PREDICTING VARIABLE ANNUITY PLAN PURCHASES

VARIABLE UNDERSTANDING AND ASSUMPTIONS

BLUE TEAM 18 - SENCE CONSULTING

Sanket Sahasrabudhe
Ethan Scheper
Noah Johnson
Charis Williams
Elizabeth Surratt

SEPTEMBER 1ST, 2022

Table of Contents

PREDICTING VARIABLE ANNUITY PLAN PURCHASES

Overview

Commercial Banking Corporation (the Bank) is seeking to identify a customer base likely to purchase a variable rate annuity product. The Bank has hired SENCE Consulting to target these customers and assist with predictive modeling. Logistic regression modeling provides an interpretable, effective way of predicting the probability of the customer buying the product. The focus of this report is a preliminary step in the model-building process: understanding the data and assessing assumptions. We found that 29 of 47 variables were statistically significant, and six of these significant variables had missing values. Of the 13 continuous, significant variables, only one had a linear relationship with the log odds. Furthermore, six pairs of significant variables accomplish the same purpose and are thus potentially redundant pairs. Therefore, we recommend continuing variable selection by alleviating the concerns of nonlinear, largely missing, and redundant predictors through transformation, imputation, and elimination. Thorough variable selection should ensure model interpretability and simplicity.

Methodology and Analysis

Predictor Variable Classification and Significance

The Bank provided our team with 47 predictor variables that describe various attributes of their customer base. Leading into variable selection, significance tests were run with each predictor individually on the target variable **INS**, a binary variable indicating whether a customer bought the variable annuity plan.

For each ordinal and binary predictor variable, a Mantel-Haenszel chi-square test was used. Similarly, for the nominal predictor variable **BRANCH**, a Pearson chi-square test was used. Additionally, a Likelihood Ratio Test was performed on the logistic regression model with each continuous variable. We treated the number of insufficient fund issues (**NSF**) as ordinal because having more than two such instances is possible. The p-values for all 47 predictors are found in Table 4 in the Appendix. Table 1 below summarizes the data types and p-values of the 29 significant variables, assuming the given $\alpha = 0.002$.

Table 1: Significant variables ranked by descending significance (ascending p-value)

Variable Description	p-Value	Type
Indicator for certificate of deposit account (**CD**)	$2.9699 * 10^{-78}$	Binary
Indicator for checking account (**DDA**)	$5.5047 * 10^{-70}$	Binary
Indicator for money market account (**MM**)	$7.4505 * 10^{-57}$	Binary
Indicator for savings account (**SAV**)	$1.4427 * 10^{-39}$	Binary
Indicator for retirement account (**IRA**)	$4.8917 * 10^{-37}$	Binary
Indicator for credit card (**CC**)	$2.4388 * 10^{-32}$	Binary
Indicator for ATM interaction (**ATM**)	$1.5945 * 10^{-29}$	Binary
Indicator for investment account (**INV**)	$1.0511 * 10^{-21}$	Binary
Indicator for direct deposit (**DIRDEP**)	$3.6668 * 10^{-11}$	Binary
Indicator for safety deposit box (**SDB**)	$4.3052 * 10^{-10}$	Binary
Indicator for local address (**INAREA**)	$7.4503 * 10^{-7}$	Binary

Variable Description	p-Value	Type
Savings account balance (**SAVBAL**)	$1.9126 * 10^{-79}$	Continuous
Certificate of deposit account balance (**CDBAL**)	$1.2981 * 10^{-62}$	Continuous
Money market account balance (**MMBAL**)	$2.7019 * 10^{-50}$	Continuous
Checking deposits (**DEP**)	$5.1044 * 10^{-41}$	Continuous
Checking account balance (**DDABAL**)	$6.9936 * 10^{-32}$	Continuous
Number of telephone banking interactions (**PHONE**)	$1.8986 * 10^{-25}$	Continuous
Retirement account balance (**IRABAL**)	$2.3642 * 10^{-19}$	Continuous
Value of home (**HMVAL**)	$5.9878 * 10^{-13}$	Continuous
Number of checks written (**CHECKS**)	$6.4089 * 10^{-12}$	Continuous
Total ATM withdrawal amount (**ATMAMT**)	$1.3374 * 10^{-9}$	Continuous
Number of point of sale interactions (**POS**)	$1.6224 * 10^{-7}$	Continuous
Amount of insufficient funds (**NSFAMT**)	$1.4608 * 10^{-5}$	Continuous
Total amount deposited (**DEPAMT**)	$6.8045 * 10^{-5}$	Continuous
Branch of bank (**BRANCH**)	$5.4880 * 10^{-14}$	Nominal
Number of money market credits (**MMCRED**)	$8.4698 * 10^{-16}$	Ordinal
Number of insufficient fund issues (**NSF**)	$4.9349 * 10^{-11}$	Ordinal
Number of credit card purchases (**CCPURC**)	$1.0666 * 10^{-10}$	Ordinal
Number of cash back requests (**CASHBK**)	$7.0585 * 10^{-4}$	Ordinal

Logistic Regression and Assumptions

We assessed two assumptions before logistic regression modeling. Based on the Bank's data collection method, we assumed independence of the observations. Afterward, generalized additive modeling was used to evaluate the linearity assumption of all continuous variables with the target variable. Table 2 summarizes the linearity assumptions for only the significant continuous variables. Of the 13 significant continuous variables, only the customer's home value had a linear relationship with the target variable. Table 5 in the Appendix displays the entirety of the continuous variables tested for linearity.

Table 2: Linearity of significant continuous variables

Linear	Non-Linear		
HMVAL	**SAVBAL**	**DDABAL**	**ATMAMT**
	CDBAL	**PHONE**	**POS**
	MMBAL	**IRABAL**	**NSFAMT**
	DEP	**CHECKS**	**DEPAMT**

Odds Ratios

We used odds ratios to quantify the strength of the association between our significant binary predictor variables and the target. These are shown in Table 3, arranged by descending odds ratio magnitude.

Table 3: Odds ratios for binary predictors concerning the purchase of a variable rate annuity product

Variable Description	Odds Ratio
Indicator for investment account (**INV**)	3.4720
Indicator for certificate of deposit account (**CD**)	3.4272
Indicator for retirement account (**IRA**)	3.1848
Indicator for money market account (**MM**)	2.8503
Indicator for checking account (**DDA**)	0.3751
Indicator for savings account (**SAV**)	1.8312
Indicator for credit card (**CC**)	1.7813
Indicator for local address (**INAREA**)	0.5746
Indicator for ATM interaction (**ATM**)	0.5930
Indicator for safety deposit box (**SDB**)	1.5497
Indicator for direct deposit (**DIRDEP**)	0.7119

As shown in Table 3, customers with an investment account were 3.472 times more likely to purchase a variable annuity product compared to those who do not have an investment account. This is the strongest association among all significant binary variables. Considering that a variable annuity product is an investment and customers who already invest have a greater propensity to continue investing, this finding seems logical. Similarly, customers with a certificate of deposit account and those with a retirement account were also over three times as likely to purchase a variable annuity product.

Additional Data Considerations

Missing Values

Our next step was identifying missing values. Of the 47 predictor variables, 15 have missing values, six of which are statistically significant: **HMVAL**, **PHONE**, **INV**, **CC**, **CCBAL**, and **CCPURC**. Hence, over 20% of all significant variables (6 out of 29) contained at least one missing value, meaning many of our predictors will need imputation in the future. The other nine insignificant variables were customer age (AGE), customer income (INCOME), length of residence (LORES), an indicator of home ownership (HMOWN), number of point of sales interactions (POS), total amount for point of sale interactions (POSAMT), investment account balance (INVBAL), age of oldest account (ACCTAGE), and credit score (CRSCORE). The percentages of missing values for all 15 variables were quantified and visualized in descending order in the bar graph found in Figure 1.

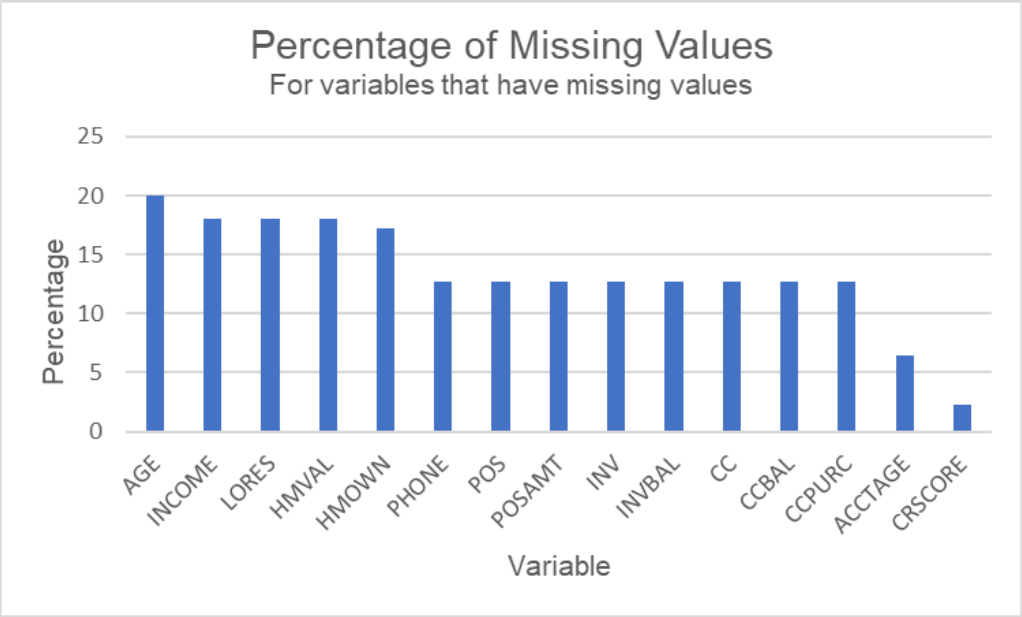

Figure 1: Percentage of values that are missing for variables that have any missing values

Redundant Variables

There were six pairs of significant variables where one is a quantity, and the other is an indicator for the existence of that quantity. Overall, these convey redundant information because we know that a zero value for the indicator corresponds to a zero value for the quantity. These pairs were CD(BAL), DDA(BAL), MM(BAL), SAV(BAL), IRA(BAL), and ATM(AMT).

Results and Recommendations

After preliminary variable exploration, we found that 29 of the 47 provided variables were significant, with most being binary or continuous. Initial results show that only one of the 13 significant continuous variables was linear to the log odds of the target variable. SENCE Consulting recommends ultimately transforming nonlinear significant variables to determine if a linear relationship with the target variable is possible. There are five odds ratios among the 11 significant binary variables with a magnitude greater than two, meaning these variables dictate strong associations with the target variable. In future reports, we propose expanding the variable selection process to include other methods, such as stepwise selection. Furthermore, we will implement imputation to address significant predictors that have missing values. To remove redundancy between variables, we recommend only selecting one predictor from each redundant pair to limit multicollinearity in our final model.

Conclusion

Statistical and practical significance were established through both statistical tests and odds ratio calculations for the binary predictors. While the linearity assumption failed for nearly all significant continuous variables, it highlights the need to transform them eventually. Finally, taking missing values and redundancies into consideration allows us to simplify the future model. We will implement many of our new insights through variable selection techniques to confirm which variables will appear in our final logistic regression model.

Appendix

Table 4: All variables ranked by descending significance (ascending p-value) [significant variables bolded]

Variable Description	p-Value	Type
Indicator for certificate of deposit account (**CD**)	$2.9699 * 10^{-78}$	Binary
Indicator for checking account (**DDA**)	$5.5047 * 10^{-70}$	Binary
Indicator for money market account (**MM**)	$7.4505 * 10^{-57}$	Binary
Indicator for savings account (**SAV**)	$1.4427 * 10^{-39}$	Binary
Indicator for retirement account (**IRA**)	$4.8917 * 10^{-37}$	Binary
Indicator for credit card (**CC**)	$2.4388 * 10^{-32}$	Binary
Indicator for ATM interaction (**ATM**)	$1.5945 * 10^{-29}$	Binary
Indicator for investment account (**INV**)	$1.0511 * 10^{-21}$	Binary
Indicator for direct deposit (**DIRDEP**)	$3.6668 * 10^{-11}$	Binary
Indicator for safety deposit box (**SDB**)	$4.3052 * 10^{-10}$	Binary
Indicator for local address (**INAREA**)	$7.4503 * 10^{-7}$	Binary
Recent address change (MOVED)	0.23707	Binary
Indicator for line of credit (LOC)	0.4995	Binary
Indicator for mortgage (MTG)	0.52811	Binary
Indicator for home ownership (HMOWN)	0.91961	Binary
Savings account balance (**SAVBAL**)	$1.9126 * 10^{-79}$	Continuous
Certificate of deposit account balance (**CDBAL**)	$1.2981 * 10^{-62}$	Continuous
Money market account balance (**MMBAL**)	$2.7019 * 10^{-50}$	Continuous
Checking deposits (**DEP**)	$5.1044 * 10^{-41}$	Continuous
Checking account balance (**DDABAL**)	$6.9936 * 10^{-32}$	Continuous
Number of telephone banking interactions (**PHONE**)	$1.8986 * 10^{-25}$	Continuous
Retirement account balance (**IRABAL**)	$2.3642 * 10^{-19}$	Continuous
Value of home (**HMVAL**)	$5.9878 * 10^{-13}$	Continuous
Number of checks written (**CHECKS**)	$6.4089 * 10^{-12}$	Continuous
Total ATM withdrawal amount (**ATMAMT**)	$1.3374 * 10^{-9}$	Continuous
Number of point of sale interactions (**POS**)	$1.6224 * 10^{-7}$	Continuous
Amount of insufficient funds (**NSFAMT**)	$1.4608 * 10^{-5}$	Continuous
Total amount deposited (**DEPAMT**)	$6.8045 * 10^{-5}$	Continuous
Credit card balance (CCBAL)	0.00221	Continuous
Age of oldest account (ACCTAGE)	0.00822	Continuous
Number of teller visit interactions (TELLER)	0.00972	Continuous

Variable Description	p-Value	Type
Investment account balance (INVBAL)	0.01625	Continuous
Installment loan balance (ILSBAL)	0.02882	Continuous
Mortgage balance (MTGBAL)	0.05865	Continuous
Total amount for point of sale interactions (POSAMT)	0.11463	Continuous
Age (AGE)	0.21902	Continuous
Income (INCOME)	0.25759	Continuous
Credit score (CRSCORE)	0.39323	Continuous
Length of residence in years (LORES)	0.85125	Continuous
Line of credit balance (LOCBAL)	0.91151	Continuous
Branch of bank (**BRANCH**)	$5.4880 * 10^{-14}$	Nominal
Number of money market credits (**MMCRED**)	$8.4698 * 10^{-16}$	Ordinal
Number of insufficient fund issues (**NSF**)	$4.9349 * 10^{-11}$	Ordinal
Number of credit card purchases (**CCPURC**)	$1.0666 * 10^{-10}$	Ordinal
Number of cash back requests (**CASHBK**)	$7.0585 * 10^{-4}$	Ordinal
Area classification (RES)	0.23426	Ordinal

Table 5: Linearity of all continuous variables (significant variables bolded)

Linear		Non-Linear		
HMVAL	AGE	**SAVBAL**	**PHONE**	**DEPAMT**
MTGBAL	CRSCORE	**CD**	**IRABAL**	ILSBAL
CCBAL	LOCBAL	**CDBAL**	**CHECKS**	INVBAL
INCOME	ACCTAGE	**MMBAL**	**ATMAMT**	POSAMT
LORES		**DEP**	**POS**	TELLER
		DDABAL	**NSFAMT**	DEPAMT

Homework Report Checklist

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

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

Body Sections

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

Conclusion

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

Structure

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

Visuals

Introduction, Discussion, and Captions

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

Visual Design

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

Document Design

Title Page Design

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

Table of Contents Design

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

Document Design for Entire Report

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

Writing Style and Mechanics

Spelling and Capitalization

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

Grammar and Punctuation

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

Writing Style

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