Banking Insurance Product Analysis

CATHY TRAN
GRANT CLARK
EVAN LI
PRICE BURNETT
SUFYAN SHAHIN

September 25, 2019

Table of Contents

Banking Insurance Product Analysis

Executive Summary

We recommend the Bank focus on marketing to the top 20% of predicted customers, since they are twice as likely to purchase a variable rate annuity product compared to a random sample of equal size. Team Orange 2 developed a logistic model to predict which customers will buy these products. After addressing separation concerns and imputing missing values, the model had 69.91% accuracy.

Results

With 69.91% accuracy on the validation dataset, Team Orange 2 has identified a model to predict 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 P-Values

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

Specifically, the sensitivity and specificity of this model are displayed in the final confusion matrix (Table 2). The sensitivity statistic indicates that, out of all the customers from the dataset that did purchase the annuity product, our model captured 77.6% of them. The specificity statistic denotes that, out of all the customers that did not purchase the insurance product, our model predicted 65.8% of them to not purchase the product.

Table 2. Final Confusion Matrix

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	909	473
Actual Positive	166	576

Specificity	65.8
Sensitivity	77.6

The predictive power of this model is further illustrated by the Lift chart (Figure 1). After scoring the validation dataset, the lift chart indicates that the top 20% percent of the customers, based on predicted probability, are approximately two times more likely to buy an insurance product compared to targeting a randomly selected sample of 20% of customers.

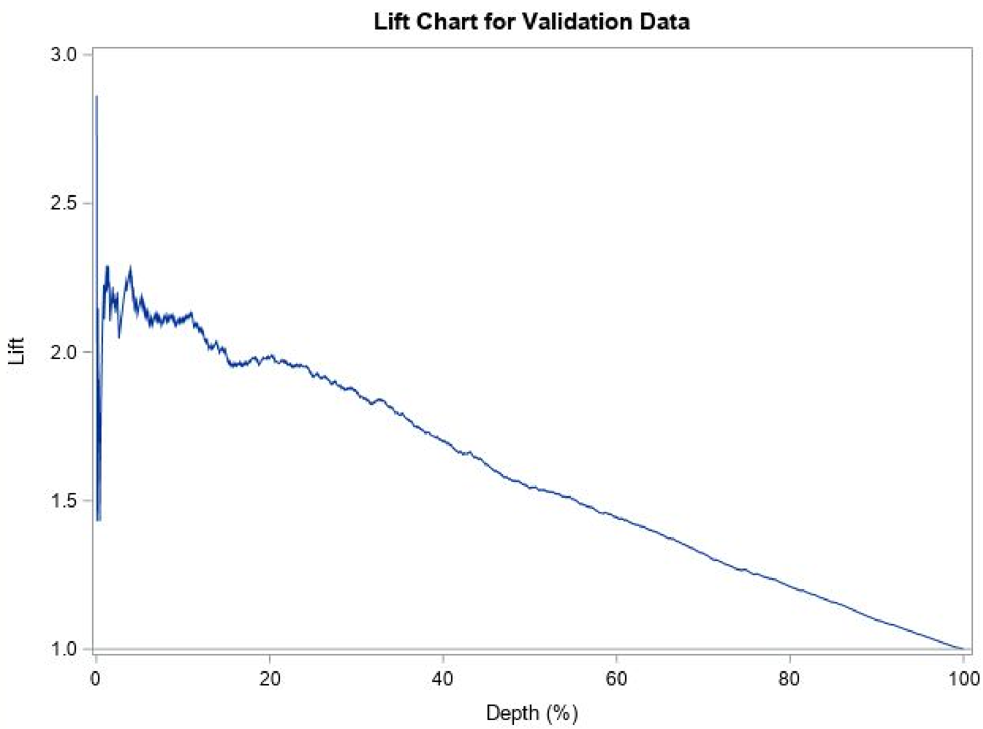

Figure 1. Lift Chart for Validation Data

Recommendations

Moving forward, Team Orange 2 would recommend building models that are subsets of the current model. Doing so may allow for more explainability within the model at the cost of some predictability. In addition, the team would recommend taking a more recent subset of data and scoring the model on it. This would show if the model still predicts well or if the model needs to be adjusted.

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, and the validation dataset contained 2124 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*' column five was condensed to column three. Both variables were re-tested and no separation concerns remained. The three significant interaction terms from our final model were also checked for separation, and no issues were present.

Threshold Selection

Utilizing the Kolmogrov-Smirnov (KS) test statistic, the team chose a cutoff of 0.3 for determining if the predicted probability of an observation would identify it as an event (purchasing a variable rate annuity product) or a non-event (not purchasing). The confusion matrix and accuracy of the validation data are based on the assumption that this is the best cutoff.

Goodness of Fit

The area under the ROC curve (AUC) of 80.80% is modeled in Figure 2, demonstrating the relationship between the True Positive Rate along the y-axis and the False Positive Rate along the x-axis. This sets a baseline value for any models derived from the one developed by Team Orange 2 should variables need to be removed in forming a new model. The AUC also represents the percent of concordant pairs within the model as no pairs were tied. Thus 80.80% of events to non-events, when compared to one another, were appropriately identified by the model.

With a discrimination slope of 0.2548, the model appears to more effectively identify non-events than events, as seen in Figure 3. The histogram on the top represents the distribution of those predicted to purchase a variable rate annuity product. This shows a closer to normal distribution around 0.6 as opposed to the prefered left-skewed distribution which would indicate more effective prediction of those who will make a purchase. The histogram on the bottom represents the distribution of those predicted not to make a purchase. The right-skewed distribution centered around 0.2 suggests the model may effectively predict these non-events.

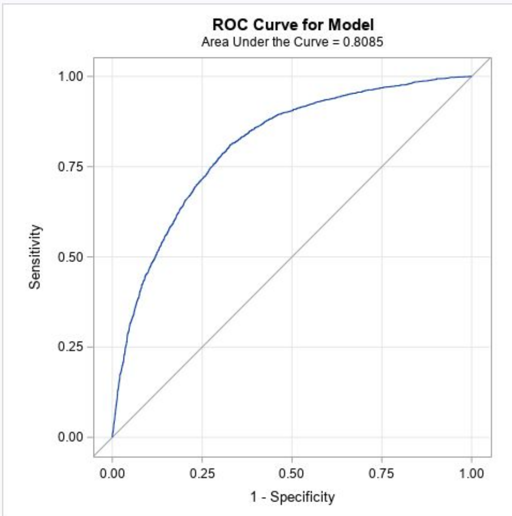

Figure 2. ROC Curve representing the AUC.

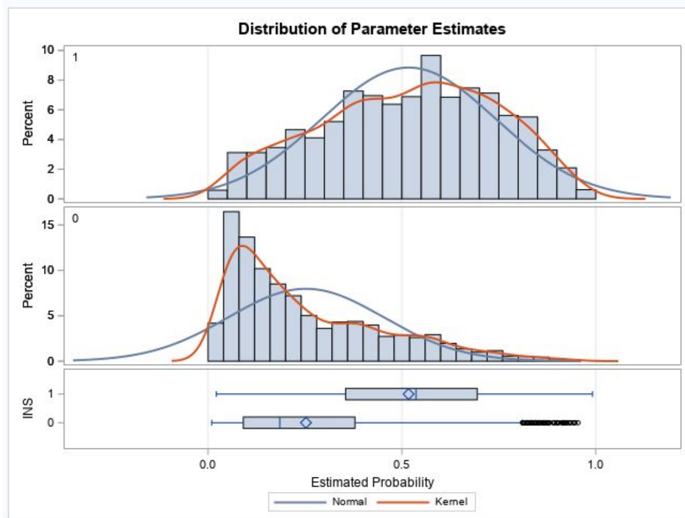

Figure 3. Discrimination Slope, showing the model likely predicts non-purchases better than purchases.

Conclusion

Our model suggested that if the Bank targets their top 20% of customers, they will receive two times more responses as compared to targeting a random sample of the same size. Moving forward, we recommend the Bank continue gathering more data and applying the model on the new data. This would enable the Bank to predict their future customers' behavior and maximize the utility of the model.

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