BANKING INSURANCE PRODUCT – PHASE 1

BLUE 18

RYAN HEGGIE
ISHI GUPTA
LUCY LIU
AUGUST MAJTENYI
AARYAN SHARMA

NOVEMBER 8, 2024

Table of Contents

Banking Insurance Product – Phase 1

Overview

The Commercial Banking Corporation (the "Bank") sought proposals to predict customers' likelihood of purchasing a variable rate annuity, hereafter called the product. Our team previously predicted the likelihood of buying a product with logistic regression.

In this report, we evaluate two models. One model uses the enhanced adaptive regression through hinges (EARTH) package, and the other uses a generalized additive model (GAM). Both models produced very similar goodness-of-fit metrics. For the GAM model, which makes more sense in this business context, we found an optimal cutoff of 0.31 and an area under the curve (AUC) of 0.8005. This model's AUC was better than the previous logistic regression model's AUC of 0.797.

Our team found that individuals owning older accounts with high liquidity – potentially measured by checking account balance and number of checks written – are more likely to purchase the product. Therefore, the bank should direct more marketing efforts to these customers.

Methodology

The following section describes the data used for analysis and how the machine learning models were created.

Data Used

The Bank provided data about customers who were offered the product. The Bank included a training data set with 8,495 observations and 37 predictor variables relating to the customer's account. We checked for missing values. We used median imputation to fill null values for continuous variables and mode imputation for categorical variables.

EARTH and GAM Models

To predict the probability that a customer will purchase the product, we implemented the EARTH algorithm for ranking variable importance and a GAM model intended to support variable selection. For the GAM model, we applied an alpha of 0.002 to find statistically significant predictors. We trained both models using the training dataset, including all variables, ensuring comprehensive analysis. We evaluated the models' AUCs on the receiver operating characteristic (ROC) curves.

Analysis

The following section showcases model formation and performance.

EARTH Model

The selection process for the EARTH model found 17 significant variables in predicting whether the product was purchased. As listed in Table 1, these variables were used to create the EARTH model.

Table 1: Variables ranked by residual sum of squares (RSS) value

Rank	Variable Name	RSS
1	Saving account balance	100.0
2	Certificate of deposit balance	67.9
3	Having a checking account	67.3
4	Checking account balance	67.3
5	MM balance	48.4
6	Having a certificate of deposit account	40.7
7	Age of oldest account	38.3
8	Number of checks written	33.6
9	Having an investment account	31.4
10	Number of teller visit interactions	29.2
11	Total ATM withdrawal amount	27.0
12	CD Balance	24.6
13	Branch 15	22.2
14	Branch 14	20.1
15	Branch 16	17.2
16	IRA balance	13.9
17	Having a savings account	7.8

Table 1 showed that the amount of money in the savings account was the most important variable, and its absence significantly impacted the fit of the EARTH model. Savings account balance was also at least 32% better than all other predictor variables.

After investigating variable importance, we evaluated the model's classification power on the training set. Figure 1 shows the ROC curve for this model.

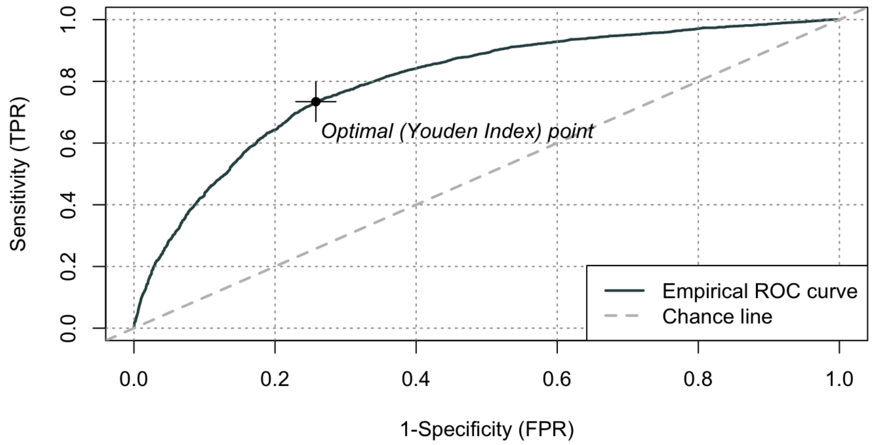

Figure 1: ROC Curve from the EARTH Model

As shown by the curve above the diagonal line, our EARTH model's classification performance was better than guessing at random. The AUC was 0.8009, and the optimal value yielded a cut-off of 0.304. For future classification, a customer with a probability of 0.304 or higher is more likely to buy the product.

GAM

Our team ran a GAM model to capture the complex relationship showcased by the variables in the data. Using our GAM model and an alpha level of 0.002, we determined the significant variables displayed in Table 2.

Table 2: Significant variables ranked by ranked by p-value

Rank	Variable Name	P - Values
1	Age of oldest account	< 2e-16
2	Checking account balance	< 2e-16
3	Number of checks written	< 2e-16
4	Number of teller visit interactions	< 2e-16
5	Saving account balance	< 2e-16
6	Total ATM withdrawal amount	< 2e-16
7	Having a checking account	< 2e-16
8	Branch 14	2.23e-06
9	Having an MM account	4.28e-06
10	Having a credit card account	5.03e-06
11	Having an investment account	7.15e-05
12	Branch 16	2.87e-05
13	Branch 15	4.22e-05
14	Certificate of Deposit balance	4.72e-04
15	Having a retirement account	2.92e-04
16	Having a certificate of deposit account	1.99e-04

Table 2 showed that variables related to high liquidity—such as checking account balance and number of checks written—were important factors in predicting product purchases. Additionally, we found that branches 14, 15, and 16 played a significant role in determining product purchases.

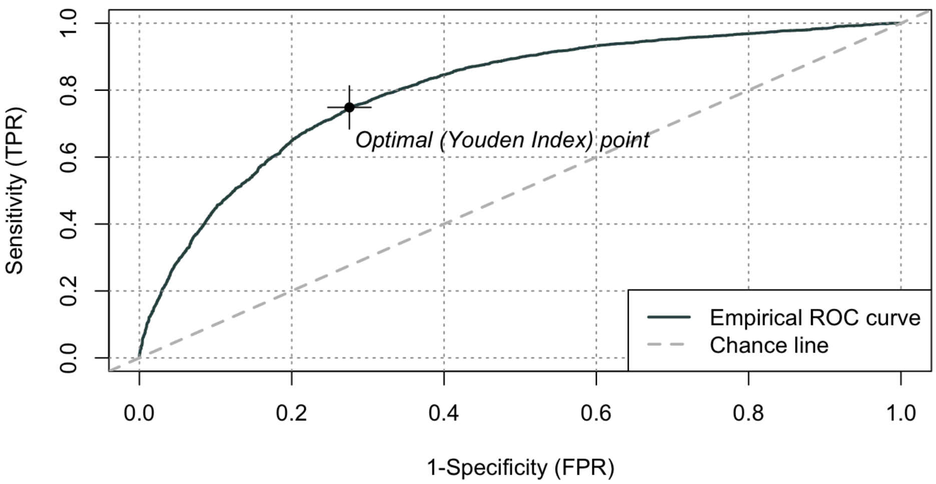

Figure 2: ROC Curve from the GAM Model

Our GAM model performed similarly to the EARTH model's ROC curve in accurately predicting which customers in the training data set would purchase the product. We obtained an AUC of 0.8005 and found an optimal cut-off of 0.31. For future classifications, customers with a probability of 0.31 or higher were more likely to buy the product.

Results & Recommendations

Although both models had similar AUC values, we recommend using the GAM model. The predictor variables in this business context have a nonlinear relationship with the product, allowing us to capture more complex relationships. We recommend the Bank evaluate its potential customers by:

- **Marketing Material**: Sending promotional material to customers with liquid accounts or savings-related accounts balances or liquid accounts above a cutoff could increase the sales of the product
- **More Nuanced Website Experience**: Customers who are likely to buy the product should have more visibility to this part of the website. This can increase sales and improve user experience.
- **Branch 16 Marketing:** The bank should push out more marketing material to customers who frequently visit branch 16. These customers are more likely to buy the product.

Conclusion

In this report, we compared two different models created from the same training data. Although both models had similar goodness-of-fit metrics, we recommend the Bank use the GAM model to capture nonlinear, complex relationships. For the GAM, we found an optimal cutoff of 0.31 and an AUC of 0.8005.

Our team identified significant variables – such as age of account, checking account balance, and number of checks written – that were important for predicting product purchases. Therefore, the Bank should direct more marketing efforts to these customers.

Appendix

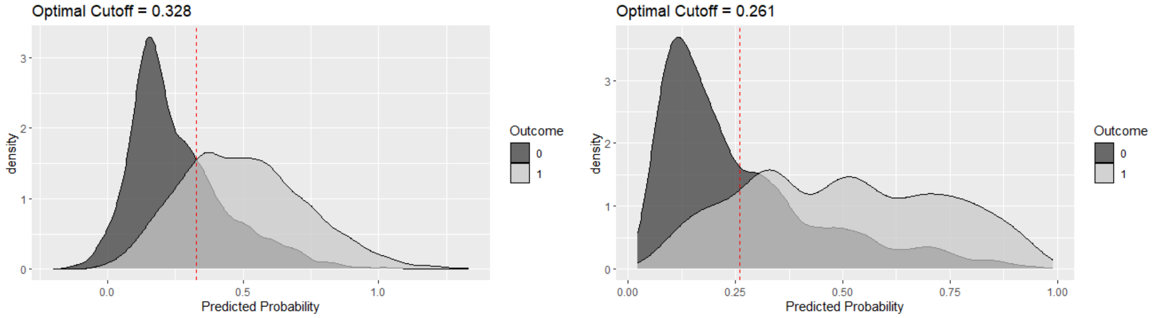

Figure 3. Side-by-side comparison of discrimination plots. GAM model (left) and logistic regression (right)