CMC Banking Phase 1: MARS and GAMs

COMMERCIAL BANKING, CORP

Orange 1
Madison Burns
Macy Faw
Phil McDonough
Jacob Stanley
Jason Wang

November 8, 2024

Table of Contents

CMC Banking Phase 1: MARS and GAMs

Overview

The Commercial Banking Corporation (hereafter "the Bank") has introduced a variable rate annuity product and engaged our team to predict which customers will likely purchase this product. Accurate prediction of prospective buyers will allow the Bank to optimize marketing efforts, target high-potential customers, and allocate resources efficiently.

We implemented two advanced analytical models to address this business challenge: the Multivariate Adaptive Regression Splines (MARS) algorithm and the Generalized Additive Model (GAM). We chose the MARS algorithm—for its flexibility in handling nonlinear relationships and interactions and GAM—to incorporate smooth splines on continuous variables for better fit and predictive power. Both models are evaluated based on their classification accuracy and the area under the ROC curve (AUC).

Although the GAM model achieved a marginally higher AUC than the MARS model, both models showed comparable predictive performance. We recommend implementing the GAM model as the primary predictive tool for broader targeting, with the potential to combine it with MARS in an ensemble approach to improve accuracy further. This strategy will enable the Bank to enhance its outreach to potential customers, potentially increasing annuity sales by more effectively identifying prospective buyers.

Methodology & Analysis

This section of the report covers key aspects of our study, including the data used, an assessment of the MARS algorithm, and an evaluation of our GAM model.

Data Used

We received a customer dataset from the Bank with 8,495 observations and 38 variables—37 predictors and one target INS, which takes a value of 1 if they bought the annuity and 0 if they did not buy. To address missing values, we imputed the median for continuous variables and the mode for categorical variables to preserve the data's overall pattern.

MARS (EARTH)

We applied the MARS algorithm utilizing the earth package in R to predict customers likely to purchase a variable rate annuity product. This algorithm provides the capability to handle nonlinear relationships and interactions effectively. Additionally, since our goal for this project is binary classification, we specified our EARTH model to incorporate a binomial generalized linear model. Finally, EARTH automatically performs variable selection to prune less impactful variables. Table 1 lists the 17 most important variables for predicting our target variable, INS.

Table 1: Variable importance ordered for EARTH model

Order of Importance	Variable
1	Savings account balance
2	Certificate of deposit balance
3	Indicator for checking account
4	Checking account balance
5	Money market balance
6	Indicator for credit card (yes)
7	Age of oldest account
8	Number of checks written
9	Indicator for investment account
10	Number of teller visit interactions
11	Total ATM withdrawal amount
12	Credit card balance
13	Bank branch 15
14	Bank branch 14
15	Bank branch 16
16	Individual retirement account balance
17	Indicator for a savings account

Table 1 presents the variable importance ranking for the EARTH model used to predict the likelihood of purchasing a variable rate annuity. The model identifies key predictors, with higher importance placed on account balances and account-related indicators, such as savings and certificate of deposit balances, checking account indicators, and the age of the oldest account. While factors like bank branch also contribute, they are relatively less influential compared to account balance and transaction-based features.

GAM

A GAM was employed to predict the likelihood of customers purchasing the Bank's variable rate annuity. We incorporated all relevant predictors and used smoothing splines for continuous variables to capture nonlinear effects.

An iterative selection process refined the GAM model, starting with splines for all continuous variables. Variables with p-values above 0.5 were then removed to retain only those that meaningfully contributed to the target variable. The removal process involved eliminating three variables at a time based on the

highest p-values, with model performance re-evaluated after each step. The final selected variables are listed in Table 2 in no particular order.

Table 2: Variables included in the GAM model (unranked)

Variables		
Checking deposits	Age of oldest account	Indicator for an investment account
Savings account balance	Credit card balance	Indicator for retirement account
Checking account balance	Indicator for an investment account	Number of insufficient fund issues
Certificate of deposit balance	Total amount for point-of-sale interactions	Number of telephone banking interactions
Money market balance	Indicator for checking account	Number of credit card purchases
Number of checks written	Total amount deposited	Indicator for direct deposit
Total ATM withdrawal amount	Length of residence in years	Indicator for savings account
Indicator for ATM interaction	Indicator for certificate of deposit	Indicator for money market account
Money market balance	Indicator for safety deposit box	Indicator for local address
Number of teller visit interactions	Value of home	Branch of bank

The variables in Table 2 were retained in the final GAM model after our variable selection and included customer data and key financial and behavioral variables.

Results

To assess classification performance, we used the ROC curve and K-S statistic on the training data to set optimal cutoffs (0.304 for MARS and 0.301 for GAM). Figure 2 displays the ROC curves for the training data set using MARS (left) and GAM (right).

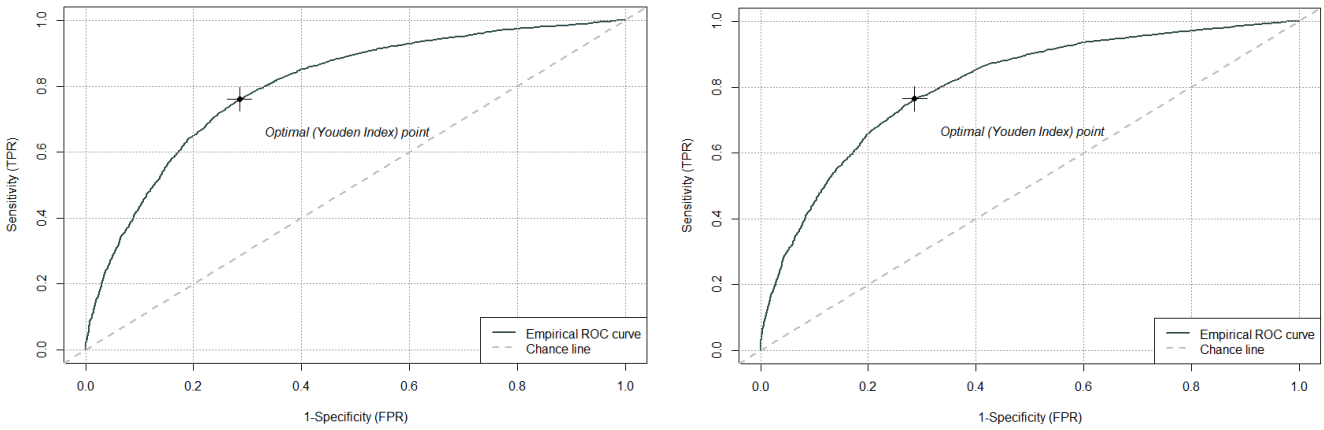

Figure 1: ROC curves for training dataset using MARS (left) and GAM (right)

In Figure 1, both models show similar performance, with the GAM model achieving a slightly higher AUC compared to the MARS model (0.805 vs. 0.801), suggesting that the GAM model captures finer nuances in the data. Overall, both models demonstrate strong predictive performance.

Recommendations

After modeling with both approaches, we have the following recommendations :

1. **Implement the GAM model to target customers for marketing efforts**: The GAM model outperformed MARS by 4% in AUC, correctly ranking customers who would purchase 80.5% of the time. The ROC curve identified 0.304 as the optimal cutoff, balancing purchase predictions while minimizing non-purchases, assuming equal cost. This cutoff should be adjusted based on the bank's budget and priorities. Implementing the GAM model will increase purchase conversions by reaching a broader customer base.

2. **Ensemble GAM and MARS for marketing:** Both the GAM and MARS models identified purchases about 30% better than random guessing, with GAM being more lenient in distinguishing between customers who purchased and those who did not. Using GAM for broad targeting and MARS for high-impact campaigns could offer savings. An ensemble of both models may outperform individual models by smoothing biases with averaged predictions.

3. **Variable selection using the MARS Algorithm:** The MARS algorithm captures variables differently than traditional methods like stepwise or forward/backward selection by identifying complex nonlinear relationships and interactions. Revisiting past models with these newly identified variables could improve interpretations and predictive power.

We recommend implementing an ensemble of GAM and MARS while enhancing the model's performance through an iterative process of discovery and refinement.

Conclusion

Our analysis shows that both the GAM model and MARS algorithm offer similar predictive capabilities for identifying purchasers of the Bank's variable rate annuity. Although the GAM model demonstrated slightly better performance, the similarity in AUC scores suggests that a combined approach could maximize predictive accuracy. By adopting an ensemble model strategy and fine-tuning marketing thresholds, the Bank can enhance its marketing efficiency, ultimately boosting customer acquisition for the annuity product. We recommend ongoing model refinement and variable exploration to continue improving the accuracy and impact of these predictions, ensuring that the Bank remains competitive in targeting the right customers effectively.

Homework Report Checklist

As instructed by Dr. Egan Warren, 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 initial checklist attached.

Sections & Structure

Overview

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

Body Sections

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

Conclusion

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

Structure

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

Visuals

Introduction, Discussion, and Captions

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

Visual Design

MF	Do figures/tables use audience-friendly labels rather than variable names?
MF	Are the visuals easy to interpret?
MF	Are the visuals appropriately sized?
MF	Do tables appear on one page (*not split between 2 pages*)?
MF	Are legends and axis labels included for figures?
MF	Are numbers in tables right aligned?

MF	Are the visuals designed well (*ex*: *re-created in Word or Excel*, *not blurry or stretched*,…)?

Document Design

Title Page Design

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

Table of Contents Design

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

Document Design for Entire Report

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

Writing Style and Mechanics

Spelling and Capitalization

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

Grammar and Punctuation

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

Writing Style

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