CMC Banking Phase 3: Model Interpretation

COMMERCIAL BANKING, CORP

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

November 26, 2024

Table of Contents

CMC BANKING PHASE 3: MODEL INTERPRETATION

Overview

The Commercial Banking Corporation ("The Bank") aims to optimize marketing for its new variable rate annuity product by identifying likely buyers. We evaluated various machine learning models, including XGBoost and a neural network model. XGBoost outperformed all models, achieving an area under the receiver operating characteristic (ROC) curve of 0.862 on training and 0.794 on validation, compared to 0.783 for the neural network model. We recommend deploying XGBoost for its superior accuracy to target buyers, prioritize marketing, boost sales, and allocate resources efficiently.

Methodology and Analysis

This section of the report covers key aspects of our study, including the data used, an assessment of our neural network, an evaluation of our XGBoost model, and a comparison of model accuracy.

Data Used

The Bank provided a training dataset of 8,495 observations and a validation dataset of 2124 observations, each with 38 variables and a target variable Indicating the purchase of an insurance product (INS). We used median imputation for continuous variables and mode imputation for categorical variables to preserve patterns. We retained all variables, as variable selection negatively impacted model accuracy. Cross-validation was incorporated to ensure robustness and mitigate overfitting.

To ensure consistency in model training and evaluation, we standardized data processing across the training and validation datasets. We transformed categorical features into dummy variables using a unified approach, while numeric features were standardized based on the training set's mean and standard deviation. This ensured all numeric features had a consistent scale, enabling reliable and accurate model predictions.

Neural Network

We developed an initial neural network model with five hidden layer nodes and a logistic output. Performance was assessed using the area under the ROC curve (AUC) on the training dataset. We conducted a grid search with 10-fold cross-validation to optimize the model, identifying the optimal configuration as four hidden layer nodes and a decay parameter of one. The final model was evaluated on the validation dataset with its AUC compared to the XGBoost model for performance benchmarking.

XGBoost

We previously explored an Extreme Gradient Boosting (XGBoost) model 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, a max depth of 7, and a learning rate of 0.25. Table 1 shows the feature importance of the top five variables in the final XGBoost model based on the mean decrease in impurity.

Table 1: Feature Importance of Top 5 Variables in XGBoost Model Using Mean Decrease in Impurity

Variable	Importance
Savings account balance	0.2708
Checking account balance	0.1088
Certificate of deposit balance	0.0816
Does not have a checking account	0.0757
Has a money market account	0.0675

In Table 1, the final XGBoost model identifies a customer's savings account balance as the most influential predictor of purchasing the annuity product. Checking account balance and certificate of deposit balance follow in importance, further emphasizing the role of account balances in a customer's decision to purchase the annuity product.

ROC Curve for XGBoost

To assess classification performance, we used a ROC curve to visualize the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across different classification thresholds. Figure 2 displays the ROC curve for the validation dataset using the final XGBoost model.

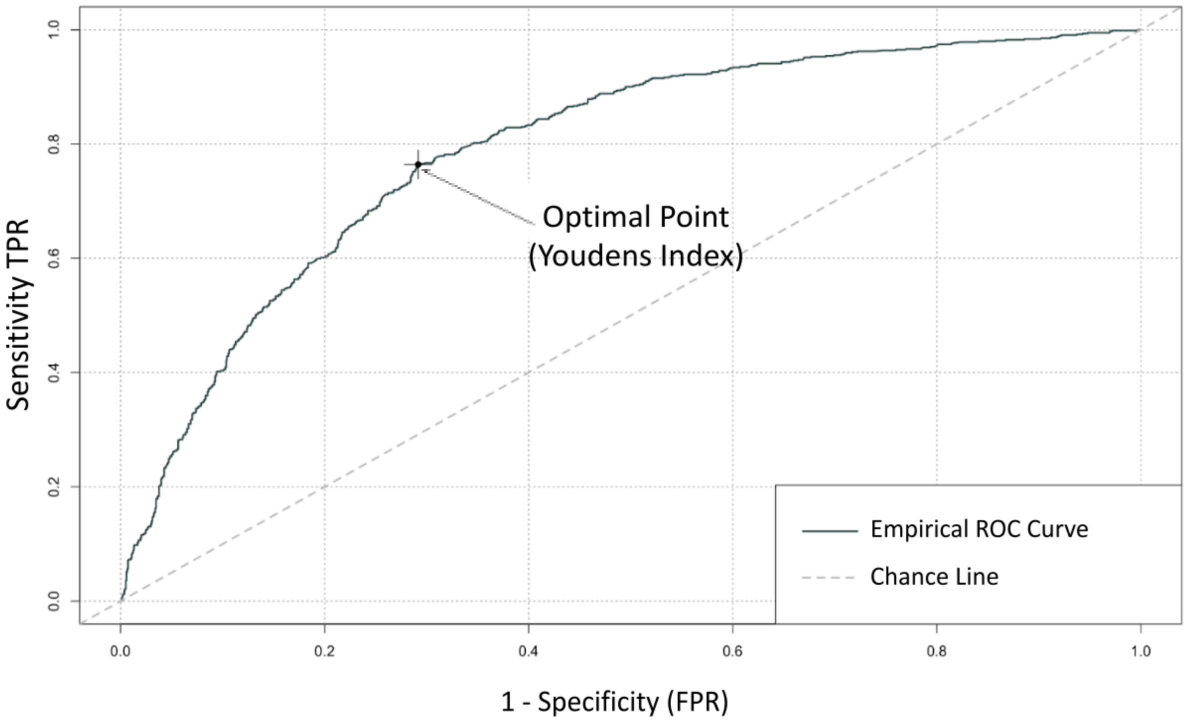

Figure 1: ROC Curve for XGBoost Model Using Validation Dataset

In Figure 1, the final XGBoost model demonstrates better true positive rates for the same or lower false positive rates, demonstrating strong predictive performance.

AUC Comparison

To compare all models' performance on the Bank's annuity product data, we used AUC. Table 2 lists this metric for each model using the training dataset, as we did not test the logistic regression, MARS, GAM, or random forest models on the validation dataset.

Table 2: AUC Comparison for All Models

Model	Training AUC	Validation AUC
XGBoost	0.862	0.794
Neural network	0.814	0.783
GAM	0.805	*NA*
MARS	0.801	*NA*
Logistic regression	0.800	*NA*
Random forest	0.794	*NA*

In Table 2, the XGBoost model demonstrated the highest classification performance with an AUC of 0.862 on the training dataset, meaning it correctly distinguishes between buyers and non-buyers 86.2% of the time. The GAM and MARS models also performed well, with AUC values of 0.805 and 0.801, respectively, indicating discriminatory ability.

Results

We prioritized models that handle complex, non-linear relationships and ranked feature importance for customer insights. XGBoost outperformed others, achieving the highest AUC (0.862 on training via cross-validation and 0.794 on validation), demonstrating predictive accuracy and generalization. Its performance and reliability make XGBoost the most suitable model.

Global Importance

To assess the global impact of account age on purchasing the new variable rate annuity product, we used a Partial Dependence Plot (PDP) with the XGBoost model.

The relationship between account age and annuity purchases shows a decreasing trend, with spikes around 3 and 7.5 years and stabilization after 8 years. Customers are less likely to purchase the variable rate annuity product as their account age increases. Accounts 3 years or younger seem optimal when looking for customers more likely to purchase. Older account holders may already own similar products or be risk-averse, so targeting customers with accounts under 3 years will likely yield the best results.

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 1.1% over the neural network on the validation set, correctly ranking 79.43% of buyers. The ROC curve identified 0.327 as the optimal cutoff, balancing predictions and false positives. We recommend adjusting this threshold to align with the Bank's specific marketing goals, such as focusing on a particular target audience or controlling for cost-effectiveness in outreach efforts.

- **Explore outliers and interesting observations:** We recommend leveraging Shapley local interpreters to explore individual observations in greater depth. For example, the Bank could analyze customers such as observation 732, a long-tenured customer, or observation 1720, which has the largest savings account balance. The Bank may be interested in exploring observation 1720 specifically as savings account balance ranked at the top for variable importance of predicting the purchase of a variable rate annuity. Exploring these cases would provide a more detailed understanding of how predictor variables impact specific observations.

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

Conclusion

Our analysis confirms that the XGBoost model is the most effective approach for predicting purchases of the Bank's variable rate annuity product. Its robust performance, highlighted by its superior AUC compared to alternative models, demonstrates its ability to distinguish likely buyers from non-buyers accurately. This information empowers the Bank to focus its marketing efforts on high-probability customers, ultimately driving sales and improving resource allocation.

Looking ahead, we recommend leveraging explainable tools like Shapley values to gain deeper insights into individual customer behaviors. This will allow the Bank to refine its marketing strategies and maintain a competitive edge by enhancing customer understanding and engagement.

Appendix

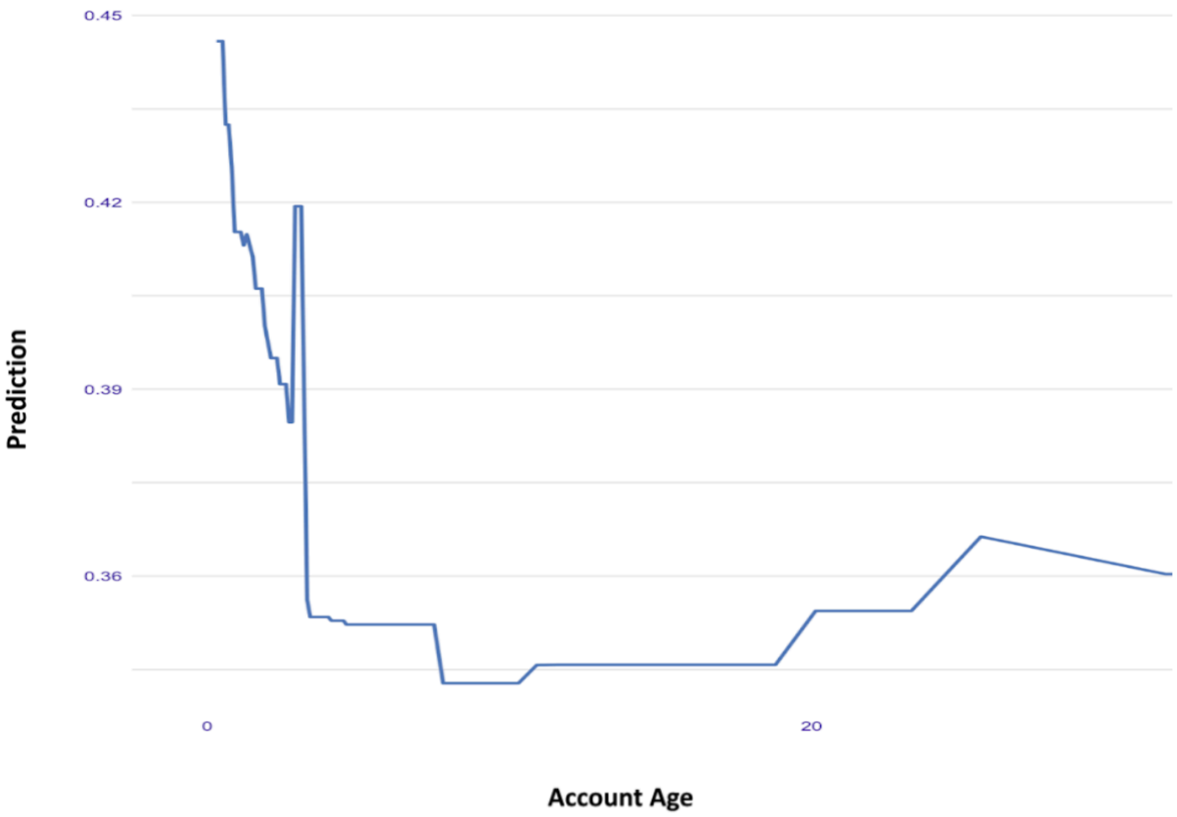

Figure 2: Partial Dependence Plot for Age of Oldest Account (ACCTAGE)

Table 3: Variable Importance in XGBoost Model Ranked by Importance

Variable	Importance
Savings account balance	0.250778014
Checking account balance	0.11505336
CD balance	0.073271247
Indicator for checking account	0.067210717
Indicator for money market account	0.050865311
Age of oldest account	0.040719459
Total ATM withdrawal amount	0.031057819
Total amount deposited	0.029291048
MM balance	0.029282606
Number of checks written	0.028692123
CC balance	0.0237352
Value of home	0.023593156
Credit score	0.02233461
Age	0.019819975
Number of teller visit interactions	0.019414555
Income	0.018562485
IRA balance	0.017203517
Length of residence in years	0.017065811
Branch of bank 15	0.01245836
Checking deposits	0.011029916
Indicator for credit card	0.009872512
Indicator for certificate of deposit account	0.009642055
Branch of bank 14	0.009267228
Amount of NSF	0.008873045
Branch of bank 16	0.008562737
Indicator for investment account	0.007826351
Total amount for point of sale interactions	0.006828257
Indicator for direct deposit	0.006024098
Indicator for retirement account	0.005518232

Variable	Importance
Number of credit card purchases	0.004861985
Number of telephone banking interactions	0.003676677
Branch of bank 4	0.002621848
Branch of bank 17	0.00254009
Indicator for safety deposit box	0.002083868
Number of point of sale interactions	0.001718004
Indicator for ATM interaction	0.001438444
Indicator for local address	0.001382469
Indicator for savings account	0.001296953
Branch of bank 12	0.001048179
Number of money market credits	0.000967612
Branch of bank 10	0.000756428
INV balance	0.000490899
Branch of bank 9	0.0004824
Branch of bank 2	0.00044085
Branch of bank 13	0.000339493

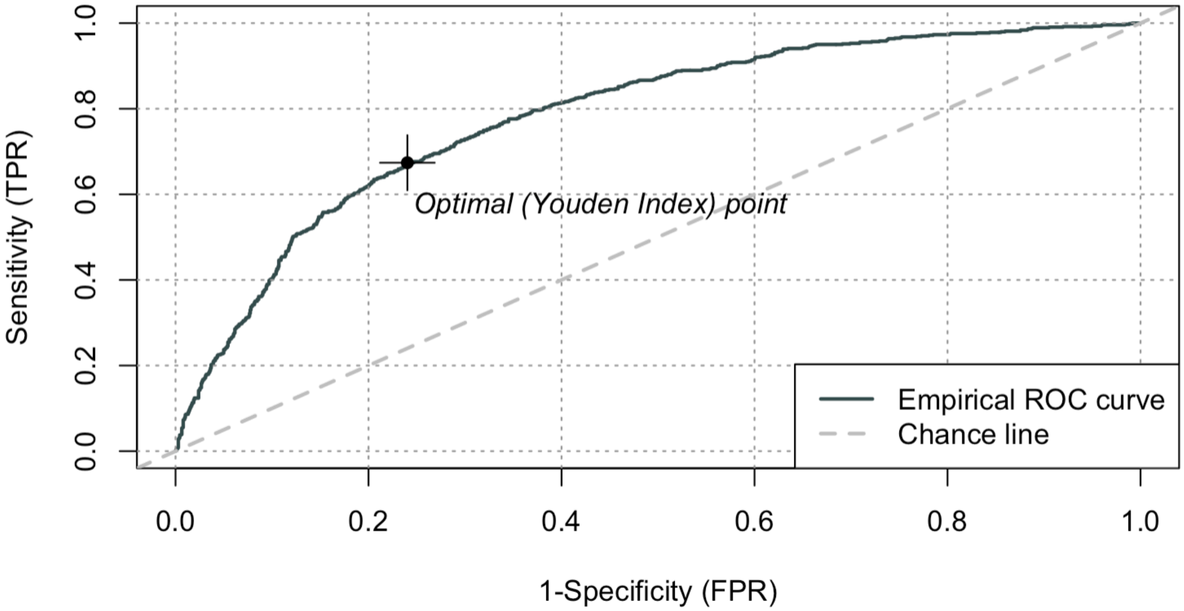

Figure 3: ROC Curve for Neural Network Model Using Validation Dataset

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?