# Banking Insurance Product – Phase 2

Blue 18

Ryan Heggie
Ishi Gupta
Lucy Liu
August Majtenyi
Aaryan Sharma

November 18, 2024

# Table of Contents

# Banking Insurance Product – Phase 2

## Overview

The Commercial Banking Corporation (the "Bank") sought proposals to predict customers' likelihood of purchasing a variable rate annuity, hereafter called the product. Our team previously predicted the likelihood of buying the product with logistic regression and generalized additive models.

In this report, we evaluated two models. One model used Random Forest (RF), and the other used eXtreme Gradient Boosting (XGBoost). After tuning both models and evaluating variable importance, we assessed the models' goodness-of-fit metrics using their receiver operating characteristic (ROC) curves. The XGBoost model had an area under the curve (AUC) of 0.801, while the RF had an AUC of 0.886.

## Methodology

The following section describes the data used for analysis and how the machine-learning models were created.

### *Data Used*

The Bank provided data about customers who were offered the product. The Bank included a training data set with 8,495 observations and 37 predictor variables relating to the customer's account. We checked for missing values. We used median imputation to fill null values for continuous variables and mode imputation for categorical variables.

### *Random Forest*

Our team created an RF model to capture the variability of a customer buying the product. To obtain true estimates without overfitting our model, we tuned the RF with different variable subsets. We also looked at the importance of our variables and built a final model based on this information.

### *XGBoost*

Our team created an XGBoost model to help minimize the error in predicting the probability of a customer buying the product. To reduce over-fitting our model, we tuned attributes relating to the number of tree nodes, the proportion of subsamples to cross-validate, and the optimal learning rate. Furthermore, we looked at variable importance and built a final model based on this information.

## Analysis

The following section showcases model formation and performance.

### *Random Forest*

After running the initial model and performing variable selection, we identified three key explanatory variables that provide the most predictive value for whether a customer will purchase the product. These variables were selected based on their strong importance in the model, indicating that they offered more relevant information than random variables in predicting customer behavior. Our results can be found in Table 1 below:

**Table 1: Variable importance ordered by increased node purity**

| Variable Name | Percent Increase in MSE | Increase Node Purity |
|---|---:|---:|
| Saving account balance | 87.766% | 801.495 |
| Checking account balance | 35.801% | 736.937 |
| Branch of Bank | 16.564% | 260.969 |

The table above shows that the selected variables—saving account balance, checking account balance, and branch of bank—performed better than the random variable. This indicates that these variables positively contributed to the model's predictive power. The "Increase Node Purity" column reflects the importance of these variables in improving the decision tree's ability to make better splits, leading to increased model accuracy.

The "Percent Increase in MSE" column further confirms that excluding these variables would positively affect the model's performance. Notably, the saving account balance variable has the most positive impact on the model, highlighting its critical role in ensuring predictive accuracy.

We evaluated the tuned RF model with variable selection on our training data. The AUC of this model was 0.886. Figure 1 demonstrates this value graphically.

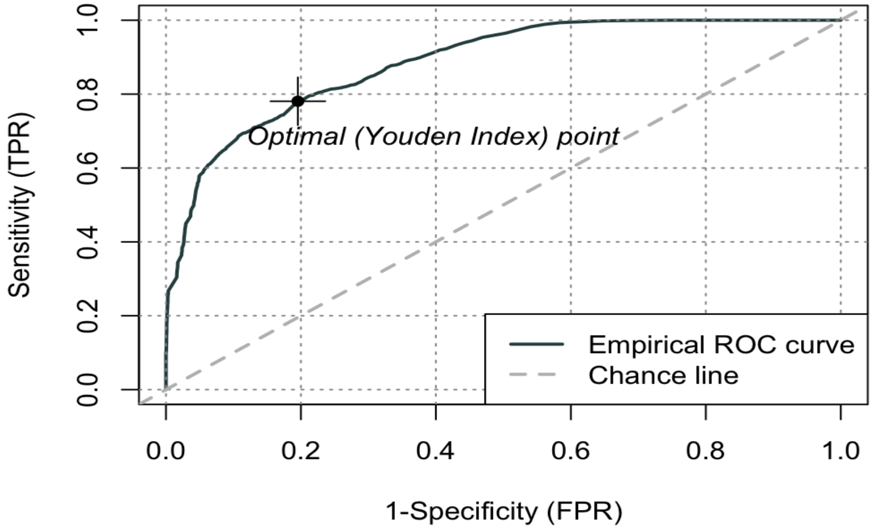


**Figure 1: ROC curve of the RF after variable selection and tuning**

The curve above along the diagonal line shows that the RF model's classification performance outperformed random guessing. The AUC achieved was 0.886, with an optimal cut-off of 0.284.

## *XGBoost*

After variable selection and initial tuning, we found a final AUC of 0.801. We analyzed variable importance in model building and ranked the top variables by gain. Our results can be found in Table 2.

**Table 2: Variable importance by gain in the tuned XGBoost.**

| Variable Name | Gain |
|---|---:|
| Saving account balance | 0.284 |
| Checking account balance | 0.133 |
| Having a checking account | 0.080 |
| CD Balance | 0.071 |
| MM Balance | 0.056 |

The variables above were used in the final XGBoost model. Table 2 demonstrates more gain than a random variable when added to the training data. All other variables were clustered with the random variable and did not contribute to our model. We removed these variables from the model and tuned the XGBoost over multiple iterations.

To tune our XGBoost model, we tested different ranges of learning rate, max depth, and subsample. We determined the optimal combination of these metrics based on AUC. The different tuning parameters can be found in Table 3.

**Table 3: Tuning Parameters tested in the XGBoost model**

| Tune Parameter | Lower Bound | Upper Bound | Best Tune |
|---|---:|---:|---:|
| Learning rate | 0.01 | 0.3 | .04 |
| Subsample | 0.25 | 1 | .43 |
| Max depth | 1 | 10 | 4 |
| Cross-validation folds | 10 | 10 | 10 |
| Number of rounds | 100 | 100 | 100 |
| All other parameters | default | default | default |

Table 3 shows the optimal values of learning rate 0.04, max depth 4, and subsample 0.43 for our final model. Exploring alternative values for the number of rounds and cross-validation folds may result in a different final model.

We evaluated the final XGBoost model on our training data. The AUC of this model was 0.801. Figure 2 demonstrates this value graphically.

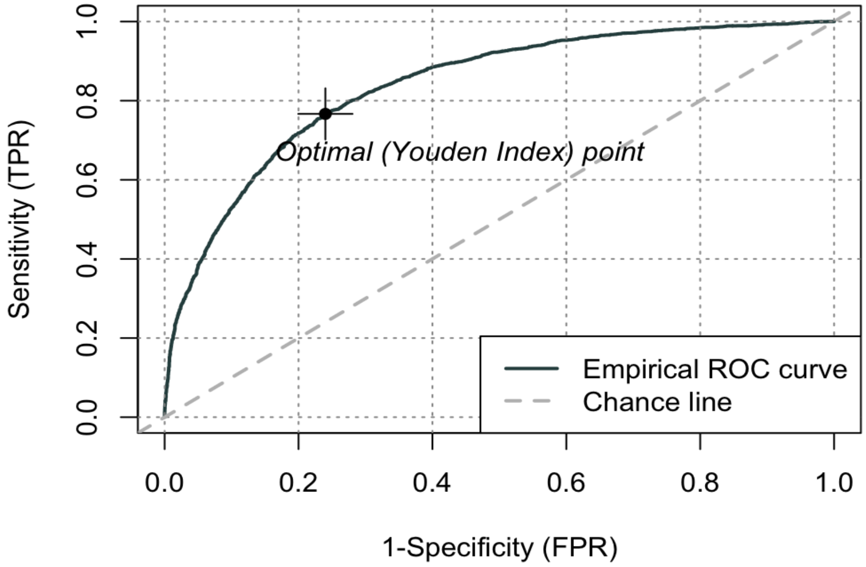


**Figure 2: ROC curve of XGBoost after variable selection and tuning**

The optimal cutoff point for this model was 0.346. That said, the appropriate cutoff value for this model should be changed based on business context and cost considerations.

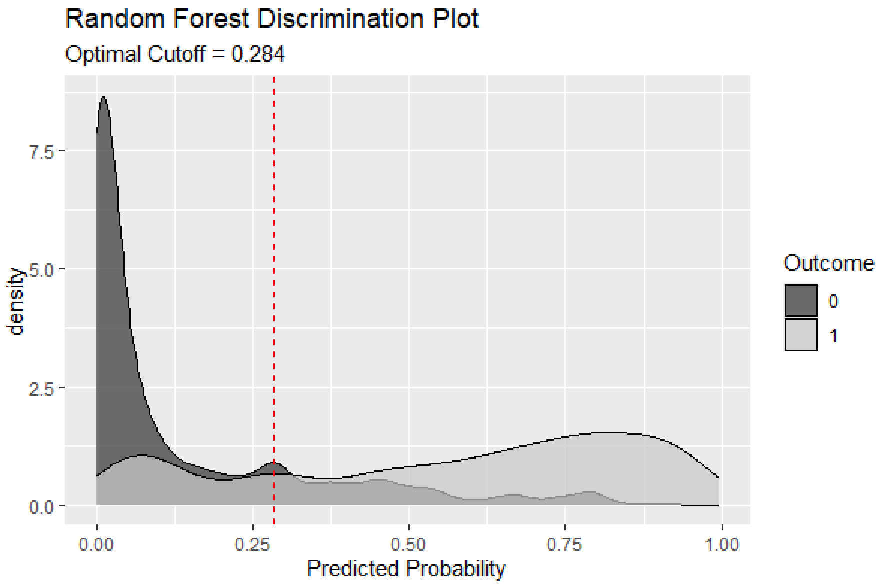
## Results & Recommendations

Since our RF model had a higher AUC than the XGBoost, we recommend using the RF model to predict whether a customer bought the product. Since two of the three most important variables relate to account balance, we recommend the Bank further investigate how these account balances impact product purchases. Due to limitations in RF, we cannot reasonably explain these relationships. In the future, we could utilize more explainable models to better describe how the Bank could target potential customers.

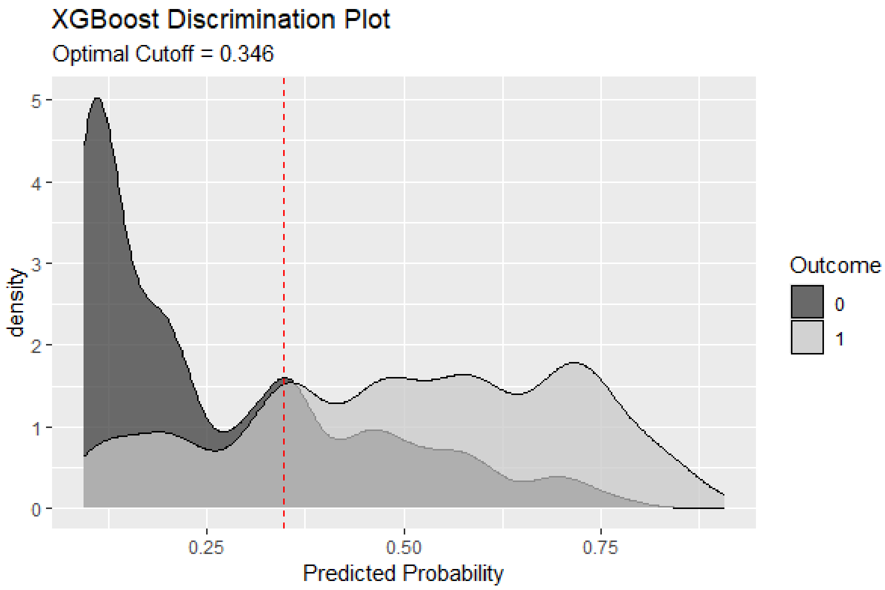
## Conclusion

In this report, we compared two different models: an RF and XGBoost. Since the RF outperformed the XGBoost model, we recommend the Bank use the RF model to capture the complex relationships in predicting product purchases. For the RF, we found an optimal cutoff of 0.284 and an AUC of 0.886. Our team identified significant variables – such as branch of bank and balances of various accounts – that were important for predicting product purchases. Therefore, the Bank should investigate how these variables relate to product purchases.

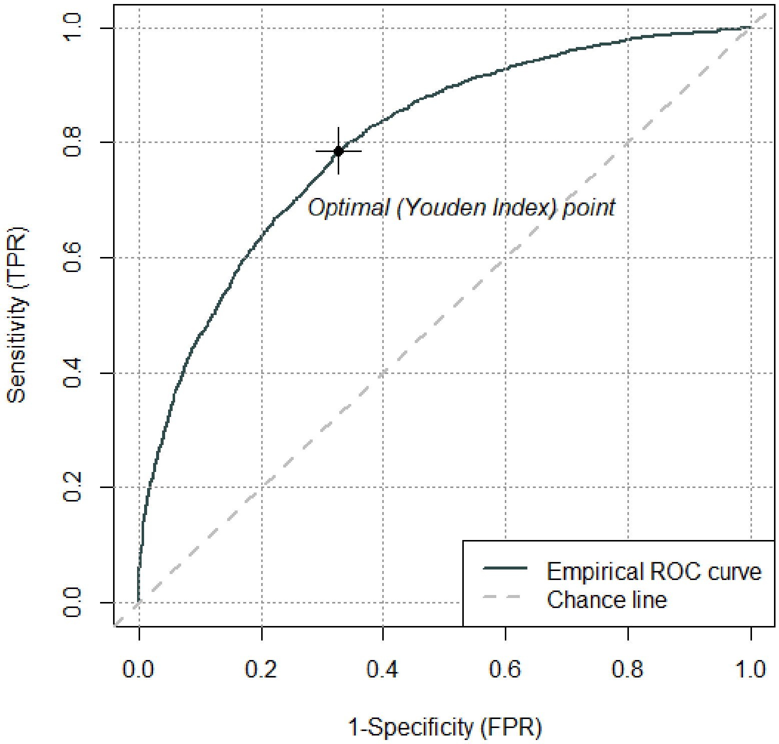
# Appendix



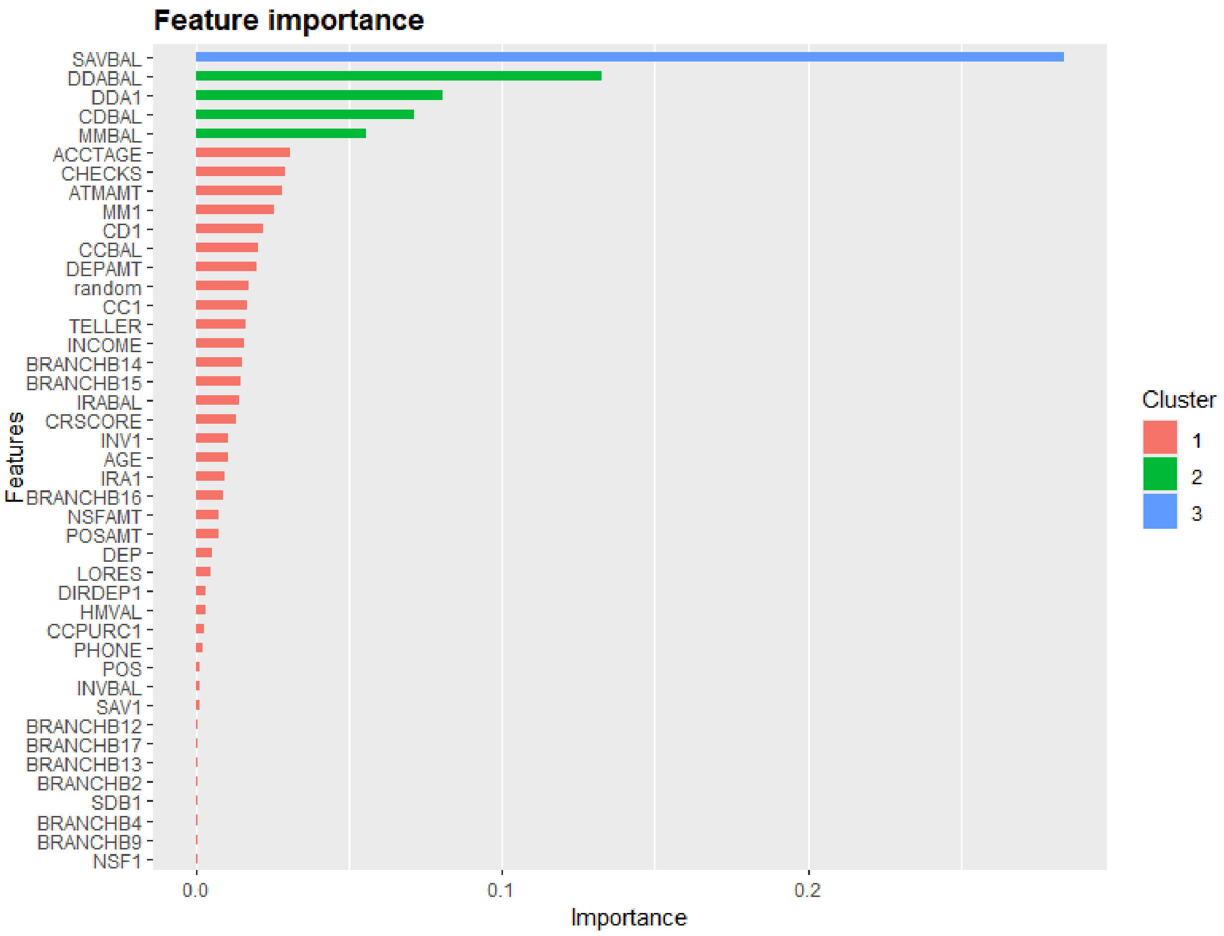
**Appendix Figure 1: Discrimination Plot of Random Forest**



**Appendix Figure 2: Discrimination Plot of Random Forest**



**Appendix Figure 3: ROC curve of XGBoost before variable selection**



**Appendix Figure 4: Feature importance of a tuned XGBoost model with the inclusion of a random variable**