|  | **Variable Annuity Product Analysis : Phase One**  Orange Team 21  Lev El-Askari, Tim Milowic, Gagan Namburi, Alex Raum, Ishanee Rudra    September 2, 2021 |
| --- | --- |

**Table of Contents**

[**Overview………………………………………………………………………………………………….1**](#_heading=h.e1zakx2idzk)

[**Methodology**](#_heading=h.la6peraii8re) **and Analysis……………………………………………………………………………**[**1**](#_heading=h.la6peraii8re)

[Data Used…………………………………………………………………………………………….**1**](#_heading=h.xg53hvut1rdu)

Determining Variable Significance………………………………………………………………....**1**

Odds Ratios…………………………………………………………………………………………..[**2**](#_heading=h.1noogksu7tvm)

[**Results**](#_heading=h.w9t3vktuhyd)**…………………………………………………………………………………………………....2**

Significant Variables………………………………………………………………………………....**2**

Odds Ratios…………………………………………………………………………………………..**4**

Linearity Assumption………………………………………………………………………………...**4**

Data Considerations………………………………………………………………………………....**4**

[**Recommendations**](#_heading=h.umcj18p01y9q)**……………………………………………………………………………………...4**

[**Conclusion**](#_heading=h.ayi1k68qg46t)**……………………………………………………………………………………………….4**

**Annuity Product Analysis: Phase 1**

# **Overview**

We have been tasked with predicting which customers will buy a variable rate annuity product from the Commercial Banking Corporation (referred to as the “Bank”). For the purpose of this report, we are primarily focusing on identifying which variables had a significant impact on variable rate annuity purchase decisions. We established the following takeaways:

1. We identified 13 variables of concern that had over 1000 missing values.
2. Twenty-eight predictor variables were significantly related to the purchase of the variable rate annuity product.
3. The indicator variable for investment account had the highest odds ratio (3.472) of all binary predictors.
4. Ten of the continuous variables met the assumption for linearity.

# **Methodology and Analysis**

## *Data Used*

The data contained demographic information for customers of the Bank. We used the training data set containing 8,495 observations of 47 predictor variables. The data were checked for missing values to preclude any potential problems with the analysis. The variables with over 1000 missing values are shown below in Figure 1.

Chart

Description automatically generated

**Figure 1: A bar graph showing the 13 variables with a high number (> 1000) of missing values**

## *Determining Variable Significance*

The variables were first typified as Continuous, Nominal, Ordinal, or Binary. The continuous variables were then evaluated for significance by building individual logistic regression models containing the continuous variable of interest as the only predictor. The ordinal and binary variables were evaluated using the Mantel-Haenszel Chi-Square test of association, and the nominal variables were evaluated with the Pearson Chi-Square test. Each test was used with a significance level of 0.0008 based on the recommendations of Adrian Raftery for datasets of comparable size (Raftery, 1995).

## *Odds Ratios*

Below, Table 1 contains all of the Odds ratios we calculated for the binary variables in order to quantify their impact. The table was then sorted by magnitude of the odds ratios. The highest odds ratio is interpreted in results. The (i) next to a variable indicates that that variable is an indicator variable

**Table 1: Sorted Odds Ratios for Binary Predictors**

| **Variable Name** | **Odds Ratio** |
| --- | --- |
| Investment Account (i) | 3.472 |
| Certificate of Deposit Account (i) | 3.427 |
| Retirement Account (i) | 3.185 |
| Money Market Account (i) | 2.850 |
| Savings Account (i) | 1.831 |
| Credit Card (i) | 1.781 |
| Safety deposit box (i) | 1.550 |
| Mortgage (i) | 1.069 |
| Line of credit (i) | 1.065 |
| Home ownership (i) | 1.005 |
| Recent address change (i) | 0.840 |
| Installment loan (i) | 0.745 |
| Direct deposit (i) | 0.712 |
| ATM interaction (i) | 0.593 |
| Local address (i) | 0.575 |
| Number of insufficient fund issues | 0.555 |
| Checking account (i) | 0.375 |

# **Results**

## Significant Variables

After we completed association testing and preliminary modeling for each variable in the data set, we ranked the 28 significant variables (at the 0.0008 significance level) by decreasing level of significance and labeled them by class (binary, ordinal, nominal, continuous). Table 2 includes the significant variables ranked by their p-value, the test used to ascertain significance, and whether or not the continuous variables met the assumption of linearity for logistic regression (see appendix B for full table).

**Table 2: Significant Variables Ranked by p-value**

| Variable Description | P-Value | Class | Test | Linearity Assumption |
| --- | --- | --- | --- | --- |
| Certificate of deposit indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Checking account indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Money market account indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Money market account balance | < 0.000001 | Continuous | Logistic Regression | Met |
| Savings account indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Number of checking deposits | < 0.000001 | Continuous | Logistic Regression | Met |
| Retirement account indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Certificate of deposit balance | < 0.000001 | Continuous | Logistic Regression | Met |
| Credit card indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| ATM interaction indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Checking account balance | < 0.000001 | Continuous | Logistic Regression | Met |
| Investment account indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Number of telephone transactions | < 0.000001 | Continuous | Logistic Regression | Met |
| Number of money market credits | < 0.000001 | Ordinal | Mantel-Haenszel | Not Applicable |
| Branch of bank | < 0.000001 | Nominal | Chi Squared | Not Applicable |
| Value of home | < 0.000001 | Continuous | Logistic Regression | Not Met |
| Number of checks written | < 0.000001 | Continuous | Logistic Regression | Met |
| IRA balance | < 0.000001 | Continuous | Logistic Regression | Met |
| Direct deposit indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Number of insufficient fund issues | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Number of credit card purchases | < 0.000001 | Ordinal | Mantel-Haenszel | Not Applicable |
| Safety deposit box indicator | < 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Total ATM withdrawal amount | < 0.000001 | Continuous | Logistic Regression | Met |
| Number of point-of-sale transactions | 0.000001 | Continuous | Logistic Regression | Met |
| Local address indicator | 0.000001 | Binary | Mantel-Haenszel | Not Applicable |
| Amount of NSF | 0.000135 | Continuous | Logistic Regression | Met |
| Total amount deposited | 0.000372 | Continuous | Logistic Regression | Met |
| Number of cash back requests | 0.000706 | Ordinal | Mantel-Haenszel | Not Applicable |

## Odds Ratios

As displayed in Table 1 we calculated and ranked the odds ratios for each of the 17 binary predictor variables. Row one of the table shows that the investment account indicator had an odds ratio of 3.472, indicating that customers with an investment account are approximately three and a half times more likely to purchase an annuity compared to customers without an investment account. It also appears that customers who actively save or invest have higher odds of purchasing an annuity than customers that do not.

## Linearity Assumption

Testing of the linearity assumption for continuous variables revealed that, of the 25 continuous variables, ten variables met the assumption of linearity. The results of these tests for significant variables are included in Table 2. A full summary of these tests for all continuous variables is included in the table in Appendix B.

## Data Considerations

We also identified several redundant variables among the indicators for account types and their corresponding balances. Other paired redundancies were home ownership/mortgage balance and credit card/line of credit*.* We also detected a nearly one-to-one relationship between the mortgage balance and credit card balance variables, which had a correlation coefficient of approximately 0.95.

# Recommendations

We suggest using the variables that were significantly related to the purchase of an insurance product. Due to the large sample size (~ 8500), we advise an alpha level of 0.0008 be used rather than the current alpha level of 0.002. This more conservative approach will provide greater confidence that the findings from your data are reliable.

To address the redundant variable issue, we recommend removing the indicator variables: certificate of deposit, installment loan, and money market account from the dataset. There are other pairs of indicator and balance variables that do include additional information, but we suggest that the Bank keep these variables or consider combining the pairs into a single variable.

# Conclusion

From our analysis, we concluded that 28 out of the 47 variables in the Commercial Banking Corporation’s dataset significantly impact whether an individual will purchase variable rate annuity products. In particular, we discovered that customers with an investment account are approximately 3 and a half times more likely to purchase an annuity than customers without an investment account. The continuous variables that did not meet the linearity assumption will require closer examination. The missing data and redundant variables present in the dataset could be a cause for concern and warrant further investigation.

# Appendix A

In assessing the linearity assumption of the continuous variables, we used spline estimation to fit a nonlinear function to the continuous variables. Since the null hypothesis is that a spline estimation provides no value to our analysis beyond a linear relationship, a p-value greater than our significance level of 0.0008 indicates a lack of evidence for a nonlinear relationship.

# Appendix B

A full list of all 47 predictor variables, their significance, their type, and whether or not they meet the assumption of linearity is provided below in Table 3. The table is sorted in descending order of significance.

**Table 3: Full List of Variables Ranked By Significance**

| **Variable Description** | **Type** | **Significance (p-value)** | **Linearity Assumption** |
| --- | --- | --- | --- |
| Certificate of Deposit Account (i) | Binary | < 0.000001 | Not Applicable |
| Checking Account (i) | Binary | < 0.000001 | Not Applicable |
| Money Market Account (i) | Binary | < 0.000001 | Not Applicable |
| Savings Account Balance | Continuous | < 0.000001 | Not Met |
| Money Market Account Balance | Continuous | < 0.000001 | Not Met |
| Savings Account (i) | Binary | < 0.000001 | Not Applicable |
| Number of Checking Deposits | Continuous | < 0.000001 | Not Met |
| Retirement Account (i) | Binary | < 0.000001 | Not Applicable |
| CD Balance | Continuous | < 0.000001 | Not Met |
| Credit Card (i) | Binary | < 0.000001 | Not Applicable |
| ATM Interaction (i) | Binary | < 0.000001 | Not Applicable |
| Checking Account Balance | Continuous | < 0.000001 | Not Met |
| Investment Account (i) | Binary | < 0.000001 | Not Applicable |
| Number of Phone Banking Interactions | Continuous | < 0.000001 | Not Met |
| Number of Money Market Credits | Ordinal | < 0.000001 | Not Applicable |
| Bank Branch | Nominal | < 0.000001 | Not Applicable |
| Home Value | Continuous | < 0.000001 | Met |
| Number of Checks Written | Continuous | < 0.000001 | Not Met |
| Retirement Account Balance | Continuous | < 0.000001 | Not Met |
| Direct Deposit (i) | Binary | < 0.000001 | Not Applicable |
| Number of Insufficient Funds Issues | Binary | < 0.000001 | Not Applicable |
| Number of Credit Card Purchases | Ordinal | < 0.000001 | Not Applicable |
| Safety Deposit Box (i) | Binary | < 0.000001 | Not Applicable |
| Total ATM Withdrawal Amount | Continuous | 0.000001 | Not Met |
| Number of POS Interactions | Continuous | 0.000001 | Not Met |
| Local Address (i) | Binary | 0.000001 | Not Applicable |
| Total Insufficient Fund Amount | Continuous | 0.000135 | Not Met |
| Total Amount Deposited | Continuous | 0.000372 | Not Met |
| Number of Cash Back Requests | Ordinal | 0.000706 | Not Applicable |
| Credit Card Balance | Continuous | 0.003215 | Met |
| Installment Loan (i) | Binary | 0.007265 | Not Applicable |
| Account Age | Continuous | 0.007871 | Met |
| Number of Teller Interactions | Continuous | 0.009309 | Met |
| Installment Loan Balance | Continuous | 0.031314 | Not Met |
| Investment Account Balance | Continuous | 0.039032 | Not Met |
| Mortgage Balance | Continuous | 0.059459 | Met |
| Amount from POS Interactions | Continuous | 0.119230 | Not Met |
| Age | Continuous | 0.218974 | Met |
| Area Classification | Nominal | 0.234264 | Not Applicable |
| Recent Address Change (i) | Binary | 0.237067 | Not Applicable |
| Income | Continuous | 0.256665 | Met |
| Credit Score | Continuous | 0.393260 | Met |
| Line of Credit (i) | Binary | 0.499495 | Not Applicable |
| Mortgage (i) | Binary | 0.528115 | Not Applicable |
| Length of Residence (Years) | Continuous | 0.851265 | Met |
| LOC Balance | Continuous | 0.911210 | Met |
| Home Ownership (i) | Binary | 0.919609 | Not Applicable |

# References

Raftery, A. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology,* *25*, 111-163. doi:10.2307/271063