Ohm Kundurthy

Customer Propensity

National City Bank

1. Objective

The research and modeling team at National City Bank is tasked with developing a customer propensity model to identify the next 100 customers for a pilot line of credit product secured by a household’s used car. Using historical data from 4,000 previous calls and mailings, along with supplemental data, the team will create and evaluate the model to predict the highest probability customers to contact. The findings, including variable importance and customer profile insights, will be presented to the bank's chief product officer for approval. Once approved, marketing efforts will proceed, with the deliverables including code, a PowerPoint presentation, a written supplemental, and a CSV of the top 100 customers by predicted probability.

1. Data Summary

Table below summarizes data dictionary with some additional insights on the attributes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Variable** | **Source** | **DataType** | **Remarks** |
| 1 | dataID | internal | Unique ID | 4000 Unique values |
| 2 | HHuniqueID | internal | Unique ID | 4000 unique values |
| 3 | Communication | internal | Categorical | 2 Values |
| 4 | LastContactDay | internal | Categorical | 31 distinct values |
| 5 | LastContactMonth | internal | Categorical | 11 values representing different months |
| 6 | NoOfContacts | internal | Continuous | 35 unique values |
| 7 | DaysPassed | internal | Continuous | Several |
| 8 | PrevAttempts | internal | Continuous | 20 unique values |
| 9 | Outcome | internal | Categorical | 3 Distinct values |
| 10 | CallStart | internal | TimeStamp | Many distinct values |
| 11 | CallEnd | internal | TimeStamp | Many distinct values |
| 12 | Y\_AccetpedOffer | internal | Categorical | Binary Target Variable |
| 13 | headOfhouseholdGender | 3rd party | Categorical | 2 distinct values |
| 14 | annualDonations | 3rd party | Continuous | Numerical |
| 16 | EstRace | 3rd party | Categorical | Many distinct values |
| 17 | PetsPurchases | 3rd party | Binary |  |
| 18 | DigitalHabits\_5\_AlwaysOn | 3rd party | Categorical | 5 Distinct Values |
| 19 | AffluencePurchases | 3rd party | Binary |  |
| 20 | Age | 3rd party | Continuous | Between 18 and 95 |
| 21 | Job | 3rd party | Categorical | 11 Distinct occupations |
| 22 | Marital | 3rd party | Categorical | 3 Distinct values |
| 23 | Education | 3rd party | Categorical | 3 Distinct values |
| 24 | carMake | 3rd party | Categorical | Many distinct values |
| 25 | carModel | 3rd party | Categorical | Many distinct values |
| 26 | carYr | 3rd party | Numeric | Many distinct values |
| 27 | DefaultOnRecord | 3rd party | Binary |  |
| 28 | RecentBalance | 3rd party | Continuous |  |
| 29 | HHInsurance | 3rd party | Binary |  |
| 30 | CarLoan | 3rd party | Binary |  |

1. ModelDevelopment

The SEMMA methodology (Sample, Explore, Modify, Model, and Assess) was employed to develop a propensity model. This methodology was adapted with a particular focus on ensuring both parsimony and model validity, allowing for a computationally efficient yet reliable model.

* 1. Sample

The first step in the SEMMA process was the selection and preparation of data. Historical data from 4,000 previous calls and mailings was collected, along with any supplemental data available for customer profiling. The data was carefully examined to ensure that the relevant features were included for predictive modeling. Key variables were selected based on initial exploratory analysis, which included demographic information, past interactions, and behavioral data, such as the number of contact attempts, communication type, and financial behavior indicators like car loans and household insurance status.

* 1. **Explore (Exploratory Data Analysis)**

Exploratory Data Analysis (EDA) played a critical role in understanding the distribution and relationships between variables. In this phase, statistical measures such as correlations, distributions, and summary statistics were examined to identify significant predictors of customer acceptance for the line of credit offer. The focus was on identifying variables that had the most predictive power, as well as those that were likely to introduce noise into the model. Visualizations and diagnostic tools, including scatter plots, histograms, and box plots, were used to detect any data issues (e.g., outliers or missing values to handle appropriately.

* 1. **Modify (Data Transformation and Feature Engineering)**

Data modification involved transforming variables and handling any issues discovered during EDA. Using functions from vtreat library, categorical variables were converted into factors, and any missing data was addressed using imputation methods. Feature engineering was also an essential part of this step, where derived variables such as "previous attempts," "days since last contact," and "car year" were created to capture more granular aspects of customer behavior. The goal of this modification phase was to refine the dataset to ensure that it captured all potentially relevant factors while remaining manageable for model fitting. Also, 90% data was included as training set and 10% as validation set.

* 1. **Model (Model Development and Selection)**

The core of the SEMMA process is model development.

* + 1. **Logistic regression**

An efficient and explainable logistic regression model was built, balancing simplicity and performance by following steps

* Initial Regression: A logistic regression model (logReg) was fit using all the input variables. AIC (2732.4) and p-value below threshnold.
* Backward Stepwise Selection: To improve the model's parsimonious nature, backward stepwise selection (step() function) was applied to eliminate less significant predictors, resulting in a refined model (bestLogReg) with AIC of 2711.2.
* Final Model: use only significant variables to form a new logistic regression model with a slightly higher AIC (3033.7), indicates a slightly less optimal fit compared to the previous model but focuses on computational efficiency
  + 1. **Classification Tree**

Below approach helped with a simple yet effective tree model by using cross-validation, optimal complexity parameter selection, and controlled splitting for robust predictive performance:

* **Complexity Parameter:** The code specifies a range of complexity parameter (cp) values to test during model training. This ensures the model is as simple as possible without sacrificing predictive power. The optimal cp of 0.005 was chosen based on the lowest Root Mean Square Error (RMSE)
* **Tree Pruning**: Used min-split = 10 and min-bucket = 5. These parameters enforce minimum requirements for node splits and final node sizes
* **Cross-Validation**: Cross-validation within caret’s training process ensures that the chosen model generalizes well to unseen data, reducing the chance of overfitting.
  + 1. **Random Forest**

Following steps were taken on the model to reduce overfitting while preserving predictive power and clarity.

* **Identifying Overfitting**: The initial model, with 100 trees and an mtry of 10, yielded overly high accuracy, signaling overfitting. This prompted a need to simplify the model to enhance generalization.
* **Reducing Model Complexity**: Through a series of iterations, adjusted the model to use only 50 trees and lowered mtry to 5, ensuring fewer features at each split. These changes made the model less dependent on specific data points, improving its robustness.
* **Evaluating Feature Importance**: By examining and plotting the top 20 influential features, we identified the key variables driving predictions. This approach enabled us to simplify the model further by focusing on essential features, achieving interpretability and reducing overengineering.

However, based on results seen the model is still overengineered to an extent given the drop in accuracy against validation set. Hence, we can conclude that there is still room for improvement based on parameter tuning.

* 1. **Assess (Model Evaluation and Validation)**

The final step in SEMMA is model assessment, where the model’s performance is rigorously evaluated. The parsimonious model was assessed using several metrics, including accuracy, precision, sensitivity, specificity, and F1 score. A key part of ensuring the model’s validity was cross-validation with a separate validation set. This was critical in ensuring that the model did not overfit the training data and could generalize well to new, unseen customer data.

The model was also evaluated using a confusion matrix, which helped identify key metrics such as the number of true positives, false positives, true negatives, and false negatives. By setting a threshold of 0.5 for classification, the model was able to predict the probability of a customer accepting the offer, for each model. A further optional step is to move the threshold up or down to improve specificity and sensitivity.

* + 1. **KPI Comparison Across Models**

See below a table with all the KPI’s to review the comparison of results from all 3 models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | Logistic Regression  Training Set | Logistic Regression  Validation Set | Classification Tree  Training Set | Classification Tree  Validation Set | Random Forest  Training Set | Random Forest Validation Set |
| Accuracy | 0.8028 | 0.8225 | 0.8322 | 0.8175 | 0.9847 | 0.82 |
| 95% Confidence Interval | (0.7894, 0.8157) | (0.7815, 0.8587) | (0.8196, 0.8443) | (0.7761, 0.8541) | (0.9802, 0.9885) | (0.7788, 0.8564) |
| No Information Rate | 0.5983 | 0.605 | 0.5983 | 0.605 | 0.5983 | 0.605 |
| P-Value [Acc > NIR] | 2.2E-16 | 2.2E-16 | 2.2E-16 | 2.2E-16 | 2.2E-16 | 2.2E-16 |
| Kappa | 0.5799 | 0.6216 | 0.6512 | 0.6177 | 0.9681 | 0.6201 |
| Mcnemar's Test P-Value | 2.57E-14 | 0.05758 | 0.8388 | 1 | 0.000001208 | 0.4094 |
| Sensitivity | 0.8825 | 0.8884 | 0.8584 | 0.8512 | 0.9958 | 0.8678 |
| Specificity | 0.684 | 0.7215 | 0.7932 | 0.7658 | 0.9682 | 0.7468 |
| Precision | 0.8062 | 0.8301 | 0.8608 | 0.8477 | 0.979 | 0.84 |
| Neg Pred Value | 0.7963 | 0.8085 | 0.7899 | 0.7707 | 0.9936 | 0.7867 |
| Prevalence | 0.5983 | 0.605 | 0.5983 | 0.605 | 0.5983 | 0.605 |
| Detection Rate | 0.5281 | 0.5375 | 0.5136 | 0.515 | 0.5958 | 0.525 |
| Detection Prevalence | 0.655 | 0.6475 | 0.5967 | 0.6075 | 0.6086 | 0.625 |
| Balanced Accuracy | 0.7832 | 0.805 | 0.8258 | 0.8085 | 0.982 | 0.8073 |
| F1 Score | 0.843 | 0.8587 | 0.8596 | 0.8495 | 0.9874 | 0.854 |

* + 1. **Final prediction and Target profile**

Final steps performed to calculate the ensemble average and identification of top 100 customers

* Combine Results: Combine predictions from the training , validation (and prospects)
* Calculate Average probability: Calculate the average of predicted probabilities from three models
* Rank Customers: Rank the customers probabilities based on the average probability in descending order.
* Select Top 100 Customers: Full Results are filtered to retain only the top 100 customers based on rank, creating a data frame of topCustomers100.

1. **Insights based on Results**

The following values for variables are noted to be strong indicators of likelihood for offer acceptance as most of the top 100 customers share the same values for these variables.

* past\_Outcome = Success
* RecentBalance > 100
* CarLoan = 0
* Duration > 200

This allows for an easier schematic to identify target demographic.

1. **Conclusion**

The SEMMA methodology provided a structured approach to developing a robust and parsimonious customer propensity model. By focusing on parsimony and validating the model through a variety of performance metrics, the team at National City Bank was able to create a model that not only accurately predicts the next 100 customers but also offers insights into customer profiles and behaviors. Optionally one would also evaluate ROC curve which was not performed here. The final model is a valuable tool for informing marketing decisions and optimizing resource allocation for the new line of credit product which should be tested and iterated further for deployment.