Banking Insurance Product Analysis

CATHY TRAN
GRANT CLARK
EVAN LI
PRICE BURNETT
SUFYAN SHAHIN

September 18, 2019

Table of Contents

Banking Insurance Product Analysis

Executive Summary

The Customer Services and New Products at the Commercial Banking Corporation is seeking Team Orange 2 to predict which customers will buy a variable rate annuity product. Using backward and forward selection, Team Orange 2 has built a model with 80.80% Area Under the ROC Curve (AUC) that will predict if a customer will buy a variable rate annuity product or not. Within our results we found that customers with a middle *CD Balance* have 1.9 times the odds of purchasing the product over those with a low balance, and customers with a high balance have 4 times the odds of purchasing the product over those with a low balance. Moving forward we would recommend the bank to:

1. Build a model using stepwise selection
2. Consider removing the interaction terms from the model

This would allow the bank to further compare the models produced by each technique and also allow for a more interpretable model.

Results

With 80.80% AUC on the training data, Team Orange 2 has identified a model which predicts whether or not a customer will buy a variable rate annuity product. The model contains 14 variables along with 3 interactions among those variables, as noted in Table 1. Most significant among these are *CD Balance*, *Branch of Bank*, and *Number of Checks Written* with the interaction between *Checking Account Balance* and *Savings Account Balance* being most significant of all.

Table 1. Final Logistic Regression Model's Variables Ranked by Significance

Variable	Test Value	P-Value
DDABAL_Bi*SAVBAL_Bin	164.3622	<.0001
CDBAL_Bin	154.7188	<.0001
BRANCH	114.3985	<.0001
CHECKS_Bin	99.0389	<.0001
SAVBAL_Bin	50.0004	<.0001
TELLER_Bin	36.6166	<.0001
ATMAMT_Bin	36.2792	<.0001
DDABAL_Bin	31.7972	<.0001
IRA	28.4354	<.0001
DDABAL_Bin*MM	27.9171	0.0002
MM	24.5479	<.0001
CC	17.4153	<.0001
NSF	17.398	<.0001
INV	12.5963	0.0004
ILS	11.666	0.0006
DDA*IRA	10.336	0.0013
DDA	6.3342	0.0118

Although some of the relationships are more complex, it can be noted that those with a greater *CD Balance* had greater odds of purchasing an annuity product. Specifically, those with a middle balance have 1.9 times the odds of purchasing an annuity product over those with a low balance; likewise, those with a high balance have over 4 times the odds of purchasing an annuity product over those with a low balance. Results such as these require further analysis, but show a promising understanding of what customer characteristics helps to explain product purchase.

The team would like to note that the variable that acts as an *Indicator for Checking Account* is not significant in the final model, but remains in order to maintain model hierarchy. This variable was significant when the first stage of the model was built with only main effects, but lost significance with the addition of interactions.

Recommendations

Besides using the backward and forward selection method, the Bank could also perform stepwise selection to build a model. This would allow the Bank to compare the models produced by each technique and evaluate which one best predicts the customers probability of buying an insurance product. Furthermore, it might be helpful for the Bank to consider building a more interpretable model by taking out the interaction terms.

Methodology & Analysis

Data Used

The dataset contained information on customers that have been offered an insurance product with a variable indicating if they bought the product or not. The training dataset that was used for this phase of the analysis contained 8,495 observations. All continuous variables from the original dataset were binned to be binary, nominal, or ordinal.

Imputation of Missing Values

In order to complete a logistic regression analysis to determine which factors lead to a customer's purchase of insurance, it was critical to check each variable for missing values. After evaluating all variables, it was determined that four variables contained missing values. The four variables were *Investment Account Indicator, Credit Card Indicator, Number of Credit Card Purchases*, and *Home Ownership Indicator*. All missing values for these variables were imputed into a new category of -1 to create a baseline for comparison.

Separation concerns

Once it was confirmed that all variables had no missing values, these variables were evaluated for separation concerns. Only two of the 47 variables appeared to have quasi-separation, *Number of Cash Back Requests* and *Number of Money Market Credits*. *Number of Cash Back Requests* was re-coded as binary and *Number of Money Market Credits's* column five was condensed to column three. Both variables were re-tested and no separation concerns remained. We also checked the three significant interaction terms from our final model for separation, and no issues were present.

Model Creation

A logistic regression model was created by utilizing backward selection to check for significant main effects of variables. The backward selection model took 31 steps, and created a model with 14 significant main effects (Table 2). These main effects were then tested for two-way interactions within a forward selection model. The forward selection model took 17 steps and resulted in a model with 17 variables. Three of these significant variables were two-way interactions (Table 3). The final model's significant variables are listed in Table 1.

Table 2. Main Effect Variables From Backward Selection Ranked by P-Values

Variable	Type of Variable	Test Value	p-Value
SAVBAL_Bin	Ordinal	543.6226	<.0001
DDABAL_Bin	Ordinal	283.7405	<.0001
CDBAL_Bin	Ordinal	165.8919	<.0001
BRANCH	Nominal	118.3907	<.0001
MM	Binary	96.7047	<.0001
CHECKS_Bin	Ordinal	88.3312	<.0001
ATMAMT_Bin	Ordinal	39.8746	<.0001
TELLER_Bin	Ordinal	35.6464	<.0001
CC	Nominal	22.138	<.0001
IRA	Binary	16.5175	<.0001
DDA	Binary	15.0568	0.0001
INV	Nominal	14.6436	0.0001
ILS	Binary	14.1659	0.0002
NSF	Binary	10.5584	0.0012

Table 3. Interaction Variables from Forward Selection Ranked by P-Value

Interaction Variable	Test Value	p-Value
DDABAL_Bin*SAVBAL_Bin	164.3622	<.0001
DDABAL_Bin*MM	27.9171	0.0002
DDA*IRA	10.336	0.0013

Odds Ratios

After creating the logistic regression model using both backward and forward selection, odds ratios were calculated for all of the significant variables in the final model. The odds ratio for *Branch 14 vs. Branch 1* was determined to be significant with an estimate of 0.173. This implies that customers at branch 14 are roughly 83 percent less likely to purchase insurance than customers at branch 1. All of the odds ratios can be found in the Appendix.

CONCLUSION

Using logistic regression, Team Orange 2 identified significant variables with interactions with an 80.80% AUC. Moving forward, Team Orange 2 will assess the performance and accuracy of the logistic regression model and predict which customers will buy a variable rate annuity product. Specifically, the team will evaluate the goodness of fit of the model by interpreting the probability metrics including the discrimination slope. The ROC curve and the K-S statistics will also be explored on the training data to determine the threshold for classification. Additionally, the team will examine the predictive power of the model with the interpretation of Lift and accuracy statistics.

APPENDIX

Table 4. Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Estimate	95% Confidence Limits	
NSF 0 vs 1	0.631	0.509	0.785
CHECKS_Bin 2 vs 1	0.991	0.803	1.225
CHECKS_Bin 3 vs 1	0.907	0.725	1.133
CHECKS_Bin 4 vs 1	0.489	0.398	0.601
TELLER_Bin 2 vs 1	1.291	1.125	1.482
TELLER_Bin 3 vs 1	1.729	1.443	2.071
CDBAL_Bin 2 vs 1	1.91	1.567	2.33
CDBAL_Bin 3 vs 1	4.035	3.164	5.181
ATMAMT_Bin 2 vs 1	1.009	0.887	1.148
ATMAMT_Bin 3 vs 1	1.852	1.496	2.296
BRANCH B10 vs B1	1.066	0.62	1.819
BRANCH B11 vs B1	1.317	0.69	2.509
BRANCH B12 vs B1	1.424	0.907	2.218
BRANCH B13 vs B1	1.164	0.758	1.78
BRANCH B14 vs B1	0.173	0.105	0.281
BRANCH B15 vs B1	0.232	0.153	0.348
BRANCH B16 vs B1	0.516	0.373	0.711
BRANCH B17 vs B1	1.224	0.846	1.766
BRANCH B18 vs B1	0.483	0.287	0.806
BRANCH B19 vs B1	0.421	0.216	0.806
BRANCH B2 vs B1	0.938	0.751	1.174
BRANCH B3 vs B1	1.09	0.846	1.405
BRANCH B4 vs B1	1.042	0.838	1.297
BRANCH B5 vs B1	0.962	0.747	1.238
BRANCH B6 vs B1	1.098	0.815	1.48
BRANCH B7 vs B1	0.94	0.694	1.271
BRANCH B8 vs B1	1.192	0.877	1.62
BRANCH B9 vs B1	1.18	0.775	1.789
INV 0 vs -1	0.573	0.419	0.777
ILS 0 vs 1	1.553	1.209	2.005
CC 0 vs -1	0.778	0.692	0.876

Table 5. Description of All Variables.

Variable Name	Description
ACCTAGE_Bin	Age of oldest account
DDA	Indicator for checking account

DDABAL_Bin	Checking account balance
DEPAMT_Bin	Total amount deposited
CASHBK	Number of cash back requests
CHECKS_Bin	Number of checks written
DIRDEP	Indicator for direct deposit
NSF	Number of insufficient fund issues
NSFAMT_Bin	Amount of NSF
PHONE_Bin	Number of telephone banking interactions
TELLER_Bin	Number of teller visit interactions
SAV	Indicator for savings account
SAVBAL_Bin	Savings account balance
ATM	Indicator for ATM interaction
ATMAMT_Bin	Total ATM withdrawal amount
POS_Bin	Number of point of sale interactions
POSAMT_Bin	Total amount for point of sale interactions
CD	Indicator for certificate of deposit account
CDBAL_Bin	CD balance
IRA	Indicator for retirement account
IRABAL_Bin	IRA balance
LOC	Indicator for line of credit
LOCBAL_Bin	LOC balance
INV	Indicator for investment account
INVBAL_Bin	INV balance
ILS	Indicator for installment loan
ILSBAL_Bin	ILS balance
MM	Indicator for money market account
MMBAL_Bin	MM balance
MMCRED	Number of money market credits
MTG	Indicator for mortgage
MTGBAL_Bin	MTG balance
CC	Indicator for credit card
CCBAL_Bin	CC balance
CCPURC	Number of credit card purchases
SDB	Indicator for safety deposit box
INCOME_Bin	Income
HMOWN	Indicator for homeownership
LORES_Bin	Length of residence in years
HMVAL_Bin	Value of home
AGE_Bin	Age
CRSCORE_Bin	Credit score
MOVED	Recent address change
INAREA	Indicator for local address

INS	Indicator for purchase of insurance product
BRANCH	Branch of bank
RES	Area classification