PREDICTING VARIABLE ANNUITY PLAN PURCHASES

VARIABLE SELECTION AND MODEL BUILDING

BLUE TEAM 18 - SENCE CONSULTING

Sanket Sahasrabudhe
Ethan Scheper
Noah Johnson
Charis Williams
Elizabeth Surratt

SEPTEMBER 14TH, 2022

Table of Contents

PREDICTING VARIABLE ANNUITY PLAN PURCHASES
VARIABLE SELECTION AND MODEL BUILDING

Overview

Commercial Banking Corporation (the Bank) is seeking to target a customer base likely to purchase a variable rate annuity product and has hired SENCE Consulting to identify these customers and assist with predictive modeling. The Bank has strategically binned all the continuous variables. This transformation satisfies all assumptions of logistic regression, and model building can proceed. Four variables contained missing values, so we imputed them with a “missing” category. Two variables displayed quasi-complete separation with the response, and both instances were remedied by condensing the problematic categories. SENCE Consulting proposes a logistic regression model with only 14 variables and one additional two-way interaction term. This effectively cuts the 29 variables reported previously in half. Based on results from our final model, we recommend that the Bank markets the variable rate annuity product to customers with expendable funds and those that have displayed a previous tendency to save or invest.

Methodology and Analysis

Data Considerations

SENCE Consulting was provided with a dataset with the same 47 variables from the previous RFP. However, each variable previously classified as continuous has now been strategically binned, making all variables categorical and model building more manageable. Missing values still existed in four predictor variables: the indicator for a customer having a credit card (CC), the number of credit card purchases (CCPURC), the indicator for having an investment account (INV), and the indicator for home ownership (HMOWN). We corrected for the missing values by creating a “missing” category.

Next, we explored the potential complete and quasi-complete separation of predictor variables with the response variable. Although there was no complete separation between any variables, there were two ordinal variables that displayed quasi-complete separation: the number of cash back requests (CASHBK) and the number of money market credits (MMCRED). In both instances, the largest category for these variables only corresponded to a non-event, as shown in Tables 7 and 8 in the Appendix. To correct for this, we collapsed the largest two categories to create the distributions shown in Tables 1 and 2.

Table 1: Number of cash back requests vs. whether a customer purchased a variable annuity product

	Did not purchase product	Did purchase product
0 cash back requests	5473	2891
1+ cash back requests	104	27

Table 2: Number of money market credits vs. whether a customer purchased a variable annuity product

	Did not purchase product	Did purchase product
0 money market credits	5409	2713
1 money market credit	130	153
2 money market credits	33	47
3+ money market credits	5	5

Originally, there were two customers with two cash back requests and one customer with five money market credits who did not purchase the variable rate annuity product. With the new collapsed categories ("1+" category for CASHBK and "3+" category for MMCRED), these quasi-complete separation issues are resolved, and we can move forward with variable selection and model building.

Variable Selection and Model Building

We built a binary logistic regression model to predict the probability of a Bank customer purchasing a variable rate annuity product. The first step was to select the main effects for future model construction steps.

Main Effects

To accomplish this, we employed backward selection with all 46 predictor variables, using p-value selection criteria at a significance level of α = 0.002. This revealed that 14 variables were significant. Table 3 shows the p-values obtained from the Likelihood Ratio Test performed on these variables.

Table 3: Main effects ordered by descending significance (ascending p-value)

Main Effect Description	p-Value
Binned savings account balance (SAVBAL_Bin)	$2.4617 * 10^{-126}$
Binned checking account balance (DDABAL_Bin)	$5.9920 * 10^{-60}$
Binned certificate of deposit account balance (CDBAL_Bin)	$2.6414 * 10^{-39}$
Indicator for money market account (MM)	$7.7428 * 10^{-23}$
Binned number of checks written (CHECKS_Bin)	$2.0846 * 10^{-19}$
Binned total ATM withdrawal amount (ATMAMT_Bin)	$2.8653 * 10^{-9}$
Binned number of teller visit interactions (TELLER_Bin)	$1.2064 * 10^{-8}$
Indicator for credit card (CC)	$1.2001 * 10^{-7}$
Indicator for checking account (DDA)	$8.0990 * 10^{-6}$
Indicator for retirement account (IRA)	$2.6362 * 10^{-5}$
Indicator for investment account (INV)	$6.6289 * 10^{-5}$
Indicator for installment loan (ILS)	$7.6437 * 10^{-5}$
Indicator for mortgage (MTG)	$8.4693 * 10^{-4}$
Number of insufficient fund issues (NSF)	$1.5755 * 10^{-3}$

Interestingly, the three most significant overall variables were binned variables, which identifies the binned account balances as important. After testing for potential interaction effects, we further explored the relationship between bin levels and the likelihood of purchasing a variable annuity product.

Interactions

Following the identification of significant main effects, we explored every two-way combination of these variables for significant interaction terms. Using forward selection, once again with p-value selection criteria at a significance level of α = 0.002, we found one significant two-way interaction between the indicator for checking account (DDA) and the indicator for retirement account (IRA). This finding, along with its accompanying p-value, can be seen in Table 9 in the Appendix.

Results

We included the 14 significant main effects and one significant interaction effect in the final binary logistic regression model to predict the probability of a customer purchasing a variable rate annuity product. These variables and their accompanying p-values are listed in Table 5 in the Appendix.

The most significant variables were the binned account balances for savings, checking, and certificate of deposit accounts. This indicates that the amount of money a customer has is one of the most important predictors in determining whether or not they will purchase a variable rate annuity product. To further investigate, we found the odds ratios of each variable using the coefficients from the final model. A subset of these odds ratios are shown in Table 4, with the full list in Table 6 in the Appendix.

Table 4: Variables with odds ratio magnitudes greater than two, ordered by descending magnitude

Variable	Odds Ratio	Magnitude
Bin 8 of checking account balance (DDABAL_Bin8)	8.7229	8.7229
Bin 7 of savings account balance (SAVBAL_Bin7)	5.7445	5.7445
Indicator for having a checking account (DDA1)	0.1856	5.3879
Bin 7 of checking account balance (DDABAL_Bin7)	4.7330	4.7330
Bin 3 of certificate of deposit balance (CDBAL_Bin3)	4.0974	4.0974
Bin 6 of savings account balance (SAVBAL_Bin6)	3.7553	3.7553
Bin 6 of checking account balance (DDABAL_Bin6)	3.6723	3.6723
Bin 5 of checking account balance (DDABAL_Bin5)	2.8686	2.8686
Bin 5 of savings account balance (SAVBAL_Bin5)	2.4907	2.4907
Interaction of DDA1 and IRA1	2.2476	2.2476
Bin 4 of checking account balance (DDABAL_Bin4)	2.2184	2.2184
Indicator for having a money market account (MM1)	2.2035	2.2035

As expected from their low p-values in Table 5, the bins corresponding to the highest account balances had the largest odds ratios in Table 4. For example, customers in the highest checking account balance bin were 8.72 times more likely to purchase a variable rate annuity product than those in the lowest balance bin. Similarly, customers in the highest balance bin for savings account balance and certificate of deposit account balance were also over four times as likely to purchase a variable annuity than their counterparts in the lowest balance bin. This leads us to a simple conclusion: people with more money are more likely to invest. When customers have expendable income, they can set it aside and allow it to grow using a product like a variable rate annuity.

As a whole, customers with a checking account were 81.44% less likely to purchase a variable annuity than those without a checking account. On the other hand, customers with an investment account were 1.84 times more likely to purchase a variable annuity than those without an account. This indicates that customers with long-term holdings like investment or savings accounts are more likely to purchase a variable rate annuity product than customers with short-term holdings like checking accounts.

Two variables in the model, the missing level for the credit card indicator (CCM) and the missing level for the investment account indicator (INVM), were perfectly aliased. When one was true, the other was also true, so they made identical predictions. This led us to believe that all branches that do not offer investment accounts also do not offer credit cards. Due to this distinction, variable rate annuity products should be marketed differently to their customers than customers at other branches.

Recommendations

After finalizing a binary logistic regression model with 15 effects, SENCE Consulting recommends that the Bank do the following when targeting customers for the purchase of a variable rate annuity product:

- Market most aggressively to customers with large account balances.
- Target customers that already have long-term holdings accounts such as savings accounts, certificate of deposit accounts, retirement accounts, or investment accounts.
- Focus almost exclusively on customers in the highest account balance bins for customers that only possess short-term accounts like checking accounts.
- Recommend the variable rate annuity product as a preferred way to invest for customers who already invest with the Bank in some capacity.
- Tailor the marketing approach for branches that do not offer certain types of accounts, such as investment accounts or credit cards. For example, different indicators like savings accounts should be used in place of investment accounts to identify customers of focus.

In general, when marketing the variable rate annuity product, the Bank should focus on customers with expendable funds and/or customers with a previous tendency to save or invest. The model identified these individuals as the most likely to follow through in purchasing the product.

Conclusion

SENCE Consulting performed data cleaning, variable selection, and model building by utilizing a fully categorical dataset. Four variables had at least one missing value replaced with a missing category. We condensed the levels of two variables displaying quasi-complete separation with the response. Next, we implemented backward p-value selection on all 46 predictors. The result was a model with 14 main effects. Then, we used forward p-value selection with all possible two-way interactions, and one such interaction was chosen. The magnitudes of the odds ratios are the largest among bins corresponding to high account balances. The negative effect of having a checking account on purchasing the product is counteracted only when the account balance is high. SENCE Consulting thus recommends targeting customers with large long-term account balances, marketing the variable rate annuity product as an investment with the bank, and tailoring strategies to branches depending on the account types they offer. Given this final model, our next step is to assess and gain insight into the model’s performance.

Appendix

Table 5: Final regression model variables ordered by descending significance (ascending p-value)

Variable Description	p-Value
Binned savings account balance (SAVBAL_Bin)	$1.5882 * 10^{-126}$
Binned checking account balance (DDABAL_Bin)	$4.8841 * 10^{-59}$
Binned certificate of deposit account balance (CDBAL_Bin)	$2.0279 * 10^{-39}$
Indicator for money market account (MM)	$3.0722 * 10^{-23}$
Binned number of checks written (CHECKS_Bin)	$6.6320 * 10^{-20}$
Binned total ATM withdrawal amount (ATMAMT_Bin)	$2.6162 * 10^{-9}$
Binned number of teller visit interactions (TELLER_Bin)	$1.1731 * 10^{-8}$
Indicator for credit card (CC)	$1.2884 * 10^{-7}$
Indicator for checking account (DDA)	$3.5133 * 10^{-6}$
Indicator for installment loan (ILS)	$8.1925 * 10^{-5}$
Indicator for investment account (INV)	$8.2438 * 10^{-5}$
Indicator for checking account (DDA), indicator for retirement account (IRA)	$2.5171 * 10^{-4}$
Indicator for mortgage (MTG)	$8.0252 * 10^{-4}$
Number of insufficient fund issues (NSF)	$1.3239 * 10^{-3}$
Indicator for retirement account (IRA)	$8.6667 * 10^{-1}$

Table 6: All odds ratios of variables in the final model ordered by descending magnitude

Variable	Odds Ratio	Magnitude
Bin 8 of checking account balance (DDABAL_Bin8)	8.7229	8.7229
Bin 7 of savings account balance (SAVBAL_Bin7)	5.7445	5.7445
Indicator for having a checking account (DDA1)	0.1856	5.3879
Bin 7 of checking account balance (DDABAL_Bin7)	4.7330	4.7330
Bin 3 of certificate of deposit balance (CDBAL_Bin3)	4.0974	4.0974
Bin 6 of savings account balance (SAVBAL_Bin6)	3.7553	3.7553
Bin 6 of checking account balance (DDABAL_Bin6)	3.6723	3.6723
Bin 5 of checking account balance (DDABAL_Bin5)	2.8686	2.8686
Bin 5 of savings account balance (SAVBAL_Bin5)	2.4907	2.4907
Interaction of DDA1 and IRA1	2.2476	2.2476
Bin 4 of checking account balance (DDABAL_Bin4)	2.2184	2.2184
Indicator for having a money market account (MM1)	2.2035	2.2035
Bin 2 of certificate of deposit balance (CDBAL_Bin2)	1.9624	1.9624
Bin 4 of number of checks written (CHECKS_Bin4)	0.5373	1.8612
Indicator for having an investment account (INV1)	1.8380	1.8380
Bin 3 of ATM withdrawal amount (ATMAMT_Bin3)	1.7206	1.7206
Bin 3 of number of teller interactions (TELLER_Bin3)	1.7132	1.7132
Bin 2 of savings account balance (SAVBAL_Bin2)	0.5983	1.6714
Indicator for missing investment account data (INVM)	0.5986	1.6706
Indicator for having an installment loan (ILS1)	0.6104	1.6383
Bin 3 of checking account balance (DDABAL_Bin3)	1.6367	1.6367
Indicator for having a mortgage (MTG1)	0.6607	1.5135
Indicator for having an insufficient fund issue (NSF1)	1.4100	1.4100
Indicator for having a credit card (CC1)	1.3711	1.3711
Bin 3 of savings account balance (SAVBAL_Bin3)	0.7572	1.3207
Bin 4 of savings account balance (SAVBAL_Bin4)	1.3067	1.3067
Bin 2 of number of teller interactions (TELLER_Bin2)	1.2551	1.2551
Bin 2 of checking account balance (DDABAL_Bin2)	1.2460	1.2460
Bin 2 of ATM withdrawal amount (ATMAMT_Bin2)	0.8982	1.1133
Bin 2 of number of checks written (CHECKS_Bin2)	1.0503	1.0503
Bin 3 of number of checks written (CHECKS_Bin3)	0.9579	1.044
Indicator for having a retirement account (IRA1)	0.9714	1.0294

Table 7: Number of cash back requests vs. whether a customer purchased a variable annuity product

	Did not purchase product	Did purchase product
0 cash back requests	5473	2891
1 cash back requests	102	27
2 cash back requests	2	0

Table 8: Number of money market credits vs. whether a customer purchased a variable annuity product

	Did not purchase product	Did purchase product
0 money market credits	5409	2713
1 money market credit	130	153
2 money market credits	33	47
3 money market credits	4	5
5 money market credits	1	0

Table 9: Interaction effects ordered by descending significance (ascending p-value)

Interaction Description	p-Value
Indicator for checking account (DDA), indicator for retirement account (IRA)	$2.5171 * 10^{-4}$

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>
SS	Does it briefly address your team's work, quantifiable results, and recommendations? <Action>
SS	Does it offer audience-centered reasons for recommendations? <Context>

Body Sections

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

Conclusion

CW	Does the report have a conclusion?
NJ	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

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

Visual Design

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

Document Design

Title Page Design

NJ	Does it include a descriptive title?
NJ	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?
EAS	Do the page numbers and sections in the Table of Contents match the report?

Document Design for Entire Report

CW	Is a standard typeface (*Calibri*, *Arial*, *etc.*) used?
SS	Is the size of the body text between 10-12 pt.?
SS	Are headings and subheadings used to organize information?
NJ	Are distinctive text styles (*bold*, *italic*, *etc.*) used to distinguish between heading levels?
NJ	Are text styles for headings used consistently (*ex*: *all level-one headings are bold*)?
EAS	Are all paragraphs an appropriate length (*fewer than 12 lines*)?
EAS	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

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

Writing Style

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