Banking Insurance Product: Phase 2

Blue 10

Alison Bean de Hernández
Austin Minihan
Ishi Gupta
Patrick Costa

September 11, 2024

Table of Contents

Banking Insurance Product: Phase 2

Overview

The Department of Customer Services and New Products at the Commercial Banking Corporation, hereinafter called 'the Bank,' partnered with Blue Team 10 to create a model that predicts which customers will buy their variable-rate annuity product, hereinafter called the 'product.'

Table 1 provides an overview of how the variable count changed during each step of the modeling process.

Table 1: Count of main effects and interaction terms in each modeling step

Categories	Full model with all variables	Main effects model	Final forward selection model
Number of main effects variables	47	14	14
Number of interaction terms	0	0	1

Our final model contained one interaction term and 14 main effects variables, as shown in Table 1. Our team found a significant interaction between the indicators for having a retirement account and having a checking account. We recommend including this interaction in the final prediction model.

Our team recommends that the Bank gear its marketing efforts to people who (1) have multiple types of accounts at the bank and (2) have more money in their accounts. Specifically, people with higher checking accounts, savings accounts, and certificate of deposit (CD) balances positively impacted the likelihood of a customer buying the product. Furthermore, customers with checking accounts are 2.16 times more likely to purchase the product than those without a checking account, given that they already have a retirement account at the Bank.

Methodology

The following section explains the steps we took to clean and prepare the data and the methods we used for model selection.

Data Cleaning and Preparation

The data contained information about bank customers who were offered the product. The complete dataset contained 10,619 observations. It was split into a train and validation set with 8,495 observations and 2,124 observations, respectively. The dataset included 47 attributes (variables) about the customer and an indicator variable indicating if the customer purchased the product. The Bank binned continuous variables to transform them into categories.

Blue team 10 checked for missing values. If a value was missing, 'M' was imputed to create a 'missing' category. We assessed the complete and quasi-separation of each variable using cross tables.

Model Selection

Our team used a combination of model selection techniques to create a logistic model for predicting the purchase of the product, given the provided data. We used p-value-based backward selection with the

provided significance level of $\alpha = 0.002$ to create an initial set of significant predictor variables for the target. P-value-based forward selection was subsequently performed to add significant interaction terms to the initial model, resulting in a final recommended model.

Analysis & Results

The following subsections explore which variables we included in the main effects model and the final interaction model. Likewise, we look at some interpretations of the odds ratios from the final model selected.

Data Cleaning and Preparation

Our analysis found that the number of cash-back requests and the number of money market (MM) credits had a quasi-complete separation issue. Since the people who requested cash back twice did not have the product, we solved this issue by creating a "one or more" category. Similarly, when the number of MM credits exceeded three, there was insufficient data related to the product. Thus, a new category labeled "three or more" was created to address this convergence issue.

Main Effects

From the backward selection process, the response variable features 14 variables selected as significant predictors. As listed in Table 2, six of these variables were binary with updated missing value codings, while eight were continuous variables converted to categorical through predetermined cutoffs.

Table 2: Main effects table ranked by significance selected from backward selection

Variable	Type	P-value
Savings account balance bin	Categorical Bin	8.01E-129
Checking account balance bin	Categorical Bin	5.63E-60
CD balance bin	Categorical Bin	2.58E-39
MM balance bin	Categorical Bin	2.37E-23
Number of checks written bin	Categorical Bin	5.76E-20
Withdrawal amount bin	Categorical Bin	9.09E-10
Number of teller visit interactions bin	Categorical Bin	1.93E-08
Indicator for CC	Binary	1.53E-07
Indicator for checking account	Binary	1.12E-05
ILS balance bin	Categorical Bin	5.64E-05
Indicator for investment account	Binary	1.15E-04
Indicator for mortgage	Binary	6.62E-04
Number of insufficient fund issues	Binary	9.57E-04
Indicator for IRA account	Binary	8.80E-01

Table 2 shows that after backward selection our main effects model contained 14 variables, which are ranked by p-value.

Final Model

From the forward selection process, one interaction term was included alongside the 14 original main effect variables. This new term featured an interaction between the indicator variable for checking accounts and the indicator variable for retirement accounts. Table 3 ranks all variables in the final recommended logistic regression model by significance.

Table 3: Final logistic regression model's variables ranked by significance

Variable	P-value
Indicator for checking account	1.12E-05
Number of insufficient fund issues	9.57E-04
Indicator for IRA account	8.80E-01
Indicator for investment account	1.15E-04
Indicator for mortgage	6.62E-04
Indicator for CC	1.53E-07
Checking account balance bin	5.63E-60
Number of checks written bin	5.76E-20
Number of teller visit interactions bin	1.93E-08
Savings account balance bin	8.01E-129
Withdrawal amount bin	9.09E-10
CD balance bin	2.58E-39
ILS balance bin	5.64E-05
MM balance bin	2.37E-23
Interaction between checking account and IRA account	3.13E-04

Table 3 shows variables in the final logistic regression model ranked by their significance.

Odds Ratio

Table 4 shows selected odds ratios based on account balances of different account types. The general trend we found was that people with higher account balances were more likely to buy the product. For example, when the person had a checking account larger than $6,126.24, they were 7.93 times more likely to buy the product than those with an account balance of $0.10. Additionally, when the savings balance was greater than or equal to $8,334.97, the person was 5.97 times more likely to purchase the product than those with a balance of $0.01.

Table 4: Selected odd ratios based on account balances

Variable	Odds Ratio
Checking account balance > $6,126.24	7.93
Savings account balance > $8,334.97	5.97
Certificate of deposit balance > $9,200	4.09
Interaction: Checking account given IRA account	2.16

Table 4 displays chosen odds ratios based on account balances. Another group of individuals who were saving were people with retirement accounts and checking accounts. Given that a customer had a retirement account, they were 2.16 times more likely to buy the product if they also had a checking account.

Recommendations

Based on the results of Blue Team 10's final model, we recommend that marketing strategies center around the following groups:

- Customers who already have retirement accounts, specifically if those individuals also have a checking account.
- Customers with higher account balances as outlined in Table 4.

Conclusion

Using multiple variable selection methods, Blue Team 10 found 14 main effects variables and one interaction term in our final prediction model. This significant interaction was between the indicators for having a retirement account and having a checking account. Moving forward, we recommend including this interaction in the final model. Finally, recommendations were given to increase marketing to two groups of customers who were more likely to buy a product than the rest.