BANKING INSURANCE PRODUCT: RANDOM FOREST AND XGBOOST

BLUE 15

VANSHIKA BHARDWAJ
GARRETT CURRAN
SUYEON KIM
KEVIN LLOYD
ERIC MILLER

NOVEMBER 18, 2024

Table of Contents

BANKING INSURANCE PRODUCT: RANDOM FOREST AND XGBOOST

Overview

The Commercial Banking Corporation, hereinafter referred to as 'the bank,' partnered with Blue 15 to create predictive models using random forest and extreme gradient boosting (XGBoost) to identify customers likely to purchase a variable-rate annuity product, hereinafter referred to as 'the product.'

We suggest the bank thoroughly investigate data collection practices at branches 14, 15, 18, and 19. This review aims to address the significant amount of missing data observed across eight key variables, which could impact the accuracy and reliability of our analysis. Both the random forest and XGBoost models are likely overfit, so we will evaluate their performance using the test dataset in the next phase.

Methodology & Analysis

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

Data Used

The 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.

For the continuous variables, we imputed the median for missing values and made an imputation indicator column. Though the bank typically imputes mode for categorical variables, we decided to explore whether missingness was impactful in the prediction of our target variable. Thus, we imputed "M" for categorical variables to represent missingness. For the money market credits variable, we binned values of three or more to address linear separation.

Random Forest

We built a random forest model to classify customers on whether or not they would purchase the product. We optimized the number of trees in this model by finding a point at which additional trees were not increasing our model's performance. We performed variable selection by introducing a randomly generated variable and removing predictors that improved the model's accuracy less than the random variable. We also extracted variable importance using the mean decrease in model accuracy when excluding each variable.

XGBoost

We built an XGBoost model to predict which customers would purchase the product. We optimized the sample proportion of the data for each iteration, the algorithm's learning rate, the maximum depth of the decision trees, and the number of boosting rounds. We performed optimization using ten-fold cross-validation. We performed variable selection using the same method as the random forest model. After reducing the variables in the model, we re-optimized the model parameters.

Results & Recommendations

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

Random Forest Results

Initially, we built a random forest model with all 43 of the provided predictor variables. After performing variable selection, our final model utilized 39 of the predictors. Table 1 shows the top 10 out of the 39 variables we included in order of importance to our model.

Table 1: Random forest top ten variable importance rankings

Rank	Variable Name	Relative Importance
1	Savings balance	100%
2	Checking acct balance	73%
3	Total amount deposited	58%
4	CD balance	52%
5	Number of checks written	48%
6	Indicator for certificate of deposit account	43%
7	Total ATM withdrawal amount	42%
8	Investment acct indicator	36%
9	MM balance	34%
10	Number of credit card purchases	31%

In Table 1, we can see four of the top five important variables have to do with the quantity of money in a customer's account, including the top variable, which is a customer's savings account balance. From this, we recommend that the bank segment customers based on account balances and appropriately market to the different customer segments.

Figure 1 shows the receiver-operating characteristic (ROC) curve with the optimal Youden's Index, which is the optimal cutoff value for the random forest model.

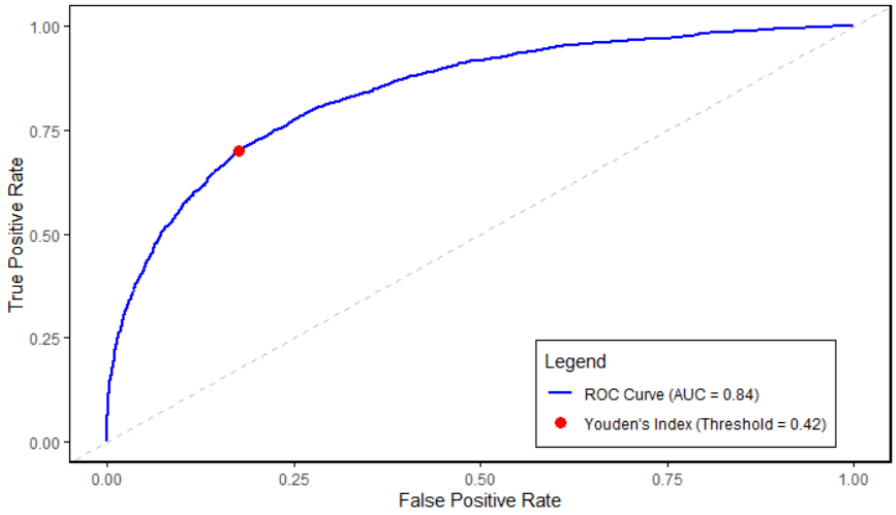

Figure 1: ROC curve for the random forest model

From Figure 1, we utilized Youden's Index of 0.422 to select the optimal threshold to classify a customer on whether or not they would likely purchase the product. Our random forest model achieved an area under the curve (AUC) of 0.84, indicating that our model correctly distinguishes customers who did and did not purchase the product 84% of the time.

XGBoost Results

We built an XGBoost model with 77 total variables, including dummy variables. After conducting variable selection, our final model utilized 51 total variables. Table 2 shows the 10 out of 51 variables retained in the final XGBoost model sorted by their importance measured by gain.

Table 2: XGBoost top ten variable importance rankings

Rank	Variable Name	Relative Importance
1	Checking acct indicator	100%
2	MM balance	28%
3	Savings balance	26%
4	Imputation flag for number of telephone banking interactions	25%
5	IRA acct indicator	19%
6	CD balance	18%
7	Investment acct indicator - 1	14%
8	Checking acct balance	14%
9	Branch 14	14%
10	Branch 15	12%

In Table 2, we see that the most important indicator in the XGBoost model was the checking account indicator, which was almost four times as important as the next highest variable, money market balance.

When comparing the two models, five of the top 10 important variables were the same. These were the MM balance, savings balance, CD balance, investment acct indicator, and checking acct balance. These variables all have to do with investment accounts or account balances.

Figure 2 shows the ROC curve with the optimal Youden's Index point, which was used to recommend the optimal cutoff value for the XGBoost model.

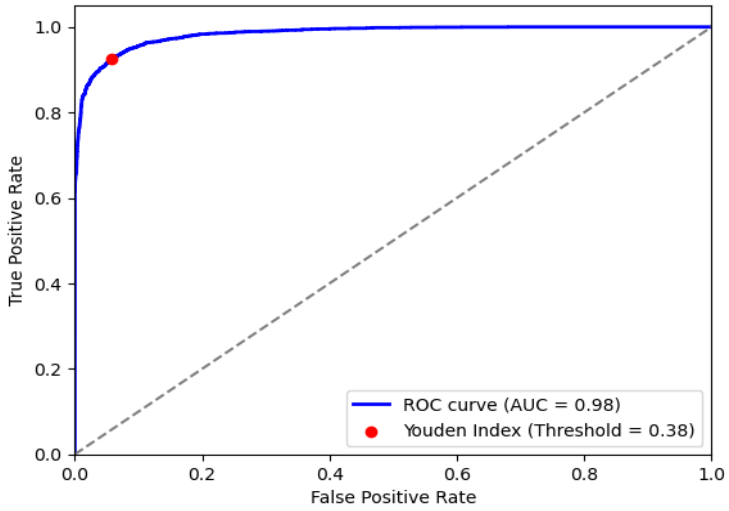

Figure 2: ROC curve for the XGBoost model

Figure 2 shows the optimum threshold in predicted probability to classify a customer on whether they would purchase the product. The optimum cutoff value for distinguishing between customers who would or would not be likely to purchase the product was 0.394. The AUC for this value was 0.98, indicating that our model is 98% accurate in distinguishing between customers who did and did not purchase the product. A predictive value this high is likely due to the model being overfit to the training data.

Recommendations

We recommend that the bank take the following actions:

- Investigate patterns in the missingness of the number of telephone banking interactions.

  Missingness was the fourth top predictor in the XGBoost model. Branches 14, 15, 18, and 19 had 1074 missing values; branches 14 and 15 are also highly predictive. Better data collection at these branches could improve the model and boost sales.

- Wait for the results of our next analysis before making any drastic marketing decisions.

  Both the random forest and XGBoost models are likely overfit to the training data. The next phase of our analysis will use the test dataset to assess model performance, providing a more generalized understanding of model performance.

Conclusion

The bank contacted Blue 15 to develop predictive models using the random forest algorithm and XGBoost to identify customers likely to purchase their variable annuity product. We used the AUC to evaluate the performance of each model. The random forest model yielded an AUC of 0.84, while the XGBoost model achieved an AUC of 0.98.

We recommend the bank to review data collection processes in branches 14, 15, 18, and 19 to address substantial missing data in eight key variables. Additionally, the random forest and XGBoost models seem to be overfitting the training data. Therefore, we recommend waiting for the results from the next phase, where we will use the test dataset to obtain a more generalized model performance.

Appendix

Table 3: All random forest variables ranked by importance

Rank	Variable Name	Importance (mean decrease in accuracy)
1	Savings balance	28.431
2	Checking acct balance	20.757
3	Total amount deposited	16.357
4	CD balance	14.813
5	Number of checks written	13.702
6	Indicator for certificate of deposit account	12.179
7	Total ATM withdrawal amount	12.038
8	Investment acct indicator	10.222
9	MM balance	9.588
10	Number of credit card purchases	8.794
11	Branch	8.637
12	Retirement account balance	7.434
13	Credit card indicator	7.33
14	Number of checking acct deposits	7.267
15	Indicator for retirement account	6.693
16	Indicator for money market account	6.635
17	Checking acct indicator	6.63
18	Income	6.176
19	Credit card balance	6.133
20	Investment account balance	5.758
21	Age of oldest account	5.47
22	Home value	5.351
23	IMP.1075	4.969
24	Amount of POS interactions	4.764
25	Indicator for ATM interaction	4.489
26	Indicator for savings account	4.181
27	Number of telephone banking interactions	3.872
28	Number of POS interactions	3.869

Rank	Variable Name	Importance (mean decrease in accuracy)
29	Number of telephone banking interactions	3.097
30	MM credits	2.397
31	Indicator for direct deposit	2.209
32	Safety deposit box indicator	1.913
33	IMP.1537	0.511
34	Imputation flag for account age	0.334
35	Number of insufficient funds issues	0.292
36	Amount of NSF	0.06
37	Credit score	-0.219
38	Local address indicator	-1.597
39	Imputation flag for credit score	-1.885

Table 4: All XGBoost variables ranked by importance

Rank	Variable Name	Relative Importance
1	Checking acct indicator	100%
2	MM balance	28%
3	Savings balance	26%
4	Imputation flag for number of telephone banking interactions	25%
5	IRA acct indicator	19%
6	CD balance	18%
7	Investment acct indicator - 1	14%
8	Checking acct balance	14%
9	Branch 14	14%
10	Branch 15	12%
11	Credit card indicator - 1	12%
12	Branch 16	11%
13	Branch 9	11%
14	Branch 6	10%
15	Amount of NSF	10%
16	Number of teller visit interactions	9%
17	Investment acct indicator - 0	9%
18	IRA balance	9%
19	Branch 3	9%
20	Number of checks written	8%
21	Credit card balance	8%
22	Amount of POS interactions	8%
23	MM credits - 1	8%
24	Acct age	8%
25	Total ATM withdrawal amount	8%
26	Credit Card Purchases - 1	7%
27	Total amount deposited	7%
28	Indicator for credit card - 0	7%

Rank	Variable Name	Relative Importance
29	Indicator for savings acct	7%
30	MM credits - 2	7%
31	Number of telephone banking interactions	7%
32	Branch 17	7%
33	Branch 5	7%
34	Number of POS interactions	7%
35	Safety deposit box indicator	7%
36	Direct deposit indicator	7%
37	Income	7%
38	Home value	6%
39	Credit score	6%
40	Investment acct balance	6%
41	Customer age	6%
42	Credit Card Purchases - 0	6%
43	Length of residence (years)	6%
44	Local address indicator	6%
45	Imputation flag for income	5%
46	Credit Card Purchases - 3	4%
47	Credit Card Purchases - 2	3%
48	Credit Card Purchases - 4	3%
49	MM credits - 3+	1%
50	Missingness flag for credit card indicator	0%
51	Missingness flag for number of credit card purchases	0%

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

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

Body Sections

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

Conclusion

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

Structure

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

Visuals

Introduction, Discussion, and Captions

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

Visual Design

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

Document Design

Title Page Design

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

Table of Contents Design

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

Document Design for Entire Report

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

Writing Style and Mechanics

Spelling and Capitalization

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

Grammar and Punctuation

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

Writing Style

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