CMC Banking Phase 2: Tree Based Models

COMMERCIAL BANKING, CORP

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

November 18, 2024

Table of Contents

CMC Banking Phase 2: Tree Based Models

Overview

The Commercial Banking Corporation ("the Bank") launched a variable rate annuity product and engaged us in predicting likely buyers to optimize marketing and resource allocation. We developed two models for this purpose: a random forest model and an Extreme Gradient Boosting (XGBoost) model. While both models identify similar features impacting purchase of the annuity product, such as savings and checking account balances, the XGBoost model correctly ranks 86.23% of buyers compared to 79.41% for the random forest model. We recommend the Bank consider adopting the XGBoost model to improve targeting accuracy, optimize outreach efforts, and potentially increase product sales.

Methodology and Analysis

The Bank provided a dataset of 8,495 observations and 38 variables, with a target, INS; the full variable definitions can be found in the data dictionary located in the Appendix. We used median imputation for continuous variables and mode imputation for categorical variables to preserve patterns.

Variable Selection

We retained all variables as variable selection negatively impacted model accuracy. Cross-validation was incorporated to ensure robustness and mitigate overfitting.

Random Forest

We used a random forest algorithm to predict the likelihood of a customer purchasing the variable rate annuity product. The model was initially configured at 500 trees and tuned from there to select the optimal number of trees and variables at splits. Our random forest model was finalized with 375 trees and 6 variables at each split; Table 1 presents variables ranked in order of mean decrease in accuracy (MDA).

Table 1: Random Forest Variable Importance

Variable	MDA	*Variable*	MDA	*Variable*	MDA	*Variable*	MDA
SAVBAL	52.74	*IRABAL*	15.51	*POSAMT*	10.45	*AGE*	2.51
DDABAL	41.56	*DEP*	14.99	*INCOME*	10.15	*CRSCORE*	1.63
CDBAL	27.23	*IRA*	14.34	*SAV*	9.80	*NSF*	0.58
DEPAMT	25.30	*BRANCH*	14.03	*POS*	8.95	*NSFAMT*	0.17
CHECKS	20.40	*C CBAL*	13.19	*PHONE*	6.59	*SDB*	-0.02
ATMAMT	20.20	*DDA*	11.49	*CCPURC*	6.55	*LORES*	-0.63
CD	18.58	*INV*	11.24	*ATM*	6.44	*INAREA*	-1.73
MMBAL	16.92	*HMVAL*	11.05	*TELLER*	5.21		
CC	16.79	*INVBAL*	10.74	*DIRDEP*	4.38		
MM	15.99	*ACCTAGE*	10.64	*MMCRED*	3.79		

In Table 1, three types of balances–savings, checking, and certificate of deposit–are shown to be most important in determining likelihood of purchase. Understanding the key variables can support strategy through helping target product offerings to relevant customer segments.

XGBoost

We also explored XGBoost for predicting the likelihood of purchasing the variable rate annuity product. The initial model we started with used a subsample ratio of 0.5 and 50 boosting rounds. The final model was tuned to a subsample ratio of 0.75, 16 boosting rounds, max depth of 7, and a learning rate of 0.25. Figure 1 shows variables grouped by importance as determined by the XGBoost model.

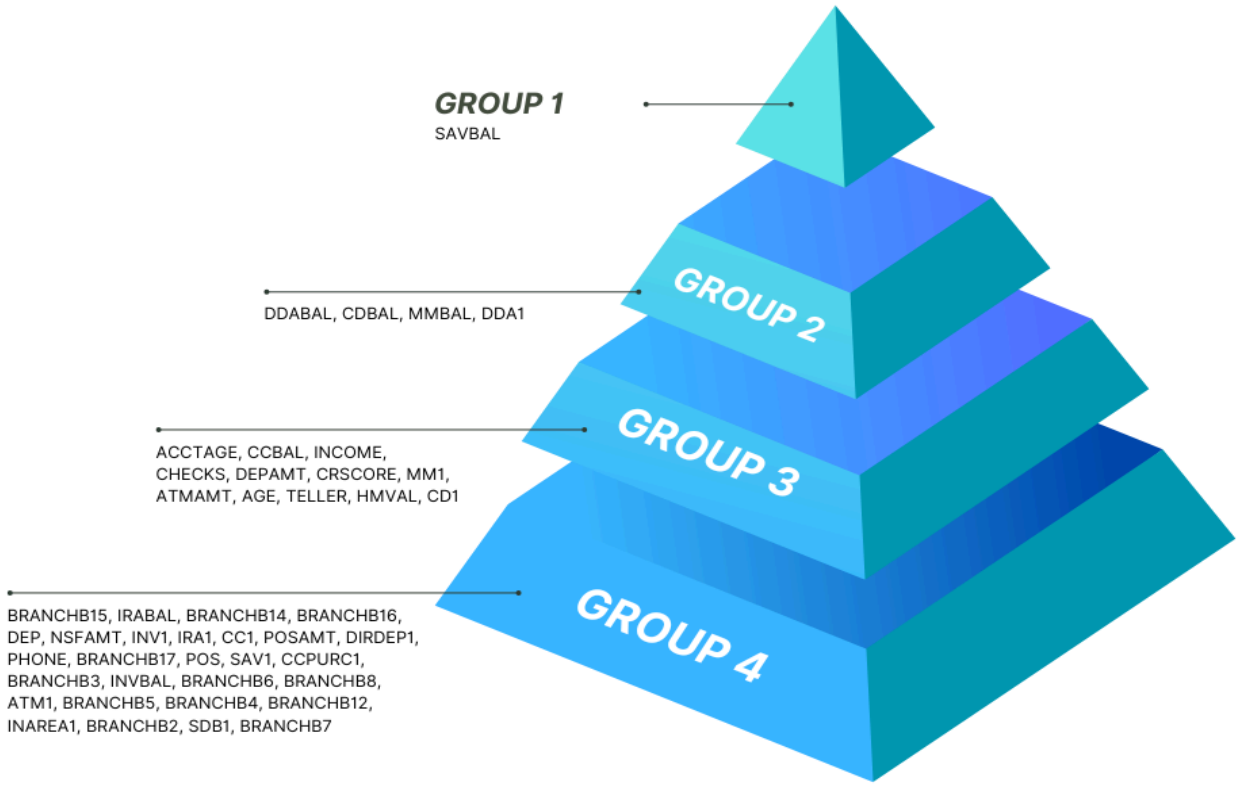

Figure 1: Variable Grouping for XGBoost Model

Figure 1 demonstrates which variables held a similar level of importance in the XGBoost algorithm.

Results

We evaluated the predictive capabilities of random forest and XGBoost models on the data; these models were chosen for their robustness in handling complex, non-linear relationships and the ability to rank features by importance, providing valuable insights into customer attributes.

Model Comparison

The random forest model was initially trained with 500 trees. Post tuning, the optimal configuration was determined at 375 trees and 6 variables at each split. The model's performance measured by the area under the ROC curve (AUC), was 0.7941.

The XGBoost model, after fine-tuning through cross-validation, achieved an optimal configuration with 16 boosting rounds, a learning rate of 0.25, and a maximum tree depth of 7. The AUC for the final XGBoost model was 0.8623, indicating an improvement in performance compared to the random forest model.

Slight differences in feature importance were seen between the two models, as shown in Figure 2.

Top 5 Important Features: Random Forest	**Top 5 Important Features: XGBoost**
1. Savings Balance 2. Checking Account Balance 3. Certificate of Deposit Balance 4. Total Amount Deposited 5. Number of Checks Written	1. Savings Balance 2. Checking Account Balance 3. Certificate of Deposit Balance 4. Money Market Balance 5. Indicator for Checking Account

Figure 2: Model Feature Importance

In Figure 2, we highlight the influence of financial indicators such as account balances and credit score in predicting purchases.

Figure 3 shows ROC curves for both models. The random forest model has an AUC of 0.7941, indicating moderate discrimination between customers who purchased and those who did not. The XGBoost model has an AUC of 0.8623, outperforming the random forest model, particularly at higher sensitivity thresholds.

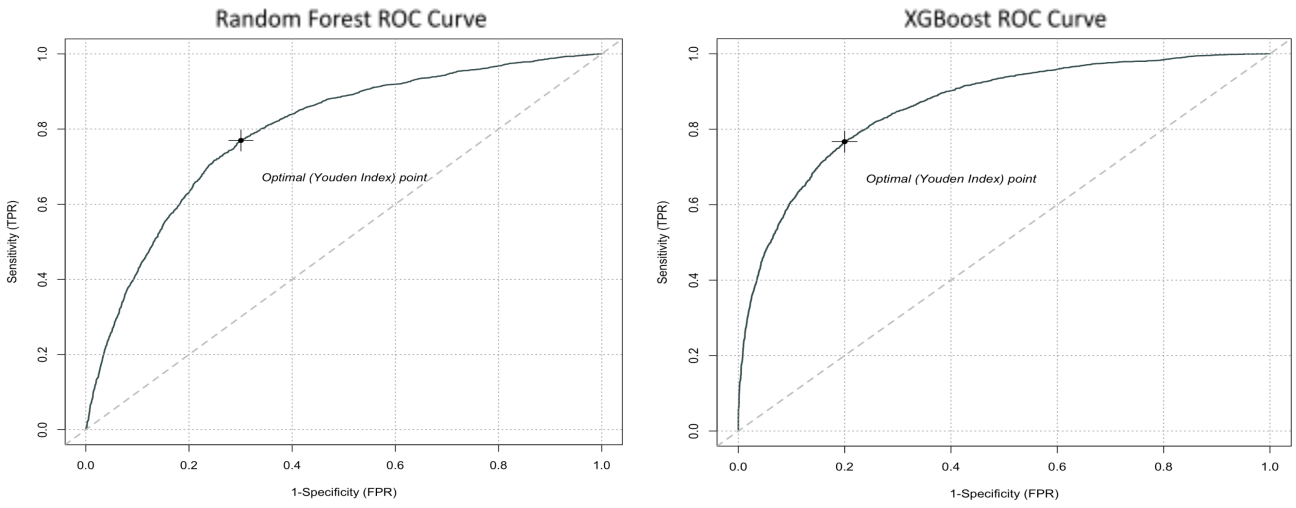

Figure 3: ROC Curve for Both Models

Figure 3 shows that the XGBoost model demonstrates superior true positive rates for the same or lower false positive rate when compared to the random forest. Both models prioritized savings balance as the most important predictor of annuity purchases.

Recommendations

After modeling with both approaches, we recommend the following:

- **Deploy the XGBoost model for predicting purchases of the variable rate annuity product:** The XGBoost model for predicting annuity purchases improved AUC by 6.82% over random forest, correctly ranking 86.23% of buyers. The ROC curve identified 0.368 as the optimal cutoff, balancing predictions and false positives. Adjusting this threshold to align with the Bank's goals is recommended.

- **Explore more explainable models:** Our XGBoost model focuses primarily on predictability. To better interpret the variables used for predictions, we recommend exploring other models such as an explainable boosting machine.

Deploying the XGBoost model will improve the Bank's ability to target customers and drive annuity purchases. Pairing this implementation with efforts to explore interpretable models will enhance decision-making and provide deeper insights into factors related to customer purchases of the annuity product.

Conclusion

Our analysis reveals that both the random forest and XGBoost models identify similar features, such as savings and checking account balances, which indicate increased likelihood of purchasing the Bank's variable rate annuity. However, the XGBoost model outperformed the random forest model, correctly identifying 86.23% of buyers compared to 79.41%, respectively.

We recommend that the Bank consider implementing the XGBoost model to predict purchases of the annuity product and explore more explainable models to enhance prediction accuracy and ensure the Bank remains competitive in targeting the right customers effectively.

Appendix

Table 2: Data Dictionary

Name	Description	Name	Description
ACCTAGE	Age of oldest account	IRA	Indicator for retirement account
DDA	Indicator for checking account	IRABAL	IRA balance
DDABAL	Checking account balance	INV	Indicator for investment account
DEP	Checking deposits	INVBAL	INV balance
DEPAMT	Total amount deposited	MM	Indicator for money market account
CHECKS	Number of checks written	MMBAL	MM balance
DIRDEP	Indicator for direct deposit	MMCRED	Number of money market credits
NSF	Number of insufficient fund issues	CC	Indicator for credit card
NSFAMT	Amount of NSF	CCBAL	CC balance
PHONE	Number of telephone banking interactions	CCPURC	Number of credit card purchases
TELLER	Number of teller visit interactions	SDB	Indicator for safety deposit box
SAV	Indicator for savings account	INCOME	Income
SAVBAL	Savings account balance	LORES	Length of residence in years
ATM	Indicator for ATM interaction	HMVAL	Value of home
ATMAMT	Total ATM withdrawal amount	AGE	Age
POS	Number of point of sale interactions	CRSCORE	Credit Score
POSAMT	Total amount for point of sale interactions	INAREA	Indicator for local address
CD	Indicator for certificate of deposit account	INS	Indicator for purchase of insurance product
CDBAL	CD balance	BRANCH	Branch of bank

Table 3: XGBoost Variable Importance

Name	Description	Name	Description
SAVBAL	0.2486986432	NSFAMT	0.0088083255
DDABAL	0.1164450929	INV1	0.0070179587
CDBAL	0.072458034	IRA1	0.0064707624
MMBAL	0.0655689514	CC1	0.0057690447
DDA1	0.0598709558	POSAMT	0.0052628355
ACCTAGE	0.0421135257	DIRDEP1	0.0034739366
CCBAL	0.0294902716	PHONE	0.0034724366
INCOME	0.0292934852	BRANCHB17	0.0027414325
CHECKS	0.027756004	POS	0.002658536
DEPAMT	0.0270453897	SAV1	0.0022776867
CRSCORE	0.0269343769	CCPURC1	0.0020509809
MM1	0.0246998196	BRANCHB3	0.0018990865
ATMAMT	0.0233049641	INVBAL	0.0018845097
AGE	0.0199195774	BRANCHB6	0.0017419858
TELLER	0.0183057115	BRANCHB8	0.001662892
HMVAL	0.0172886211	ATM1	0.0015615616
CD1	0.0168106546	BRANCHB5	0.0012497232
LORES	0.0141384439	BRANCHB4	0.0010587582
BRANCH 15	0.0138882052	BRANCHB12	0.0010346431
IRABAL	0.0115332079	INAREA1	0.0007332583
BRANCHB14	0.010377758	BRANCHB2	0.0006927908
BRANCHB16	0.0098511063	SDB1	0.0006174861
DEP	0.0095267598	BRANCH 7	0.0005398088

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?