BANKING INSURANCE PRODUCT - PHASE 2

SUPREET DESHPANDE, NICK STERLING, HEQING SUN
CHELSEA THOMAS, CHESANEY WYSE

SEPT 18, 2019

TABLE OF CONTENTS

BANKING INSURANCE PRODUCT - PHASE 2

EXECUTIVE SUMMARY

The department of Customer Services and New Products at Commercial Banking Corporation is seeking to predict which customers would buy a variable rate annuity product. In this phase of the process, we began variable selection and model building. To proceed, we first needed to tackle the concerns of missing variables and linear separation. To correct the four variables with missing observations, we created a missing category, accounting for each missing value. In the case of the two quasi-separation concerns, we condensed the variable categories to ensure the existence of the maximum likelihood estimates.

Then, we began developing our model. Through backward selection, 30% of the original 46 variables were determined to be significant. After seeing a model with just the main effects, we tested the relationships between them. We identified some significant interactions, which led us to a final model with main effects and interaction terms. As seen in Table 1 below, this method produced a hybrid model with 17 significant variables, of which 14 were main effects variables, and three were interaction terms.

TABLE 1: MODEL SUMMARY

Main Effect Variables	14
Interaction terms	3
Total Variables	17

METHODOLOGY

We used a data set consisting of 46 customer feature variables, and one target variable representing whether or not a customer purchased a variable rate annuity product. Four predictor variables contained missing values: INV (indicator for investment account), CC (indicator for credit card), CCPURC (number of credit card purchases), and HMOWN (indicator for homeownership). Table 2 displays these four variables ordered by the number of missing values.

TABLE 2: VARIABLES WITH MISSING VALUES

Variable	# Missing Values
INV	1075
CC	1075
CCPURC	1075
HMOWN	1463

Two variables exhibited quasi-complete separation: CASHBK (number of cashback requests) and MMCRED (number of money market accounts). For CASHBK, levels '1' and '2' were combined into a '1+' level. Similarly, all levels '2' and above were collapsed into a '2+' level for MMCRED. Tables 3A and 3B show the frequency tables for CASHBK and MMCRED before and after these modifications.

TABLE 3: CASHBK AND MMCRED BEFORE AND AFTER LEVEL COLLAPSING

3A) *Before*

		CASHBK		
		0	1	2
INS	0	5473	102	2
	1	2891	27	0

After

		CASHBK	
		0	1+
INS	0	5473	104
	1	2891	27

3B) *Before*

		MMCRED				
		0	1	2	3	5
INS	0	5409	130	33	4	1
	1	2713	153	47	5	0

After

		MMCRED		
		0	1	2+
INS	0	5409	130	38
	1	2713	153	52

Backward Selection method was used to determine the main effects from the 46 predictor variables. Additionally, second-order interaction terms were constructed from significant main effects using forward selection. A p-value of 0.002 was the criterion for the backward and forward selection techniques employed. The final model contains the 17 significant variables identified, which includes 14 main effects and 3 interactions terms. The odds ratios for each variable in the final model was also calculated for further insight.

ANALYSIS

BACKWARD ELIMINATION

All the predictor variables entered a backward selection procedure which selected the main effects for the model basis. Table 4 shows the 14 significant main effects variables ordered by their chi-square statistic.

TABLE 4: BACKWARD ELIMINATION

Variable	Chi-Sq Statistic	p value
SAVBAL_Bin	543.62	<.0001
DDABAL_Bin	283.74	<.0001
CDBAL_Bin	165.89	<.0001
MMBAL_Bin	96.70	<.0001
CHECKS_Bin	88.33	<.0001
BRANCH	87.81	<.0001
ATMAMT_Bin	39.87	<.0001
TELLER_Bin	35.65	<.0001
CC	22.14	<.0001
IRA	16.52	<.0001
DDA	15.06	0.0001
INV	14.64	0.0001
ILS	14.17	0.0002
NSF	10.56	0.0012

It is interesting to note the presence of the pair DDA & DDABAL_bin. The DDA variable is a binary indicator for the presence of a direct deposit account. DDABAL_Bin is a binned variable which corresponds to the amount in the direct deposit account. It would be interesting to consider the relationship of these variables to each other and the target variable.

FORWARD SELECTION (WITH INTERACTION EFFECTS)

We used forward selection to build the final model, including any significant second-order interactions. Table 5 displays all significant main effects variables in the final model ordered by their chi-square statistic.

TABLE 5: MAIN EFFECTS

MAIN EFFECT		
Variable	Chi-Sq Statistic	p value
SAVBAL_Bin	930.6537	<.0001
DDABAL_Bin	578.0557	<.0001
CDBAL_Bin	220.5619	<.0001
MMBAL_Bin	131.6624	<.0001
INV	98.9123	<.0001
CHECKS_Bin	54.5841	<.0001
ATMAMT_Bin	33.6197	<.0001
TELLER_Bin	35.5527	<.0001
BRANCH	57.3068	<.0001
DDA	21.026	<.0001
IRA	18.308	<.0001
NSF	15.0359	0.0001
CC	12.7293	0.0004
ILS	11.7417	0.0006

Additionally, Table 6 demonstrates the final interaction variables ranked by chi-square statistic. Of all possible second-order interaction effects, the forward selection method found three significant terms.

TABLE 6: INTERACTION EFFECTS

INTERACTION EFFECT		
Variable	Chi-Sq Statistic	p value
DDABAL_Bin*SAVBAL_Bin	159.9314	<.0001
DDABAL_Bin*MMBAL_Bin	32.5023	<.0001
DDA*IRA	10.4047	0.0013

ODDS RATIOS

Table 7 shows the three highest magnitude odds ratios from the final model. For individuals with a certificate of deposit account, people with a higher deposit account balance are approximately four times more likely to purchase the insurance product when compared to people with a low certificate of the deposit account balance. Additionally, individuals affiliated with branch 1 are more likely on average to purchase the insurance product, when compared to individuals affiliated with branch 14.

TABLE 7: ODDS RATIOS

Variable	Odds Ratio
CDBAL_Bin 3 vs 1	4.03
BRANCH B1 vs B14	3.86

RESULTS AND RECOMMENDATIONS

RESULTS

Through backward selection for main effects and forward selection for interaction terms, we were able to narrow our model down to a total of 17 variables. This model achieved an 80.8 percent concordance using the training dataset.

NEXT STEPS

We recommend you consider using the variables from the final model to predict whether or not an individual would purchase the variable-rate annuity product. Moving forward, we would ensure the accuracy of this model by evaluating various model assessment metrics.

CONCLUSION

Through further variable analysis, we were able to build a potential model for predicting whether or not a customer would buy Commercial Banking Corporation's variable rate annuity product. We began this process by taking care of the four missing variables in our data and correcting the two variables that exhibited separation concerns.

We addressed both of these issues to ensure model integrity. We tracked missing values by creating a missing category. By condensing the categories of both variables of concern, we were able to deal with the quasi-separation. From there, we moved on to the first step of developing our model using backward selection to pick out the main effects. In doing so, we produced a model with 14 significant variables which represent 30% of the original variables.

We wanted to further assess these significant variables and the relationships between them. To do so, we built a model using the forward selection method, which looked at the previously found significant variables plus the interaction terms between them. This method produced a hybrid model with 17 significant variables, of which 14 were main effects variables, and 3 were interaction terms.

Now that we have a potential model built, in moving on to the next phase, we would like to assess the accuracy and strength of our model.