Variable-Rate Annuity Purchase Prediction: Phase Two

Orange Team 11

Alex Raum
Elliott Reece
Andrew Tammaro
Rachel Tan
Landon Wilson

November 19, 2021

Table of Contents

Overview **1**

Methodology & Analysis **1**

- Data Used 1
- Random Forest 1
- XGBoost 2

Results **2**

- Random Forest 2
- XGBoost 3

Recommendations **4**

Conclusion **4**

Appendix **5**

Variable Annuity Purchase Prediction: Phase Two

Overview

The Commercial Banking Corporation (the Bank) requested proposals to build models that predict which customers will purchase a variable rate annuity product. Continuing from Phase 1 of the proposal, our team built and tuned two random forest models and two gradient boosted (XGBoost) models seen in Table 1 below. We determined that the XGBoost model using all predictors had the highest area under the receiver operating characteristic (AUROC) curve at 0.8414.

Table 1: Random Forest and XGBoost Models Evaluated

Model Name	AUROC
Random Forest Model 1: All predictors	0.7865
Random Forest Model 2: Predictors with a positive mean decrease in accuracy from Random Forest Model 1	0.7915
XGBoost Model 1: All predictors	0.8414
XGBoost Model 2: Predictors selected in comparison to random variable	0.8376

Methodology and Analysis

Data Used

The data for Phase 2 of this project consisted of 8,495 observations and 38 variables. The data contained continuous and categorical variables to be used to predict the target variable. In this dataset, the target variable was whether or not customers purchased an annuity product. The other variables consisted mainly of financial product indicators and balances. A number of variables contained missing values. This was addressed by imputing the median of the column containing the missing values while including an indicator variable for the imputed, continuous variables. We created a separate “missing” category to represent missing values for categorical variables. Table 2 in the Appendix shows the number of missing observations per variable imputed in the dataset. Additionally, we added a random variable in order to compare the relative importance of the predictor variables.

Random Forest

Per the Bank’s request, our team generated a random forest model to provide more accurate predictions of annuity purchases. Using all of the predictors in the dataset, we tuned the model parameters for the number of trees and the number of variables considered for each split. We found that the error flattened out around 200 trees and that the ideal number of variables considered for each split was six.

We built Random Forest Model 1 using these parameters and all of the predictor variables including the indicator variables for imputed data and the random variable added to assess variable importance. We then examined variable importance using the mean decrease in accuracy metric for the variables in this model. Additionally, we visualized the AUROC as shown in Figure 4 in the Appendix.

In examining variable importance for our initial random forest model, we noted that six variables had a negative mean decrease in accuracy: the random variable, the indicator for imputed credit score, the indicator for imputed age of oldest account, the indicator for local address, the length of residence in years, and the indicator for safety deposit box. We built Random Forest Model 2 without these variables, using the same parameters as our initial model. Next, we looked at variable importance for this model using the mean decrease in accuracy metric and visualized the AUROC.

XGBoost

Our team also examined the prediction performance of two XGBoost models in identifying customers likely to buy the annuity product. We first used cross validation with XGBoost to find the number of trees that maximized the test AUC. Next, we used grid search to tune the eta value, maximum depth of each tree, and subsample size. After visualizing the XGBoost grid search results, we selected 11 trees, an eta value of 0.25, a maximum depth of five, and a subsample size of one as the parameter values to minimize the error rate.

Using these parameters, we built XGBoost Model 1 with all of the predictor variables including the indicator variables for imputed data and the random variable added to assess variable importance. We then examined variable importance using the metric of gain. We also compared the importance of each variable to the random variable. Finally, we visualized the AUROC for this model.

Next, we built XGBoost Model 2 including only the variables with a higher variable importance, measured by gain, than the random variable and using the same parameters as XGBoost Model 1. We examined the variable importance for the variables in this model measured by gain and visualized the AUROC for this model. Refer to Figure 5 in the Appendix for details of the AUROC.

Results

Random Forest

Our Random Forest Model 1, containing 200 trees and splitting on six variables, had an AUROC of 0.7865.

We examined variable importance for all of the variables in Random Forest Model 1 and then performed variable selection for Random Forest Model 2 by removing all variables from the model that had a negative mean decrease in accuracy. Our team ranked each variable in Random Forest Model 2 by the mean decrease in accuracy; this ranking is displayed in Table 3 of the Appendix. After making these adjustments, our Random Forest Model 2 had an AUROC of 0.7915, displayed in Figure 1.

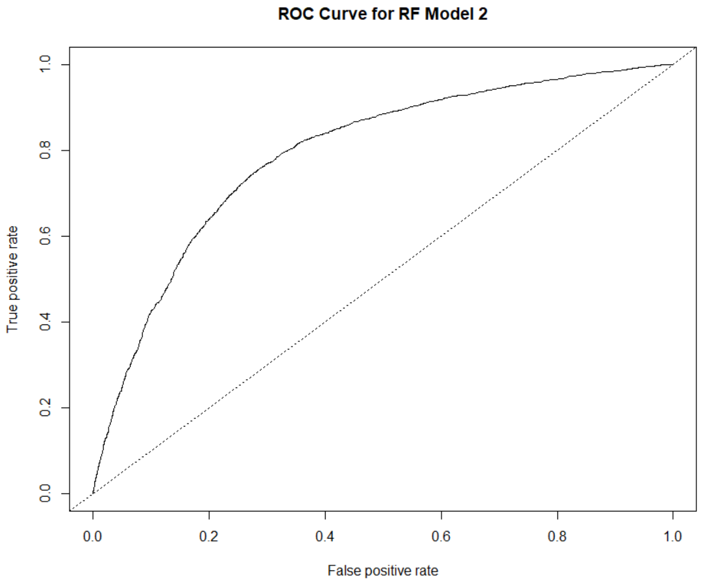

Figure 1: Random Forest Model 2 ROC Curve

XGBoost

Our XGBoost Model 1, containing all predictor variables and our tuned parameters, resulted in an AUROC of 0.8414, an improvement from our random forest models.

Upon assessing variable importance on our XGBoost Model 1, we determined that there were a total of 17 variables identified as being more important than the random variable included in the model. These variables and their corresponding importance are displayed below in Figure 2.

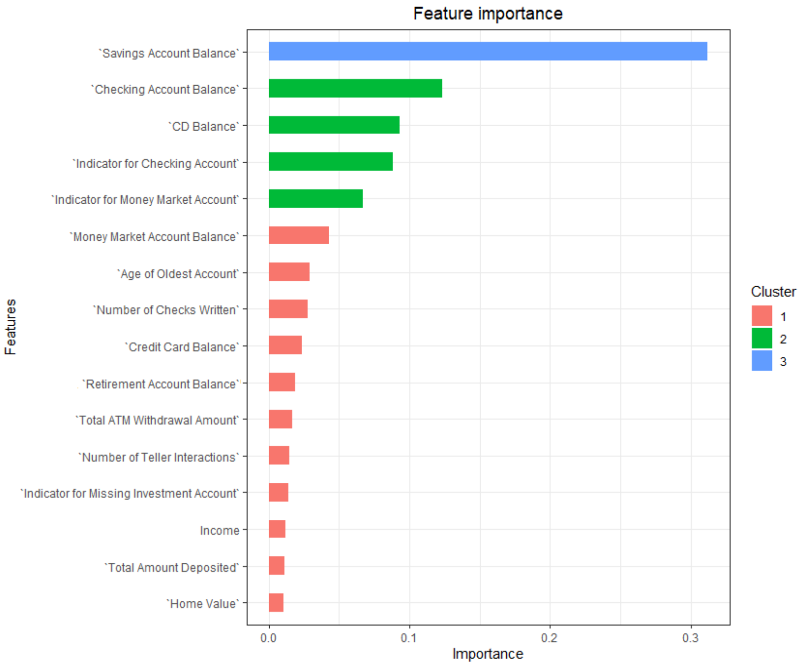

Figure 2: XGBoost Model 1 Variable Importance Relative to Random Variable

We determined XGBoost Model 1, containing all predictor variables, to have an AUROC of 0.8414. The plotted ROC curve for this model is displayed in Figure 3. In addition, the individual importance, measured by gain, for each variable in this model is included in Table 4 of the Appendix.

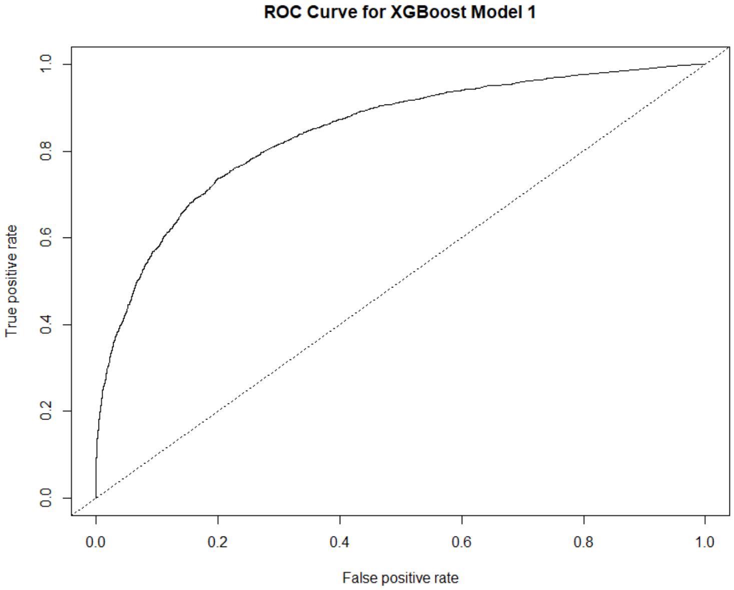

Figure 3: XGBoost Model 1 ROC Curve

Recommendations

Based on our analysis of the random forest and XGBoost models, we recommend using the XGBoost Model 1 to predict which customers will purchase the annuity product.

When considering all previous models we have developed (GAM and MARS), we also recommend using the XGBoost Model 1, assuming that predictive power is the most important feature of the model.

Conclusion

The Bank requested models that would provide predictions for which customers would purchase the annuity product. We ran two random forest models and two XGBoost models during this phase. The XGBoost Model 1 with all predictor variables had an AUROC of 0.8414, which was the highest of the four models ran during this phase. It was also higher than the GAM and MARS models considered during Phase 1. As a result, we recommend that the Bank use XGBoost Model 1 to predict which customers will purchase the annuity product.

Appendix

Table 2: Number of Imputed Variables

Variable	Missing Observations	Variable Type
Age of Oldest Account	546	Continuous
Telephone Banking Interactions	1075	Continuous
Point of Sale Interactions	1075	Continuous
Total Amount for Point of Sale Interactions	1075	Continuous
Indicator for Investment Account	1075	Categorical
Investment Account Balance	1075	Continuous
Indicator for Credit Card	1075	Categorical
Credit Card Balance	1075	Continuous
Number of Credit Card Purchases	1075	Categorical
Income	1537	Continuous
Length of Residence in Years	1537	Continuous
Home Value	1537	Continuous
Age	1702	Continuous
Credit Score	195	Continuous

Table 3: Random Forest Model 2 Variable Importance

Variable Name	Mean Decrease in Accuracy
Savings account balance	35.3384
Checking account balance	30.5596
CD balance	19.5320
Total amount deposited	17.8749
Total ATM withdrawal amount	15.5309
Bank branch	14.8996
Number of checks written	14.7720
Money market account balance	14.1496
Indicator for certificate of deposit account	13.8233
Number of checking deposits	13.8216
Indicator for credit card	13.1304
Indicator for investment account	12.1074
Indicator for retirement account	11.6966
Retirement account balance	11.4483
Indicator for money market account	11.0882
Credit card balance	10.4852
Number of credit card purchases	9.5843
Indicator for checking account	8.6944
Age of oldest account	8.2844
Home value	8.0998
Indicator for savings account	7.7055
Total amount for point of sale interactions	6.8945
Income	6.6800
Investment account balance	6.3833
Indicator for imputed total amount for point of sale interactions	6.2335
Indicator for imputed value for number of phone interactions	5.8753
Indicator for ATM interaction	5.8716
Number of teller visit interactions	5.8246
Number of point of sale interactions	5.7373

Indicator for imputed age	5.0989
Indicator for imputed credit card balance	4.7304
Indicator for imputed investment account balance	4.2983
Number of phone interactions	3.4202
Number of money market credits	3.2988
Indicator for imputed number of point of sale interactions	3.2837
Indicator for imputed income	2.5342
Age	1.9249
Indicator for imputed home value	1.4588
Indicator for direct deposit	1.1342
Indicator for imputed length of residence	0.8171
Number of insufficient fund issues	0.7888
Credit score	-0.1287
Amount of insufficient fund issues	-0.6044

Table 4: XGBoost Model 1 Variable Importance

Variable Name	Gain
Savings account balance	0.3077
Checking account balance	0.1272
CD balance	0.0902
Indicator for checking account	0.0850
Indicator for money market account	0.0666
Money market account balance	0.0447
Age of oldest account	0.0331
Number of checks written	0.0274
Credit card balance	0.0190
Retirement account balance	0.0190
Total ATM withdrawal amount	0.0168
Number of teller interactions	0.0158
Indicator for investment account (Missing)	0.0156
Income	0.0120
Total amount deposited	0.0109
Home value	0.0105
Random variable	0.0102
Indicator for credit card (1)	0.0101
Credit score	0.0085
Indicator for investment account (1)	0.0079
Age	0.0079
Number of checking deposits	0.0078
Length of residence	0.0062
Bank branch (16)	0.0055
Indicator for retirement account	0.0047
Bank branch (15)	0.0035
Number of credit card purchases (1)	0.0033
Total amount of point of sale interactions	0.0027
Bank branch (6)	0.0027

Bank branch (14)	0.0021
Investment account balance	0.0018
Number of insufficient fund issues	0.0016
Bank branch (12)	0.0014
Bank branch (18)	0.0014
Amount of insufficient fund issues	0.0014
Indicator for imputed age of oldest account	0.0013
Indicator for certificate of deposit account	0.0013
Bank branch (17)	0.0012
Indicator for direct deposit	0.0011
Indicator for imputed income	0.0007
Indicator for security deposit box	0.0007
Number of credit card purchases (2)	0.0006
Number of money market credits	0.0004
Number of phone interactions	0.0003
Bank branch (2)	0.0001

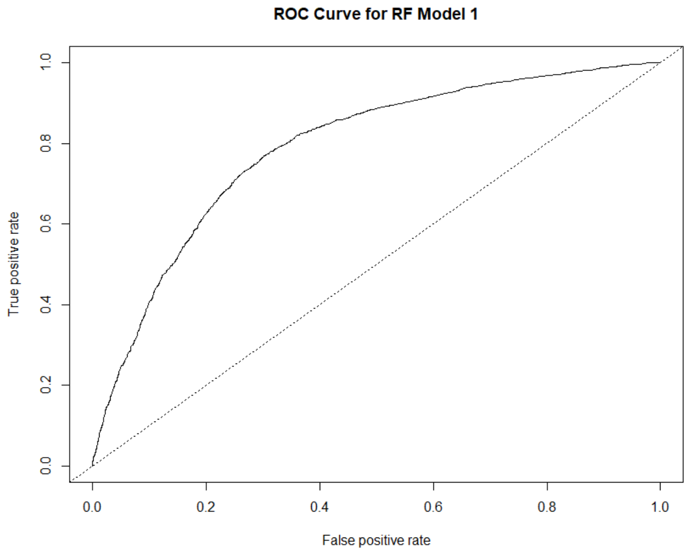

Figure 4: Random Forest Model 1 ROC Curve

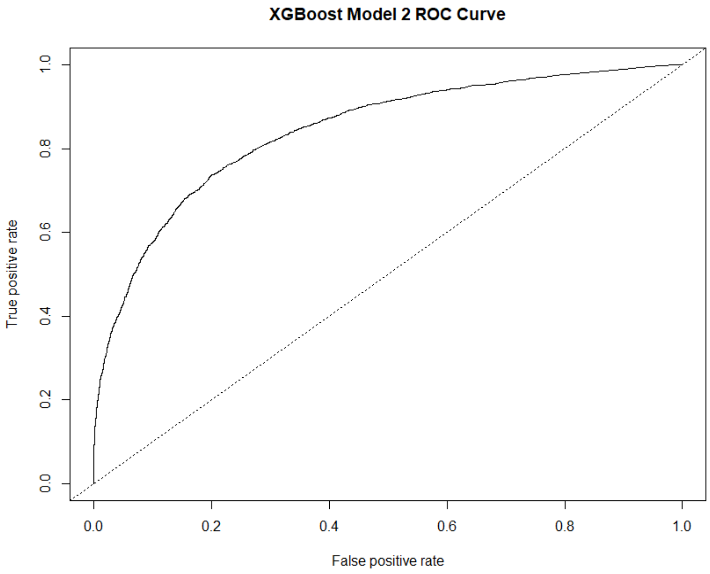

Figure 5: XGBoost Model 2 ROC Curve

Homework Report Checklist

The team member(s) responsible for checking each item should enter their initials in the field next to each question. All items should be addressed before submitting the assignment with the signed checklist attached.

Sections & Structure

Overview

ER	Is the overview concise?
ER	Does it provide context about the business problem?
ER	Does it briefly address your team's work, quantifiable results, and recommendations?
ER	Does it offer audience-centered reasons for recommendations?

Body Sections

ER	Does the report body include information on methods, analysis, quantifiable results, and recommendations?
ER	Is content grouped into appropriate sections (*methodology*, *analysis*, *results*, *recommendations*)?

Conclusion

ER	Does the report have a conclusion?
ER	Does the conclusion sum up the report and emphasize relevant takeaways?

Structure

ER	Does each major section have a heading?
ER	Are sections, subsections, and paragraphs organized logically for easy navigation?

Visuals

Introduction, Discussion, and Captions

AR	Is each visual introduced in the text before it appears?
AR	Is each visual close to where it is introduced?
AR	Does each visual include a title with the following information: type (*table* or *figure*), number, and a descriptive caption?
AR	Is each visual discussed and interpreted in the text?
AR	Are figures and tables numbered separately?
AR	Are table captions above the table? Are figure captions below the figure?

Visual Design

AR	Do figures/tables use audience-friendly labels rather than variable names?
AR	Are the visuals easy to interpret?
AR	Are the visuals appropriately sized?
AR	Do tables appear on one page (*not split between 2 pages*)?
AR	Are legends and axis labels included for figures?
AR	Are numbers in tables right aligned?
AR	Are the visuals designed well (*ex*: *re-created in Word or Excel, not blurry or stretched,…*)?

Document Design

Title Page Design

DT	Does it include a descriptive title?
DT	Does it state the team name, team members' names, and the submission date?

Table of Contents Design

DT	Does it list all the major sections of the report with corresponding page numbers?
DT	Do the page numbers and sections in the Table of Contents match the report?

Document Design for Entire Report

DT	Is a standard typeface (*Calibri*, *Arial*, *etc.*) used?
DT	Is the size of the body text between 10-12 pt.?
DT	Are headings and subheadings used to organize information?
DT	Are distinctive text styles (*bold*, *italic*, *etc.*) used to distinguish between heading levels?
DT	Are text styles for headings used consistently (*ex*: *all level-one headings are bold*)?
DT	Are all paragraphs an appropriate length (*fewer than 12 lines*)?
DT	Is white space used to indicate paragraph breaks?
DT	Are bullet lists used for a series of items and numbered lists to show a hierarchy?

Writing Style and Mechanics

Spelling and Capitalization

LW	Are spelling errors located and corrected?
LW	Is spelling consistent throughout (*no switching between acceptable spellings*)?
LW	Is capitalization used appropriately (*proper nouns*, *etc*.)?
LW	Is capitalization of words consistent throughout the report?

Grammar and Punctuation

LW	Are verb tenses used appropriately?
LW	Are marks of punctuation used appropriately?
LW	Is subject-verb agreement used in every sentence?
LW	Is the grammar checker updated and are underlined grammar issues addressed?

Writing Style

RT	Are all sentences in the report easy for your audience to understand quickly?
RT	Are most sentences written in active voice?
RT	Are idioms and vague words (*there*, *here*, *etc*.) eliminated from the report?
RT	Are acronyms introduced before being used?
RT	Are well-written topic sentences included at the beginning of each paragraph?
RT	Are lists parallel?
RT	Is the appropriate point of view used when addressing your audience or describing team actions?