Propensity Model for Used Car Line of Credit

By: Tom Herman
National City Bank
November 11, 2024

Primary Objective: Identify the Next 100 Customers from a Prospective Customer List to
Secondary Objective: Provide Customer Profiles for Both Current and Prospective Customers

Summary Findings: Used SEMMA workflow to build three classification models, of which a decision tree exhibited the best characteristics. Decision tree yielded predicted probabilities for prospective customers accepting our new line of credit offer. Prospective customers were then arranged into the top 100 with the highest likelihood of accepting the offer. Our marketing team should begin with these top 100 prospects

YouTube Video Link: https://youtu.be/NXIQVTwUI5I

Sample:

Current customers data was enriched with various additional third-party data including credit and vehicle information. Race was removed as a factor for prediction. N/A values were replaced with "other," "unknown," or other replacement values. Some outliers were removed (less than 1% of the data), for large number of contacts, too many days passed, too many previous attempts or too high of a recent balance. Prospect data was aligned with the current customers data (same columns, order, etc.).

10% of the data was used for vtreat rows, with the remainder being split 80% and 20% for training data and validation data, respectively.

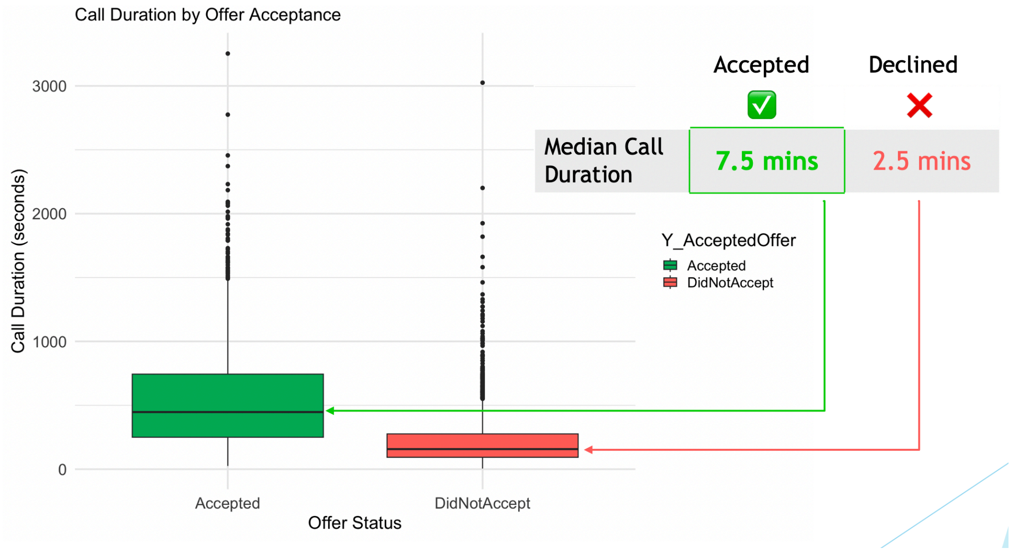
Explore:

The training data was explored for any interesting observations related to age, gender and call times. Most of our current customers are millennials and gen X-ers, with ages falling between the inter-quartile range of 32 and 49, with a median of 39. Interestingly, customers in the first and fourth quartiles (younger and older clients) have a much higher rate of acceptance than those in the middle. Age was later identified as an important variable for prediction.

Genders were evenly represented in the data set and there wasn't a major difference in propensity for males or females to accept our offer.

The "Call start" and "call end" variables were not present in the prospect customers data, so they had to be removed from the current customers data. However, the call durations were calculated and explored, and it was determined that customers who accepted our offer were on the phone with us a median 3 times longer than customers who did not accept our offer.

Further, the minimum call duration for a customer who accepted our offer was 25 seconds. We should consider how our call scripts can keep our customers engaged for longer, which bodes well for our offer being accepted.



Modify:

Informative features were determined using the varImp() and step() functions in the Random Forest and GML modeling process. Following the identification of the eight most informative features, they were loaded into the vtreat plan. The outcome variable name was “Y_AcceptedOffer” with a success class of “Accepted.” The training, validation and prospect data was treated with the plan, and these data sets were used as a foundation for all three models.

Model:

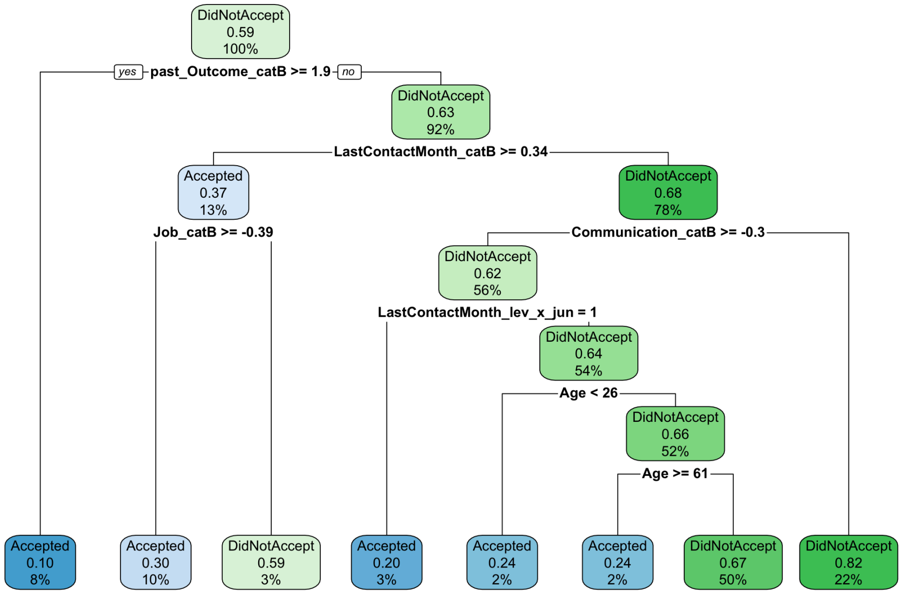
Three algorithms were used to create classification models to look at the data of our current customers, learn which patterns lead to accepted offers, and use that to predict which of our prospect customers are most likely to accept our new line of credit.

These models were made to be as parsimonious as possible, meaning certain functions (varImp and step) were used to identify and deploy only the most influential variables. This reduced noise in the models, made them more accurate and allowed them to run faster.

Accuracy of each model was evaluated by how accurate the model was at making predictions on the current customer data it learned from, but also against validation data it hasn’t seen yet. The objective is to achieve high accuracy on the training data, but also a relatively consistent accuracy when applied to the validation data, which means that it should generalize or perform just as well with unseen data.

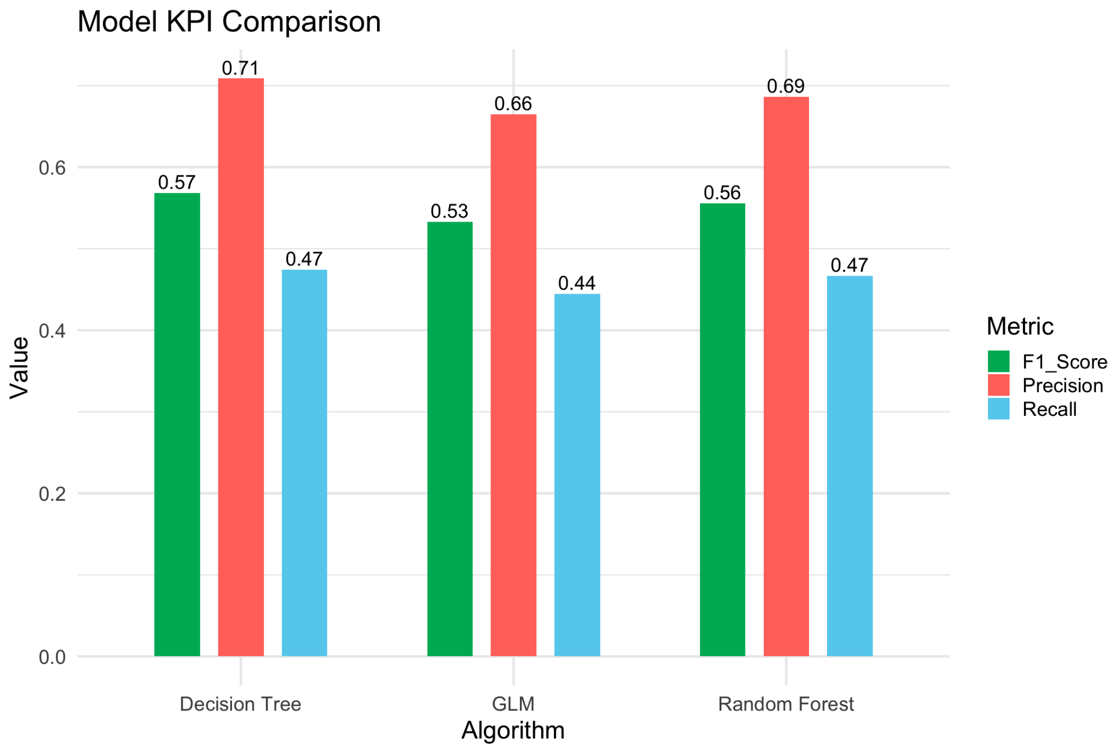
The Random Forest model had the highest accuracy (79.5%), but lost quite a bit of accuracy on validation (71.6%). The Decision Tree accuracy is much more consistent and has the highest accuracy on the validation data set (first vote in favor of using the Decision Tree model on the prospect data).

A decision tree diagram (see below) shows the full tree and the branches where it splits off and makes decisions. In this case, the "past outcome" is the most influential variable, where 8% of our current customers accepted the offer when they had a past outcome that was successful (which is intuitive). Other important branches for this decision tree also include last contact month, job, communication channel, and age. These all happen to also be variables that the variable importance function and step function identified as being highly influential in the GLM and Random Forest models.

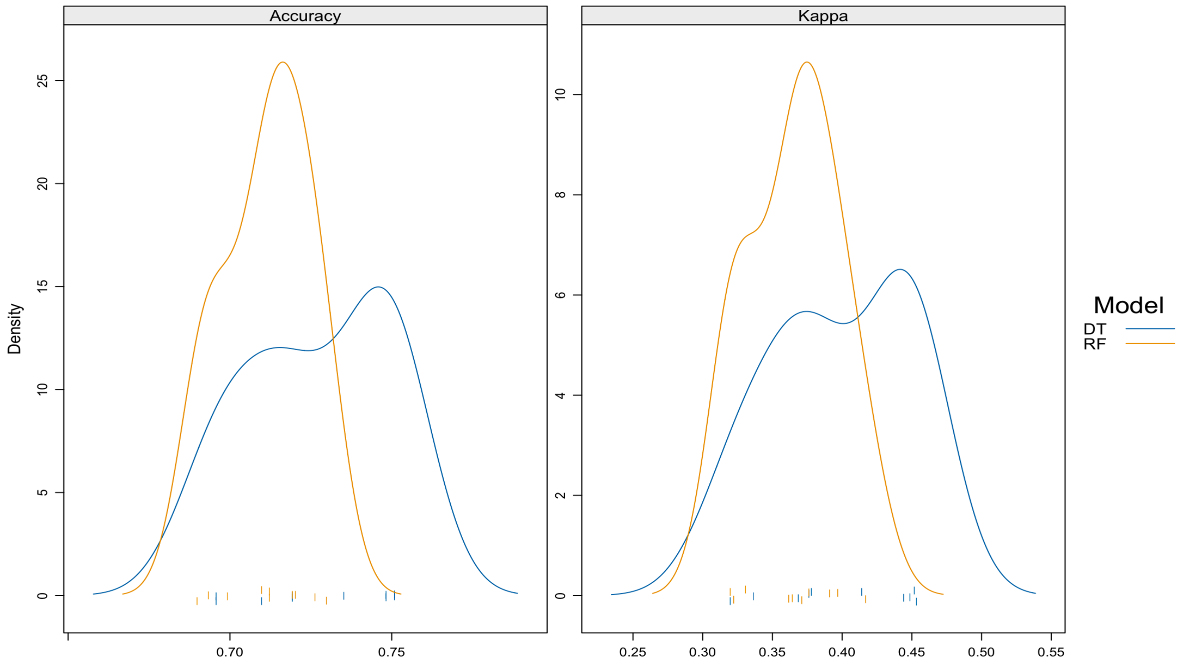


Assess:

Below (next page) we can see a comparison of all three models using three important metrics: Precision, Recall and F1 Score. In this case we want to maximize precision, which means that when our model predicts an accepted offer, it's mostly correct. This is important because we want to identify the top 100 sure bets in our prospective client data set. Recall represents the proportion of true positives to the total actual positives, and the F1 score gauges the performance of each model, using a balance of both precision and recall. An F1 score closer to one means the model is doing a good job at both minimizing false positives and capturing most of the actual positives. The decision tree exhibits the highest precision and F1 score.



Finally, the distribution charts below show the accuracy and kappa ranges for the Decision Tree versus Random Forest. The Decision Tree has a wider range of accuracy, which should generalize better to unseen prospective customer data. The Random Forest shows a much tighter band, which won't generalize as well.



Based on the accuracy, accuracy ranges and other KPIs, the Decision Tree was selected as the best model to apply to our thousand prospective customers, and determine the top 100 who should be the most likely to accept our new line of credit offer.

Output:

The Decision Tree model was applied to the prospective customer data, which yielded two columns of “Accepted” or “Did Not Accept” probabilities. This information was joined with the original Prospect Customers data set, which was then reordered and filtered to only include the top 100 prospects with the highest probability of accepting the offer. These probabilities all ranged from approximately 80% to 90%. The marketing team should consider engaging these prospects first.

Some additional EDA was performed on the top 100 prospects to identify any common themes:

- The vast majority 93% of our top prospects are cell phone users
- Most of our top prospects have only interacted with us once or twice
- 79% of our top prospects have previously had a “success” outcome with us
- Most of our prospects (82%) have a secondary or tertiary degree
- The most common jobs are “management” (27%), “technician” (15%), and “retired” (15%)
- Our top prospects are nearly 50/50 male/female

Conclusion:

Recap: The SEMMA data mining workflow was used to sample and explore the data from our current customers, modify the data for modeling, and then to assess three classification models to choose the best one, which turned out to be the Decision Tree.

The Decision Tree was used to predict the top 100 of our prospect customers who are the most likely to accept our new line of credit offer. All of the data can be found in the CSV file called “Top 100 Prospect List,” and the top prospects are ranked from top to bottom by probability of accepting the offer.

Marketing should consider engaging these top 100 prospects first, while also keeping in mind the unique characteristics of this population and what corresponding tactics can be used to engage them most effectively.