RefiJet SMS Modeling

Ben Sunshine – Data Science Intern

1/3/23

<u>Introduction</u>:

At RefiJet, our managers are constantly trying to improve the efficiency of our sales team. The recent implementation of Balto to monitor calls and provide our sales team with dynamic prompts and managers with advanced analytical toolkits will no doubt raise the level of performance of our call center. The standardization of a good call script is essential to engineering as many deals as possible. Likewise, the SMS messages we send to new leads and customers should adhere to a consistent level of standardization. Currently, LDR's and FSR's send personalized or shared SMS messages to new leads.

<u>Goal</u>:

I aim to use data from both our internal MYYESGO database and Bright Pattern to construct a statistical model. This model will utilize word combinations from initial outbound SMS messages to predict the likelihood of receiving a follow-up call after the message has been sent. This model will identify the most effective words and message types that prompt a call response from our customers after the initial message is sent.

<u>Data Preparation</u>:

The data provided from Bright Pattern consisted of all the SMS interactions between members of our sales team and customers from 1/1/2022, to 7/1/2023. Because markets change, I restricted my data to only messages sent after 3/1/2023. Similarly, from the MYYESGO database, I extracted data from the "call_detail", "applications", "applicant_numbers", "app_comments", and "app _stages" relations. This data when joined together gave a detailed chronological log for each application. Next, I filtered the dataset to isolate applications initiated with the specific outbound messaging scenario of interest. These scenarios were characterized by criteria such as applications not marked as "in progress," scenarios where no direct contact was established, scenarios not originating from direct mail leads, and instances where no voice calls, lending tree interactions, or bridge scenarios occurred prior to the first text message. I defined a "call response" scenario for SMS messages as one involving a call or bridge scenario (transfer) occurring after the SMS and lasting for sixty seconds or more. Messages failing to meet these criteria were considered as receiving no response.

When cleaning the bright pattern data, I was mainly concerned with scrubbing as much personal information about our customers and members of our sales team as possible. This included removing customer/RefiJet full names, first names, last names, years, months, days of the week, phone numbers, money amounts, locations, rates, car makes, car models, email addresses, times, and other SMS slang. This step was taken to remove bias in the future modeling stages and standardize many text message slangs.

Feature Engineering:

After performing message cleaning, I conducted sentiment analysis to explore the influence of sentiment on the likelihood of receiving a response to calls. These sentiment scores were computed using different sentiment lexicons, resulting in some variability among libraries. Additionally, I employed the "emotion()" function from the sentimentr package to identify the prevalence of emotional categories in the text. Finally, I generated variables to monitor character count, sentence count, exclamation marks, question marks, capital characters, and established ratios for each of these variables per sentence (i.e., exclamation marks per sentence).

Data Exploration:

One of the first important observations to note is the class imbalance between the target variable "call_response", which indicates a '1' for a call response and a '0' for no call response. There are 31,245 non-response messages and 2,110 messages that got a call response. Upon visualizing the distributions of the features in the data set, the majority appeared to be right skewed. I opted to use "modeling_msg", "is_double_text", "word_count", "time_since_lead_validation", "sms_sent_dow_utc", "sms_sent_time_utc", "refijet_first_name_used", "refijet_full_name_used", "car_make_model_referenced", "char_count", "exclam_mark_count", "question_mark_count", "capital_char_count" for the modeling of my data. I didn't include any of the sentiment/emotional analysis because these values were predicted and can reduce model interpretability.

Data Modeling:

First, I split my data into a 70-30 train/test split and stratified the split by the target variable, "call_response". Additionally, I split my training data into 5 folds for cross validation and resampled the data using 3 repeats to later get a picture for how well each model specification generalizes unseen data. In terms of model selection, I explored different specifications, including Lasso regression, Elastic Net regression, Support Vector Machines (SVM), and Extreme Gradient Boost (XGB). For Lasso regression, I fine-tuned the penalty parameter. In Elastic Net regression, I optimized both the penalty and mixture parameters. In the case of SVM models, I performed tuning on the penalty parameter and the width of the Radial Basis Function (RBF) kernel. Lastly for XGB models, I tuned the number of trees, tree depth, and learning rate.

Next, I formulated my tidymodels recipes with a consistent approach. Across all recipes, I implemented a series of data preprocessing steps to prepare the dataset for modeling. These steps included down sampling the data to include twice as many messages with no call response as those with call responses, tokenizing the text messages to break them into individual words, removing stop words, n-grams with tunable parameters, filtering tokens based on their frequency, and fine-tuning the maximum number of tokens. I also applied Term Frequency-Inverse Document Frequency (TF-IDF) to assign numerical importance to tokens based on their frequency. Additionally, I one-hot encoded categorical variables, removed highly correlated variables (with a tunable correlation threshold), removed variables with near-zero variance, and centered the numeric data by subtracting the mean from each value. Next, I

applied a Yeo-Johnson transformation to handle right-skewed data distributions, and lastly normalized the data to ensure that numeric features had a mean centered at 0.
In addition to these common preprocessing steps, I tailored each recipe to different model specifications, including Lasso, Elastic Net, SVM, and XGB models. These variations in recipes encompassed two distinct approaches:

1. Stemming: One recipe involved applying stemming to reduce words to their root form.
2. No stemming: This recipe preserved the original tokenized text without any stemming.

For the SVM models specifically, I employed additional down-sampling to equalize the number of instances with and without a call response. This was done because SVM models are known to be sensitive to imbalanced data. Additionally, I applied an inverse logit transformation in my SVM models to constrain numeric variables to a range between 0 and 1, which helped mitigate the model's susceptibility to variance.

I then tuned all my models using a grid search via racing, which is a more efficient than an iterative grid search because it checks whether a parameter is statistically different from another. In this tuning grid I measured each model’s, accuracy, precision, recall, F1 score, and The Area under the ROC curve.

Model Selection:

After tuning each combination of models to the 5 folds of training data I initially made, I plotted all the models to compare their ROC, which is a metric that describes a model’s ability to discriminate between positive and negative classes in binary classification problems. Upon plotting the ROCs for each model, XGB’s seemed to perform the best with slightly higher ROCs (0.64) than SVMs as seen in *Figure 1*. I then fitted this model to my testing dataset, resulting in an 81% accuracy and an ROC of 0.64. The ROC curve can be seen below in *Figure 2*.

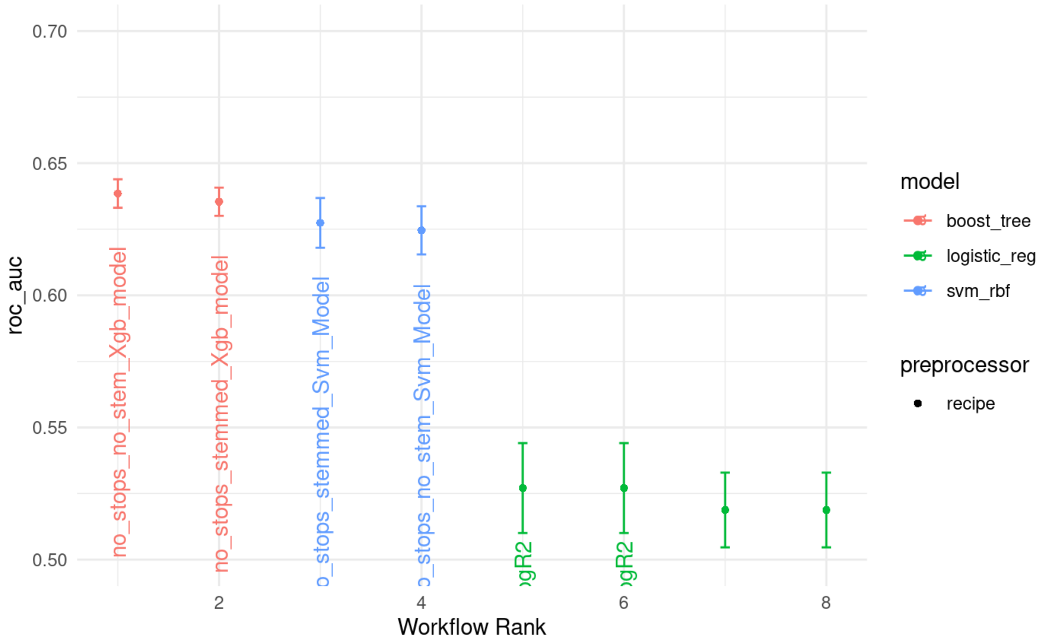

Figure 1

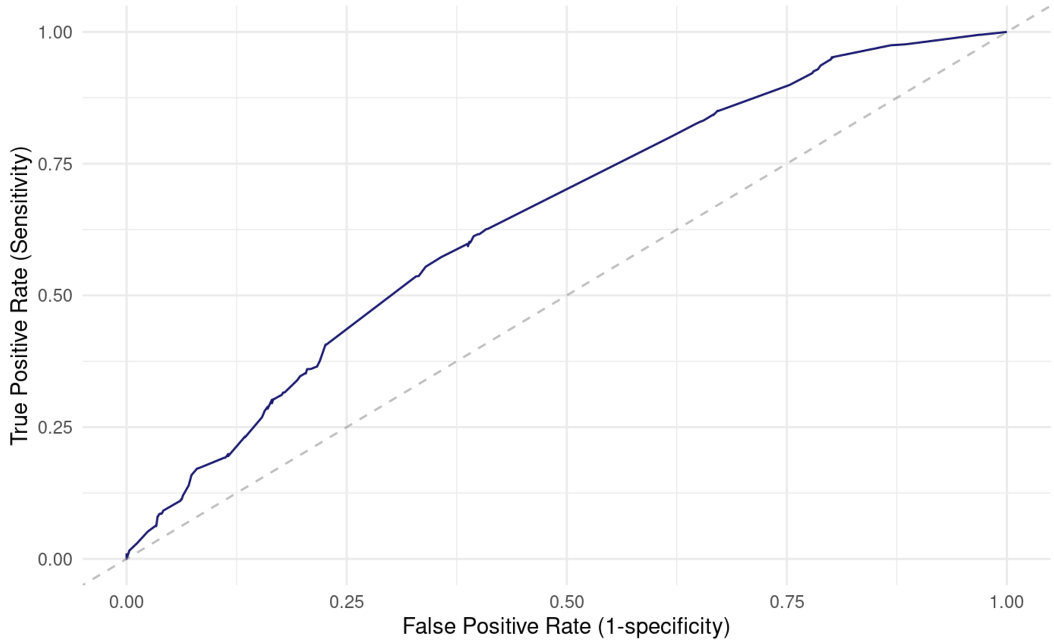

Figure 2

It's worth noting, as indicated by the curve above, that the top section of the curve (where specificity is 1 and sensitivity is 1) represents the model's accuracy in predicting the non-call response class. In this aspect, the curve exhibits a consistently flat slope, signifying the model's effectiveness in identifying messages that won't receive call responses. This is further affirmed by the model's 95% accuracy in predicting messages that, with 50% or greater certainty, should not receive a call.

On the flip side, the lower end of the curve is less consistent, which is expected. The likelihood of a lead responding to a message isn't solely determined by the message's quality; it depends on various other factors. The same message may be sent to multiple leads, but only a subset may respond due to these external factors.

Once again, the primary focus is on the messages belonging to a higher probability of receiving a call response. Therefore, I filtered the dataset to include only messages that the model predicted would receive a call response with 65% or higher confidence. Among these messages, the accuracy rate was 27.9%, with 19 true positives and 49 false positives. To gain further insights, I constructed a data frame that ranked the testing set messages in descending order of their predicted likelihood of receiving a call. This provides the most probable texts from the testing set to prompt a call response.

<u>Findings for Shared & Owned Leads</u>:

An efficient way of examining weights in XGB models is generating variable importance plots. These are ordered by the phrases and other variables that serve as the most influential variables in the model. As seen below in *Figure 3*, the FSR mentioning their phone number with an exclamation point, the time since lead validation, the sent time, and mentioning "payment" served as important aspects of the model.

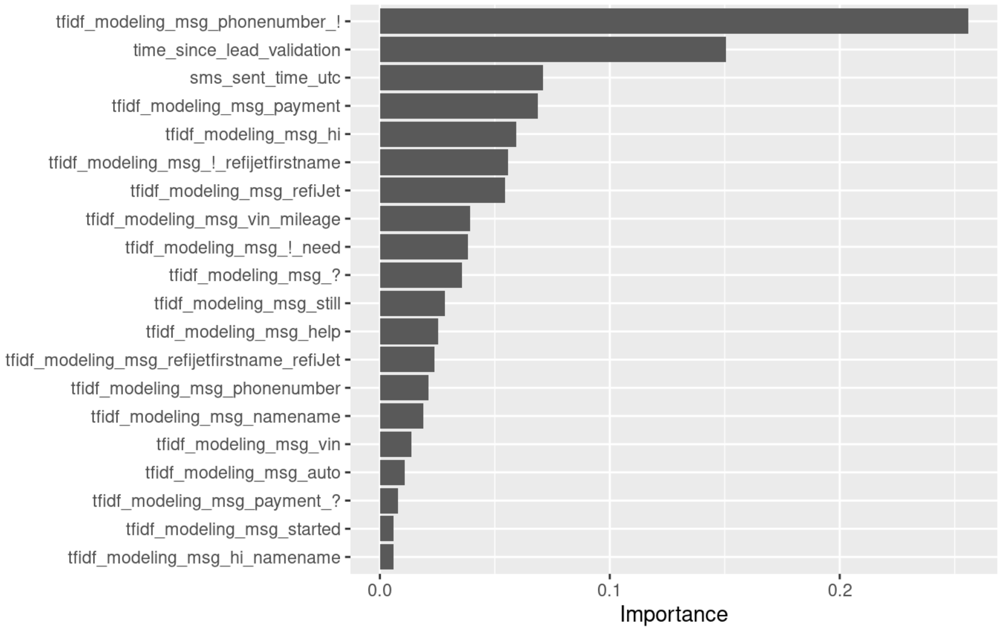

Figure 3

Since the time since lead validation and the sent time of messages served as larger weights in the model, I analyzed the distribution of time since lead validation and message sending hours in the model. This breakdown focused on messages with a high predicted probability of receiving a call response and all messages that did receive a call response. From the plots below, most messages which receive a call response lie within 0-40 days since lead validation.

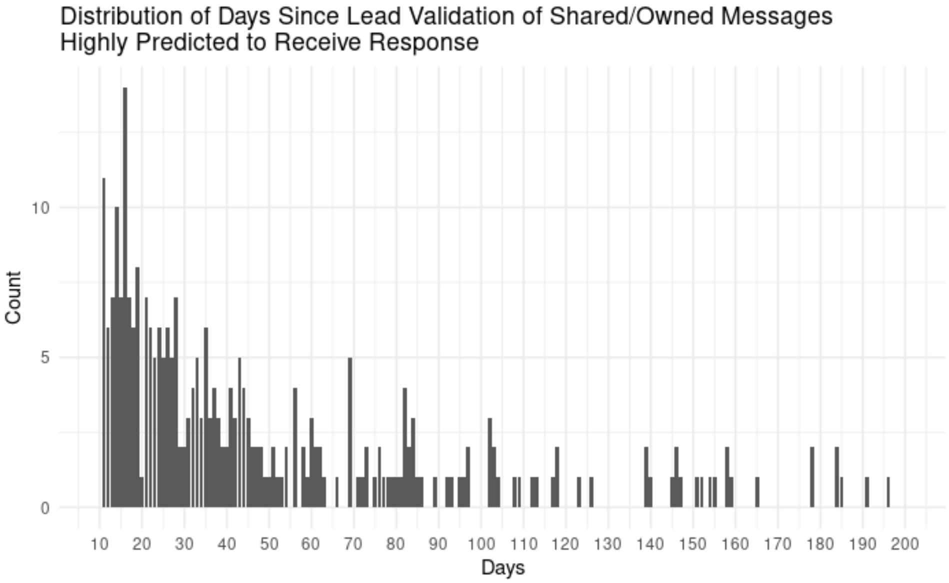

Figure 4

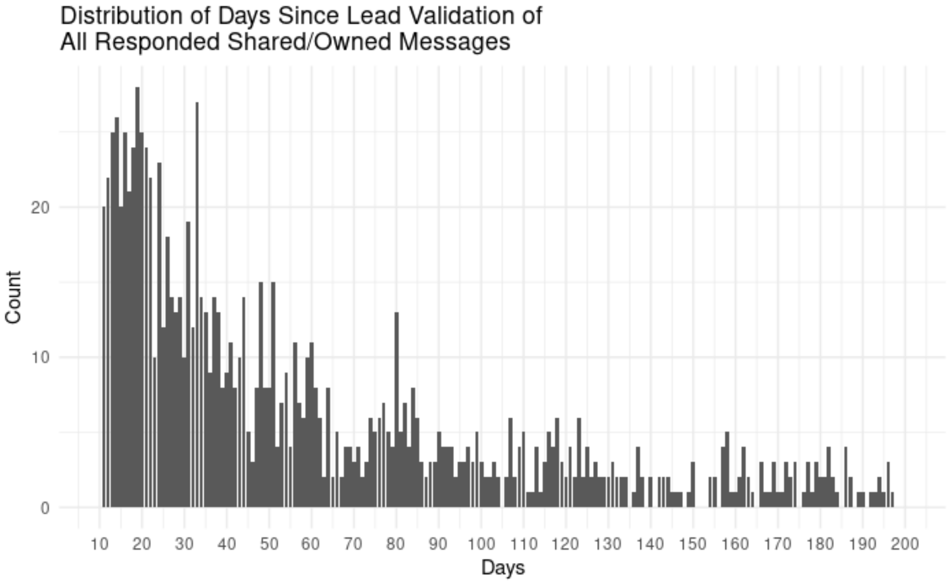

Figure 5

In terms of the best sent hour, most call responses from the overall data set, as well as the highly predicted messages from the model occurred from 8:00am-8:59am MST as seen below in both *Figures 6&7*:

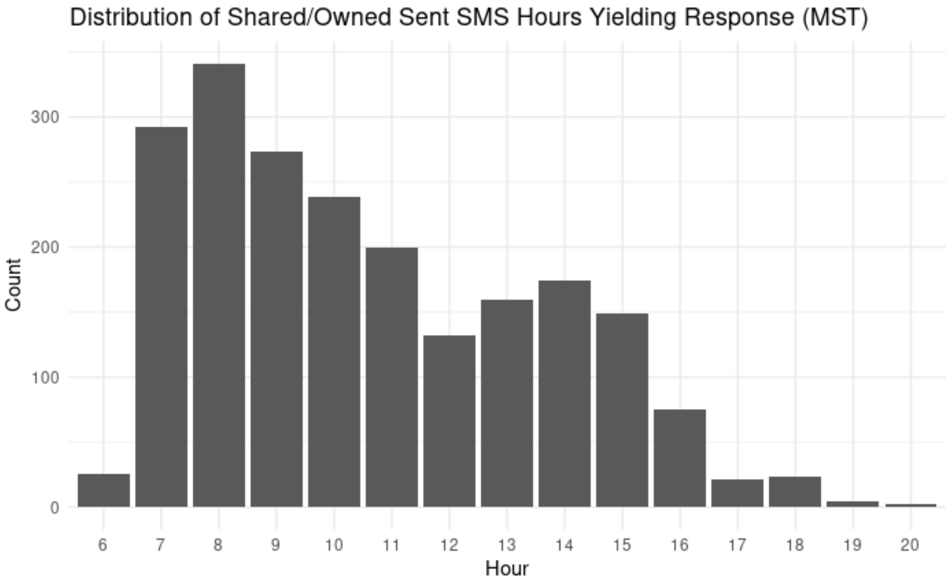

Figure 6

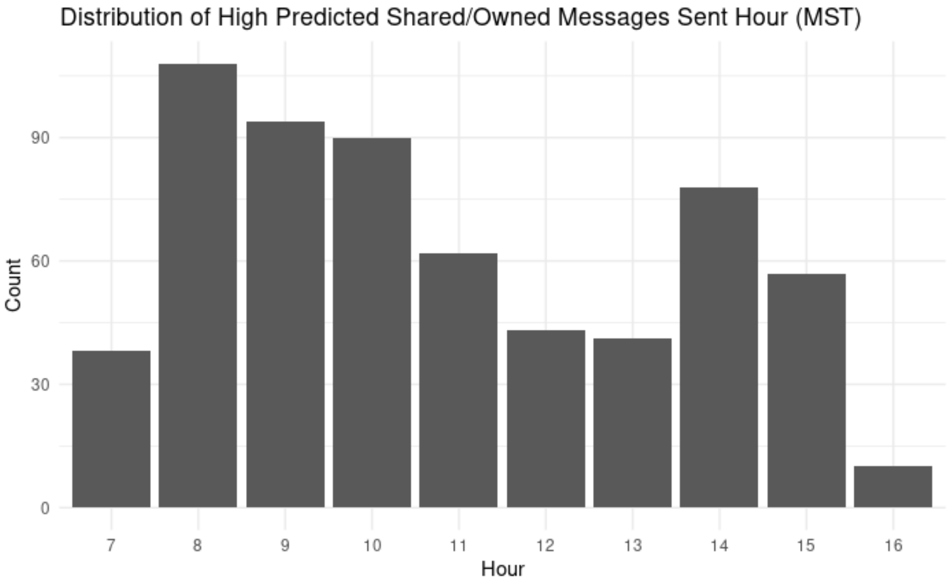

Figure 7

As a reminder, these messages underwent a transformation process, resulting in stripped-down versions. This transformation involved eliminating capitalization at sentence beginnings, anonymizing personal information, expanding slang terms, and removing periods for consistency. For reference, the term "refijetfirstname" represented the first name of the RefiJet salesperson, "phonenumber" signified a phone number exchanged via text, "monthname" indicated a specific month, and "namename" referred to a customer's name. The top three messages identified through this process are detailed below:

1. 25.7% Response Rate (42/163):
   "hi namename! refijetfirstname with refiJet here reaching out about your auto refinance are you still looking to refinance at a lower payment? i can also help you offset your first payment until monthname! we are able to get options to refinance more than one vehicle as well! all we need is your vin and mileage to get started this will only take to minutes my direct line is phonenumber!"

2. 23.1% Response Rate (393/1696):
   "hi namename! refijetfirstname with refiJet here reaching out about your auto refinance are you still looking to refinance at a lower payment? i can also help you offset your first payment until monthname! all we need is your vin and mileage to get started this will only take to minutes my direct line is phonenumber!"

3. 6.5% Response Rate (34/523):
   "hi namename! refijetfirstname with refijet here reaching out about your auto refinance are you still looking to refinance at a lower payment? i could also help you offset your first payment until monthname! all we need is your vin and mileage to get started thanks and have a great day!"

To visualize common phrases and words in messages my model predicted responses for, I produced word clouds, which increase the size of the word in the plot based on its frequency. As seen below, many of the words from the messages mentioned above are plotted in *Figure 8*. I then used all the messages which received a call response from the data set to visualize the most common phrases from all messages receiving a call response in *Figure 9*.

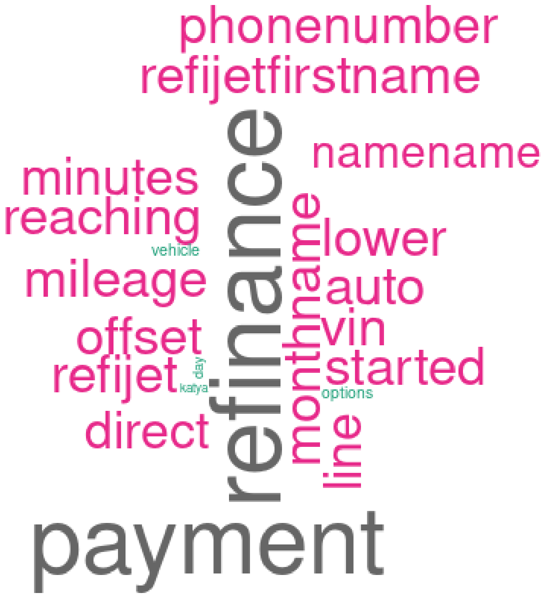

Figure 8

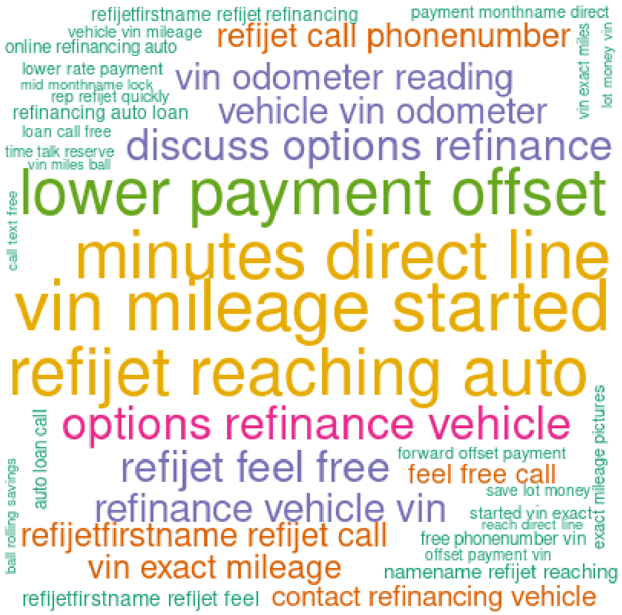

Figure 9

The reason I decided to examine the data set instead of using the highly predicted messages is because the model had lower predictive power in differentiating between messages receiving a call response and not. This led me to use the same modeling techniques but instead create different models for owned and shared leads.

Shared Leads Findings:

After fitting the same types of models to only leads which are "shared," the XGB model's ROC slightly increased to 0.68 with an accuracy of 25.9% with messages with a 70% or greater probability of receiving a response. The variable importance plot for this model can be seen below in *Figure 10* which shares many of the same important variables:

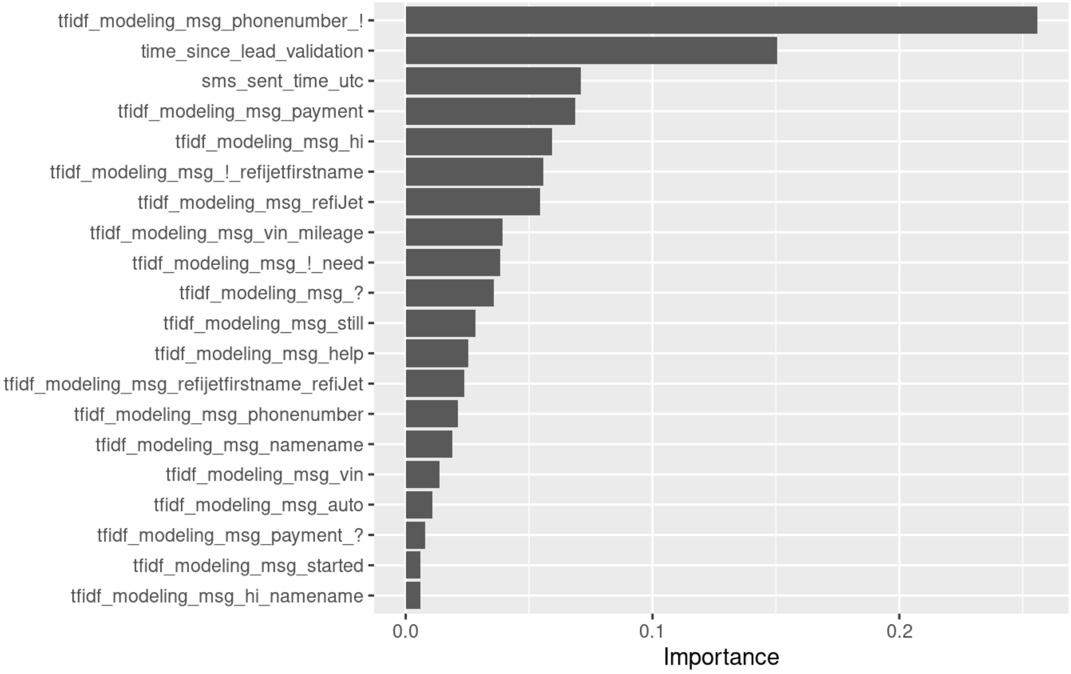

Figure 10

Like before, I broke down the distribution of time since lead validation and message sending hours in the shared model. This breakdown focused on shared lead source messages with a high predicted probability of receiving a call response (*Figure 11)* and all shared lead source messages that did receive a call response *(Figure 12).* From the plots below, most messages which receive a call response lie within 0-60 days since lead validation.

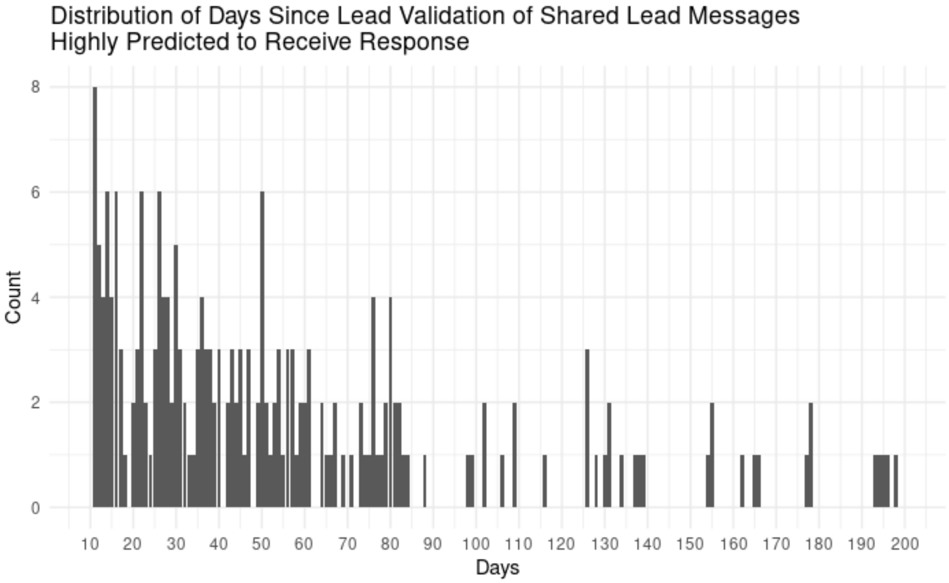

Figure 11

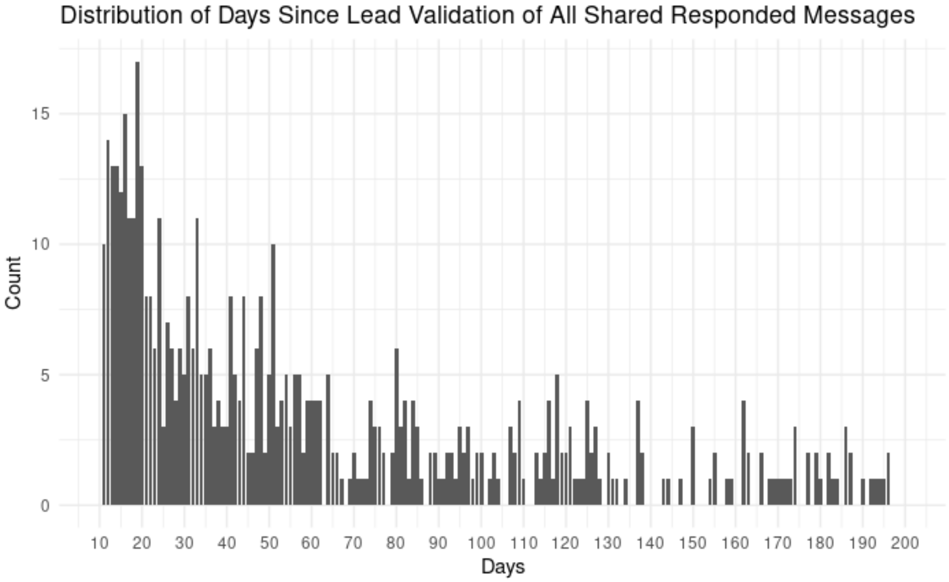

Figure 12

In terms of the best hour sent, there is a slight discrepancy (unlike the aggregate model) between the distributions of sent time for shared leads in the data set and the model. The messages from the data set follow a similar distribution to the aggregate model *(Figure 14),* but the highly predicted messages from the shared lead model are centered with the highest count of responses when SMS messages are sent from 10:00am-10:59am MST *(Figure 13).* In contrast, the most common time for sending a message to an owned lead and receiving a response is from 8:00am-8:59am

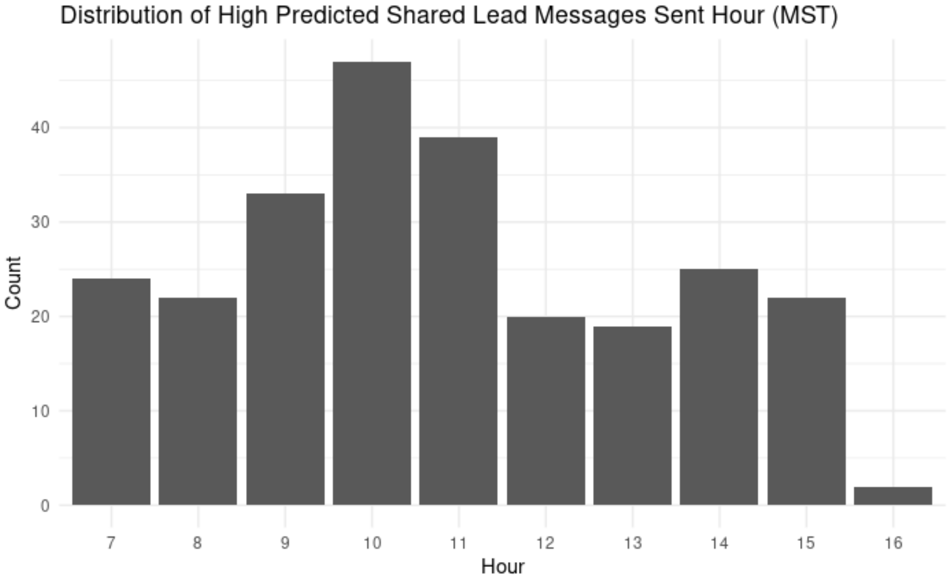

Figure 13

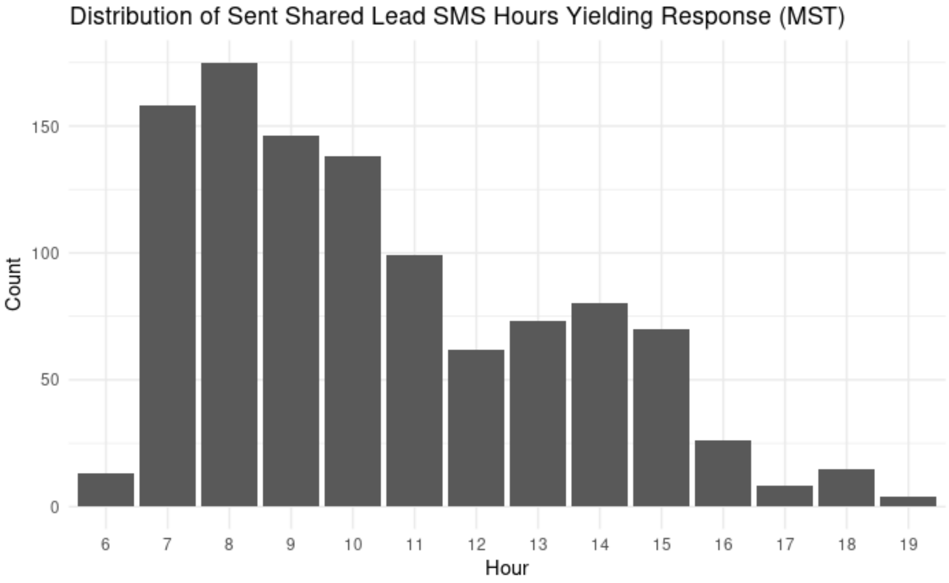

Figure 14

For the shared SMS Model, the top models and their respective conversion rates are as follows:

1. 17.2% Response Rate (15/87):
   "hi namename! refijetfirstname with refiJet here reaching out about your auto refinance are you still looking to refinance at a lower payment? i can also help you offset your first payment until monthname! we are able to get options to refinance more than one vehicle as well! all we need is your vin and mileage to get started this will only take to minutes my direct line is phonenumber!"

2. 16.4% Response Rate (166/1007):
   "hi namename! refijetfirstname with refiJet here reaching out about your auto refinance are you still looking to refinance at a lower payment? i can also help you offset your first payment until monthname! all we need is your vin and mileage to get started this will only take to minutes my direct line is phonenumber!"

3. 10% Response Rate (1/10):
   "hi there! this is refijetfirstname with refiJet we just missed one another! are you still looking to buyout your lease? all we need is your VIN and mileage to get started this will only take to minutes my direct line is phonenumber!"

4. 9.5% Response Rate (7/73):
   "hi there! this is refijetfirstname with refiJet we just missed one another! are you still looking to refinance at a lower rate or payment? i can also help you offset your first payment! all we need is your VIN and mileage to get started this will only take to minutes my direct line is phonenumber!"

5. 7.7% Response Rate (11/143)
   "hi refijetfullname with refijet i have your application for your auto refinance options available please contact me at phonenumber"

I then used word clouds to visualize the most frequently used words in shared lead source messages with a high probability of call response as seen in *Figure 15.* To give a broader perspective, I also plotted a word cloud for the top bigrams from the entire data set of shared leads that received a call response in *Figure 16 below:*

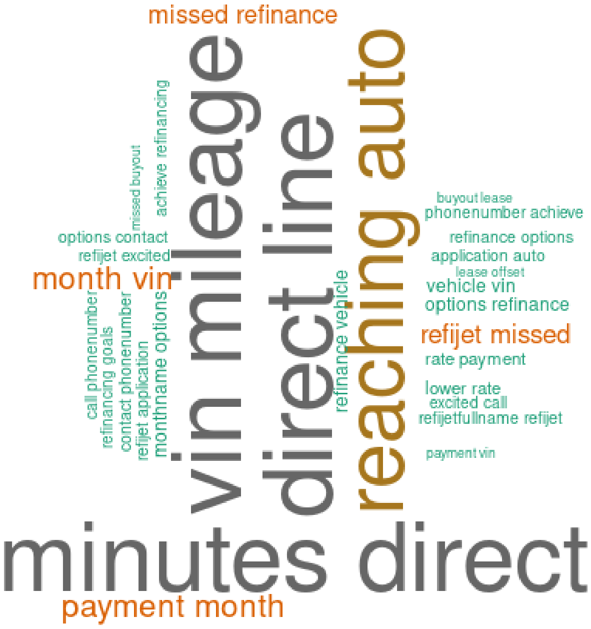

Figure 15

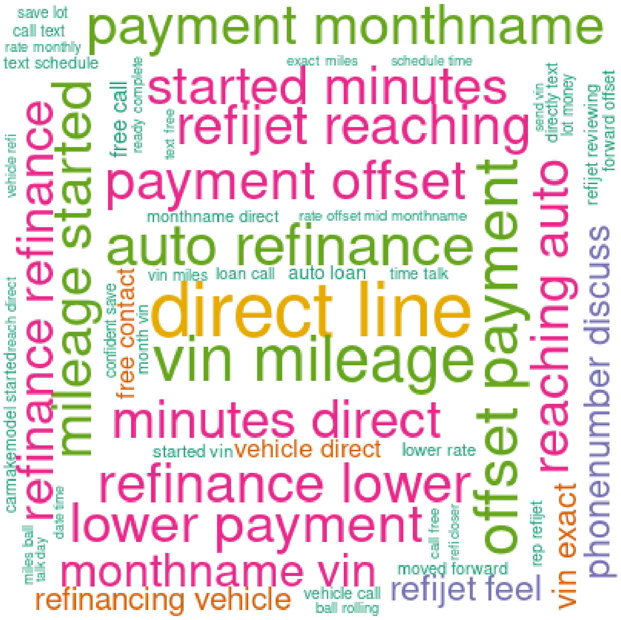

Figure 16

<u>Owned Lead Findings</u>:

After fitting the same types of models to only leads which are "owned," the XGB model's ROC slightly increased to 0.69 with an accuracy of 35.8% with messages with a 65% or greater probability of receiving a response. The variable importance plot for this model can be seen below in *Figure 17* which shares many of the same important variables:

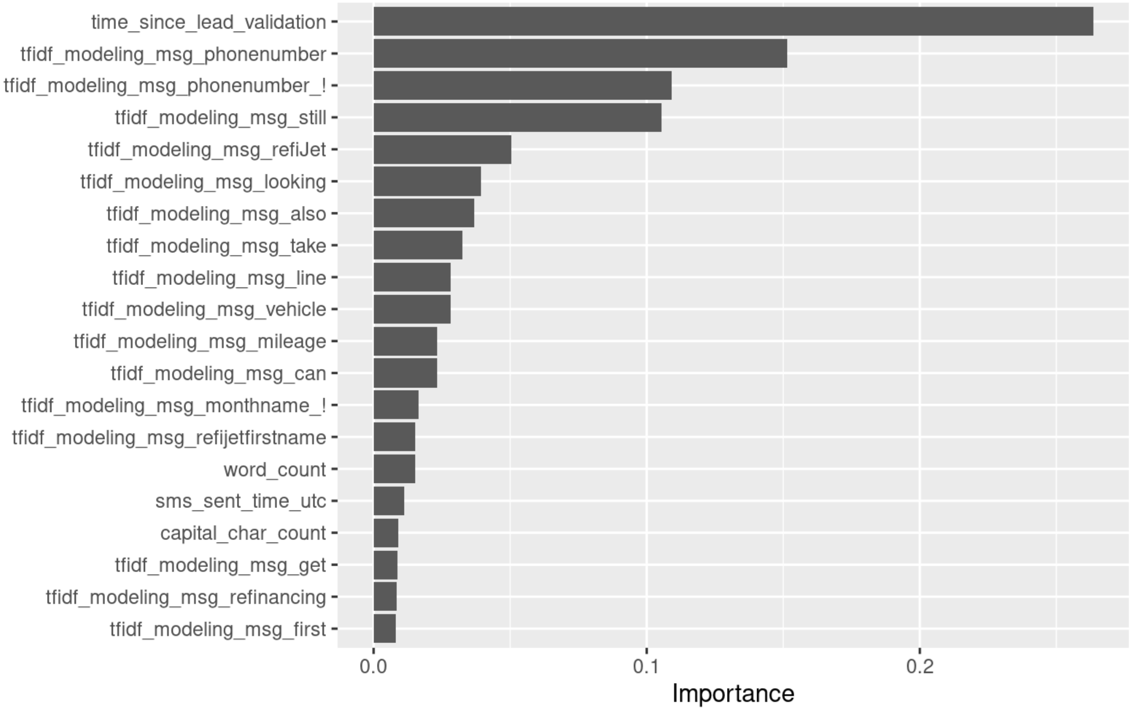

Figure 17

Again, I broke down the distribution of time since lead validation and message sending hours in the owned model. This breakdown focused on owned lead source messages with a high predicted probability of receiving a call response (*Figure 19)* and all shared lead source messages that did receive a call response *(Figure 18).* From the plots below, most owned lead source messages which receive a call response lie within 0-40 days since lead validation.

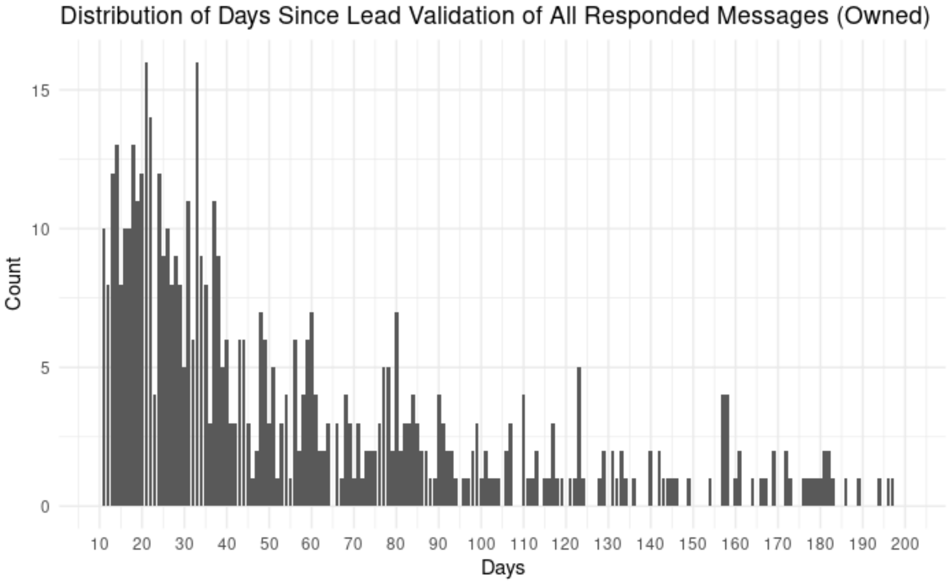

Figure 18

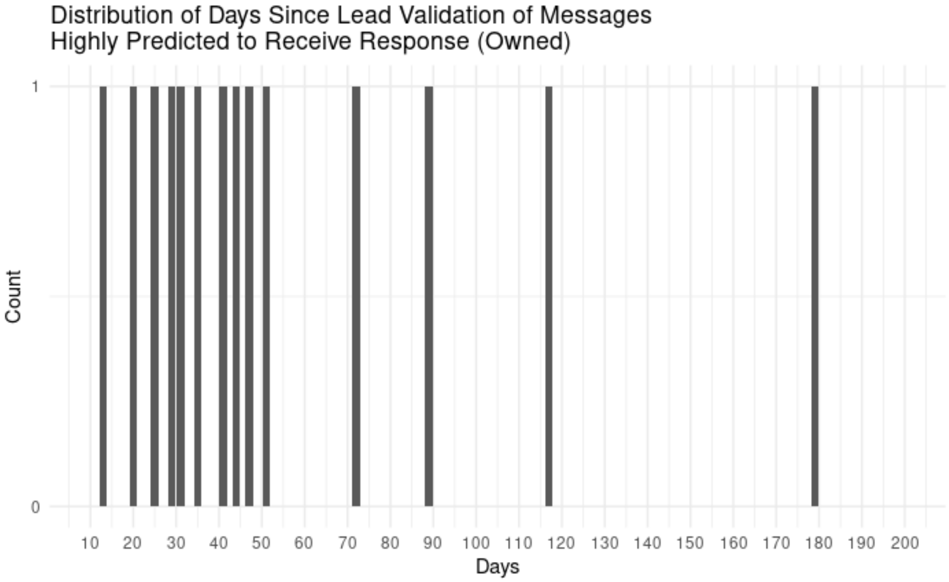

Figure 19

Analyzing the best hour sent, there is a slight discrepancy (unlike the aggregate model) between the distributions of sent time for owned leads in the data set and the owned model. The messages from the data set follow a similar distribution to the aggregate model, but the highly predicted messages from the owned lead model have no sent hours between 8:00am-8:59am *(Figure 20),* while for all owned leads which receive a response, the most common sent hour is from 8:00am-8:59am *(Figure 21).*

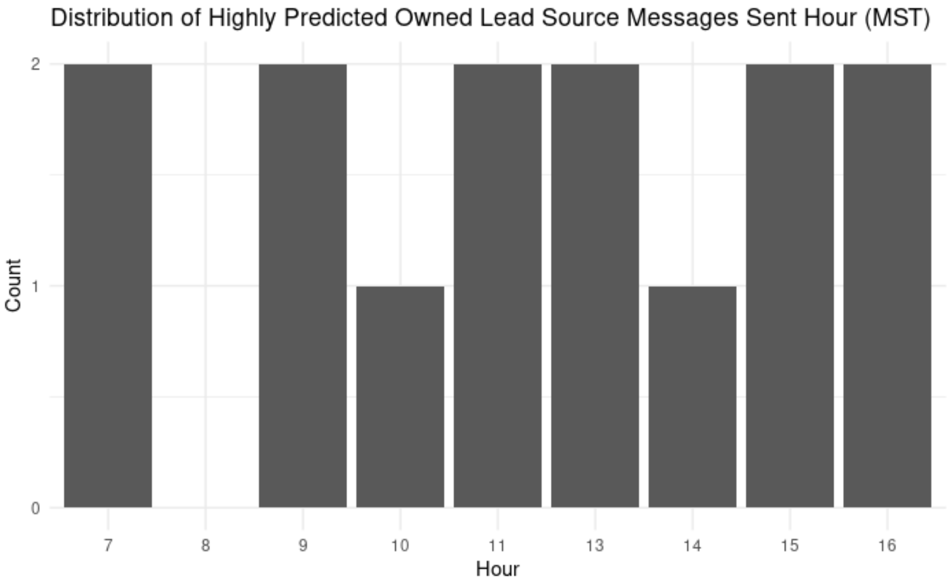

Figure 20

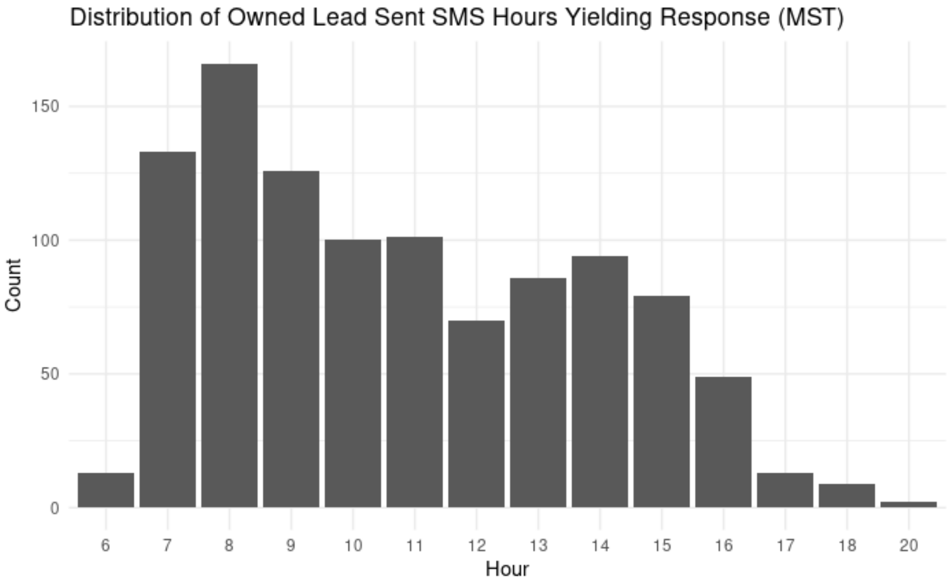

Figure 21

For the owned SMS Model, the top models and their respective conversion rates are as follows:

1. 92.8% Response Rate (13/14):
   " hi this is refijetfirstname with refijet following up my direct line is phonenumber our company direct line is phonenumber "

2. 31.2% Response Rate (168/539):
   "hi namename! refijetfirstname with refiJet here reaching out about your auto refinance are you still looking to refinance at a lower payment? i can also help you offset your first payment until monthname! all we need is your vin and mileage to get started this will only take to minutes my direct line is phonenumber!"

3. 20% Response Rate (3/15):
   "hi there! this is refijetfirstname from refijet i am excited to work with you please give me a call at phonenumber to achieve your refinancing goals"

4. 11.8% Response Rate (2/17):
   "hi there! this is refijetfirstname from refijet i am excited to work with you please give me a call at your convenience at phonenumber to help me achieve your refinance goals"

I then visualized the most frequent phrases of the messages predicted to have a high probability of receiving a call response (*Figure 22*), as well as all the owned lead source messages which received a call response (*Figure 23*):

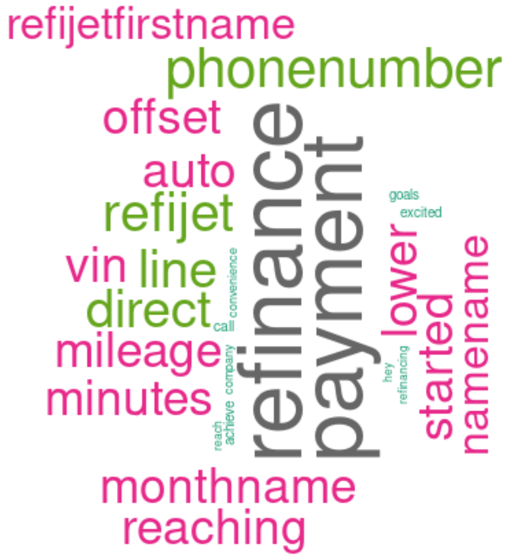

Figure 22

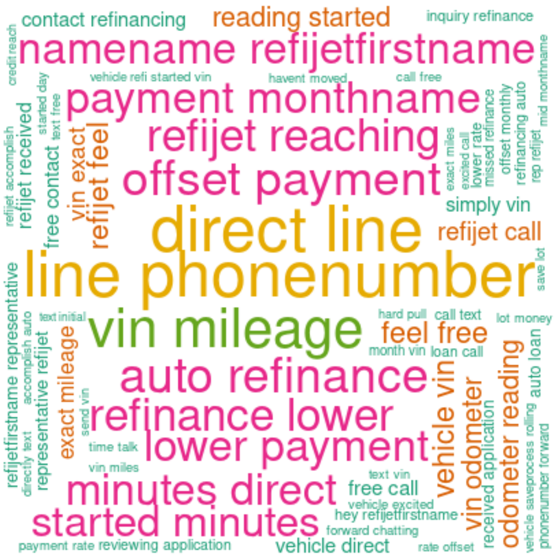
contact refinancing
reading started
inquiry refinance
namename refijetfirstname
credit reach
vehicle refi started vin
havent moved
call free
payment monthname
refijet reaching
offset payment
refijet accomplish
refijet received
started day
text free
free contact
vin exact
refijet feel
exact miles
excited call
lower rate
missed refinance
offset monthly
refinancing auto
rep refijet
mid monthname
simply vin
direct line
refijet call
line phonenumber
save lot
refijetfirstname representative
accomplish auto
representative refijet
directly text
text initial
exact mileage
send vin
vin mileage
hard pull
call text
lot money
feel free
month vin
loan call
auto refinance
refinance lower
lower payment
time talk
vin miles
vehicle vin
vin odometer
received application
odometer reading
auto loan
vehicle saveprocess rolling
phonenumber forward
text vin
free call
minutes direct
vehicle excited
hey refijetfirstname
started minutes
forward chatting
payment rate
reviewing application
vehicle direct
rate offset

Figure 23

Conclusion:

This study observed 33,355 initial SMS messages sent from RefiJet's sales team to prospective leads from 1/1/24-3/1/24. From those messages, 2,110 resulted in call response, giving a 6.3% response rate. My models were able to identify SMS messages with three times or greater response rates. Successful messages exhibit consistent patterns in their content structure, offering valuable insights for optimizing SMS communication strategies. These messages are characterized by a well-defined structure, commencing with a personalized greeting (e.g., "hi namename!"), introducing the RefiJet representative ("refijetfirstname"), and highlighting RefiJet's association. Notably, the emphasis is placed on the purpose of the message, often centered around exploring auto refinancing options. In terms of offer inclusion, successful messages frequently incorporate incentives, such as the opportunity to refinance at a lower payment or offset the first payment until a specific month. The content is clear and concise, outlining the process and requesting essential information like VIN and mileage. Also, these messages feature a direct call to action, inviting the lead to respond or contact the RefiJet representative directly ("direct line"). Contact details, including the direct line ("phonenumber"), are provided for seamless communication. Personalization proves to be a crucial component, with mentions of the lead's name ("namename") and tailored information based on the lead's situation. The use of "refijetfirstname" adds a personalized touch to the communication. Analyzing time and frequency patterns reveals that messages receiving a call response are most concentrated within 0-60 days since lead validation. Also, the timing of messages, especially from 8:00am to 8:59am MST, emerges as a critical factor influencing call responses.

While the distribution of message sent times for the overall lead model aligns with the pattern observed in all messages receiving a call response, this consistency diminishes when constructing models based on lead source categories (shared vs. owned). Notably, in the shared lead source model, the most prevalent sent hour, as indicated by the model, deviates, and occurs from 10:00am to 10:59am, while the entire dataset of shared leads receiving a response is concentrated from 8:00am to 8:59am. It is important to highlight that RefiJet serves customers across various time zones, and the message sent time represents MST from Denver, not the localized time for the customer.