Modeling Insurance Purchase Behavior

Team 1

Lainey English,
Joschka Labinsky,
Ella Moses,
Christian Nielsen,
John-Carl Wharton

9/19/2025

Table of Contents

Overview

The Commercial Banking Corporation (the Bank) seeks a predictive model to identify which customers will most likely purchase a variable rate annuity product. To achieve this, we utilized logistic regression to predict purchase probability. By targeting the Bank's marketing efforts towards the top 20% of customers ranked by their predicted purchase probability, the Bank is 1.866 times more likely to capture a buyer than if customers were selected randomly, improving future marketing campaigns' effectiveness.

Methodology

First, we describe the data, preparation steps, and modeling approach to build the logistic regression framework for predicting if a customer will purchase a variable rate annuity.

Data Used

All variables in the provided dataset were categorical customer behavior and account activity predictors. Variables that were initially continuous were transformed into binned categories for this analysis.

Data Preparation

To ensure completeness, we assigned missing values a new “Missing” category so that observations were retained and potential predictive information in missingness was not lost. During initial modeling, we identified cases in the money market credits and cash-back requests variables where certain rare categories almost perfectly predicted whether a customer would buy the insurance product, a situation known as quasi-complete separation. This created unstable coefficient estimates in the model. To address this, we collapsed these sparse edge categories in training and validation sets, stabilizing the results. We also reduced multicollinearity in the data by removing aliased predictors and those with high scaled Generalized Variance Inflation Factor values (>5).

Model Building

With the refined predictor set, we developed a series of logistic regression models using forward, backward, and stepwise selection, testing both with and without interaction terms. We then evaluated model performance by balancing fit to the training data with model parsimony, using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Nagelkerke’s pseudo-R^2 as selection criteria.

Analysis

The following section outlines the model selection process and explains the justification for the chosen cutoff value.

Model Selection

Comparing the different models we generated, our final model had a pseudo-R^2 value of 0.307 and an AIC and BIC value of 8849.6 and 9046.9, respectively. The variables contained in our final model are

listed in Table 2 in the appendix, along with their associated likelihood ratio test p-values indicating statistical significance.

The model's performance was evaluated by its coefficient of discrimination (0.239), with the discrimination density plot (Figure 1) illustrating its ability to distinguish between customers who purchase an annuity and those who do not. A value of 0 would mean no separation, while values closer to 1 reflect a stronger distinction.

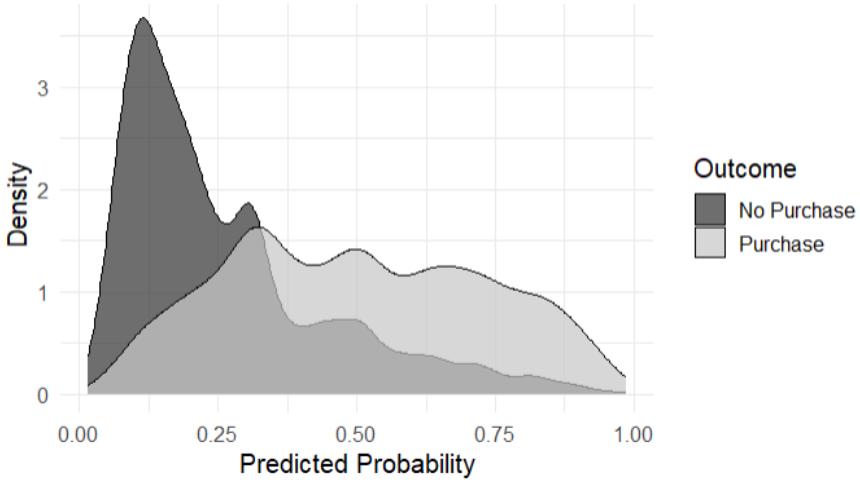

Figure 1: Discrimination Density Plot

The concordance percentage was also calculated to complement the discrimination value, resulting in a value of 83% meaning that the customers who purchased an annuity had a higher predicted probability value 83% of the time compared to the predicted probability of a customer who didn't purchase an annuity.

Cutoff Selection

Our model achieved a Kolmogorov-Smirnov (K-S) statistic of 0.467 at an optimal cutoff of 0.311 on the training data. However, the current practice of selecting the cutoff based on the optimized K-S statistic treats sensitivity and specificity equally. We based this decision on the assumption that missing a buyer and contacting a non-buyer have the exact cost.

We believe that in our business context, this assumption is unrealistic. For variable rate annuities, missing a potential customer represents a higher financial loss than the relatively low costs of contacting a non-buyer. Relying only on the optimized K-S statistics leads to underestimating the opportunity costs. A more appropriate approach is to include the trade-off between cost and revenue in our cutoff decision. This way, we maximize the profit rather than balancing statistical measures that ignore the business context.

Figure 2 shows the Receiver Operating Characteristic (ROC) curve, highlighting the trade-off between sensitivity and specificity. Increasing sensitivity necessarily results in reduced specificity. As described above, this trade-off is acceptable and desirable for our model.

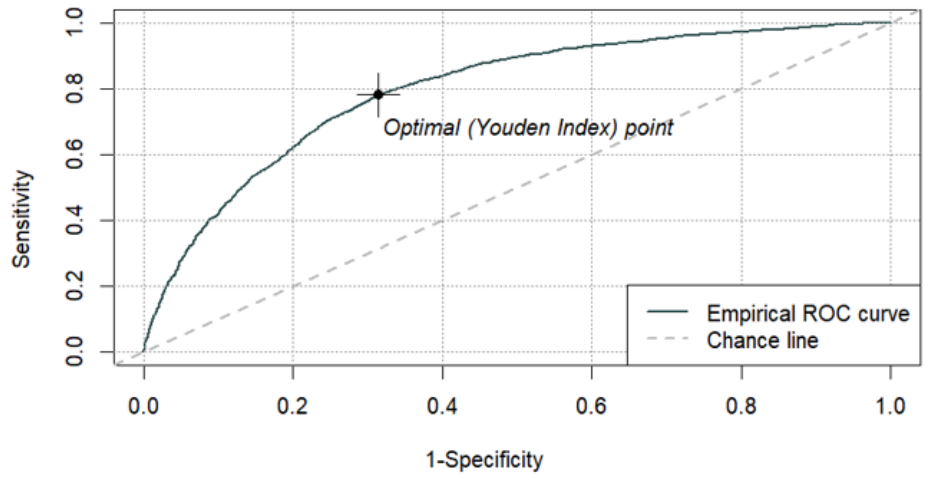

Figure 2: ROC Curve

We evaluated several cutoffs to determine which one provided the best balance between sensitivity and specificity in our business context. We selected 0.20 as the most appropriate cutoff for our final model because lowering the threshold further would have resulted in contacting too many non-buyers without generating additional value.

Results & Recommendations

This section summarizes the model evaluation results, focusing on accuracy, lift, and overall predictive performance, and provides recommendations based on these results.

Out-of-Sample Predictive Performance

When evaluating our model on the validation dataset, Table 1 shows that it correctly identifies most true buyers.

Table 1: Confusion Matrix – Insurance Product Purchase Prediction

	Predicted: No Purchase	Predicted: Purchase
Actual: No Purchase	645	737
Actual: Purchase	79	663

The model has a false negative rate of 10.6%, meaning it misses a relatively small portion of actual buyers. In an effort to reduce the number of missed buyers, the model has a higher false positive rate of 53.3%. Although this may result in additional marketing costs for customers who do not purchase

despite the model's predictions, we consider this trade-off acceptable. Overall, the model achieves an accuracy of 62%, balancing our primary objective of identifying true buyers with a necessary tolerance for false positives.

Lift

Figure 3 shows the model's lift, comparing its ability to identify buyers against random selection.

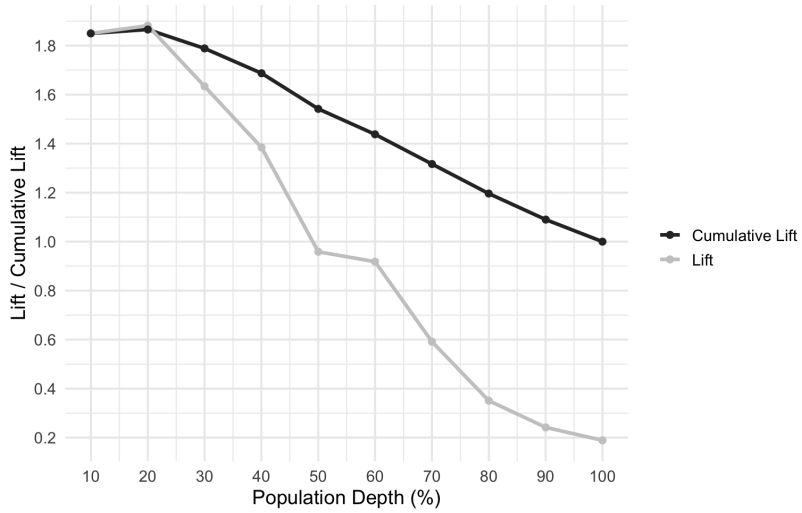

Figure 3: Model Lift by Population Depth

Both the lift and cumulative lift are maximized at a population depth of 20%. If the Bank targets the top 20% of customers ranked by predicted purchase probability, it is 1.866 times more likely to capture a buyer than if it selects 20% of customers at random.

Recommendations

We recommend using our model cutoff to capture as many potential buyers as possible, rather than relying on the KS statistic, which treats all misclassifications equally. By focusing on this cutoff, the Bank can reach more potential buyers. Marketing efforts should also prioritize the top 20% of customers identified by the model, where the likelihood of purchase is highest. This approach will help the Bank reach more potential buyers while using marketing resources efficiently.

Conclusion

We developed a logistic regression model to predict customers' purchasing probability of variable annuity products. The model has a false negative rate of 10.6%, meaning it correctly identifies most potential buyers. By targeting the top 20% of customers based on their predicted purchasing probability, the Bank can increase its success rate compared to a random selection. This approach leads to more targeted marketing efforts, increased efficiency, and a higher volume of variable annuity products sold.

Appendix

Table 2: List of Model Variables with Associated Likelihood Ratio Test p-Values

Variable	p-Value
Savings Balance	2.2×10^{-16}
Checking Account Balance	2.2×10^{-16}
Certificate of Deposit Balance	2.2×10^{-16}
Indicator for Money Market Account	2.2×10^{-16}
Indicator for Investment Account	2.2×10^{-16}
Number of Checks Written	2.2×10^{-16}
Total ATM Withdrawal Amount	1.63×10^{-10}
Number of Teller Visit Interactions	2.61×10^{-8}
Indicator for Retirement Account	1.46×10^{-5}
Indicator for Installment Loan	1.43×10^{-3}

Table 3: Lift Chart

Bucket	Obs	CObs	Depth	Resp	CResp	RespRate	CRespRate	CCapRate	Lift	CLift
1	212	212	0.1	137	137	0.646	0.646	0.185	1.85	1.85
2	213	425	0.2	140	277	0.657	0.652	0.373	1.881	1.866
3	212	637	0.3	121	398	0.571	0.625	0.536	1.634	1.789
4	213	850	0.4	103	501	0.484	0.589	0.675	1.384	1.687
5	212	1062	0.5	71	572	0.335	0.539	0.771	0.959	1.542
6	212	1274	0.6	68	640	0.321	0.502	0.863	0.918	1.438
7	213	1487	0.7	44	684	0.207	0.46	0.922	0.591	1.317
8	212	1699	0.8	26	710	0.123	0.418	0.957	0.351	1.196
9	213	1912	0.9	18	728	0.085	0.381	0.981	0.242	1.09
10	212	2124	1	14	742	0.066	0.349	1	0.189	1

Sources

McGee, C. (2024, October 10). *Average Customer Lifetime Value for Financial Services*. Focus Digital. https://focus-digital.co/average-customer-lifetime-value-for-financial-services/

Patil, S. (2025, April 5). *What Impacts Lead Gen Costs? A Guide for Businesses in 2025*. Revnew. https://revnew.com/blog/lead-generation-pricing