Final Project

November 30, 2022

Team Member Names: Madeline Witters

Project Title: Predicting Customer Churn and Identifying Attributes of At-Risk Customers

Problem Statement

One of the most important metrics of success in any business is customer retention. A business's customer churn rate can have very significant financial impacts that affect the company in a multitude of ways. A Harvard Business Review article recently stated that merely increasing a company's customer retention rates by 5% can increase profits upwards of 25%; conversely, the cost of obtaining a new customer can be anywhere from 5 to 25 times as expensive as retaining an existing one.

It therefore follows that if a company is able to predict which customers are at risk of leaving, they can use this information to better position themselves in a variety of ways, such as: creating an intervention plan for customers at risk of leaving, calculating potential loss of revenue in the next quarter, or simply better understanding their customer demographic and various market segments.

In this project, there are two central research questions that I will aim to answer:

1. Can I create a model that will predict customer churn with a reasonable accuracy rate?
    - Furthermore, does one model type outperform another in predicting customer churn?
2. What features/customer attributes (present in the dataset) are most important in predicting whether a customer will churn?
    - Additionally, what does interpretation of these attributes reveal (for example, are younger customers more at risk of churning)?

Data Source

For this project I am using the "Bank Customer Churn" dataset, sourced from Kaggle. The dataset has 10,000 data points and 12 variables. Each row represents an individual customer. Variables are both categorical and quantitative, and include demographic variables (age, gender, country), as well as industry-specific variables (balance, credit_card, etc.). The response variable "churn" is categorical, containing a 1 if the client left the bank and a 0 if they remained a customer.

Methodology

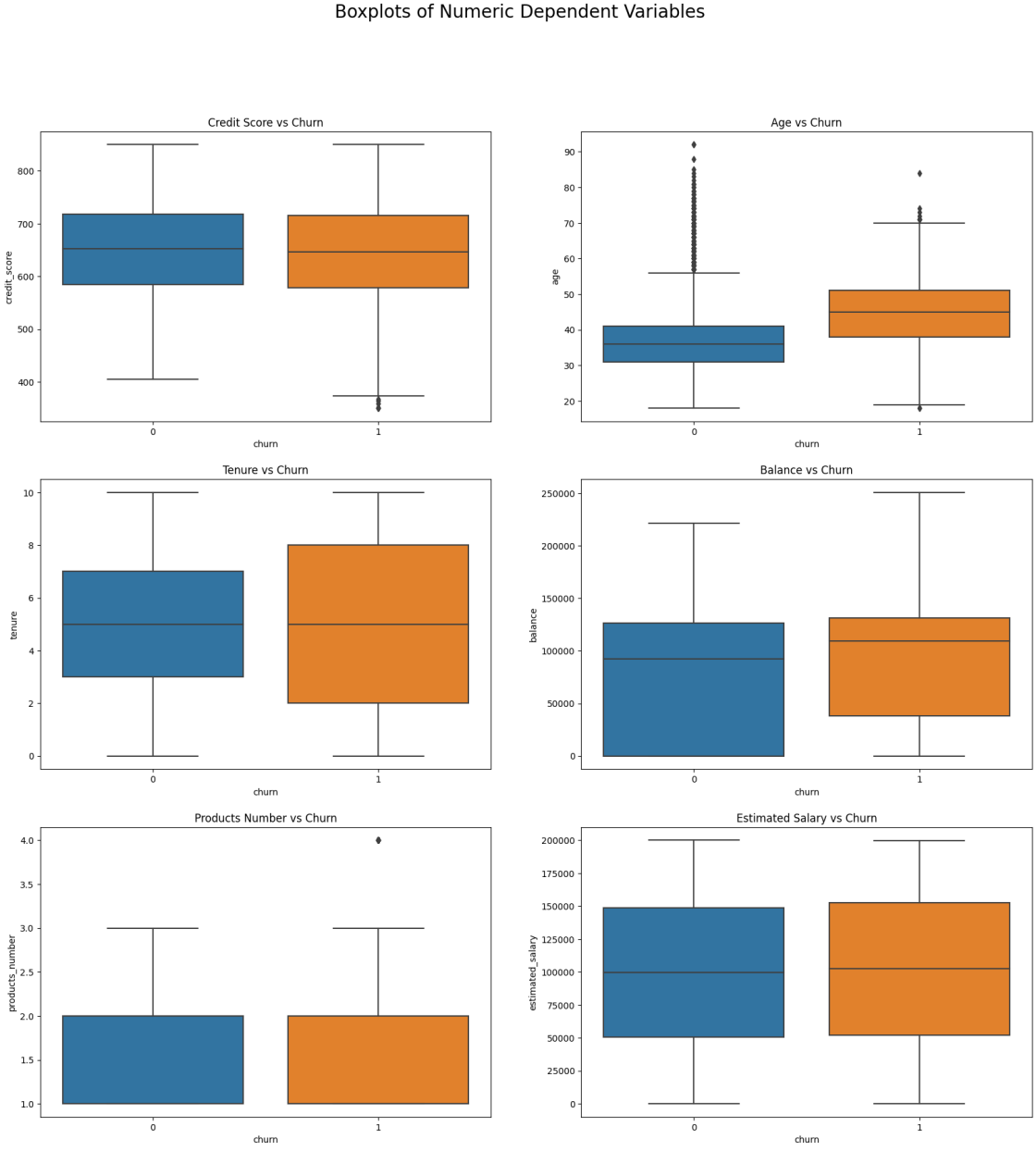
My methodology for this project consisted of three distinct parts:

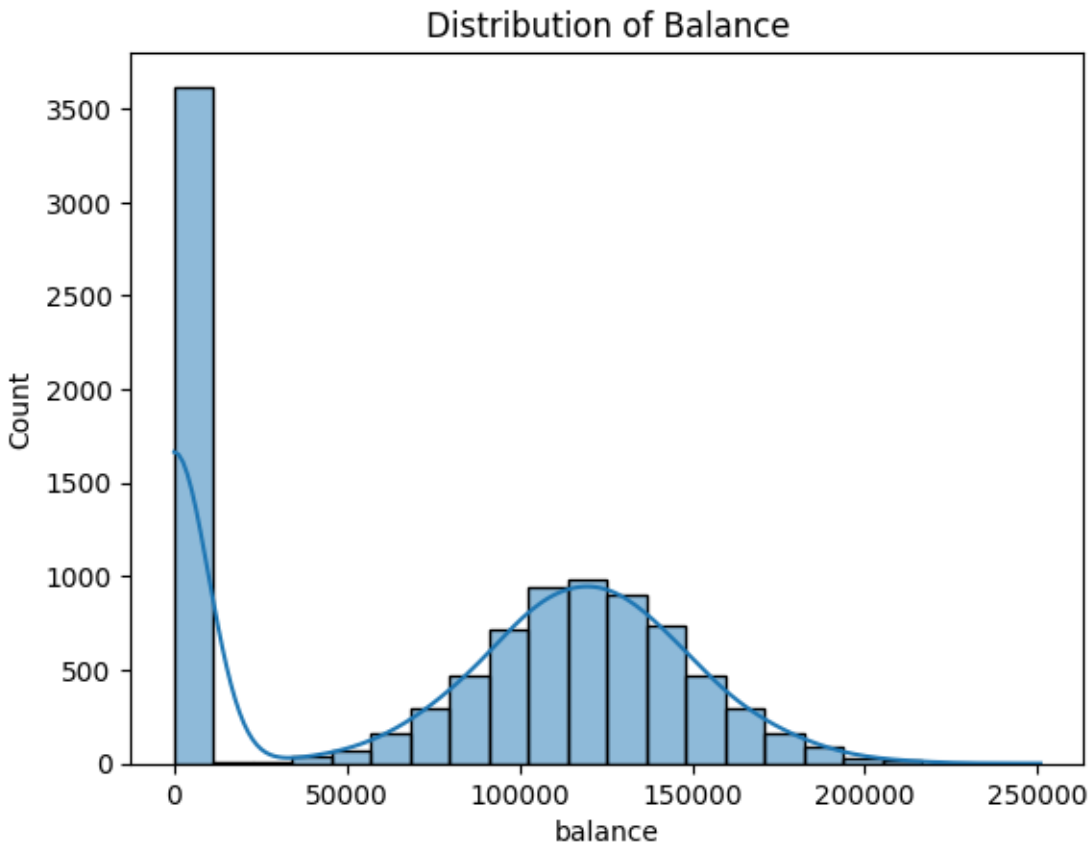
- Exploratory Data Analysis
- Variable Selection

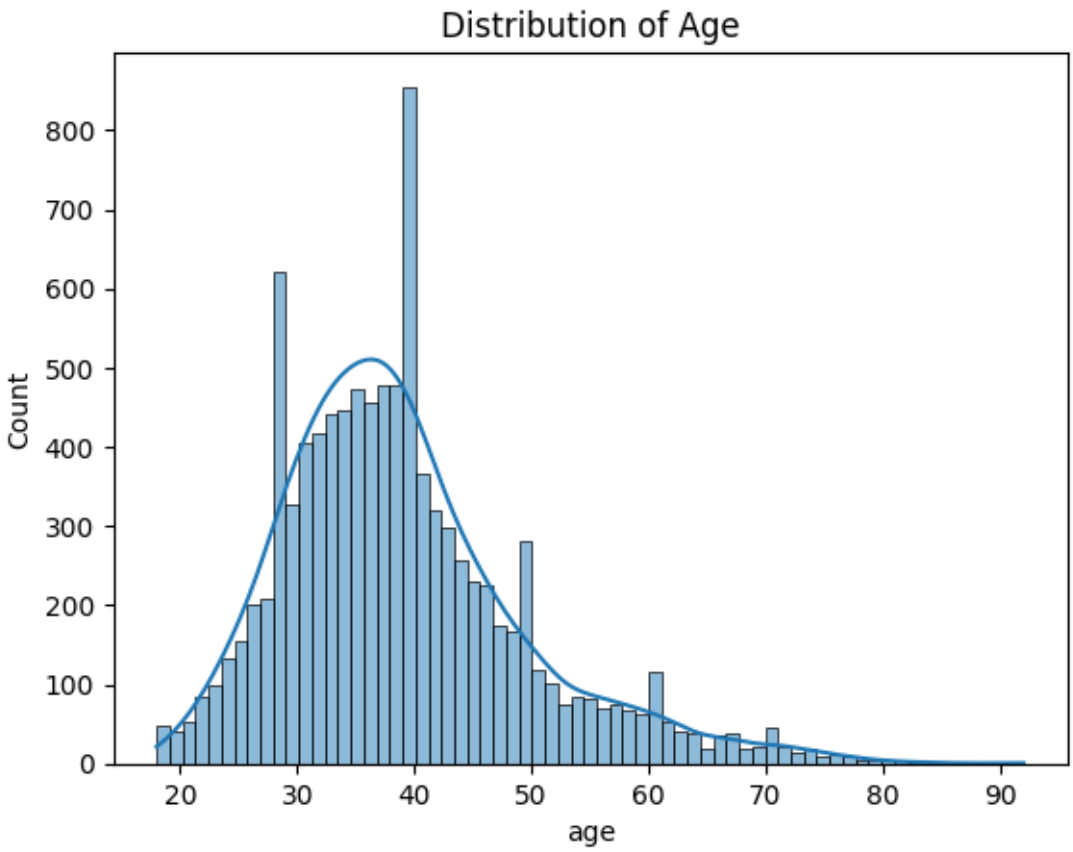
- Modeling

I will begin with the exploratory data analysis.

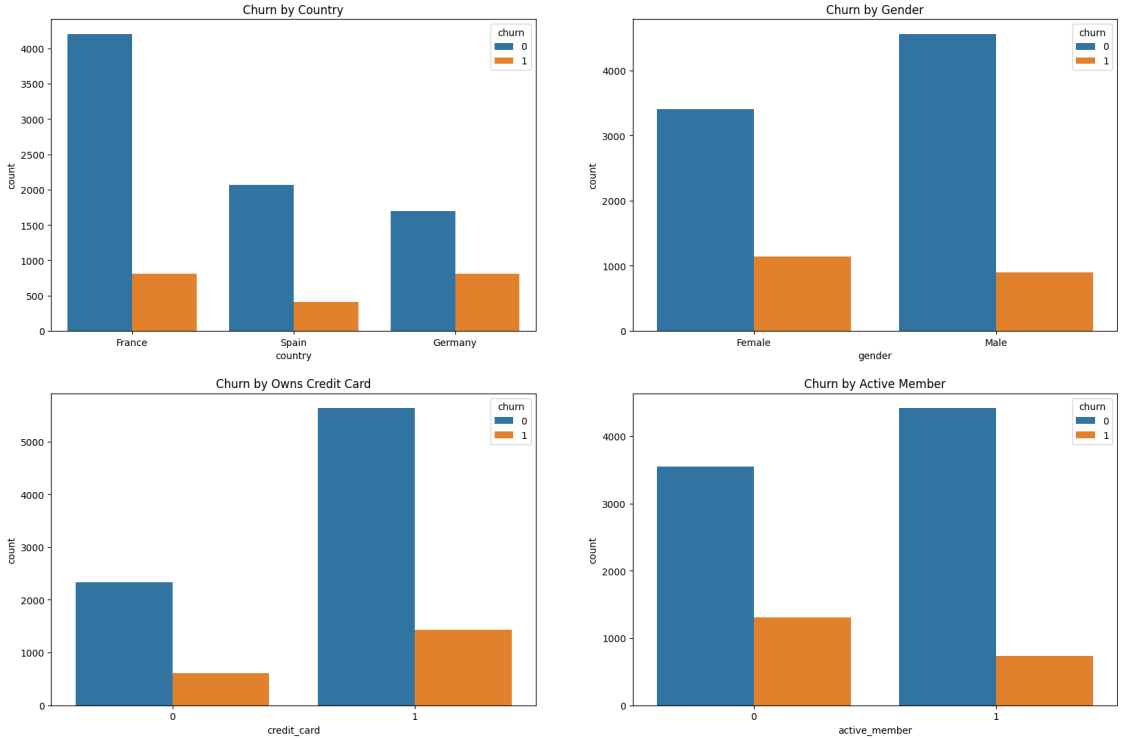
Exploratory Data Analysis



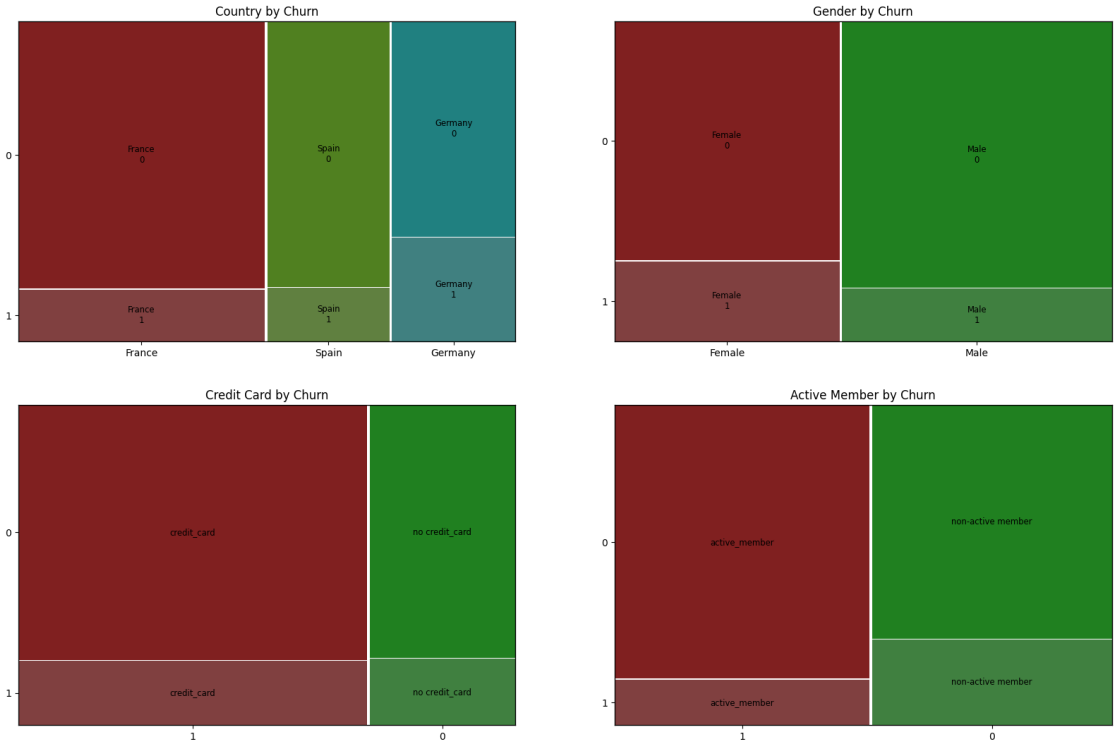

Distribution of Balance
Count
balance
0
500
1000
1500
2000
2500
3000
3500
0
50000
100000
150000
200000
250000

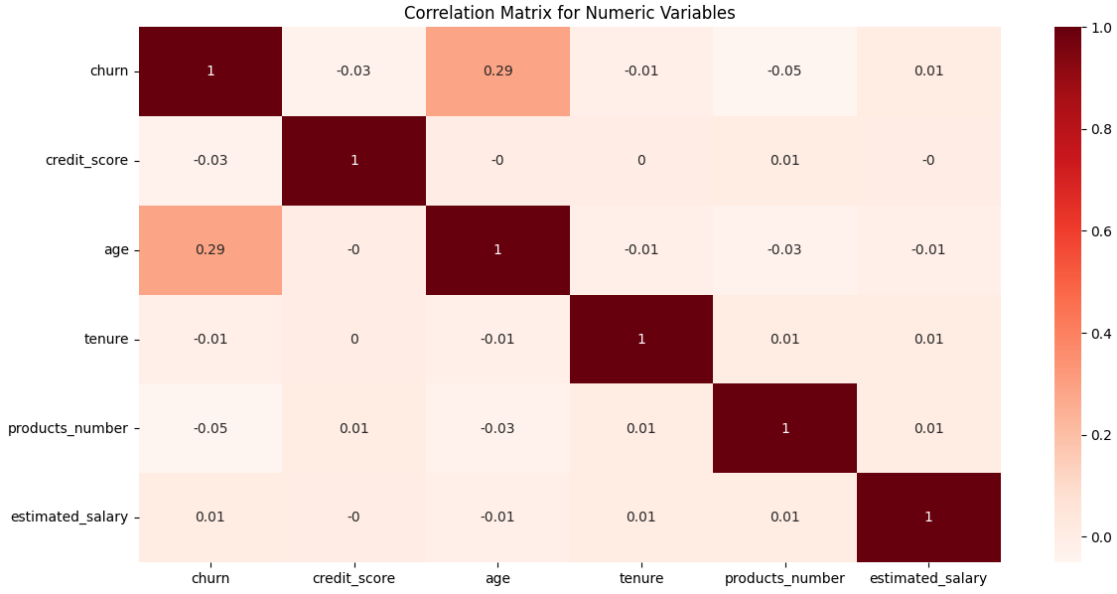

Distribution of Age
Count
age
0
100
200
300
400
500
600
700
800
20
30
40
50
60
70
80
90

Bar Charts for Categorical Dependent Variables

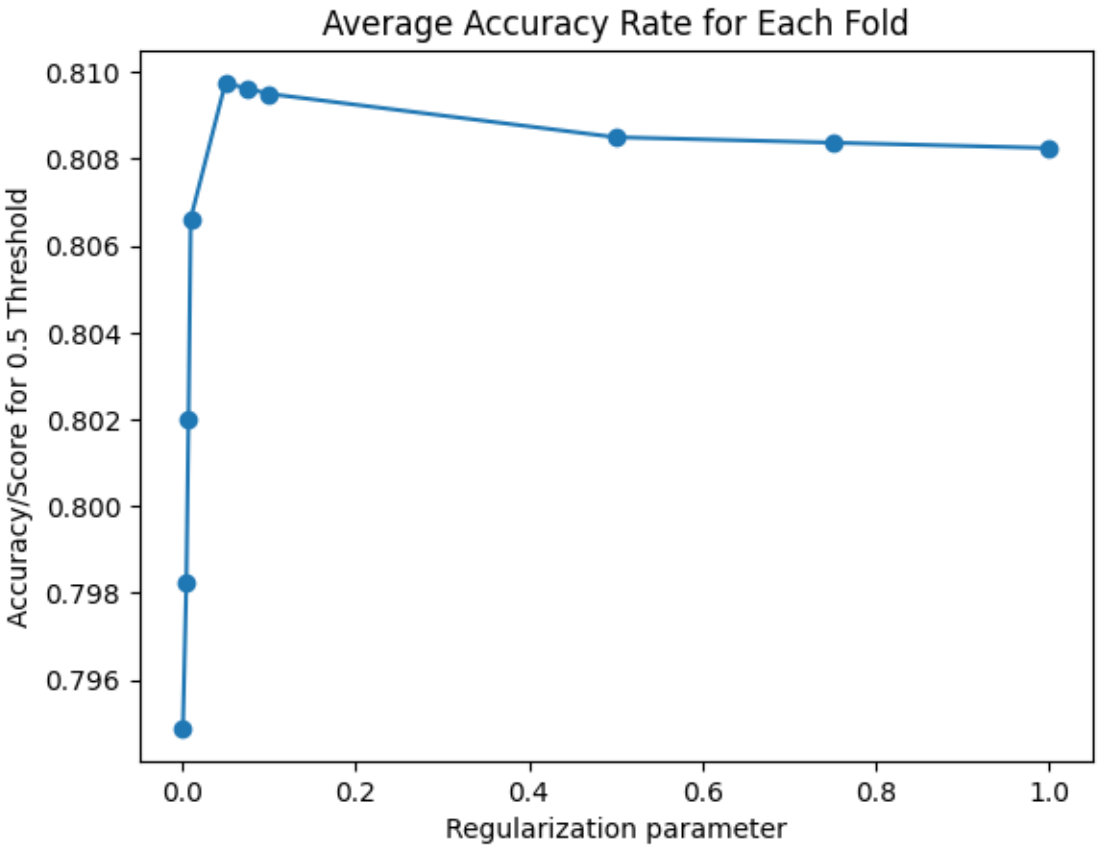


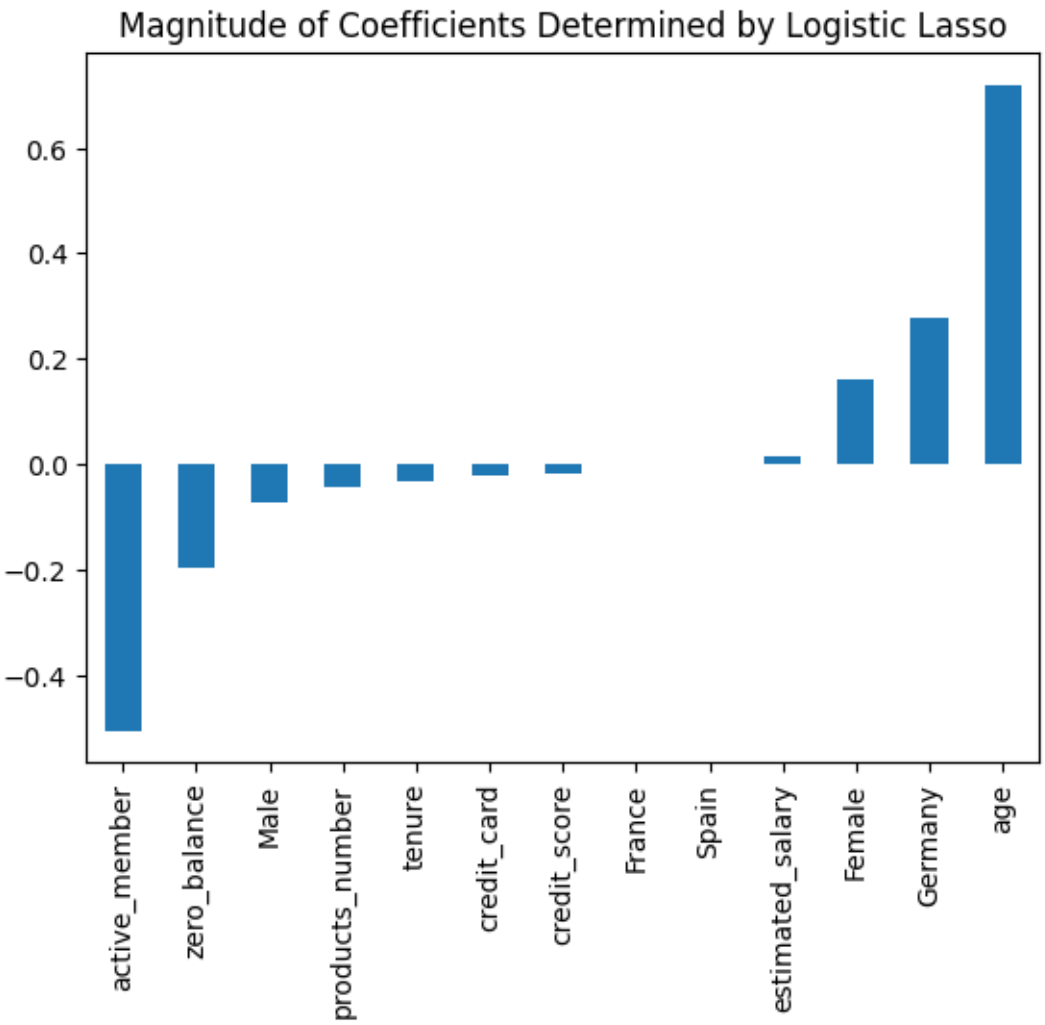
Mosaic Plots for Categorical Dependent Variables





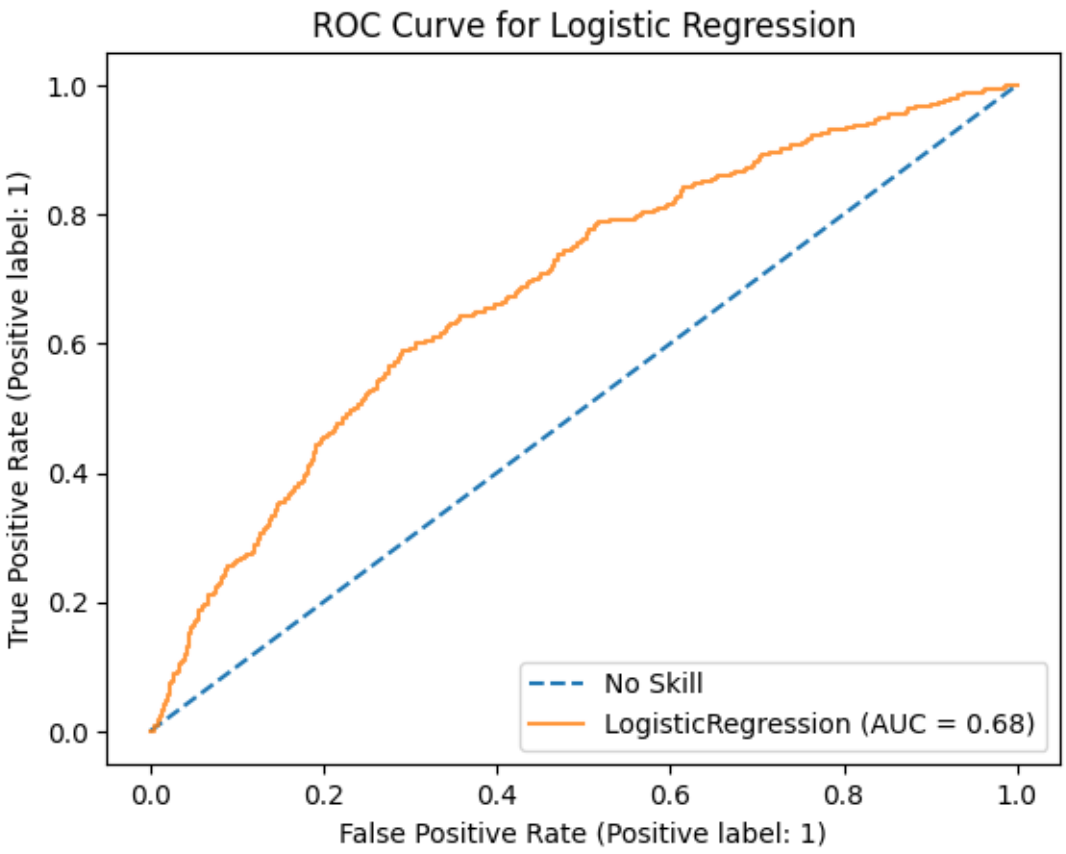
Variable Selection

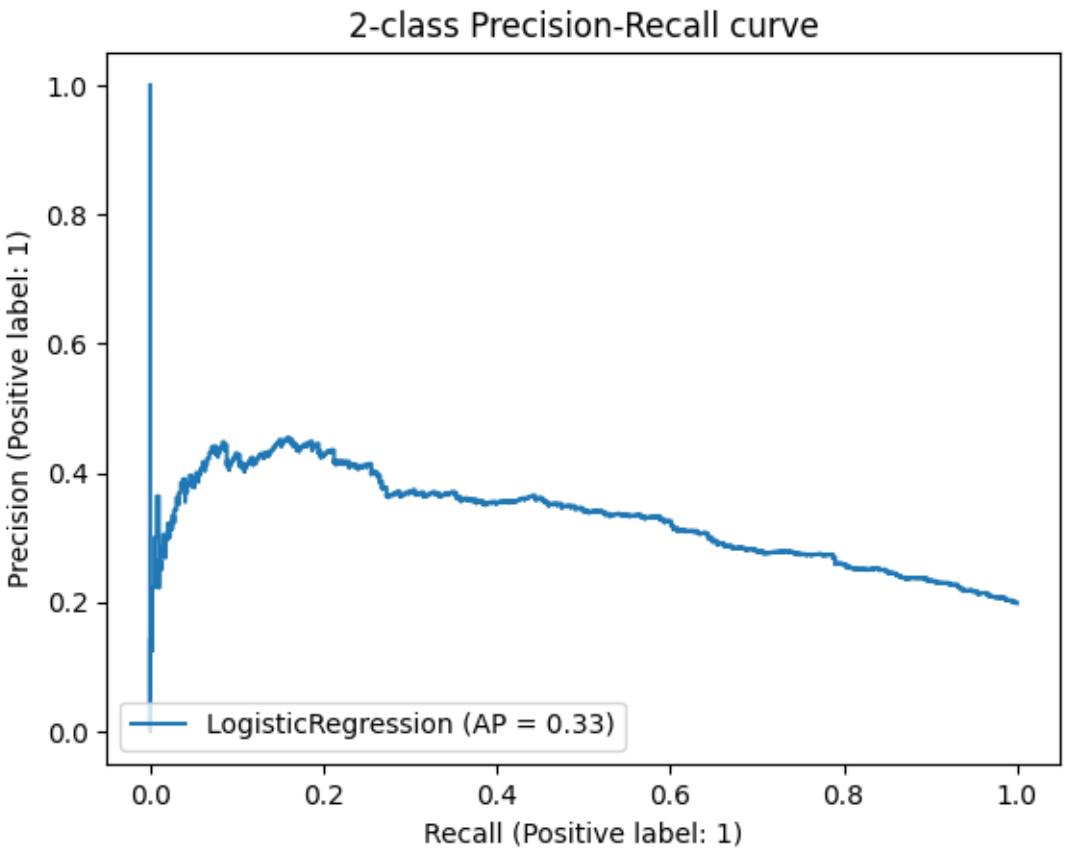




Modeling

Logistic Regression

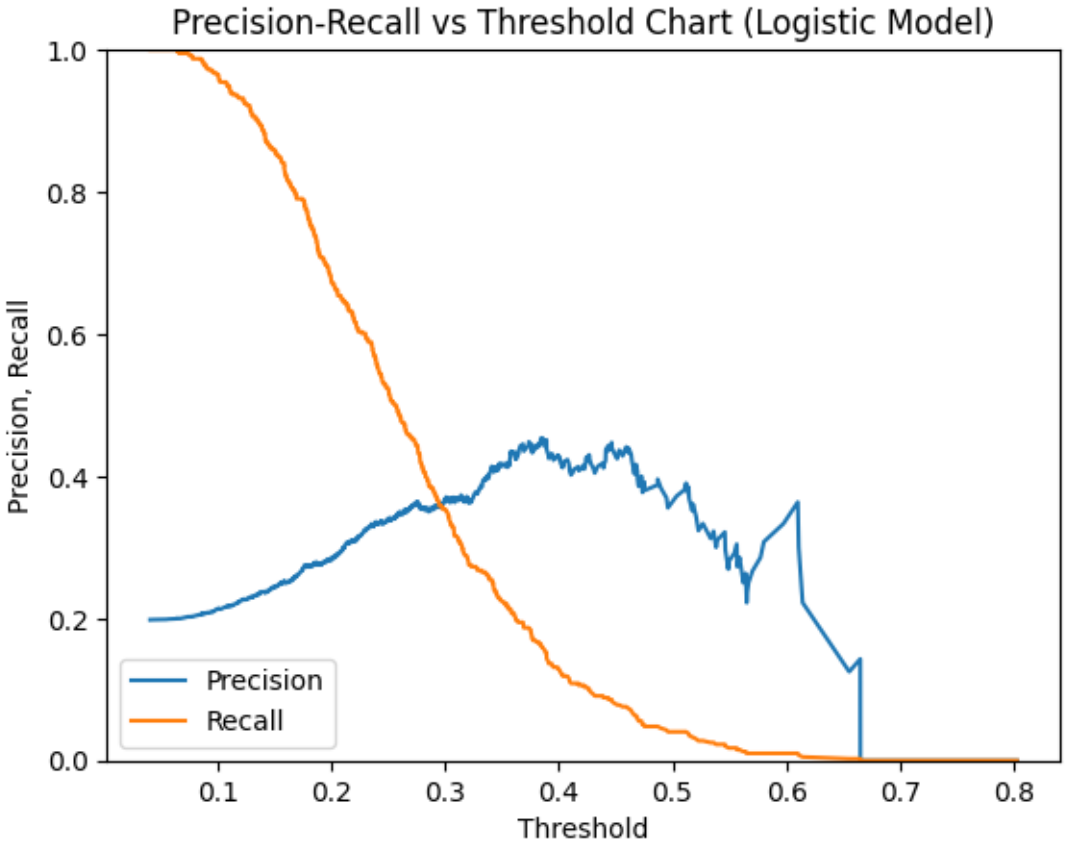

ROC Curve for Logistic Regression
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
No Skill
LogisticRegression (AUC = 0.68)




```
******** Logistic Regression model where threshold = 0.1 ******
Accuracy/Score is 0.291
Confusion Matrix:
[[ 200 1404]
 [  14  382]]
              precision    recall  f1-score   support

           0       0.93      0.12      0.22      1604
           1       0.21      0.96      0.35       396

    accuracy                           0.29      2000
   macro avg       0.57      0.54      0.29      2000
weighted avg       0.79      0.29      0.25      2000


******** Logistic Regression model where threshold = 0.25 ******
Accuracy/Score is 0.7045
Confusion Matrix:
[[1203  401]
```

```
 [ 190  206]]
              precision    recall  f1-score   support

           0       0.86      0.75      0.80      1604
           1       0.34      0.52      0.41       396

    accuracy                           0.70      2000
   macro avg       0.60      0.64      0.61      2000
weighted avg       0.76      0.70      0.73      2000


******** Logistic Regression model where threshold = 0.3 ******
Accuracy/Score is 0.7515
Confusion Matrix:
[[1363  241]
 [ 256  140]]
              precision    recall  f1-score   support

           0       0.84      0.85      0.85      1604
           1       0.37      0.35      0.36       396

    accuracy                           0.75      2000
   macro avg       0.60      0.60      0.60      2000
weighted avg       0.75      0.75      0.75      2000


******** Logistic Regression model where threshold = 0.5 ******
Accuracy/Score is 0.7965
Confusion Matrix:
[[1577   27]
 [ 380   16]]
              precision    recall  f1-score   support

           0       0.81      0.98      0.89      1604
           1       0.37      0.04      0.07       396

    accuracy                           0.80      2000
   macro avg       0.59      0.51      0.48      2000
weighted avg       0.72      0.80      0.72      2000


******** Logistic Regression model where threshold = 0.6 ******
Accuracy/Score is 0.8005
Confusion Matrix:
[[1597    7]
 [ 392    4]]
              precision    recall  f1-score   support
```

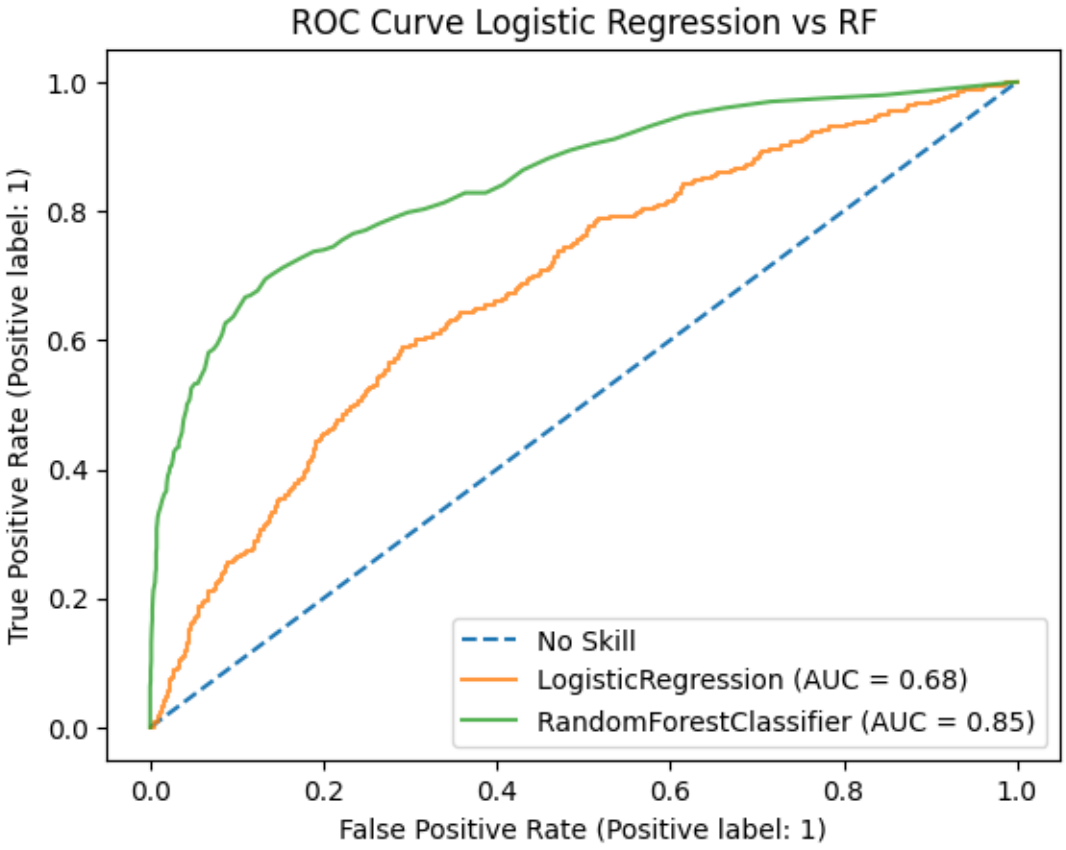
```
              0       0.80      1.00      0.89      1604
              1       0.36      0.01      0.02       396

    accuracy                            0.80      2000
   macro avg       0.58      0.50      0.45      2000
weighted avg       0.72      0.80      0.72      2000
```

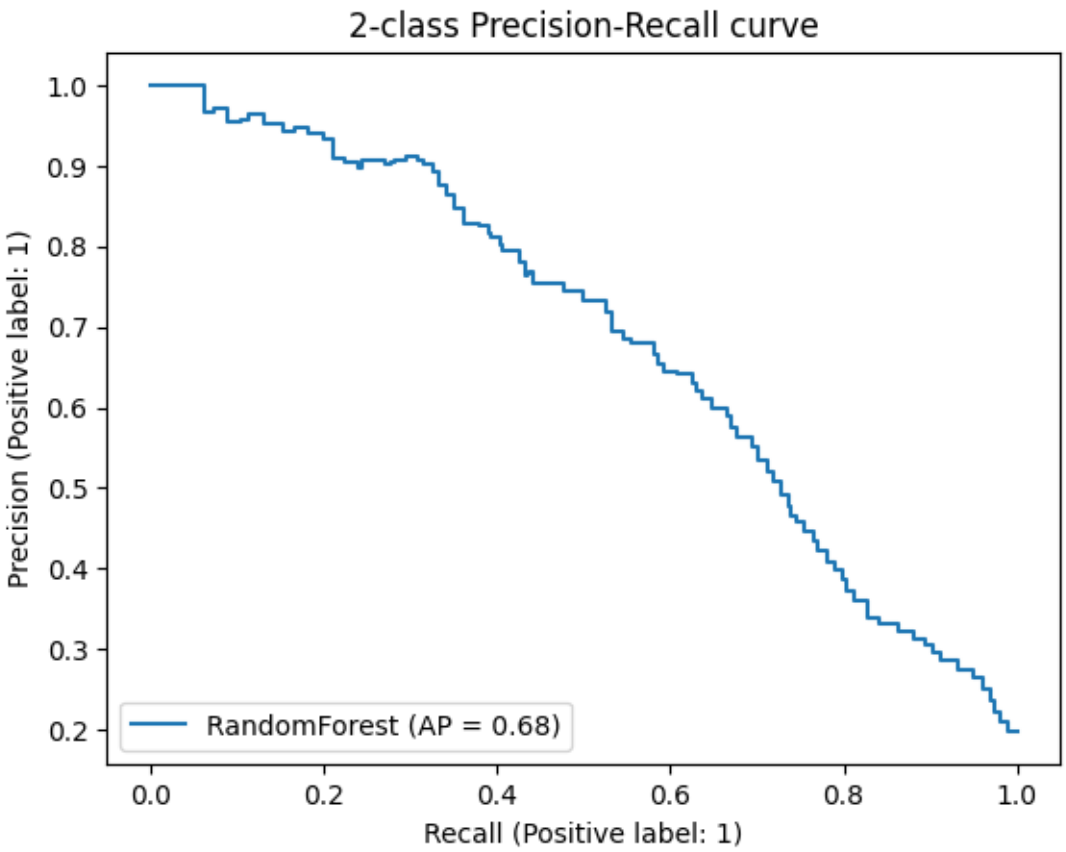


Random Forest

```
Random Forest Results
Accuracy/Score is 0.866
Confusion Matrix:
[[1537   67]
 [ 201  195]]
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      1604
           1       0.74      0.49      0.59       396

    accuracy                           0.87      2000
```

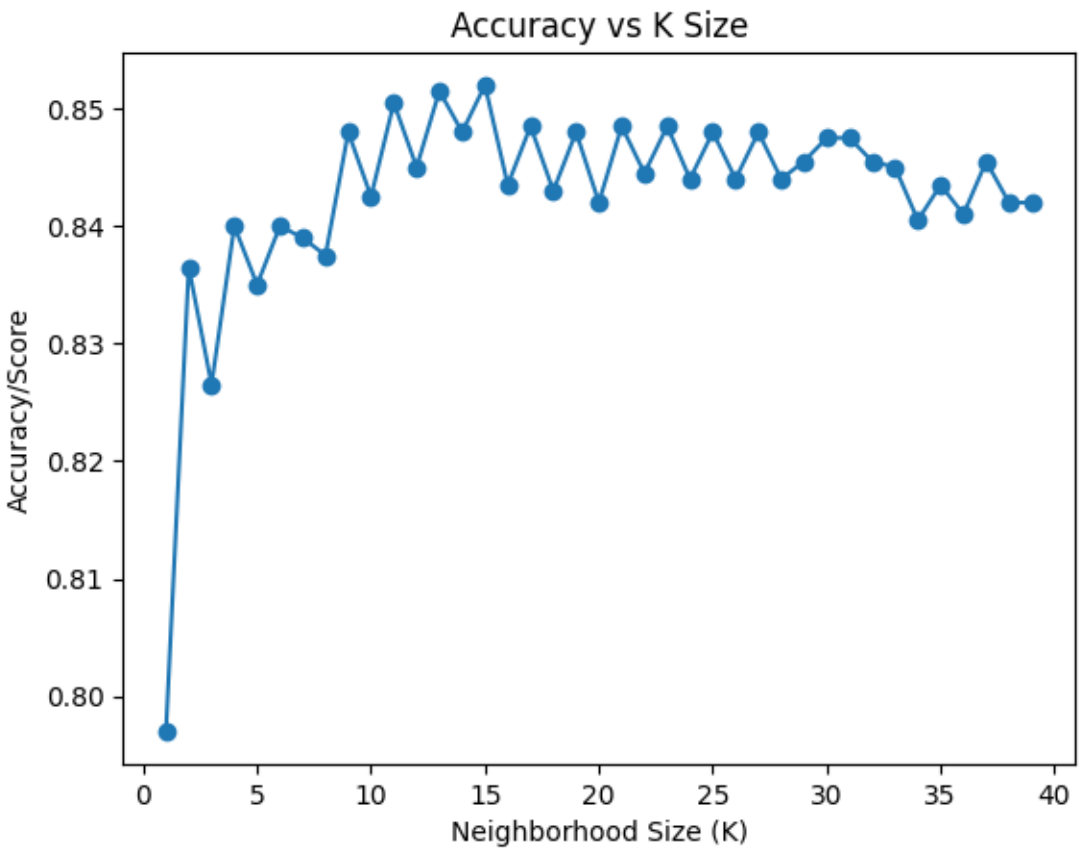
macro avg	0.81	0.73	0.76	2000
weighted avg	0.86	0.87	0.86	2000

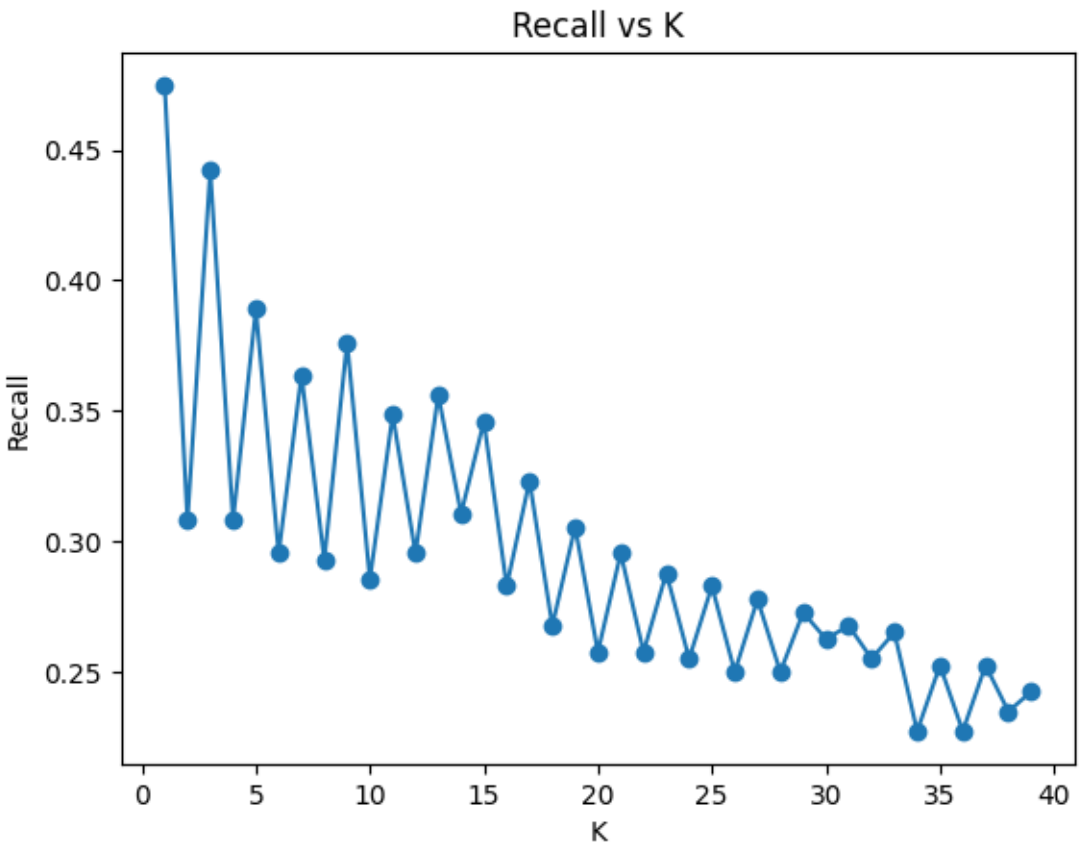




Refitting of Models

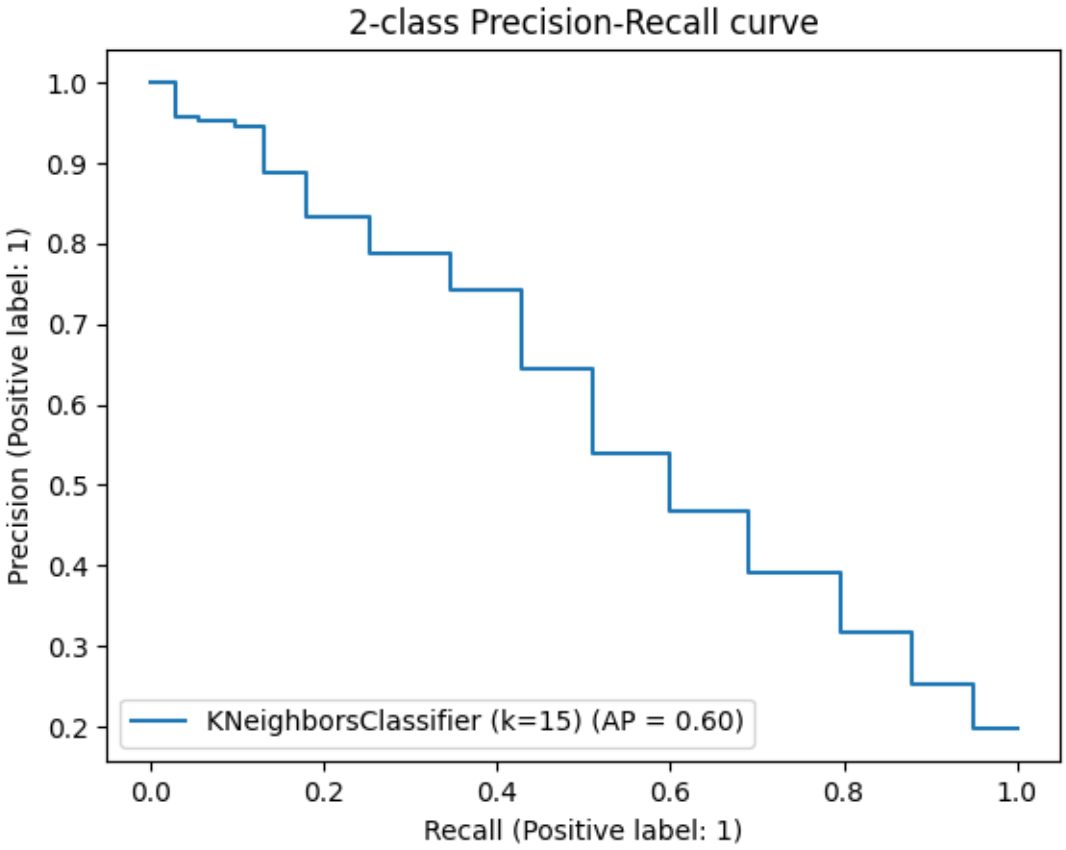
KNN Classifier

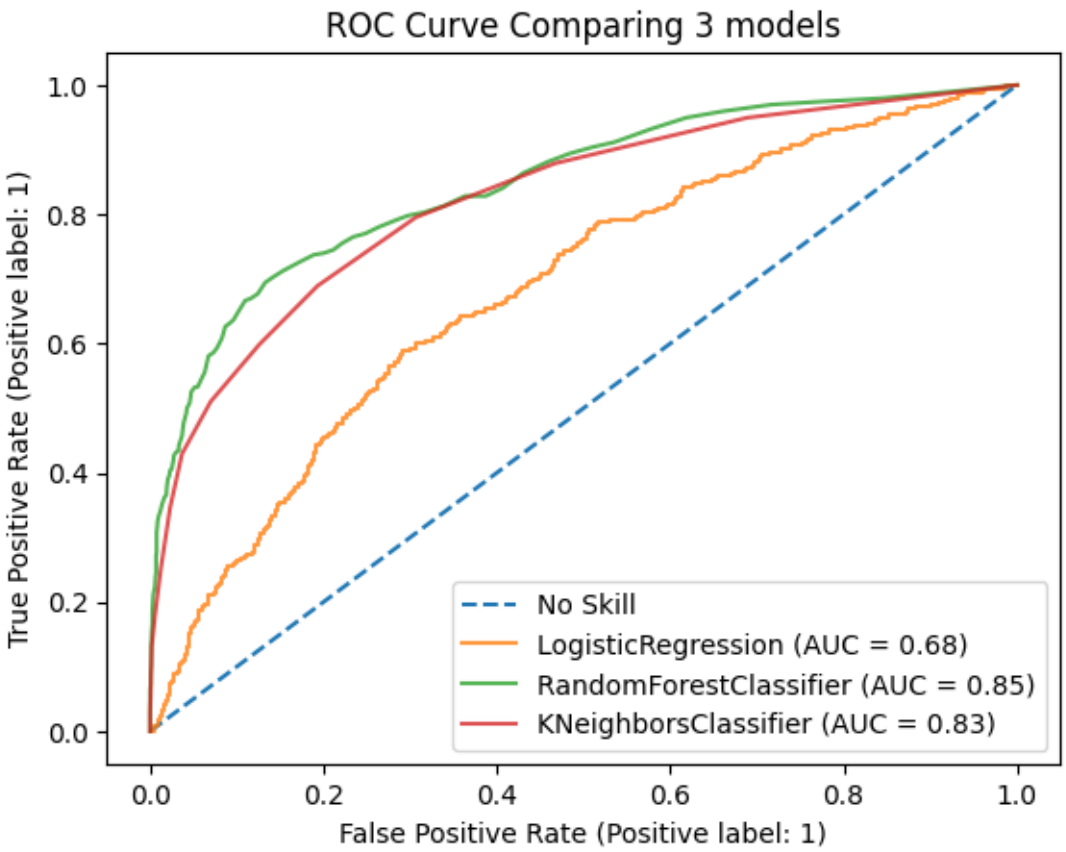

Accuracy vs K Size
Accuracy/Score
0.85
0.84
0.83
0.82
0.81
0.80
0
5
10
15
20
25
30
35
40
Neighborhood Size (K)




```
KNN Results when k=15
Accuracy/Score is 0.852
[[1567   37]
 [ 259  137]]
              precision    recall  f1-score   support

           0       0.86      0.98      0.91      1604
           1       0.79      0.35      0.48       396

    accuracy                           0.85      2000
   macro avg       0.82      0.66      0.70      2000
weighted avg       0.84      0.85      0.83      2000


KNN Results when k=1
Accuracy/Score is 0.797
[[1406  198]
 [ 208  188]]
              precision    recall  f1-score   support

           0       0.87      0.88      0.87      1604
```

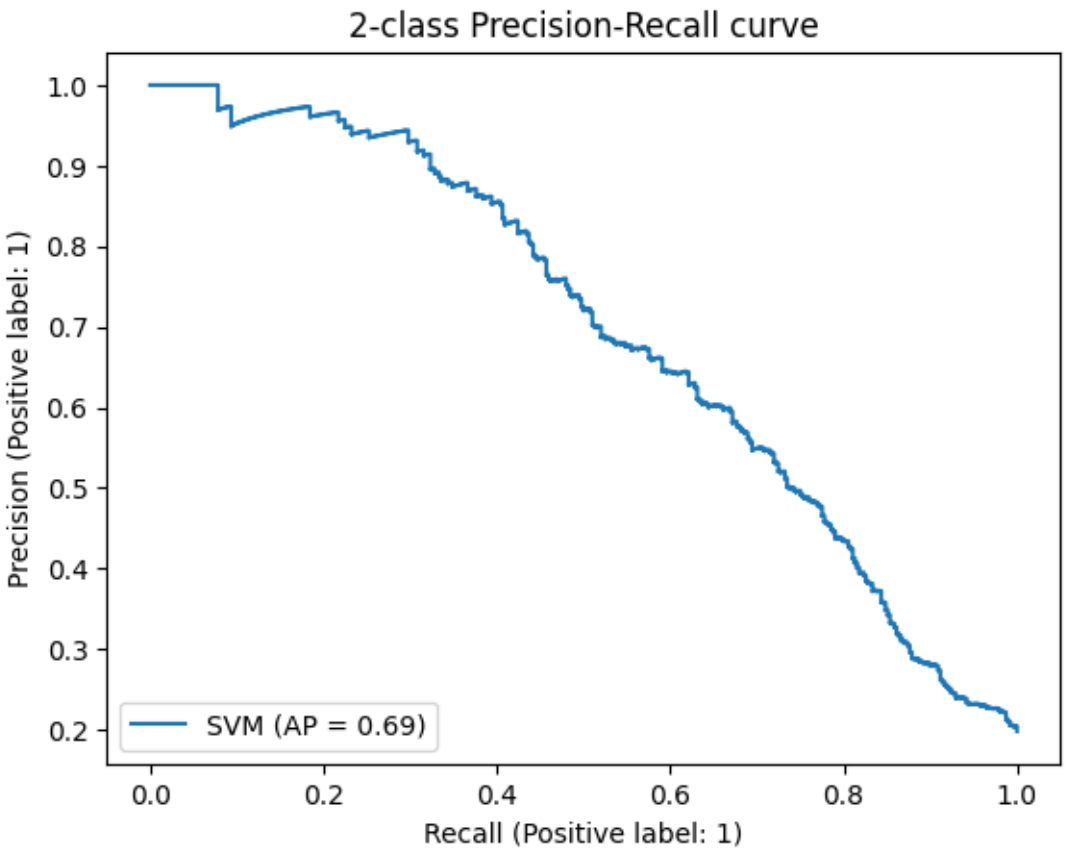
```
                1       0.49      0.47      0.48       396

         accuracy                           0.80      2000
        macro avg       0.68      0.68      0.68      2000
     weighted avg       0.80      0.80      0.80      2000
```

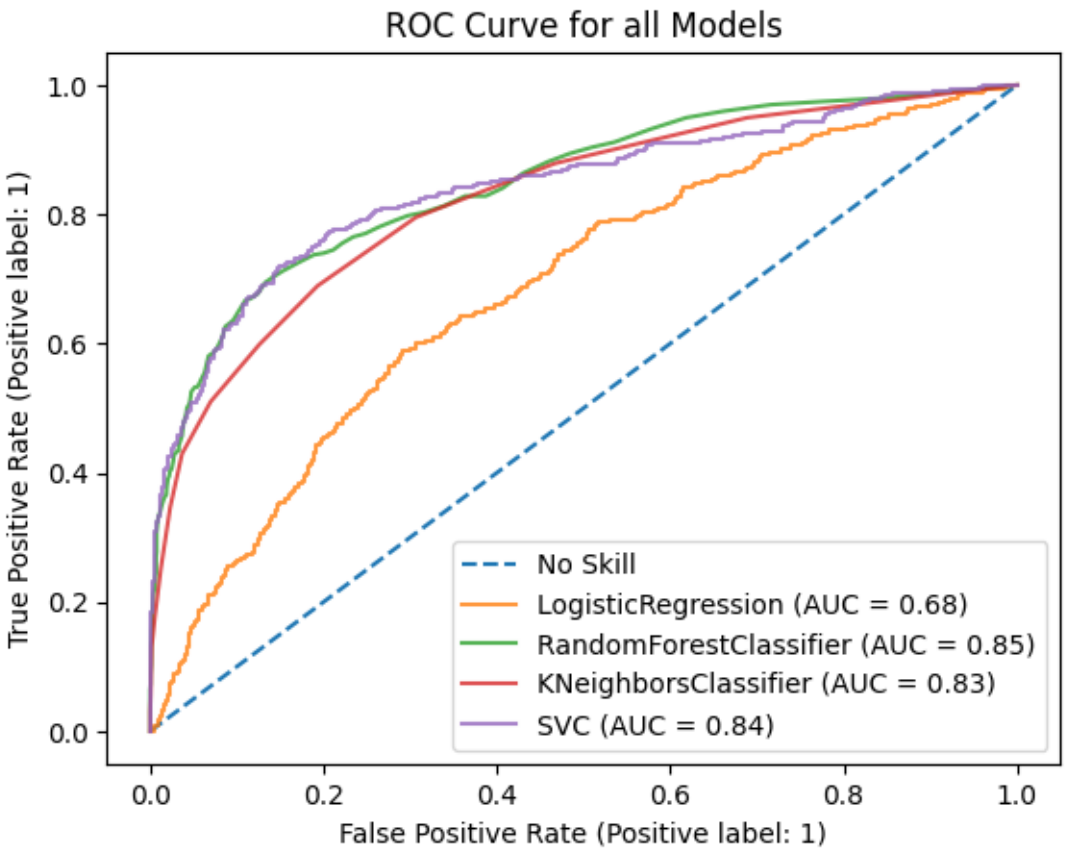




Support Vector Machine

```
SVM Results
Accuracy/Score is 0.869
[[1570   34]
 [ 228  168]]
              precision    recall  f1-score   support

           0       0.87      0.98      0.92      1604
           1       0.83      0.42      0.56       396

    accuracy                           0.87      2000
   macro avg       0.85      0.70      0.74      2000
weighted avg       0.86      0.87      0.85      2000
```


2-class Precision-Recall curve
Precision (Positive label: 1)
Recall (Positive label: 1)
SVM (AP = 0.69)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.2
0.4
0.6
0.8
1.0



Evaluation and Final Results

Conclusion

Citations

- "Bank Customer Churn", https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset
- "The Value of Keeping the Right Customers", Amy Gallo, https://hbr.org/2014/10/the-value-of-keeping-the-right-customers