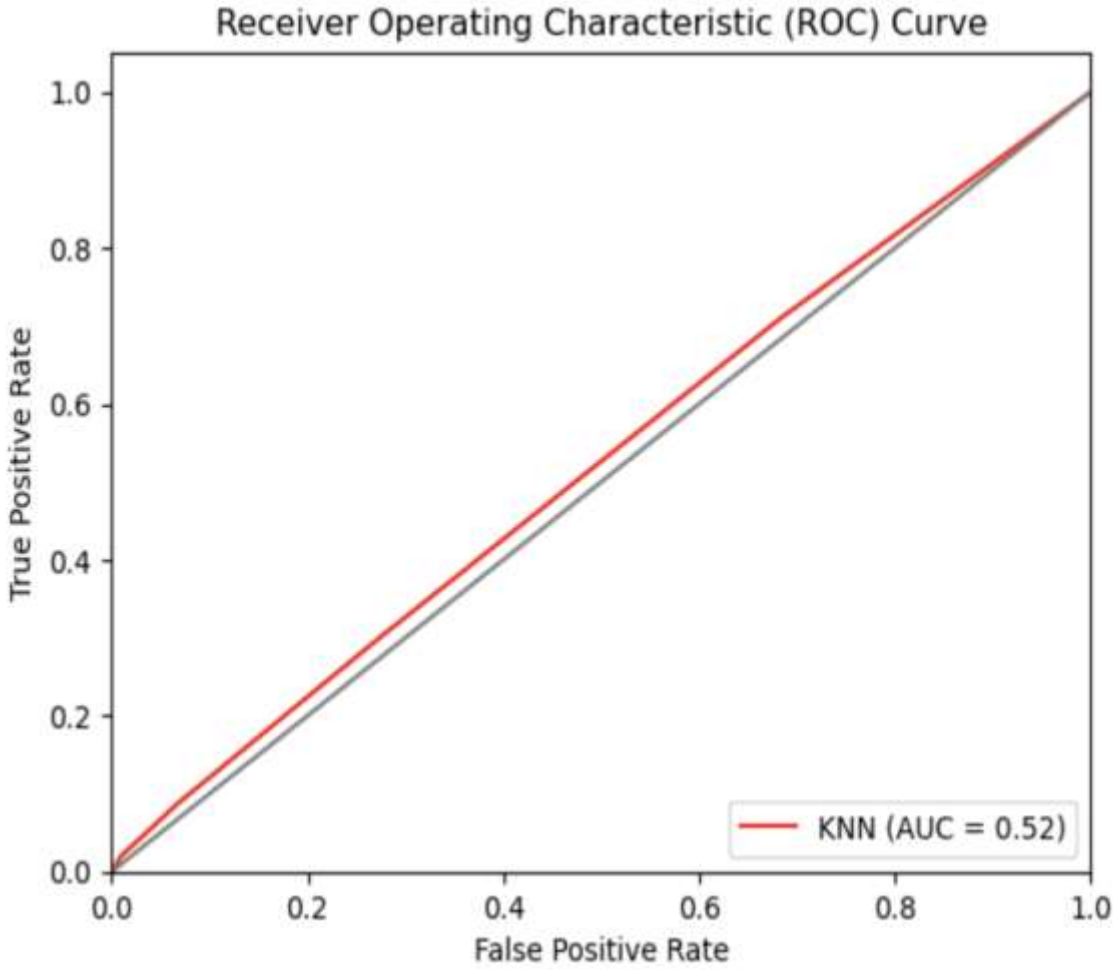




Model 2 – KNN Classifier

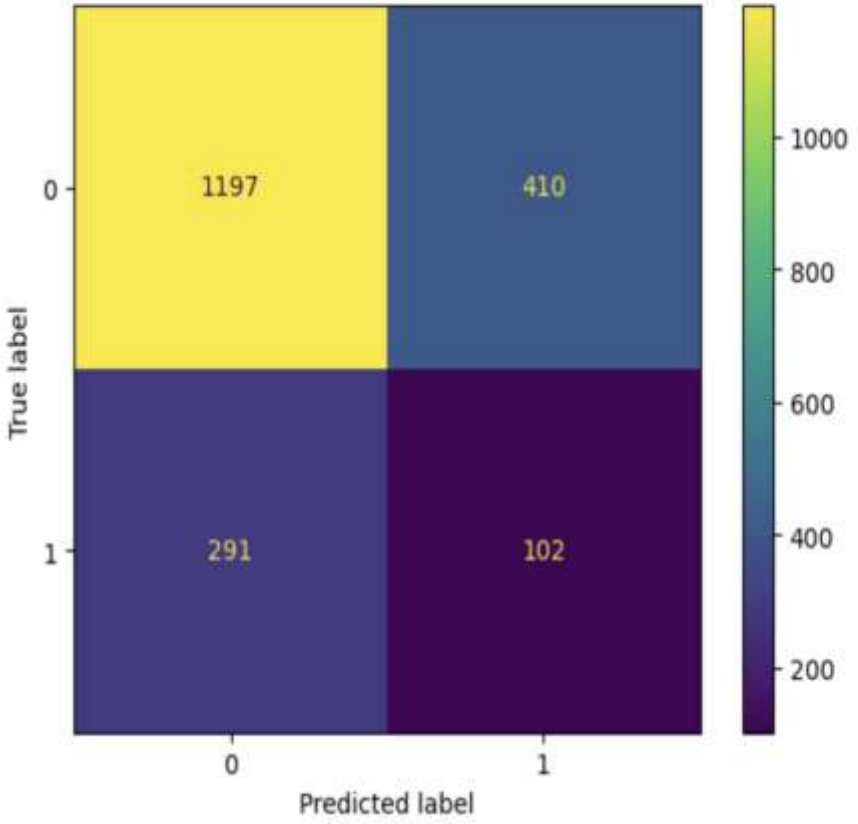
```
Confusion Matrix for KNN:
 [[1500  107]
 [ 359   34]]
```

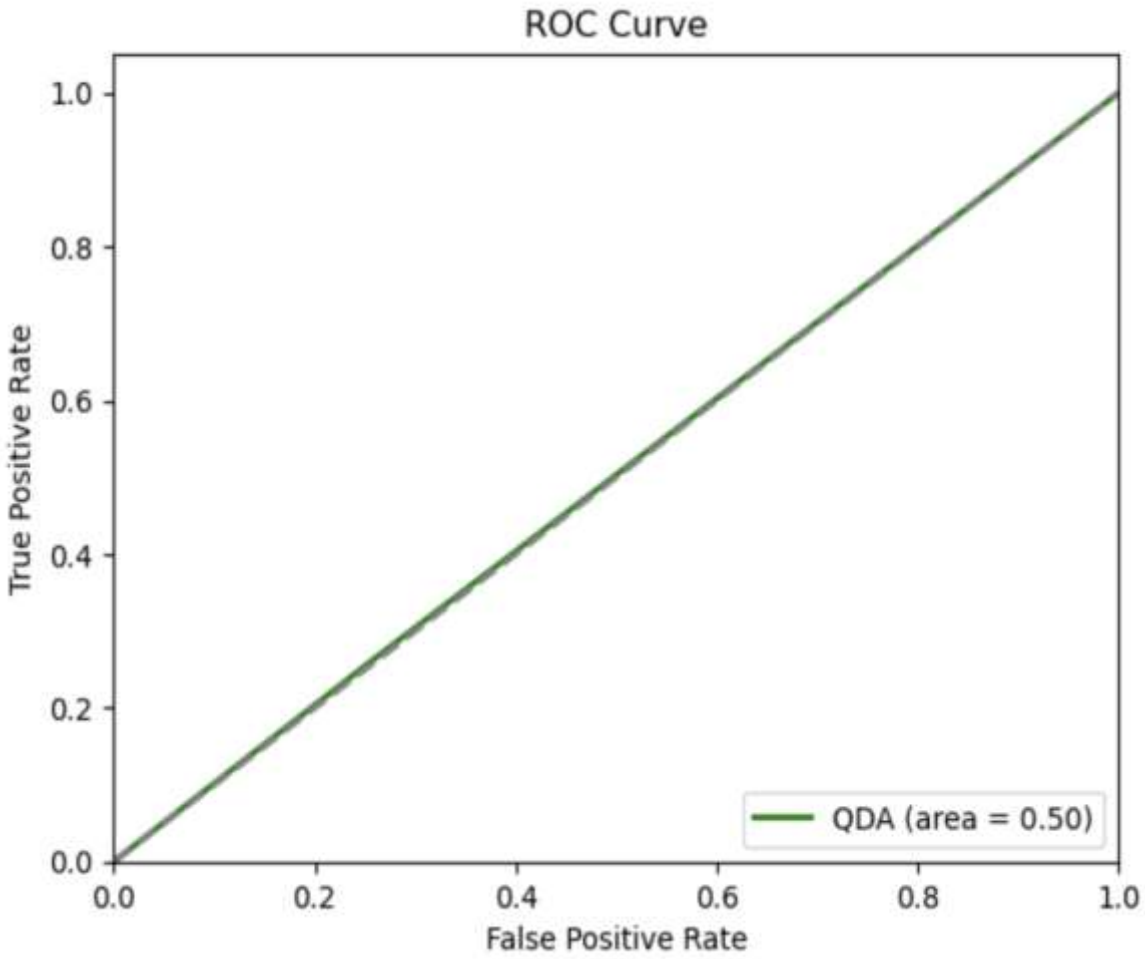




Model 3 – QDA[Quadratic Discriminant Analysis]

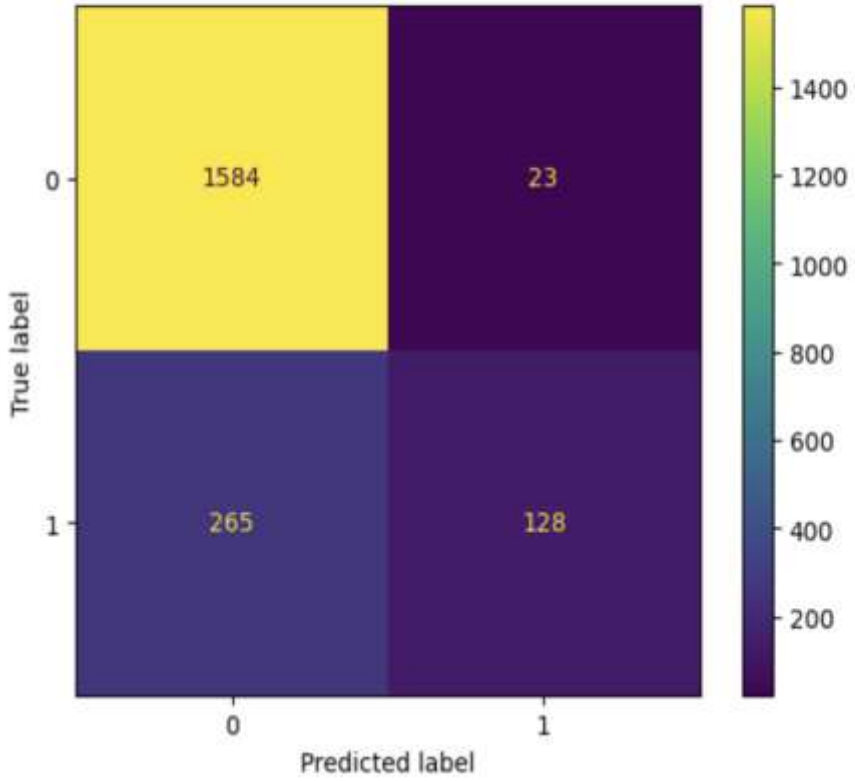
```
Confusion Matrix for QDA:
[[1197  410]
 [ 291  102]]
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x152435b10>
```



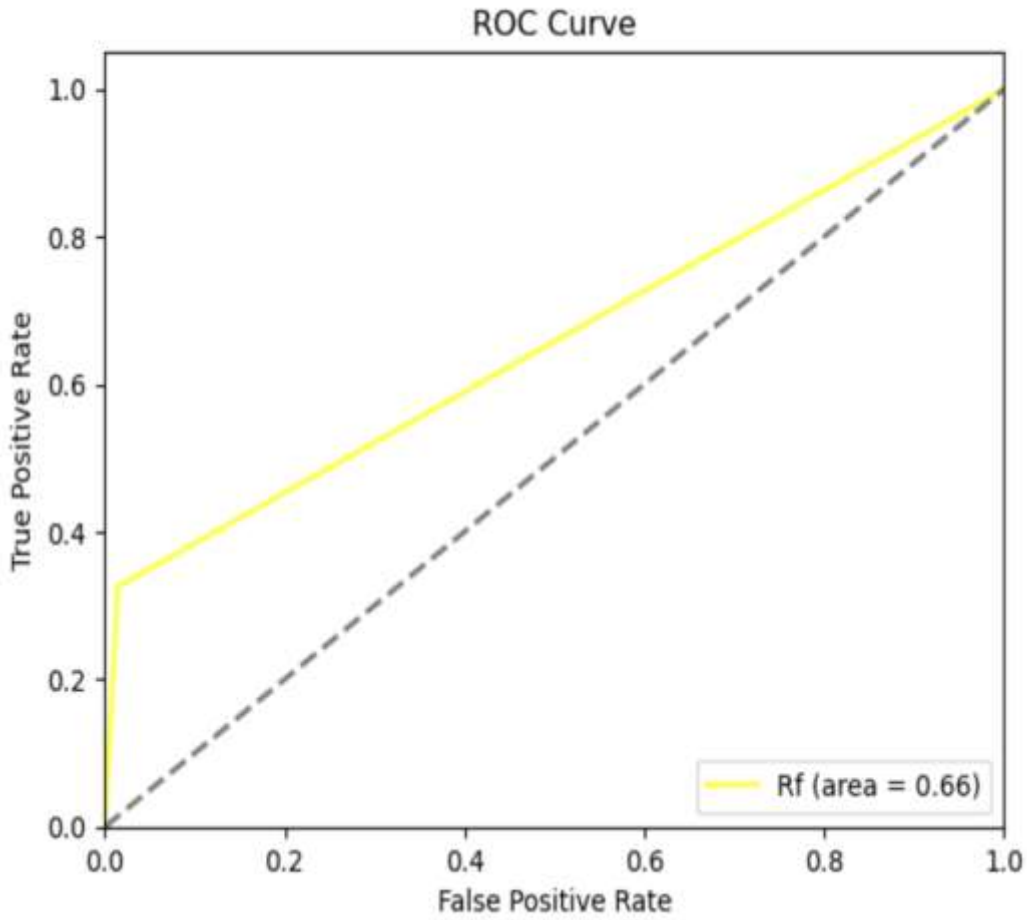


Model 4 – Random Forest

```
Confusion Matrix for Random Forest:
[[1584   23]
 [ 265  128]]
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x16de86f90>
```




```
AUC for Rf: 0.6556936811120558
```



Comparing the models

```
Logistic Regression - TPR: 0.0, FPR: 0.0
KNN - TPR: 0.08651399491094147, FPR: 0.06658369632856254
QDA - TPR: 0.2595419847328244, FPR: 0.2551337896701929
Random Forest - TPR: 0.3256997455470738, FPR: 0.01431238332296204
```

Conclusion

I will choose the model with high true positive rate and lower false positive rate as my ideal model for prediction.

As from the above four models, the Random Forest model has higher positive rate compared to the remaining models. So, I chose Random Forest model as my ideal model which has true positive rate of approximately 33%.

Since the true positive rate is too low (i.e.,33%), which means that the data needs more information.

We need to consider True Positive rate along with other factors such as Precision, Specificity, F1 Score depending on the specific goals.