Variable-Rate Annuity Purchase Prediction: Phase One

Orange Team 11

Alex Raum
Elliott Reece
Andrew Tammaro
Rachel Tan
Landon Wilson

November 12, 2021

Table of Contents

Variable Annuity Purchase Prediction: Phase One

Overview

The Commercial Banking Corporation (the Bank) is interested in predicting which customers will buy its variable rate annuity product. The Bank previously focused on better understanding the factors related to purchasing its product and is now interested in correctly predicting these customers. Our team evaluated one multivariate adaptive regression splines (MARS) model and two generalized additive models (GAM). The GAM model using the variables identified by variable importance in the MARS model had the highest area under the receiver operating characteristic curve (AUROC) value of 0.8002 (Table 1).

Table 1: MARS and GAM Models Evaluated

Model Name	AUROC
MARS Model: All Predictor Variables	0.7999
GAM Model 1: Predictors Identified in MARS Model 1 Variable Importance	0.8002
GAM Model 2: Categorical Predictors from MARS Model 1 with Spline Selection for All Continuous Predictors	0.7995

Methodology and Analysis

Data Used

The data for Phase 1 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 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 3 in the Appendix shows the number of missing observations per variable imputed in the dataset.

MARS

Per the Bank’s request, our team generated a MARS model, using piecewise regression to provide more accurate predictions. Our MARS Model included all predictors in the data set. We then assessed the ranked output of variables listed by their importance. Additionally, we visualized the AUROC for this model.

GAM

In addition to the model generated using the MARS procedure, our team also examined the prediction performance of two GAM models in identifying customers likely to buy the annuity product. We performed variable selection for the first of the two models using the variables that the MARS Model flagged as being important. The second model was fit using the same categorical variables as the first model, but with spline selection applied to all continuous variables in our dataset. We then visualized the AUROC for these models.

Results

MARS

The MARS Model we trained resulted in an AIC of 8,764 and R-squared of 0.244. Table 2 presents important variables identified by the MARS Model. MARS variable importance ranked the variables by the number of models they appeared in and by the scaled decrease in residual sum of squares (RSS) compared to the previous subset. We included the entire list of important variables identified by the MARS Model in Table 4 of the Appendix.

Table 2: MARS Model Variable Importance

Rank	Variable	Number of Models Containing this Variable	Decrease in RSS relative to the previous subset (scaled)
1	Savings Balance	22	100.0
2	CD Balance	20	67.5
3	Checking Account Indicator	20	66.9
4	Checking Account Balance	20	66.9
5	Money Market Balance	18	47.7
6	Missing Indicator for Investment Account	16	39.8
7	Age of Oldest Account	15	36.1
8	Number of Checks Written	12	31.3
9	Number of Teller Visit Interactions	11	29.7
10	Total ATM Withdrawal Amount	10	27.5

We next plotted the ROC curve for this model as visualized in Figure 1. The AUROC for this model is 0.7999.

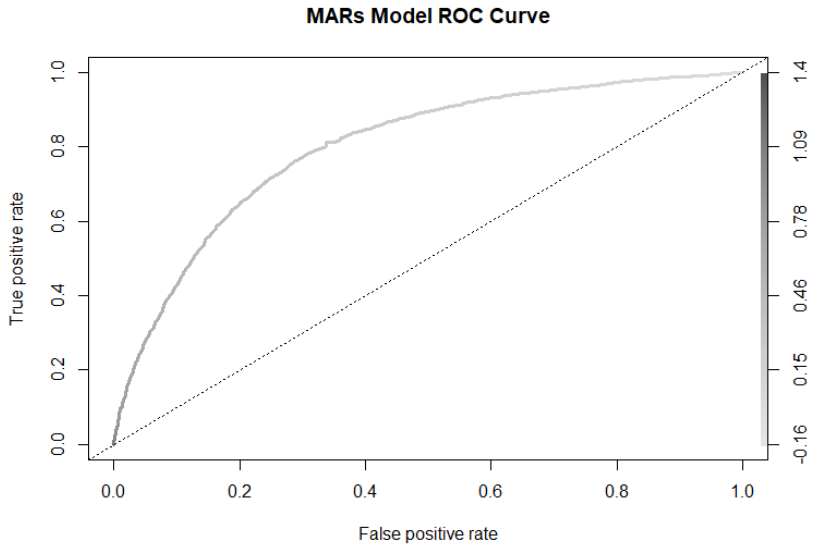

Figure 1: MARS Model ROC Curve

GAM

We trained GAM Model 1 using the 16 variables included in Table 4 in the Appendix. We plotted the ROC curve for this model as visualized in Figure 2. The AUROC for this model was 0.8002, slightly higher than the MARS Model. GAM Model 2, which included spline selection for all continuous predictor variables in the dataset, generated an AUROC of 0.7995 as presented in Figure 3 in the Appendix.

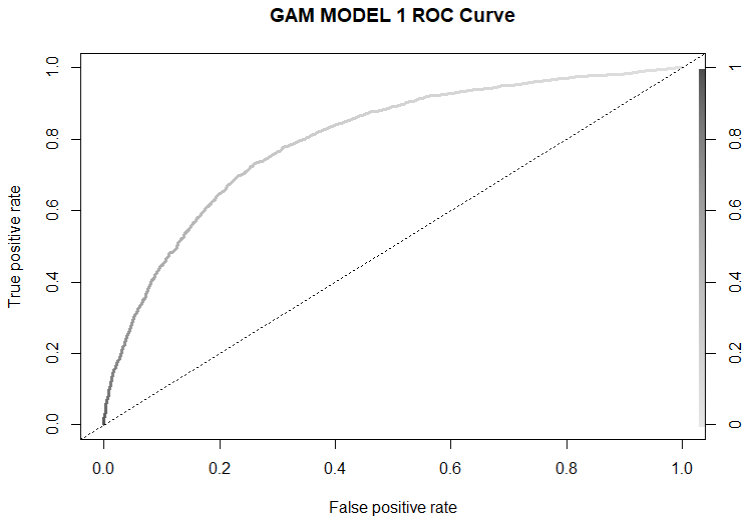

Figure 2: GAM Model 1 ROC Curve

Recommendations

Based on our analysis of the MARS and GAM models, we recommend using GAM Model 1 to predict which customers will buy the annuity product. Doing so will improve the predictive capabilities of other methods like logistic regression from previous project phases. Tree-based methods may also increase predictive capabilities as well.

We also recommend that the Bank consider using cross-validation for model building. By performing cross-validation, the final model will be more generalizable than a model created using the entire training set at once. The benefit for the Bank is that the model's predictions will be more accurate when using new unseen data.

Conclusion

The Bank requested a predictive model to predict which customers will purchase the annuity product. We investigated three different predictive models. GAM Model 1, which used the important variables identified with the MARS Model, had the highest AUROC value of 0.8002. This model is simpler than other models that included all continuous predictors and still had the highest AUROC. Overall, our team recommends using GAM Model 1 to predict which customers will buy the Bank's annuity product.

Appendix

Table 3: 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 4: Complete List of Ranked Important Variables from the MARS Algorithm

Rank	Variable	Number of Models Containing this Variable	Decrease in RSS relative to the previous subset (scaled)
1	Savings Balance	22	100.0
2	CD Balance	20	67.5
3	Checking Account Indicator	20	66.9
4	Checking Account Balance	20	66.9
5	Money Market Balance	18	47.7
6	Missing Indicator for Investment Account	16	39.8
7	Age of Oldest Account	15	36.1
8	Number of Checks Written	12	31.3
9	Number of Teller Visit Interactions	11	29.7
10	Total ATM Withdrawal Amount	10	27.5
11	Investment Account Indicator	9	25.1
12	Credit Card Indicator	8	22.4
13	Credit Card Balance	7	20.2
14	Branch 16	6	17.3
15	IRA Balance	5	14.0
16	Number of Checking Deposits	2	8.2

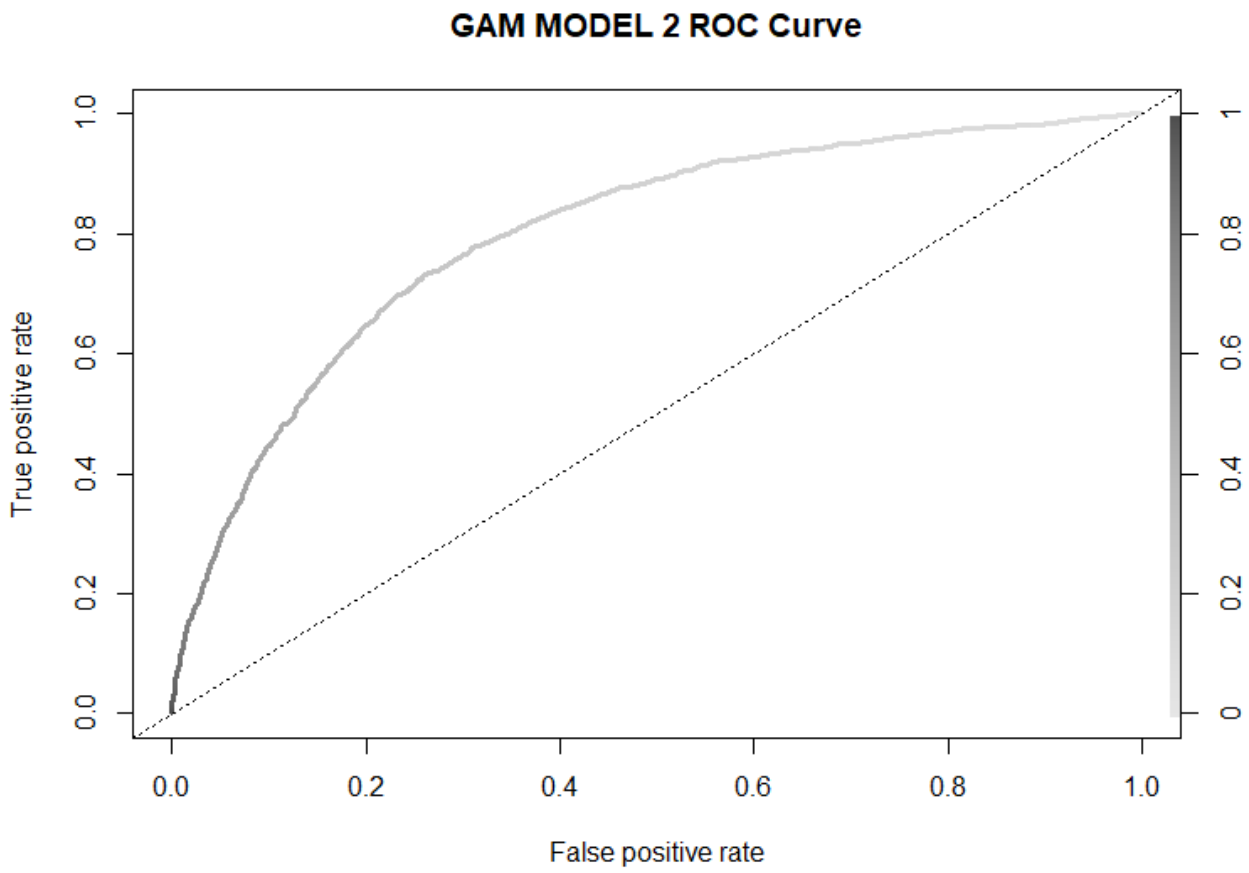

Figure 3: GAM 2 Model ROC Curve

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

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

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

Grammar and Punctuation

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

Writing Style

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