Modeling Insurance Purchase Behavior

Team 1

Lainey English,
Joschka Labinsky,
Ella Moses,
Christian Nielsen,
John-Carl Wharton

11/07/2025

Table of Contents

Modeling Insurance Purchase Behavior

Overview

The Commercial Banking Corporation (the Bank) aims to identify customers likely to purchase a variable rate annuity product. Our team built two predictive models using multivariate adaptive regression splines (MARS) and generalized additive models (GAM) to estimate each customer's likelihood of purchase. We evaluated model performance with receiver operating characteristic (ROC) curves and the area under the curve (AUC) values. Both models demonstrate strong predictive power, with the GAM model (AUC = 0.802) performing slightly better than the MARS model (AUC = 0.800). We recommend that the Bank use the final GAM model to target its marketing efforts, specifically focusing on the "high-capital" customers that both models identified as the most powerful predictor of a purchase.

Methodology & Analysis

The following section describes the dataset, data preparation steps, and the methodologies used for model building and evaluation.

Data Used

The Bank's dataset contains account and personal information on 8,495 customers, including 12 categorical and 25 continuous predictors. Fourteen variables had missing information, which we addressed using multiple imputation by chained equations (MICE), allowing for imputations based on relationships among variables. For each of the 14 variables with missing data, we added a flag variable to indicate whether imputation was performed. We also checked for multicollinearity among predictors using their adjusted generalized variance inflation factor (GVIF) and removed the 'Home Value' variable from the list of predictors, as it was the only variable with a GVIF greater than five.

Model Building

We built MARS and GAM models to predict customer purchases of the annuity product. Each of these models was initially built using all 36 predictors remaining in the data set after data preparation, as well as the 14 new columns indicating imputation. For the MARS model, we analyzed feature importance to determine what variables affect a customer's probability of purchasing the annuity product. For the GAM model, we refined the original model by removing variables that were not statistically significant (p-value > 0.05) to produce a more focused final model.

Model Evaluation

We evaluated model performance on the training data to compare the two approaches. As requested, we used the AUC to gauge how well each model distinguishes between purchasers and non-purchasers. To further compare the models, we calculated accuracy, precision, and recall. All metrics were calculated using the optimal cutoff value for each model, determined via Youden's J statistic to balance sensitivity and specificity.

Results & Recommendations

This section details the findings from both the MARS and GAM models, provides a direct comparison of their performance, and concludes with recommendations for the Bank.

MARS Results

The final MARS model is composed of 16 variables, listed in **Table 2** in the appendix and ranked in descending order by variable importance. The MARS results suggest that total available capital across multiple account types is a key factor influencing the likelihood of purchasing a variable rate annuity, as four of the five most important predictors represent account balances. In addition to these financial factors, the 'missingness indicator for telephone banking interactions' was also identified as a highly important predictor. As shown in **Figure 1**, the resulting AUC score of 0.8 for the MARS model indicates that it has strong predictive ability in distinguishing between individuals who will or will not purchase a variable rate annuity.

Figure 1 illustrates the ROC curve for the MARS model, which displays the model's discrimination performance across various thresholds.

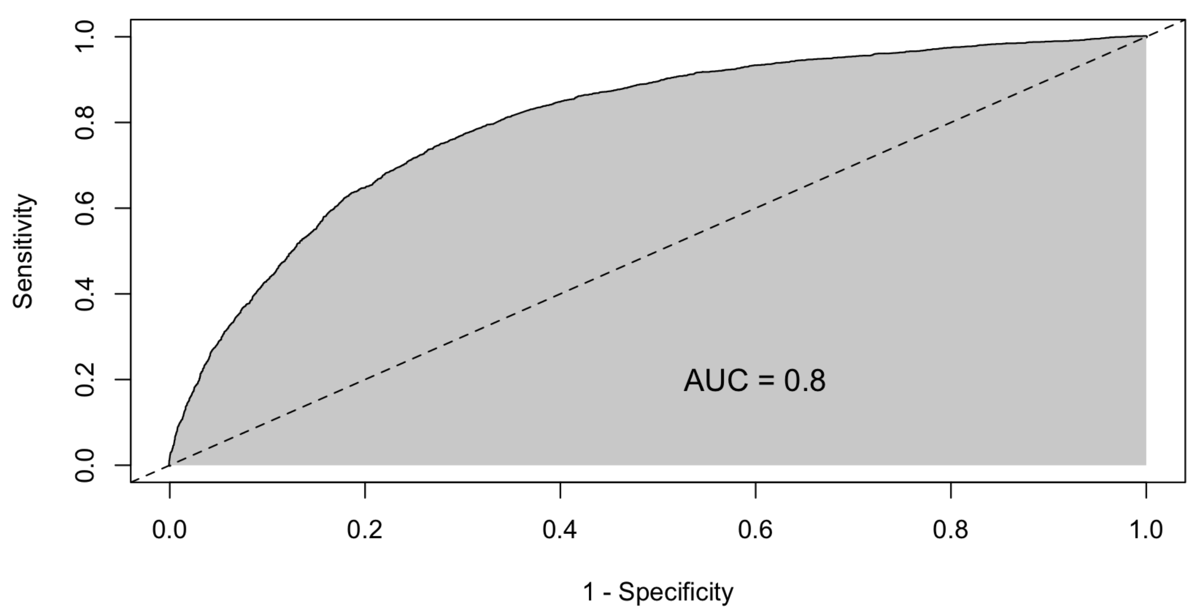

Figure 1: MARS ROC Curve

GAM Results

The final GAM model comprises 17 variables, consisting of ten continuous and seven categorical variables. All continuous variables were fit as splines, except 'Number of Insufficient Fund Issues'. Our analysis showed that this variable had too few unique values to justify a spline, so it was modeled as a linear term instead. These variables and their significance levels are shown in **Table 3** in the appendix.

Twelve of the significant variables in the GAM model are also among the most important features in the MARS model, as shown in **Table 2**. This overlap indicates that both models identified a consistent set of key predictors, reinforcing their importance in driving customer purchase likelihood. The GAM model achieves an AUC of 0.802, indicating strong predictive performance and slightly higher discrimination ability than the MARS model.

Figure 2 shows the ROC curve for the GAM model, which illustrates the trade-off between the true positive rate and false positive rate across different classification thresholds.

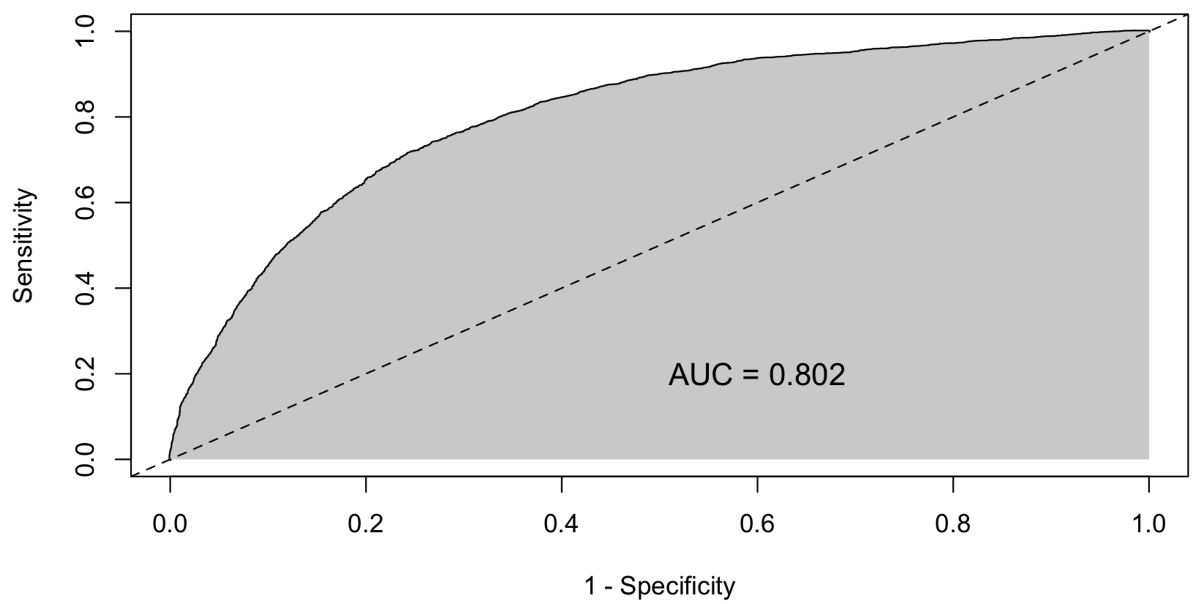

Figure 2: GAM ROC Curve

Comparison

Table 1 summarizes the performance metrics for both models. Overall, the GAM and MARS models demonstrate very similar results, with the GAM model showing marginally higher AUC, precision, and accuracy. These metrics are nearly identical, indicating comparable overall effectiveness in distinguishing purchasers from non-purchasers. All metrics were calculated using optimal cutoff values determined via Youden’s J statistic, ensuring balanced sensitivity and specificity for each model.

Table 1: Model Accuracy Metrics

Model	AUC	Optimal Cutoff	Accuracy	Recall	Precision
MARS	0.800	0.315	0.732	0.751	0.586
GAM	0.802	0.311	0.733	0.750	0.587

Recommendations

We recommend that the Bank implement the GAM model to identify customers most likely to purchase the variable rate annuity, given its stronger ability to distinguish buyers from non-buyers. While the GAM model highlights statistically significant predictors, the MARS model provides complementary insights through feature importance, identifying the variables with the greatest practical influence on purchase likelihood. Together, these results can inform targeted marketing strategies and support future data-driven decision-making. First, marketing efforts should be focused on "high-capital" customers. Both models confirm that variables such as ‘savings account balance’, ‘certificate of deposit balance’, and ‘checking account balance’ are the most powerful predictors of purchase. Second, the MARS model identified the 'missingness indicator for telephone banking interactions' as a highly important predictor. We recommend that the Bank investigate why this data is missing, as this is a potential opportunity for tailoring future marketing.

Conclusion

The GAM model demonstrated strong predictive ability with an AUC of 0.802, effectively distinguishing between purchasers and non-purchasers. Both the GAM and MARS models identified a consistent set of influential predictors, confirming the factors that drive customer purchase behavior and can help guide the Bank’s marketing efforts.

Appendix

Table 2: Variables in Final MARS Model

Variable	Type	Number of subsets	Normalized square root of generalized cross-validation	Normalized square root of residual sum of squares
Savings account balance	Continuous	22	100.0	100.0
Certificates of deposit balance	Continuous	20	66.5	67.6
Indicator for checking account	Categorical	20	65.9	67.0
Checking account balance	Continuous	20	65.9	67.0
Money market balance	Continuous	18	45.7	47.8
Missingness Indicator for telephone banking interactions	Categorical	16	37.5	39.9
Age of oldest account	Continuous	15	33.7	36.3
Indicator for investment account	Categorical	12	29.1	31.5
Number of checks written	Continuous	11	27.1	29.5
Number of teller visit interactions	Continuous	10	25.3	27.6
Total ATM withdrawal amount	Continuous	9	22.8	25.2
Indicator for credit card	Categorical	8	20.1	22.5
Indicator for branch B16	Categorical	7	17.3	19.7
Credit card balance	Continuous	6	14.3	16.9
IRA balance	Continuous	5	11.2	13.9
Checking deposits	Continuous	2	6.1	8.1

Table 3: Variables in Final GAM Model

Variable	Type	p-value
Indicator for checking account	Categorical	2.00×10^{-16}
Checking account balance	Continuous	2.00×10^{-16}
Number of checks written	Continuous	2.00×10^{-16}
Savings account balance	Continuous	2.00×10^{-16}
Total ATM withdrawal amount	Continuous	2.00×10^{-16}
Number of teller visit interactions	Continuous	2.00×10^{-16}
Branch of bank	Categorical	2.00×10^{-16}
Indicator for money market account	Categorical	2.00×10^{-16}
Indicator for credit card	Categorical	1.99×10^{-10}
Age of oldest account	Continuous	2.05×10^{-6}
Indicator for retirement account	Categorical	3.63×10^{-6}
Indicator for investment account	Categorical	4.68×10^{-5}
Indicator for certificate of deposit account	Categorical	7.56×10^{-4}
Number of telephone banking interactions	Continuous	8.96×10^{-4}
Certificate of deposit balance	Continuous	3.59×10^{-3}
Number of insufficient fund issues	Continuous	7.37×10^{-3}
Credit card balance	Continuous	1.04×10^{-2}