Comparison of Predictive Models for Variable Rate Annuity Purchases

Team 1

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

11/25/2025

Table of Contents

Comparison of Predictive Models for Variable Rate Annuity Purchases

Overview

The Commercial Banking Corporation (the Bank) aims to identify customers likely to purchase a variable rate annuity product. Our team trained five models – generalized additive model (GAM), multivariate adaptive regression spline (MARS), random forest, extreme gradient boosting (XGBoost), and explainable boosting machine (EBM) – to estimate each customer's probability of making a purchase. After evaluating the models, the random forest model proved to be the most effective at predicting a purchase, achieving an area under the receiver operating characteristic (ROC) curve (AUC) of 0.8026 and an accuracy of 71.37%. Its top-30% cumulative lift of 1.8799 indicates that targeting the top third of customers ranked by the model is more effective for reaching likely purchasers than a random approach, supporting more focused and efficient marketing efforts.

Methodology & Analysis

The following section describes the dataset, modeling approaches, and evaluation procedures used to select the best-performing model.

Data Used

The Bank's dataset contains account and personal information on 10,619 customers, including 12 categorical and 25 continuous predictors. Utilizing multiple imputation by chained equations (MICE), we addressed the missing information present in 14 variables. We selected MICE because it allows for imputations based on relationships among variables and added a flag variable to indicate if an imputation was performed for each of the 14 variables with missing data. We 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. After preprocessing, we split the data into a training set of 8,495 customers and a separate validation set of 2,124 customers to evaluate model performance.

Model Building

In previous phases, we evaluated four supervised learning models to estimate each customer's probability of purchasing the variable-rate annuity. GAM and MARS models can both model nonlinearities, with GAM resulting in smooth models and high interpretability, and MARS utilizing hinge functions that also capture interactions. Random forest and XGBoost excel on high-dimensional data, with random forest providing greater interpretability while XGBoost delivers higher predictive accuracy.

We added an EBM model to extend the set of modeling approaches used in previous phases. EBMs combine the transparency of additive models with the flexibility of boosted tree learners, allowing the model to capture nonlinear relationships while remaining highly interpretable. The model was trained using all remaining predictors after data preparation, including the 14 imputation indicator variables. After training, the model's feature importance and shape

functions were examined to understand how each predictor contributed to the purchase probability. We did not perform additional feature selection, since EBM models only select features that provide predictive power.

Model Evaluation

To compare model performance, we evaluated each model on an unseen validation dataset. This evaluation used metrics including AUC, accuracy, sensitivity, and specificity. Additionally, we calculated the top-30% cumulative lift to measure how effectively each model identifies high-value customers in the top predicted group, aligning with the bank's strategic customer targeting goals.

Results & Recommendations

This section presents the performance of all five models on the validation dataset, summarizes key findings from the evaluation metrics, and examines the feature relationships identified by the best-performing model. Based on these results, we outline the recommended modeling approach that best supports the bank’s business objectives.

Model Selection

To identify the strongest model for the Bank, we evaluated the final five models on the validation dataset. **Table 1** summarizes their performance metrics.

Table 1: Model Performance Metrics on the Validation Data

Model	AUC	Optimal Cutoff	Accuracy	Sensitivity	Specificity	Top-30% Cumulative Lift
GAM	0.7832	0.3180	0.7095	0.7453	0.6903	1.8396
MARS	0.7817	0.3055	0.7133	0.7682	0.6838	1.8485
Random forest	0.8026	0.3700	0.7137	0.7615	0.6881	1.8799
XGBoost	0.7815	0.3390	0.7067	0.7574	0.6795	1.8440
EBM	0.7961	0.3330	0.7170	0.7547	0.6968	1.8710

Based on these results, we selected the random forest model as the final model. It achieved the strongest overall performance with an AUC of 0.8026 and a top-30% cumulative lift of 1.8799. Although the EMB model showed a slightly higher accuracy, the random forest model provided the best balance of predictive strength and practical value. The model’s ROC curve is shown in **Figure 2** in the appendix, and its tuning parameters are listed in **Table 3** in the appendix.

Key Predictors

To understand the feature importance for the random forest model, we ranked the variables in descending order of mean decrease in accuracy, as shown in **Table 2**. This quickly highlights which variables contribute the most to predictive power.

Table 2: Random Forest Feature Importance

Variable	Mean Decrease Accuracy
Savings account balance	65.340
Checking account balance	51.264
Certificate of deposit account balance	34.658
Total ATM withdrawal amount	26.798
Number of checks written	25.644
Total amount deposited in checking account	25.577
Money market account balance	21.335
Indicator for certificate of deposit account	20.437
IRA balance	19.098
Branch of bank	18.336

The top predictors in the random forest model are account balances, including savings, checking, and certificates of deposit. Transaction-related variables, such as ATM withdrawals and checks written, also rank highly.

The Bank is particularly interested in understanding how account age influences product purchase, so we examined its global relationship using a partial dependence plot from the random forest model, as seen in **Figure 1**.

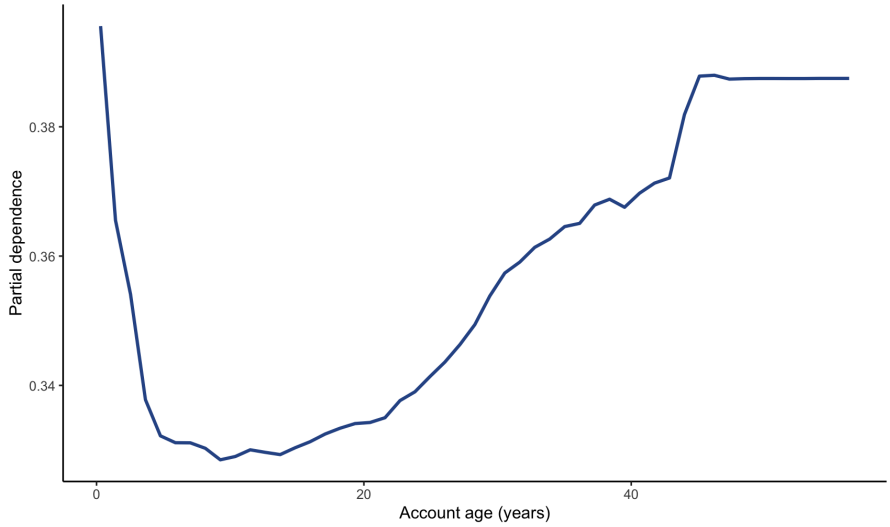

Figure 1: Partial Dependence of Purchase Probability on Account Age (Random Forest)

The results show a nonlinear pattern. Customers with very new accounts have relatively higher predicted purchase probabilities, but this likelihood declines sharply over the first few years and remains low for customers whose accounts have been open for a moderate length of time. After roughly 20 years, the predicted probability begins to rise steadily, with high likelihoods of purchase observed among customers who have had their accounts the longest. Overall, account age plays an important role, with brand-new and long-standing customers showing the strongest propensity to purchase the product.

Recommendations

Given the random forest model's clear predictive power, stable lift, and strong discrimination of purchasers and non-purchasers, we recommend using this model as the primary tool for identifying customers most likely to respond to future marketing efforts. The Bank should prioritize outreach to customers with the highest predicted probability of purchase, as these individuals demonstrate higher purchasing rates and offer the greatest return on marketing resources. If the Bank should incorporate account age into its targeting strategy, the results suggest that brand-new customers and long-tenured customers exhibit higher likelihoods of purchase, indicating that tailored messaging or differentiated offers for these groups may further enhance campaign performance.

Conclusion

Our team evaluated five predictive models to determine which customers are most likely to purchase the Bank’s variable rate annuity product. The random forest model was the most successful at distinguishing between buyers and non-buyers, with an AUC of 0.8026 and the highest top-30% cumulative lift. Feature importance results showed that account balances and transactional behavior are the most influential factors in predicting purchase likelihood, with account age also showing a meaningful nonlinear relationship. The Bank can use this model to guide targeted marketing efforts, prioritizing customers with the highest predicted probability of purchase.

Appendix

Table 3: Random Forest Hyperparameter Values

Hyperparameter	Value
Number of trees	500
Features per split	10
Node size	5
Random Seed	67

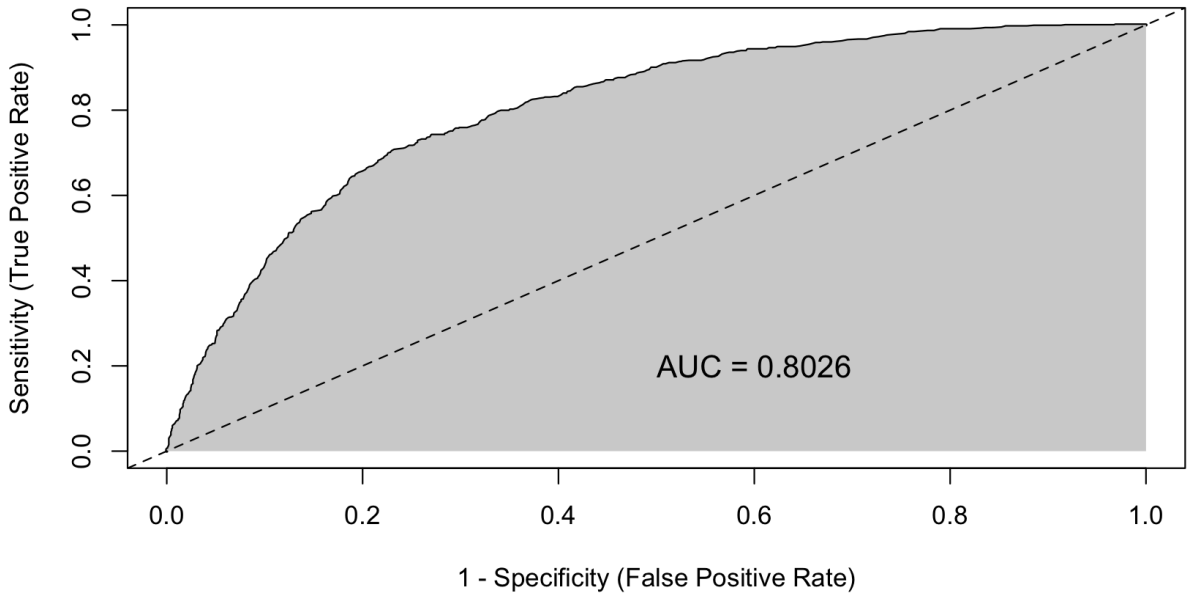

Figure 2: Random Forest ROC Curve on the Validation Set