Modeling Insurance Purchase Behavior with Tree-Based Models

Team 1

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

11/17/2025

Table of Contents

Modeling Insurance Purchase Behavior with Tree-Based Models

Overview

The Commercial Banking Corporation (the Bank) aims to identify customers likely to purchase a variable rate annuity product. Our team built random forest and extreme gradient boosting (XGBoost) models to estimate each customer's likelihood of making a purchase. We evaluated the performance of these models with receiver operating characteristic (ROC) curves, the area under the curve (AUC) values, and lift. Lift analysis revealed that targeting the top 30% of customers ranked by the XGBoost model increases the Bank's likelihood of capturing a buyer by 2.041 times compared to selecting customers at random, demonstrating that the model can effectively identify high-potential customers and directly support marketing efforts.

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. 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.

Model Building

We built random forest and XGBoost models to predict customer purchases of the annuity product. Initially, both models used all 36 predictors remaining in the dataset after data preparation, as well as the 14 new columns indicating imputation. For each model, we performed grid search tuning with five-fold cross-validation, using the AUC as the evaluation metric. After identifying the best initial models, we added a random variable to the dataset and removed all features with a lower importance than that of the random variable. We then retrained the final models on the subset of remaining features.

Model Evaluation

We evaluated the final random forest and XGBoost models to compare the results of the two approaches. We used AUC to compare how well each model distinguishes between purchasers and non-purchasers. Additionally, we calculated accuracy, precision, and recall using the optimal cutoff value for each model, determined via Youden's J statistic. We also evaluated lift to determine how effectively each model identifies high-probability buyers and to quantify the benefit of targeting the highest-probability customers over a random selection.

Results & Recommendations

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

Random Forest Results

The final random forest model achieved an AUC of 0.7975, demonstrating strong classification performance in identifying potential variable rate annuity buyers, as shown in **Figure 1**. We list the top ten most influential predictors in **Table 3** in the appendix, which shows that overall capital availability plays a major role in whether someone purchases a variable rate annuity.

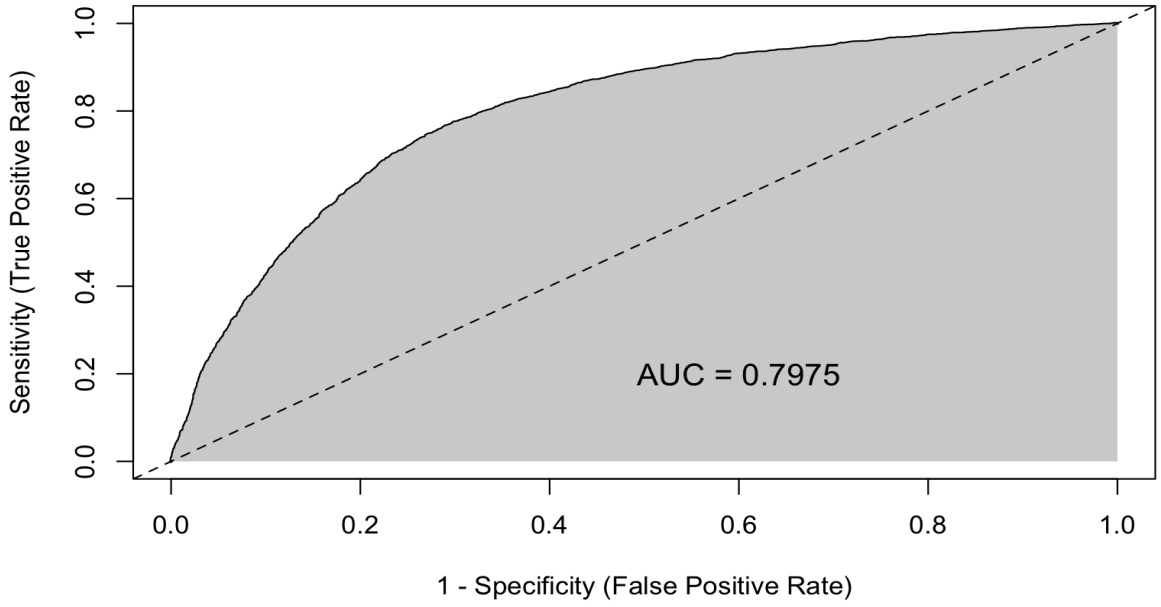

Figure 1: Random Forest ROC Curve

XGBoost Results

We began by tuning the XGBoost model, with the results summarized in **Table 4** in the appendix. After identifying the best model, we selected a subset of nine features whose gain was greater than that of a randomly generated variable. When building the final model on this

subset, only eight remained significant. **Table 5** in the appendix provides the names and gains of the eight variables that remained in the final model. The tuned model identified account balance variables, including savings, checking, certificate of deposit, and money market account balances, as the most influential features in predicting an annuity purchase. As shown in **Figure 2**, the final XGBoost model has an AUC of 0.8331, indicating a strong ability to distinguish between individuals who will or will not purchase a variable rate annuity within the training dataset.

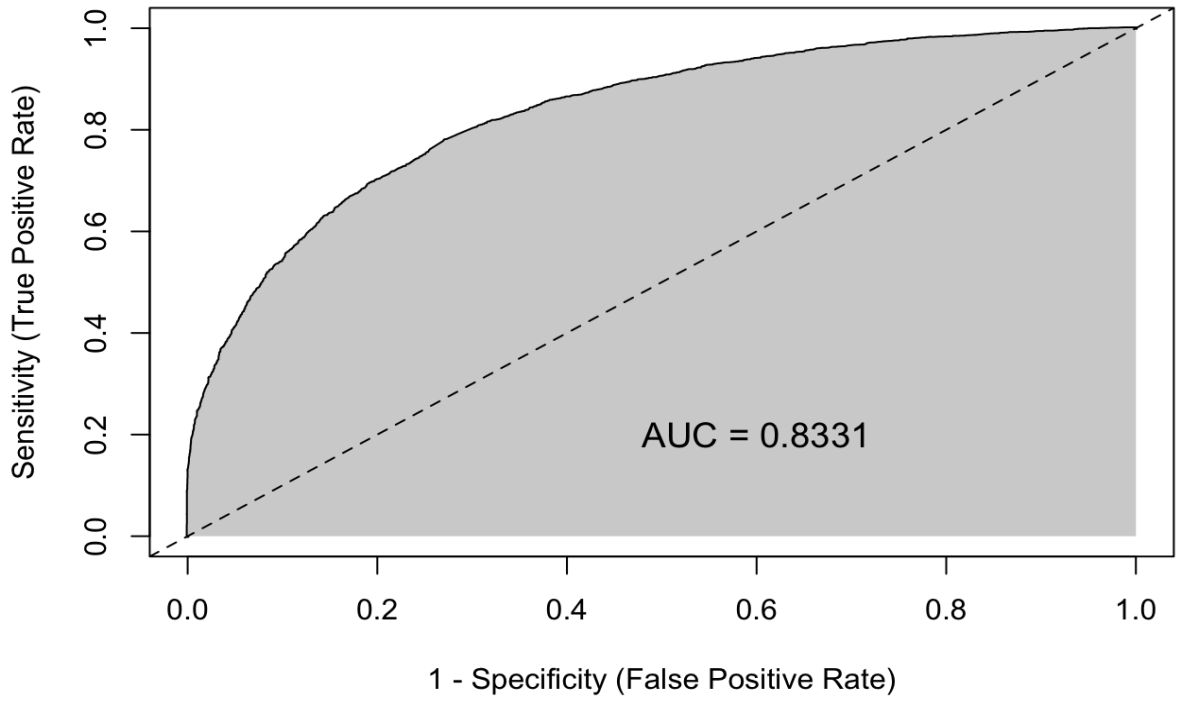

Figure 2: XGBoost ROC Curve

Comparison of Random Forest and XGBoost

To further evaluate the performance of our final models, we calculated the accuracy, recall, and precision of each model using the optimal cutoff values determined via Youden's J statistic. **Table 1** shows the calculated metrics for both models. We can see that the XGBoost model outperforms the random forest model across all metrics.

Table 1: Model Accuracy Metrics

Model	AUC	Optimal Cutoff	Accuracy	Sensitivity	Precision
Random Forest	0.7975	0.370	0.731	0.763	0.583
XGBoost	0.8331	0.339	0.746	0.780	0.601

As shown in **Tables 3** and **5** in the appendix, both models identified account balances, particularly savings, checking, and certificate of deposit balances, as key predictors of annuity purchases. Notably, six of the eight variables remaining in the XGBoost model are among the top ten most important features in the random forest model, indicating agreement between the two modeling approaches and reinforcing the predictive relevance of these account-balance variables. The other overlapping features, including the number of checks written and ATM withdrawals, highlight that transactional behavior also contributes meaningfully to prediction.

We additionally compared lift across the two models to evaluate how well each approach ranks customers by purchase likelihood. **Figures 3** and **4** in the appendix show the lift curves for the random forest and XGBoost models. The random forest model produced a lift of 1.90 in the top 30% of customers, indicating improved ranking performance. In comparison, the XGBoost model showed the strongest lift, with the top 30% of customers ranked by predicted probability being 2.041 times more likely to purchase than a randomly selected customer.

Recommendations

We recommend the Bank use our XGBoost model outputs to guide marketing efforts towards customers most likely to purchase an annuity. The lift results show that targeting the top 30% of customers ranked by predicted probability increases the Bank's likelihood of capturing a buyer by 2.041 times compared to selecting customers at random. Using the model’s ranking, the Bank can focus its outreach on high-potential customers to improve the effectiveness and efficiency of its marketing efforts.

Conclusion

Our team built random forest and XGBoost models to identify which customers of the Bank are likely to purchase a variable rate annuity product. The random forest model achieved an AUC of 0.7975, while XGBoost performed slightly better with an AUC of 0.8331, along with higher accuracy, recall, and precision. The lift analysis further confirmed its value, showing that the model effectively separates high- and low-probability customers. Feature importance results also highlighted account balances and transactional behavior as the primary drivers of purchase likelihood. The model can be used by the Bank for targeted marketing, enabling them to prioritize customers with the highest purchase probability.

Appendix

Table 2: Random Forest Tuning Result

Hyperparameter	Values Tested	Best Value
Number of trees	[200, 500, 1000]	500
Features per split	[10, 17, 25]	10
Node size	[1, 3, 5]	5

Table 3: 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

Table 4: XGBoost Tuning Result

Hyperparameter	Values Tested	Best Value
Boosting Rounds	[100, 500]	100
Columns Sampled	[0.7, 1]	1
Learning Rate	[0.05, 0.1, 0.2]	0.05
Minimum Child Weight	[1, 3]	3
Subsample	[0.5, 0.75, 1]	0.75
Tree Depth	[3, 5, 7]	5

Table 5: XGBoost Feature Importance

Variable	Gain
Savings account balance	0.325
Checking account balance	0.176
Money market account balance	0.119
Certificate of deposit account balance	0.114
Indicator for checking account	0.084
Age of oldest account	0.072
Total ATM withdrawal amount	0.061
Number of checks written	0.048

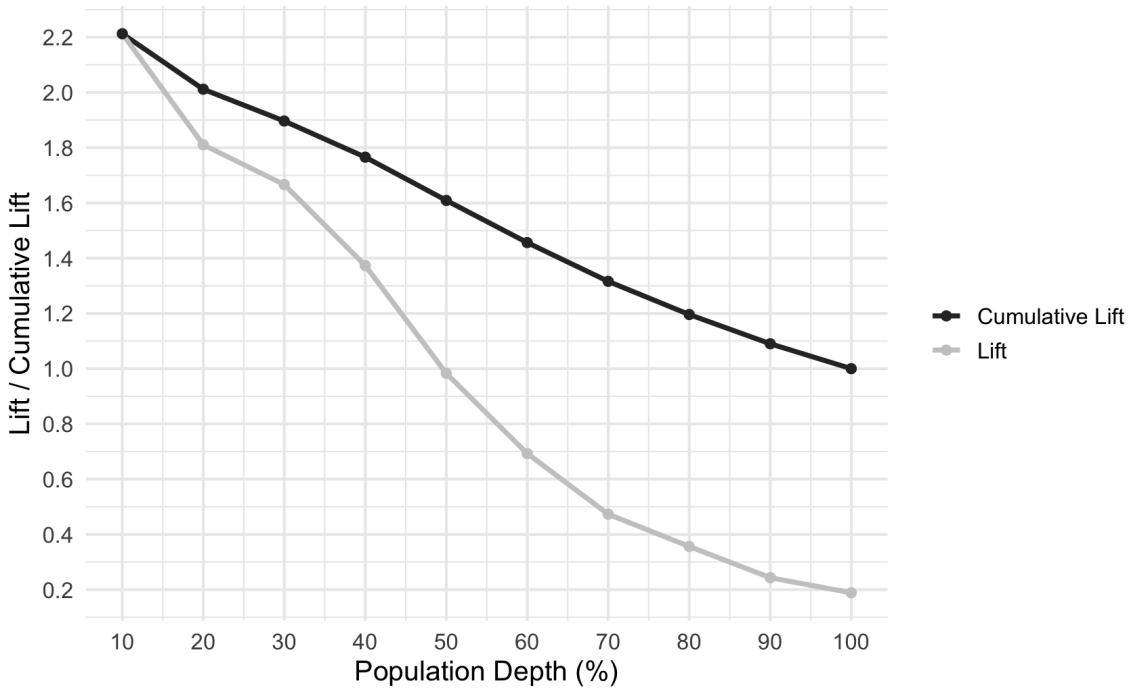

Figure 3: Random Forest Lift Curve

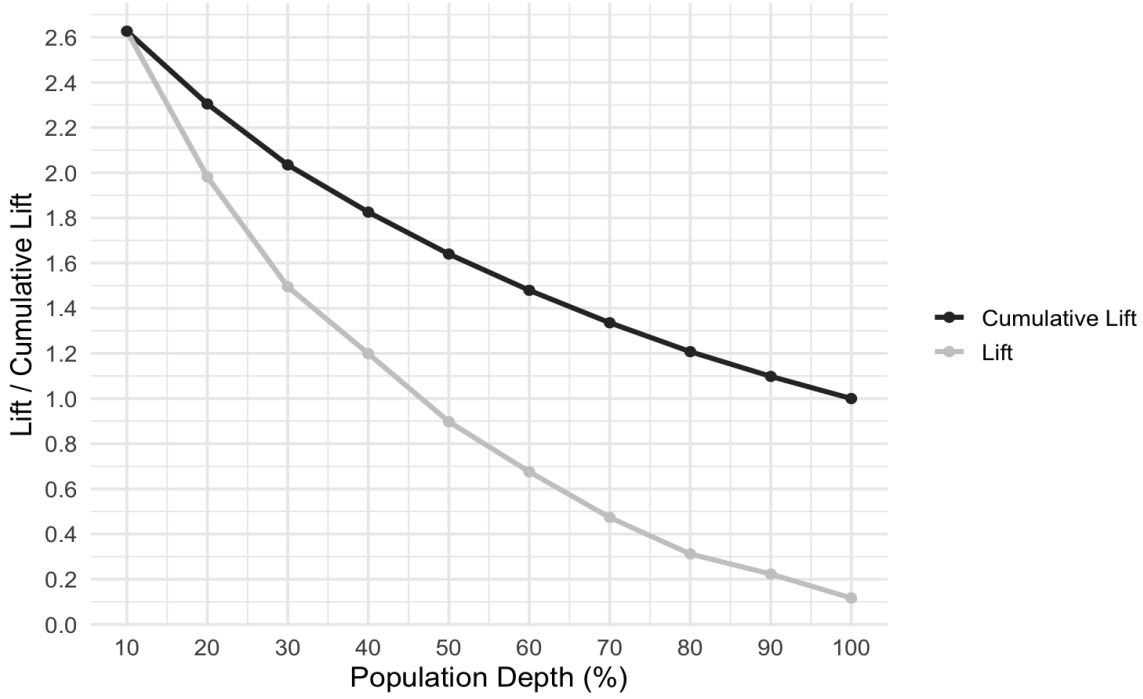

Figure 4: XGBoost Lift Curve