Banking Insurance Product: Phase 1

Blue 10

Alison Bean de Hernández

Austin Minihan

Ishi Gupta

Patrick Costa

August 30, 2024

Table of Contents

Overview 1

Methodology 1

Data Used 1

Significance 1

Test for Linearity 2

Analysis & Results 2

Missing Data 2

Significance 2

Binary Variable Interpretation 3

Linearity Assumptions 4

Recommendations & Conclusion 4

Appendix 5

Banking Insurance Product: Phase 1

Overview

The Commercial Banking Corporation, hereinafter called 'the Bank,' partnered with Blue Team 10 to create a consumer profile that predicts which customers will buy their variable-rate annuity product, hereinafter called the 'product.' Through a preliminary data evaluation, our team determined which variables are significant in relation to the Bank's product.

Table 1 displays the number of variables that could be included in the prediction model by removing potentially redundant binary variables with continuous counterparts, such as indicators for ATM use, money market (MM) account, and retirement (IRA) account.

Table 1: Number of significant variables

Categories	Total amount	Non-significant	Significant	Potential model inputs
Number of variables	48	17	28	25

Furthermore, we determined that further investigation into how the branches collect data is necessary. Identifying data collection errors could improve data quality because the missing values come from the same individuals across selected bank branches. Likewise, if certain branches are not offering select products, that would be important for future modeling.

Methodology

The following section will describe the observations and methods used to evaluate whether the Bank's customers purchased the product.

Data Used

The data contained information about bank customers who were offered the product. It included 47 attributes (variables) about the customer and an indicator variable indicating if the customer purchased the product. We checked the data for missing values and evaluated relationships with the purchase of the product. The complete data set contained 10,619 observations. The data provided by the Bank was split into a train and validation set with 8,495 observations and 2,124 observations, respectively.

Significance

To create a ranked p-value table, our team evaluated each variable type separately.[1] We assessed binary variables using a Mantel–Haenszel test. Our team evaluated ordinal variables with a Fisher's exact test because the assumptions for the chi-square test were unmet due to the small number of data points. We analyzed nominal variables using a chi-square test, while we evaluated continuous variables using a likelihood ratio test with the reduced model containing only the intercept.

[1] Following the Bank's guidance, we considered variables with p-values of 0.002 and below to be significant.

Test for Linearity

Proper assumptions based on variable type must be met to judge a logistic regression model's performance. Our team assessed linearity assumptions for categorical variables based on effect degrees of freedom (EDF) and assessment of individual generalized additive models (GAM).

Analysis & Results

Our team found trends in the missing data that could impact the prediction model. Additionally, the following section outlines relevant results from significance testing, looks at binary variable interpretation, and reports on the linearity assumption results. These findings will impact future modeling decisions.

Missing Data

Figure 1 displays the count of missing values by variable in descending order. Eight out of 15 variables have the same number of missing observations, 1,075. These specific missing observations are associated with the branches B14, B15, B18, and B19. Upon further analysis, our team concluded that those 1,075 were the same observation across the variables, indicating that specific customers in the dataset are missing notable amounts of information. Similarly, variables missing 1,537 observations had 49% of the missing values associated with branch B2.

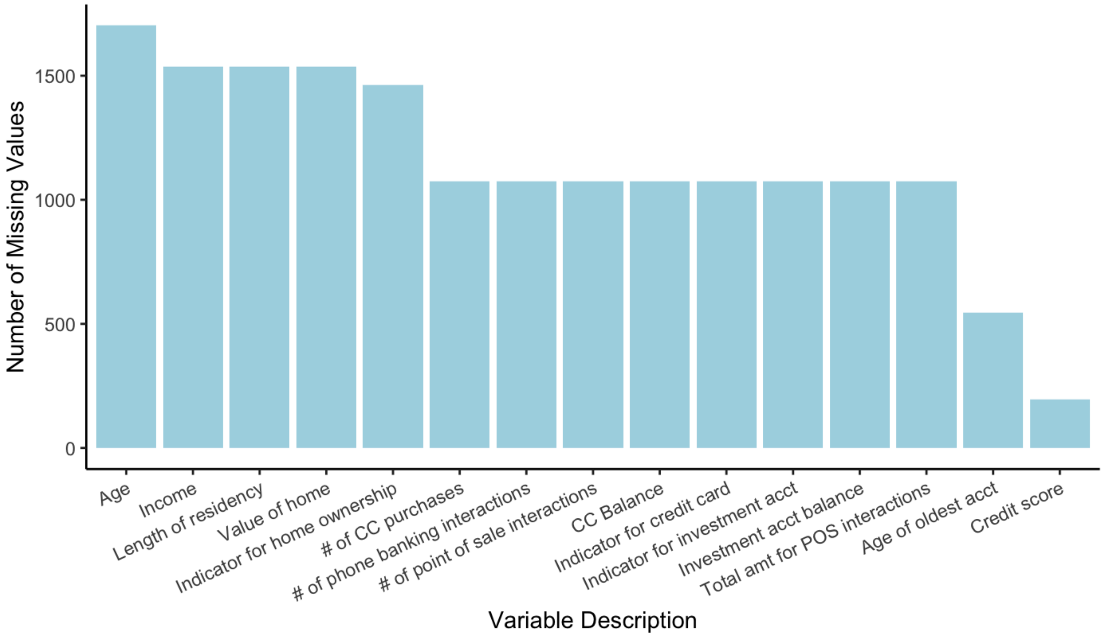

Figure 1: Missing values by variable in the provided dataset

Significance

Out of the 47 attribute variables, 28 were statistically significant. Of those, 13 were continuous, 12 were binary, three were ordinal[2], and one was nominal. Table 2 shows the significant variables ranked on the p-value.

[2] In this context, ordinal represents the non-binary ordinal variables.

Table 2: Significant variables ranked on p-value

Variable	Type	P-value
Savings acct balance	Continuous	1.91E-79
Indicator for certificate of deposit acct	Binary	2.97E-78
Indicator for checking acct	Binary	5.50E-70
CD balance	Continuous	1.30E-62
Indicator for money market acct	Binary	7.45E-57
MM acct balance	Continuous	2.70E-50
Checking deposits	Continuous	5.10E-41
Indicator for savings acct	Binary	1.44E-39
Indicator for IRA acct	Binary	4.89E-37
Indicator for CC	Binary	2.44E-32
Checking acct balance	Continuous	6.99E-32
Indicator for ATM interaction	Binary	1.59E-29
# of phone banking interactions	Continuous	1.90E-25
Indicator for investment acct	Binary	1.05E-21
IRA balance	Continuous	2.36E-19
# of MM credits	Ordinal	1.17E-15
Branch of bank	Nominal	5.49E-14
Value of home	Continuous	5.99E-13
# of CC purchases	Ordinal	4.50E-12
# of checks written	Continuous	6.41E-12
Indicator for direct deposit	Binary	3.67E-11
# of insufficient fund issues	Binary	4.93E-11
Indicator for safety deposit box	Binary	4.31E-10
Total ATM withdrawal amount	Continuous	1.34E-09
# of POS interactions	Continuous	1.62E-07
Indicator for local address	Binary	7.45E-07
Amt of NSF	Continuous	1.46E-05
Total amt deposited	Continuous	6.80E-05
# of cash back requests	Ordinal	1.19E-03

Binary Variable Interpretation

Table 3 shows the odds ratios for significant binary variables. For example, on average, customers who have an investment account are 3.47 times as likely to purchase the product than customers without an investment account. Customers who have an investment account, certificates of deposit, or IRA are all more than 3 times as likely than customers without those accounts to purchase the product.

Table 3: Odds ratios for significant binary variables

Variables	Odds Ratio
Indicator for investment acct	3.47
Indicator for certificate of deposit acct	3.43
Indicator for IRA acct	3.18
Indicator for MM acct	2.85
Indicator for savings acct	1.83
Indicator for CC	1.78
Indicator for safety deposit box	1.55
Indicator for direct deposit	0.71
Indicator for ATM interaction	0.59
Indicator for local address	0.57
Indicator for insufficient fund issues	0.56
Indicator for checking acct	0.38

Linearity Assumptions

Among all categorical variables within the provided dataset in Appendix Table 1, only the *value of home* passes linearity assumptions and features statistical significance within relevant tests.

Recommendations & Conclusion

Our team has three recommendations for the next steps related to the Bank's variable rate annuity product dataset.

First, dealing with the missing data, there are three options to consider:

- Impute the information and create a missing value binary variable for each continuous variable you impute. For categorical variables, create a missing value category.
- Explore why certain bank branches are missing more data. Do these branches not offer select services? Are there issues in the data collection process?
- Exclude the observations that are missing information across multiple variables.

Second, if variables are significant, the Bank should keep them in the dataset for future model creation. However, the Bank could consider removing some variables from the final dataset if there is potentially redundant information. Some potentially redundant information includes the following:

- *Indicator for ATM interaction; total ATM withdrawal amount*
- *Indicator for MM acct.; MM balance*
- *Indicator for IRA acct.; IRA balance*

The *indicator for ATM interaction* and *total ATM withdrawal amount* both indicate that customers used the ATM. Similarly, investment accounts such as *MM balance* and *IRA balance* would have a positive balance, which implies that the customer has an account. The Bank could consider dropping the binary counterparts and keeping the continuous data that provides more detail.

Last, we recommend that the Bank makes a logistic regression model using a combination of statistically significant variables to predict customers' likelihood of buying the product.

Appendix

Appendix Table 1: All provided variables ranked on p-value

Variable	Type	P-value
Savings account balance	Continuous	1.91E-79
Indicator for certificate of deposit account	Binary	2.97E-78
Indicator for checking account	Binary	5.50E-70
CD balance	Continuous	1.30E-62
Indicator for MM account	Binary	7.45E-57
MM acct balance	Continuous	2.70E-50
Checking deposits	Continuous	5.10E-41
Indicator for savings acct	Binary	1.44E-39
Indicator for retirement acct	Binary	4.89E-37
Indicator for CC	Binary	2.44E-32
Checking acct balance	Continuous	6.99E-32
Indicator for ATM interaction	Binary	1.59E-29
# of phone banking interactions	Continuous	1.90E-25
Indicator for investment acct	Binary	1.05E-21
IRA balance	Continuous	2.36E-19
Number of MM credits	Ordinal	1.17E-15
Branch of bank	Nominal	5.49E-14
Value of home	Continuous	5.99E-13
# of CC purchases	Ordinal	4.50E-12
# of checks written	Continuous	6.41E-12
Indicator for direct deposit	Binary	3.67E-11
# of insufficient fund issues	Binary	4.93E-11
Indicator for safety deposit box	Binary	4.31E-10
Total ATM withdrawal amt	Continuous	1.34E-09
# of POS interactions	Continuous	1.62E-07
Indicator for local address	Binary	7.45E-07
Amt of NSF	Continuous	1.46E-05
Total amt deposited	Continuous	6.80E-05
# of cash back requests	Ordinal	1.19E-03
CC balance	Continuous	2.21E-03
Indicator for installment loan	Binary	7.27E-03
Age of oldest account	Continuous	8.22E-03
# of teller visit interactions	Continuous	9.72E-03
Investment acct balance	Continuous	1.63E-02
Installment loan balance	Continuous	2.88E-02
Mortgage balance	Continuous	5.86E-02
Amt for point of sale interactions	Continuous	1.15E-01
Age	Continuous	2.19E-01
Area classification	Nominal	2.34E-01
Recent address change	Binary	2.37E-01

Variable	Type	P-value
Income	Continuous	2.58E-01
Credit score	Continuous	3.93E-01
Indicator for line of credit	Binary	4.99E-01
Indicator for Mortgage	Binary	5.28E-01
Length of residence in years	Continuous	8.51E-01
Line of credit balance	Continuous	9.12E-01
Indicator for home ownership	Binary	9.20E-01