Banking Insurance Product: Phase 3

Blue 10

Alison Bean de Hernández
Austin Minihan
Ishi Gupta
Patrick Costa

September 20, 2024

Table of Contents

Banking Insurance Product: Phase 3

Overview

The Department of Customer Services and New Products at the Commercial Banking Corporation, hereinafter called 'the Bank,' partnered with Blue Team 10 to create a model that predicts which customers will buy their variable-rate annuity product, hereinafter called the 'product.'

Our team performed a model evaluation and found a model with a concordance of 80% and accuracy of 70%. Figure 1 illustrates the lift achieved with our model.

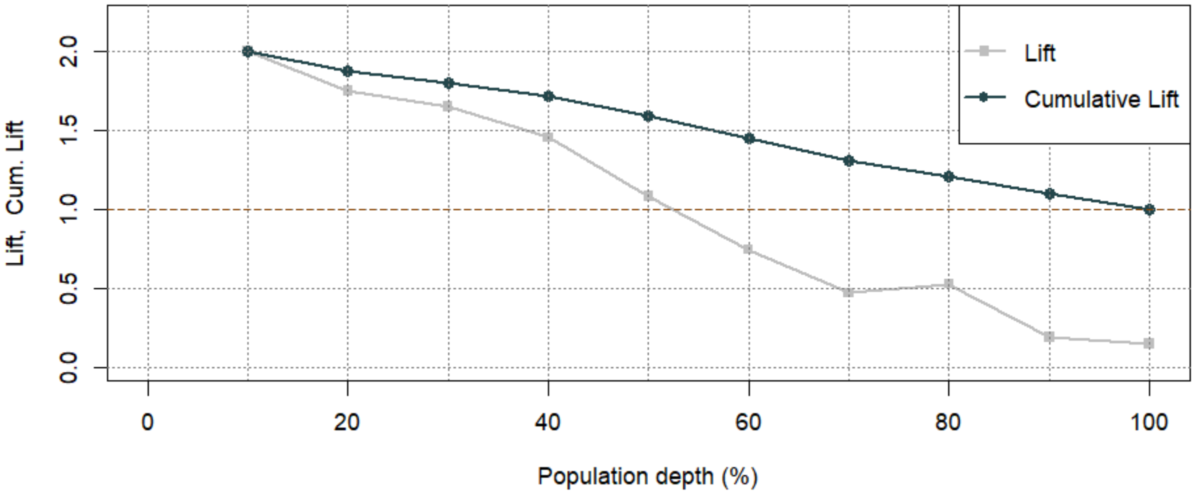

Figure 1: Lift and cumulative lift trend lines in 10% increments of the validation data set

Figure 1 demonstrates how likely customers are to purchase the product based on how the model ranked each customer's probability. For example, the top 40% of your customers, based on predicted probability, you get 1.71 times as many purchases compared to targeting a random sample of 40% of your customers.

We recommend that the Bank target marketing campaigns based on how the model ranked a customer. For example, we recommend calling the top 10%, sending marketing letters to the top 40%, and sending marketing emails to the top 80%. These recommendations could improve marketing tactics and reduce the cost of resources of the bank based on the model's customer profiles.

Methodology

The following section explains the data and the methods we used for model selection.

Data Used

The data contained information about whether or not bank customers purchased the product. The complete dataset contained 10,619 observations. The Bank split the data into a train and validation set with 8,495 observations and 2,124 observations, respectively. The dataset included 47 attributes (variables) about the customer and an indicator variable indicating if the customer purchased the product. The Bank binned continuous variables to transform them into categories.

Model Probability Metrics

Our team calculated several model probability metrics to accurately create and assess the optimal logistic regression model for predicting product purchases based on the data. We calculated the concordance percentage and the coefficient of discrimination to verify the model's ability to rank predictions accurately.

We also evaluated the model based on classification metrics. The Receiver-Operating Characteristic (ROC) curve and Kolmogorov Smirnov (KS) statistic were used to assess further the selected model's performance regarding true positive and true negative predictions on the training data.

We created a confusion matrix on the validation data to further assess predictive accuracy on a hold-out sample, with an accuracy percentage and lift plot providing insight into the validity of the selected model and the top depth percentage of our customers, respectively.

Analysis & Results

The following subsections show our final model and the probability and classification metrics we used to assess our final selected model. The first subsection details what variables were included in the model. The second subsection focuses on metrics related to the training dataset, and the final subsection shows how our model did when we predicted onto the validation dataset.

Final Selected Model

Table 1 ranks all variables in the final recommended logistic regression model by significance.

Table 1: Final logistic regression model's variables ranked by significance

Variable	P-value
Indicator for checking account	1.12E-05
Number of insufficient fund issues	9.57E-04
Indicator for IRA account	8.80E-01
Indicator for investment account	1.15E-04
Indicator for mortgage	6.62E-04
Indicator for CC	1.53E-07
Checking account balance bin	5.63E-60
Number of checks written bin	5.76E-20
Number of teller visit interactions bin	1.93E-08
Savings account balance bin	8.01E-129
Withdrawal amount bin	9.09E-10
CD balance bin	2.58E-39
ILS balance bin	5.64E-05
MM balance bin	2.37E-23
Interaction between checking account and IRA account	3.13E-04

The final model had one interaction term alongside the 14 original main effect variables. The interaction occurred between the indicator variables for checking accounts and retirement accounts.

Interpreting probability and classification metrics on the training dataset

When considering probability-based metrics, our model had a concordance of 80%, indicating that our model ranked observations correctly on 80% of all paired observations across the training dataset.

We also looked at the coefficient of discrimination to understand how well our model assigned a higher probability to those who purchased the product and a lower probability to those who did not. The coefficient of discrimination was 24.6%. The further the coefficient of discrimination is from zero, the better the discrimination. Figure 2 has two density curves, showing how well the model assigned probabilities.

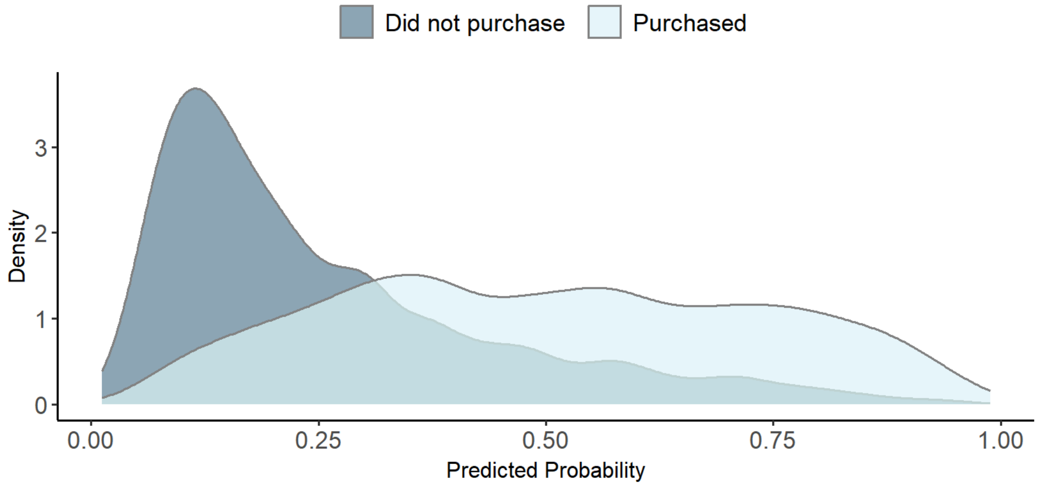

Figure 2: Density curves showing the discrimination slope

The density curve for those who did not purchase the product is skewed to the right, showing that our model does well at predicting lower probabilities for those who did not purchase the product. In contrast, the density curve for those who did purchase the product is more spread out. This spread indicates that the model was not as effective at assigning high probabilities to the customers who did purchase the product.

When looking at classification metrics for the model performance on the training dataset, our plotted ROC curve revealed a model sensitivity of 0.78 and model false positive rate of 0.34, as shown in Figure 3. Both the ROC curve and KS plot return the same values of 0.47 and recommended optimal cutoff of 0.30.

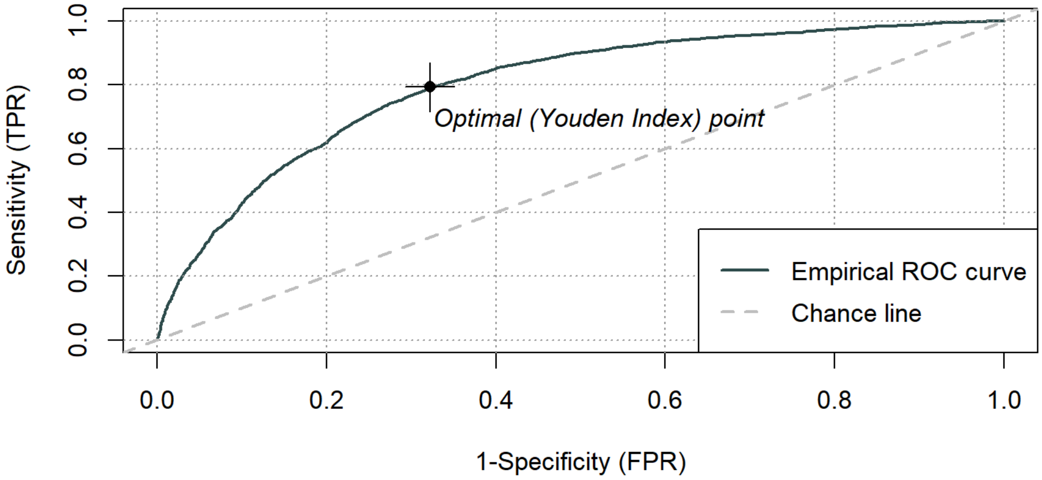

Figure 3: ROC curve with optimal Youden index point[1]

We used the optimal cutoff point of 0.30 to classify the probabilities in our model as 0s or 1s.

Evaluating model on validation data set

The validation data set was used to determine the model's predictive power. Accuracy and lift were the primary ways that the model was evaluated. Table 2 displays the confusion matrix used to assess model metrics such as accuracy.

Table 2. Confusion matrix was created on the validation data set

Categories		**Predicted**		**Total Values**
		Did not purchase	*Purchased*	*Row*
Actual	*Did not purchase*	909	473	1,382
	Purchased	160	582	742
Total Values	*Column*	1,069	1,055	2,124

Table 2 was used to find how successful our model was at determining if a customer would buy the product. In fact, our model was able to correctly predict the customer who would buy the product 70% of the time.

Furthermore, we found that the top 10% of customers in the validation data set were twice as likely to purchase the product than when 10% of people are selected at random as seen in figure 1.

Recommendations

Based on the results of our final model, we recommend that the following strategies be implemented to improve the prediction of which customer profiles will buy the product. The Bank could consider the following marketing strategies:

- Calling the top 10%, ranked on probability, to target marketing to customers twice as likely to purchase the product.
- Sending marketing letters to those in the top 40% of customers ranked on probability because these customers are 1.7 times as likely to purchase the product.
- Sending marketing emails to those in the top 80% of customers is ranked on probability because these customers are 1.2 times as likely to purchase the product.

Conclusion

Our team created a model to predict whether customers will purchase the product. The final model had a concordance of 80% and an accuracy of 70%. Considering the results of the lift analysis, we recommend that the Bank have different target marketing campaigns to selected customers based on purchase probability to save on resource use.

[1] TPR - Total Positive Rate, FPR - False Positive Rate