VARIABLE-RATE ANNUITY PURCHASE PREDICTION: PHASE THREE

ORANGE TEAM 11

ALEX RAUM
ELLIOTT REECE
ANDREW TAMMARO
RACHEL TAN
LANDON WILSON

DECEMBER 1, 2021

Table of Contents

VARIABLE ANNUITY PURCHASE PREDICTION: PHASE THREE

OVERVIEW

The Commercial Banking Corporation (the Bank) requested proposals to build models that predict which customers will purchase a variable rate annuity product. In Phase 1 of the proposal our team built generalized additive models (GAM) and multivariate adaptive regression splines (MARS) predictive models, and in Phase 2 of the proposal our team built tree-based models using random forest and XGBoost. In the final phase of our analysis, we built a neural network model to compare the predictive power against previous models as seen in Table 1. Using all predictors, the neural network model had an area under the receiver operating characteristic (AUROC) curve of 0.8145.

We determined that the best predictive model overall was the XGBoost model using all predictors, which had an AUROC of 0.8414 on the training dataset. Assessing the XGBoost model's predictive power on the validation dataset yielded an AUROC of 0.7930. Lastly, based on our analysis of the global impact of account age on the probability of annuity purchases, we recommend that bank agents focus on marketing the annuity product to customers with account ages less than three years old since newer customers are more likely to purchase the annuity product.

Table 1: Model Comparison of Training AUROC

Phase Built	Model Name	AUROC on Training
1	MARS Model: All predictor variables	0.7999
1	GAM Model: Predictors identified in MARS variable importance	0.8002
2	Random Forest Model: Predictors with a positive mean decrease in accuracy	0.7915
2	XGBoost Model: All predictors	0.8414
3	Neural Network Model: All predictors	0.8145

METHODOLOGY AND ANALYSIS

DATA USED

The training data for Phase 3 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. Several variables contained missing values. We addressed this 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 to compare the relative importance of the predictor variables.

The validation data for this phase consisted of 2,124 observations having the same 38 variables as the training dataset. We used an analogous procedure to that of the training data for imputing missing values. However, missing continuous variables in the validation data were imputed using the median value of these variables in the training dataset.

NEURAL NETWORK

Per the Bank's request, our team created an additional model. In doing so, we trained a neural network model to predict annuity purchases. We scaled each continuous variable to reduce the variance impact on the model. Using all of the predictors in the dataset, we optimized performance by tuning the model parameters for regularization and the number of nodes in the hidden layer.

FINAL MODEL - XGBOOST

In addition to the neural network model, we also reexamined models developed in prior phases. After comparing these model's performance, evaluated by AUROC on the training data, we selected the XGBoost model as our final model. The parameters of this model included 11 trees, an eta value of 0.25, a maximum depth of five, and a subsample size of one. We then fit the XGBoost model to the validation dataset and obtained the AUROC to assess its accuracy.

To assess variable importance in our final model, we added a normally distributed random variable. We compared each variable's importance measured by gain to the random variable's gain. Table 3 in the Appendix shows each of the variables in the final XGBoost model ranked by importance.

GLOBAL INTERPRETATION

After obtaining an accuracy measure for our final model, we created a partial dependence plot to examine the global relationship between the age of a customer's oldest account and the predicted probability of purchasing the annuity product.

RESULTS

NEURAL NETWORK

We created the neural network model using scaled continuous predictors and categorical predictors. This resulted in an AUROC of 0.8145 on the training dataset, which was lower than our best performing model from previous phases.

FINAL MODEL - XGBOOST

Our XGBoost model from Phase 2 had an AUROC of 0.8414 on the training dataset. The AUROC on the training dataset was 0.8414, which was the highest out of all created models, including GAM, MARS, random forest, XGBoost, and neural network models. As a result, we ran our model on the validation dataset and found an AUROC of 0.7930. The ROC curve is shown in Figure 1.

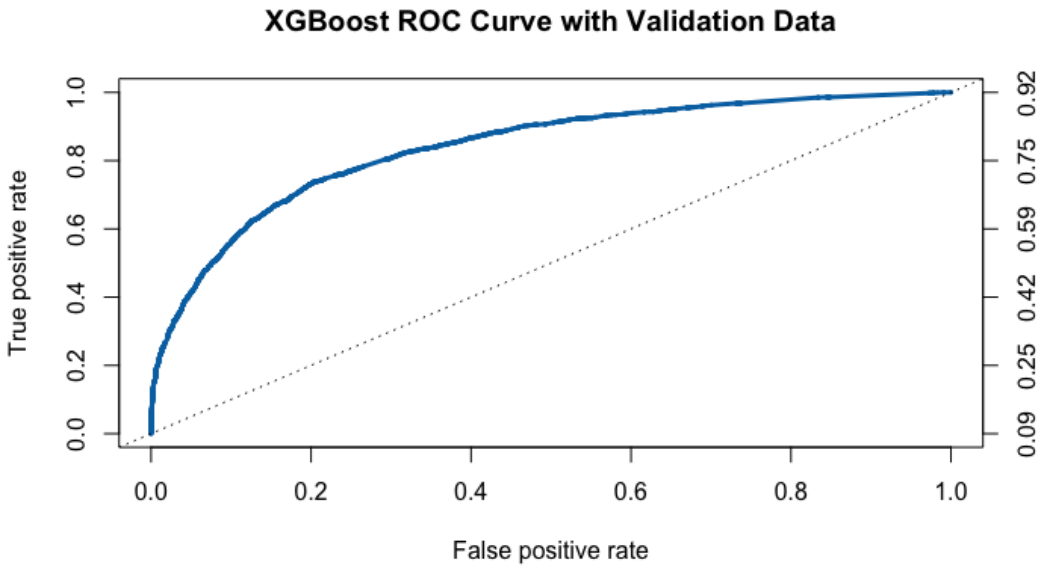

Figure 1: XGBoost Model ROC Curve for the Validation Data

The Bank requested insight on the global impact of the variable account age. Our team looked at the partial dependence plot to understand how age affected the probability of a customer purchasing the annuity product. The partial dependence plot is shown below in Figure 2. The general trend is that newer accounts have the highest probability of purchasing the annuity product. There is a sharp decline in probability around accounts that are three years old, after which the trend levels out beyond five to ten years.

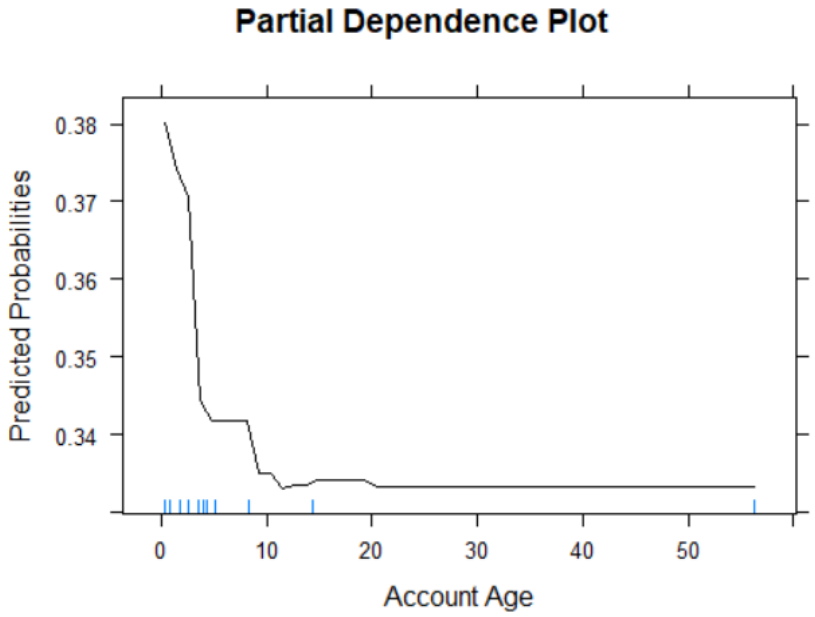

Figure 2: Partial Dependence Plot for Account Age in the XGBoost Model

RECOMMENDATIONS

After creating and tuning a neural network model during Phase 3, our team recommends using the XGBoost Model with all predictors created during Phase 2 to best predict which customers will purchase the annuity product. This model had an AUROC of 0.7930 on the validation dataset.

Additionally, based on the partial dependence plot created for the variable account age, we recommend that the Bank focus its sales efforts on customers whose oldest accounts are less than three years old. Newer customers are more likely to purchase the annuity product than older customers, as can be seen by the rapidly decreasing probability as account age increases. New customers might be more active with their accounts and financial decisions. We also noted that the majority of the Bank's current customers have oldest accounts with ages less than ten years, which supports our recommendation. Therefore, a successful plan for the Bank would be to target new account owners with annuity product marketing campaigns. Although we noted that the scale of the differences in predicted probabilities is relatively small, marketing to the newer customers will still have a marginal impact.

CONCLUSION

The Bank requested models that would provide predictions for which customers would purchase the annuity product. The XGBoost model had the highest predictive performance on the training dataset, and the model had an AUROC of 0.7930 on the validation dataset. We recommend using this model to help the Bank predict which customers will purchase the annuity product. Additionally, we assessed the global impact of account age on the probability of annuity purchases. Based on our findings, we recommend that bank agents focus on marketing the annuity product to customers with accounts that are less than three years old.

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: XGBoost Model 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

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

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

BODY SECTIONS

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

CONCLUSION

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

STRUCTURE

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

VISUALS

INTRODUCTION, DISCUSSION, AND CAPTIONS

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

VISUAL DESIGN

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

DOCUMENT DESIGN

TITLE PAGE DESIGN

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

TABLE OF CONTENTS DESIGN

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

DOCUMENT DESIGN FOR ENTIRE REPORT

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

WRITING STYLE AND MECHANICS

SPELLING AND CAPITALIZATION

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

GRAMMAR AND PUNCTUATION

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

WRITING STYLE

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