BANKING INSURANCE PRODUCT: MACHINE LEARNING MODEL INTERPRETATION

BLUE 15

VANSHIKA BHARDWAJ
GARRETT CURRAN
SUYEON KIM
KEVIN LLOYD
ERIC MILLER

NOVEMBER 26, 2024

Table of Contents

BANKING INSURANCE PRODUCT: MACHINE LEARNING MODEL INTERPRETATION

Overview

The Commercial Banking Corporation, hereinafter referred to as 'the bank,' partnered with Blue 15 to create predictive models to identify customers likely to purchase a variable-rate annuity product, hereinafter referred to as 'the product.'

We recommend the bank segment customers by their account balances to tailor and optimize marketing strategies. Additionally, we recommend focusing marketing efforts on customers whose oldest account is less than three years old.

Methodology & Analysis

In this section, we will describe the data and outline our methodology and analysis.

Data Used

The original dataset used for this report contained 8,495 observations with 38 variables. Each row represents each customer's attributes before being offered the product. The bank specified that, apart from the branch variable, all variables with more than 10 levels should be treated as continuous. When applying that standard, we found 21 continuous and 16 categorical variables along with the target indicator for whether the customer purchased the product. Imputation for continuous and categorical variables was done the same way as in previous phase of this analysis. For the money market credits variable, we binned values of three or more to address linear separation.

To optimize the performance of the neural network model, we decided to scale the quantitative variables before model training. We applied z-score standardization to all quantitative variables in the training dataset. The same standardization transformation was applied to the test dataset to ensure consistency and prevent data leakage.

Neural Network Model

We developed a neural network model to predict the likelihood of a customer purchasing the product. We optimized key parameters to enhance the model's performance, including the number of nodes in the hidden layer, the weight decay parameter for regularization, and the maximum number of iterations during model training. These optimizations were carried out using ten-fold cross-validation to ensure robust performance evaluation. Since our primary goal in this phase was to maximize predictive accuracy, we retained all available variables in the model.

Final Model Selection & Evaluation

To compare the performance of the neural network model with the previously created multivariate adaptive regression splines (MARS), generalized additive model (GAM), extreme gradient boosting (XGBoost), and random forest models, we compared the AUC of each model on the test dataset. We adjusted the XGBoost to include all predictor variables in this phase due to the focus on the predictive ability of our final selected model. The comparison between the five models allowed us to select our final model to be utilized in model evaluation and to be recommended to the bank.

To understand which variables had the largest impact on the prediction of insurance product purchases in our final model, we examined the reduction in AUC when each variable was removed from the final model. Additionally, to further understand some of the global interpretations of the final model, we created

an accumulated local effects (ALE) plot, which shows how predictions change when a variable of interest changes. As requested, our variable of interest for this was account age. Additionally, to have a more local interpretation of our final model, we analyzed the Shapley values for each variable for the customer in observation 732.

Results & Recommendations

In this section, we will discuss the results from our models and the actions we recommend based on these results.

Neural Network Results

We built a neural network model and utilized a grid-search approach to tune the relevant parameters. This model achieved an AUC of 0.79 on validation data, indicating a 79% probability of ranking a randomly chosen purchaser higher than a randomly chosen non-purchaser in terms of predicted likelihood.

Final Model Selection & Evaluation

We compared the validation AUC values for each model we built to determine our best model. Table 1 shows all of our models in order by AUC.

Table 1: All models ranked by testing AUC

Rank	Model	AUC
1	XGBoost	0.80
2	Random Forest	0.79
3	Neural Network	0.79
4	MARS	0.78
5	GAM	0.78

In Table 1, the highest AUC for the XGBoost model indicates that it was the most successful model in distinguishing customers who bought the product from those who did not. We decided that this model would be our final model, and we recommend its use for any future analysis on the insurance product. Figure 1 shows the receiver operating characteristic (ROC) curve for our final model and the corresponding AUC that was used to rank the model's performance.

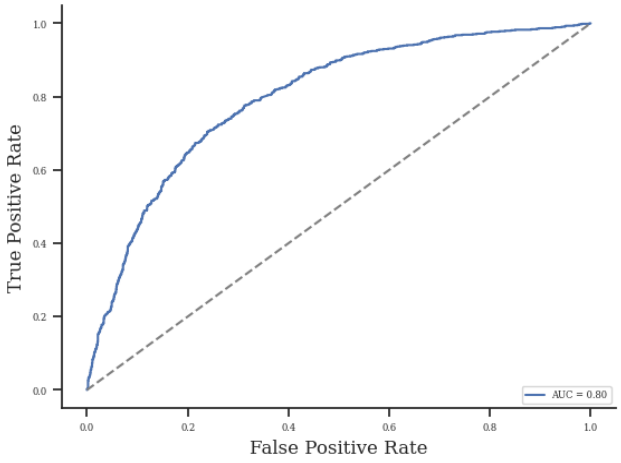

Figure 1: ROC curve for final XGBoost model

Figure 1 shows the ROC curve that gives us the final AUC performance metric of 0.80 for the XGBoost model, indicating that the model has an 80% probability of ranking a randomly chosen purchaser of the product higher than a randomly chosen non-purchaser. From this figure, we determined the XGBoost model was our best model in terms of determining if a customer would purchase the product. Table 2 shows the rank of permutation importance for the top 10 variables in our final XGBoost model.

Table 2: Permutation importance ranks for the top 10 XGBoost model variables

Rank	Variable
1	Savings acct balance
2	Checking acct balance
3	CD balance
4	Checking acct indicator
5	MM balance
6	Credit card indicator (yes)
7	Total ATM withdrawal amount
8	Number of checks written
9	IRA balance
10	Investment acct indicator (yes)

Table 2 showed us that customers' account balances and account types are highly impactful to the model's predictions. Specifically, savings balance, checking account balance, and certificate of deposit balance are the three most impactful variables, so we recommend segmenting customers based upon these variables. Figure 2 shows the ALE plot for the account age variable. This shows how the predictions for the customers' likelihood to purchase the insurance product change across different values for account age.

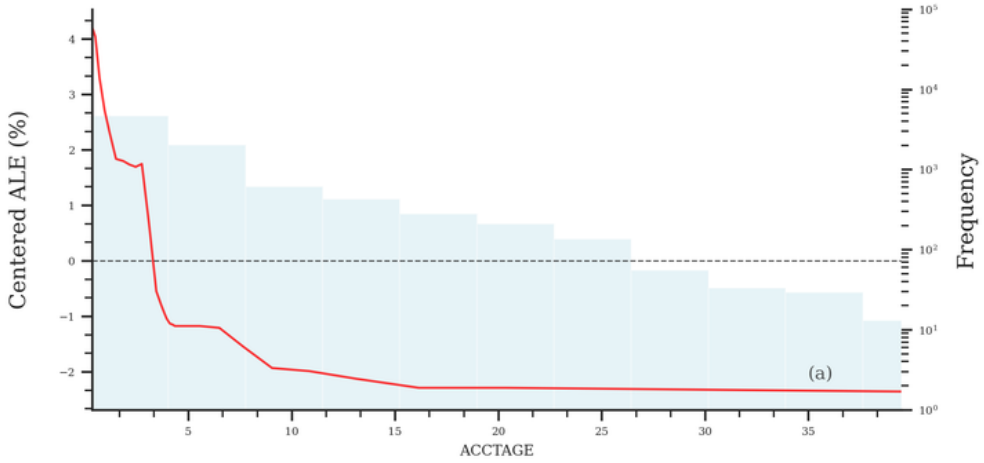

Figure 2: ALE plot for the account age variable

Figure 2 shows us that as customers have older accounts, the likelihood of purchasing the insurance product decreases. This shows us that customers with the oldest account being three years old or less are more likely than average to purchase the product.

We then looked at the Shapley values for observation 732, which was one of the most tenured customers that we were asked to look into further. For this individual, their savings balance, checking account balance, and account age were the three most impactful variables on our prediction that he/she would have a 57.5% chance of purchasing the product.

Recommendations

We recommend that the bank take the following actions:

- **Segment customers based on their account balances to focus marketing.**

  Savings, checking, and certificate of deposit balances were the three most important predictors in our final model. We recommend segmenting customers based on their account balances and values for the other top 10 predictor variables to be the most efficient in marketing to customers who are highly likely to purchase the product.

- **Focus marketing efforts on customers whose oldest account is less than three years old.**

  From our ALE plot, we can see that the impact of account age on the likelihood of purchasing the product becomes negative for accounts greater than three years old. This indicates that customers whose oldest account is less than three years old are more likely to purchase the product and could be better marketing targets.

Conclusion

The bank contacted Blue 15 to develop predictive models to identify customers likely to purchase their variable annuity product. We used the AUC to evaluate the performance of each of our five models. The XGBoost model achieved an AUC of 0.80 on the test dataset.

We recommend the bank investigate segmenting customers based on account balances, specifically in savings, checking, and certificate of deposit accounts. We also recommend the bank focus marketing efforts onto customers with an oldest account age of less than three years to target customers that are more likely to purchase the product.

Appendix

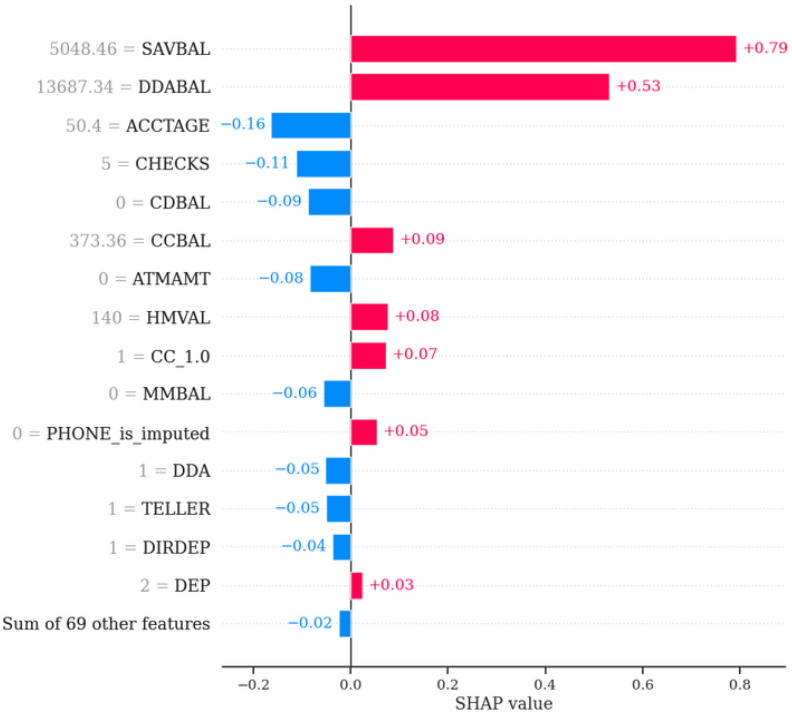

Figure 3: Full Shapley values for observation 732

Table 3: Permutation importance score for all XGBoost model variables

Rank	Variable
1	Savings acct balance
2	Checking acct balance
3	CD balance
4	Checking acct indicator
5	MM balance
6	Credit card indicator (yes)
7	Total ATM withdrawal amount
8	Number of checks written
9	IRA balance
10	Investment acct indicator (yes)
11	Retirement acct indicator
12	Number of checking deposits
13	Number of teller interactions
14	Branch B11
15	CC balance - imputation flag
16	Branch B16
17	Number of phone interactions - imputation flag
18	MM credits - 2
19	Credit score
20	Number of CC purchases - 3
21	Branch B14
22	Local address indicator
23	Amount of POS interactions - imputation flag
24	Number of phone interactions
25	Number of CC purchases - 1
26	Branch B4
27	Money market indicator
28	Number of NSF issues - imputation flag
29	Customer age
30	Credit score - imputation flag

Rank	Variable
31	Direct deposit indicator
32	Investest acct indicator - missing
33	Indicator for CD acct
34	Branch B9
35	Number of NSF issues
36	Credict card balance
37	Number of POS interactions
38	Branch B15
39	Branch B8
40	Number of checks written - imputation flag
41	Branch B10
42	Amount of POS interactions
43	Account age
44	Branch B7
45	Branch B18
46	MM balance imputation flag
47	CD balance imputation flag
48	Safety deposit box indicator
49	ATM indicator
50	Brach B12
51	Number of teller interactions - imputation flag
52	Account age - imputation flag
53	Total ATM withdrawal amount - imputation flag
54	Branch B17
55	Income - imputation flag
56	Branch B3
57	Total amount deposited - imputation flag
58	Length of residence
59	Branch B6
60	Number of checking deposits - imputation flag

Rank	Variable
61	Credit card indicator - missing
62	Number of CC purchases - 2
63	Savings acct balance - imputation flag
64	Home value - imputation flag
65	Savings acct indicator
66	Investment acct balance
67	Number of POS interactions - imputation flag
68	Investment account balance - imputation flag
69	Branch B13
70	Customer age - imputation flag
71	Total amount deposited
72	Checking acct balance - imputation flag
73	Branch B19
74	Number of CC purchases - missing
75	IRA balance - imputation flag
76	Income
77	Number of MM credits - 1
78	Length of residence - imputation flag
79	Number of MM credits - 3+
80	Number of insufficient fund issues
81	Home value
82	Branch B2
83	Number of CC purchases - 4
84	Branch B5

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 initialed checklist attached.

Sections & Structure

Overview

VB	Is the overview concise?
VB	Does it provide context about the business problem? <Content>
VB	Does it briefly address your team's work, quantifiable results, and recommendations? <Action>
VB	Does it offer audience-centered reasons for recommendations? <Context>

Body Sections

VB	Does the report body include information on methods, analysis, quantifiable results, and recommendations?
VB	Is content grouped into appropriate sections (*methodology*, *analysis*, *results*, *recommendations*)?

Conclusion

VB	Does the report have a conclusion?
VB	Does the conclusion sum up the report and emphasize relevant takeaways?

Structure

VB	Does each major section have a heading?
VB	Are sections, subsections, and paragraphs organized logically for easy navigation?

Visuals

Introduction, Discussion, and Captions

EM	Is each visual introduced in the text before it appears?
EM	Is each visual close to where it is introduced?
EM	Does each visual include a title with the following information: type (*table* or *figure*), number, and a descriptive caption?
EM	Is each visual discussed and interpreted in the text?
EM	Are figures and tables numbered separately?
EM	Are table captions above the table? Are figure captions below the figure?

Visual Design

EM	Do figures/tables use audience-friendly labels rather than variable names?
EM	Are the visuals easy to interpret?
EM	Are the visuals appropriately sized?
EM	Do tables appear on one page (*not split between 2 pages*)?
EM	Are legends and axis labels included for figures?
EM	Are numbers in tables right aligned?
EM	Are the visuals designed well (*ex*: *re-created in Word or Excel*, *not blurry or stretched*,…)?

Document Design

Title Page Design

EM	Does it include a descriptive title?
EM	Does it state the team name, team members' names, and the submission date?

Table of Contents Design

EM	Does it list all the major sections of the report with corresponding page numbers?
EM	Do the page numbers and sections in the Table of Contents match the report?

Document Design for Entire Report

VB	Is a standard typeface (*Calibri*, *Arial*, *etc.*) used?
VB	Is the size of the body text between 10-12 pt.?
VB	Are headings and subheadings used to organize information?
VB	Are distinctive text styles (*bold*, *italic*, *etc.*) used to distinguish between heading levels?
VB	Are text styles for headings used consistently (*ex*: *all level-one headings are bold*)?
VB	Are all paragraphs an appropriate length (*fewer than 12 lines*)?
VB	Is white space used to indicate paragraph breaks?
VB	Are bullet lists used for a series of items and numbered lists to show a hierarchy?

Writing Style and Mechanics

Spelling and Capitalization

VB	Are spelling errors located and corrected?
VB	Is spelling consistent throughout (*no switching between acceptable spellings*)?
VB	Is capitalization used appropriately (*proper nouns*, *etc*.)?
VB	Is capitalization of words consistent throughout the report?

Grammar and Punctuation

VB	Are verb tenses used appropriately?
VB	Are marks of punctuation used appropriately?
VB	Is subject-verb agreement used in every sentence?
VB	Is the grammar checker updated and are underlined grammar issues addressed?

Writing Style

VB	Are all sentences in the report easy for your audience to understand quickly?
VB	Are most sentences written in active voice?
VB	Are idioms and vague words eliminated from the report?
VB	Are acronyms introduced before being used?
VB	Are well-written topic sentences included at the beginning of each paragraph?
VB	Are lists parallel?
VB	Is the appropriate point of view used when addressing your audience or describing team actions?