**Executive Summary for Romob CEO**

**By: 251193433**

Overview of the Problem

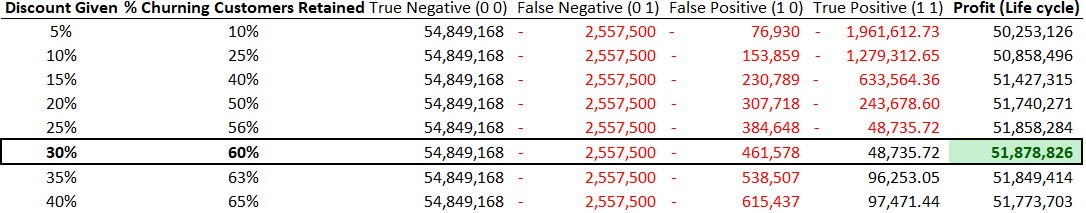
There is a big problem of customer churn at Romob and in the past we have always taken a reactive approach to this serious problem. As a start-up we have now reached a place where it is imperative to put forth better risk management practices and remedy the problem before it even begins. Many of our biggest competitors already employ predictive models to try and retain customers before they even become disgruntled and unsatisfied with our company. One such solution lies in identifying high-risk customers by looking at historical trends to determine when customers will become unhappy and will want to switch to a different carrier. At Romob, each customer churned costs us immensely as attracting a new customer can be more than 50 times as expensive as merely retaining a customer. The customer goes through three phases with our company throughout their life cycle. First, they are in the happy phase where they are satisfied with our service. Next, comes the sore phase where the customer becomes disillusioned by our service which can be caused by a myriad of reasons such as lower competitors’ rates, perceived unjust charges, and poor service quality. Lastly, the customer reaches the churn phase where the customer leaves for a competitor. Using the data provided for the past 3 months, I have created a model by classifying the first two months as a customer’s happy phase and the 3rd month as a customer entering the sore phase. Based on this, I created a model which predicts which customers will enter the churn phase next month. By using this model, we will be able to identify and prevent customers from churning. Additionally, we can determine what the appropriate discount given to unsatisfied customers should be to retain them for a longer period of time and maximize our profits.

High-Level Methodology

In simple terms, the methodology to build this model utilized two statistical methods, logistic regression as well as Naïve Bayes. Both methods were used in order to determine which one of them gave a more powerful and accurate predictive model. Many different variables were inputted into the model such as roaming minutes, network minutes, amount of data used, and how long the customer had been with us, among others. Next, I was able to discern which variables had the highest predictive capacity for customers churning. After running both logistic regression and naïve bayes techniques, I found logistic regression to produce a model which produced 93% accuracy in predicting the churn of customers. Combining this model with historical information on how discount percentages led to a certain percentage of churning customers retained enabled me to produce a table which identified what discount percentage we should offer to our sore phase customers in order to retain them.

Recommendation for the CEO

The model reveals to us that a discount of 30% given this month will maximize our profits in the long-run and lead to 60% less customers churned next month. These findings are summarized below:



Based on this information, we should offer a 30% discount immediately to all customers identified by the model as being in the sore stage this month.

**Executive Summary for Romob VP – John Wilson**

**By: 251193433**

Quantitative Problems

There are a few main questions which our company must get answers to be more proactive about this issue of customer churn. We must determine which customers should get the promotional discount so that we can target the appropriate customers and spend our limited discount budget wisely. We must also determine if the 20% marketing plan is too aggressive. Additionally, we must determine if our deep discount approach will be beneficial to the company financially in the long run. Lastly, we must determine what the risks associated with our methodology are in preventing customer churn. To be proactive in determining customers churning ahead of time, we need to implement logistic regression and naïve bayes. The reason for choosing these two statistical tools over others is because when we look at the data collected, it is clear that the churn variable will be binary and it will be impacted by several different factors. This kind of data gives credence to the idea of using logistic regression to uncover insights about the data and make predictions. Naïve bayes is another technique we should employ to observe the relationship churn has with the other customer data we have collected. There is sufficient data with over 30,000 rows which bodes well for the naïve bayes methodology since a larger database will allow us to create a large sample of data with sufficient data points.

Quantitative Methodology – Logistic Regression

For the logistic regression, the first thing I did was determine which variables I wanted to run the regression on. I realized that the 6th and 7th months for all the variables were irrelevant as predictive variables since the customers were in a happy state during these months. As a result, I removed these variables and only the 8th month variables were fed into the logistic regression model as it was during this month that customers started to get sour with the company and were on the cusp of churning the following month. Additionally, the age of customer was included as a relevant variable because how long a customer is with our company indicates how loyal the customer is to our product offering and the longer the customer is with us the more likely it is that they are satisfied with our service and have a lower desire to churn. After packing these variables into the logistic regression model, the summary statistics were ran to determine the impact of each of these variables on churn. The results of this are displayed in Exhibit 1. From Exhibit 1, all variables which had p-values which were unreasonably high (over 10%) were promptly removed from consideration of having a significant impact on customers churning. Taking just the variables which were deemed significant, summary statistics were again run as displayed in Exhibit 2.

As Exhibit 2 shows, this produced a much more compact list of variables which were relevant in predicting whether a customer would churn or not. The variables which are significant are: age of customer, network minutes of usage, off-network minutes of usage, roaming outgoing minutes, total outgoing minutes, local incoming minutes, inter-provisional incoming minutes, international incoming minutes, volume of 4g data used, and volume of 3g data used. From the coefficients of the variables, we can see that for almost all these services, the more a customer stays with us or uses our services, the less likely (negative values) they are to leave. The only two exceptions to this rule are for the roaming and total outgoing minutes. It appears our roaming outgoing phone service needs improvement or more competitive plans as customers who use this service immensely are churning at a higher rate.

Now with the relevant variables fed into the model, a confusion matrix and ROC curve was devised to determine the predictive power of this regression (Exhibit 3).

From the ROC curve in Exhibit 3, we can observe that the model had a high area under the curve of 0.87 which bodes well in telling us that the model was highly accurate. This is further confirmed by running the confusion matrix at a 0.35 threshold which gives us the confusion matrix found in Exhibit 3.

This confusion matrix has a high accuracy of 93% with 97% sensitivity and 0.49% specificity. This suggested that the logistic regression model has a high accuracy, but it has a weak point in false negatives, although this number is miniscule compared to the amount of correct predictions of the model.

Quantitative Methodology – Naïve Bayes

For the naïve bayes methodology, only the significant 13 variables from logistic regression were inputted and the model was ran with a 60% sizeable sample. This model gave a slightly less accurate result as show in Exhibit 4.

The naïve bayes model was less accurate at only 68% accuracy and although it was better than the logistic model at having fewer false negatives, it had much higher false positives which meant that our discount pricing efforts would mostly be wasted on customers who are not entering into a churn state next month anyways. This would lead to a loss of profits and a deeply ineffective discounting campaign which is why the logistic regression model is a better choice for predicting churn for customers for this particular case.

Recommendation

Using the logistic regression model, along with the discount and customer retention table, we can determine what the appropriate discount to give to customers is. This can be done by using the confusion matrix table to calculate the total profit over the life cycle of current customers. From the table displayed in the executive summary to the CEO, a discount of 30% is optimal in maximizing profits as it retains 60% of our soon-to-be-churned customers. The formulas used to calculate this are as follows:

True negative (0 0) = true negative # of customers \* average monthly gross margin \* 60 months life cycle 🡪 this gives us the profits for customers staying over their life cycle.

False negative (0 1) = false negative # of customers \* average monthly gross margin \* 60 months life cycle 🡪 this gives us the profits lost for customers we were unable to detect churning using our model.

False positive (1 0) = false positive # of customers \* discount % \* average monthly gross margin \* 60 months life cycle 🡪 this gives us the profits lost as a result of giving a discount to customers which were not going to churn.

True positive (1 1) = true positive # of customers \* customers retained \* average monthly gross margin \* customer life cycle \* (1-discount given) – (true positive # of customers \* customer not retained \* average monthly gross margin \* life cycle of customer 🡪 this calculated the net difference between the profits from customers we were able to retain as a result of the predictive power of this model and the profits which were lost in the future as a result of churned customers who left despite the discounted price offering.

For all these formulas the average monthly revenue per customer is $62 with gross margin of 55%; life cycle is 60 months (5 years). From this analysis we can conclude that giving a discount of 30% and using this predictive model will increase our profits by over $51.87m over the life cycle of current customers. The risk associated with this prediction is that customers can still churn at variable rates over time and their likelihood of retainment will not remain static at current discount levels so this model must be run every month to give updated predictions to management.

Exhibit 1:

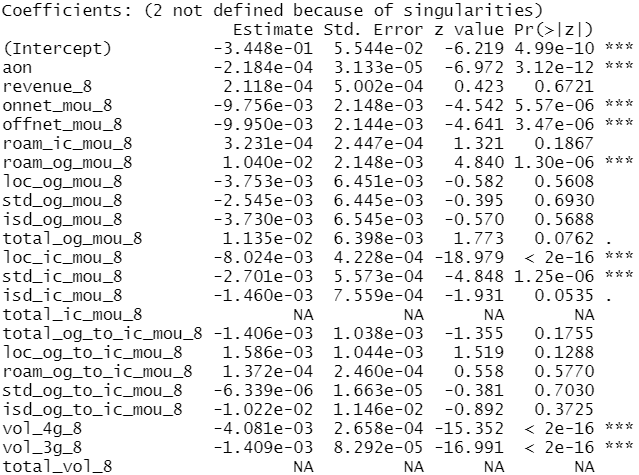


Exhibit 2:

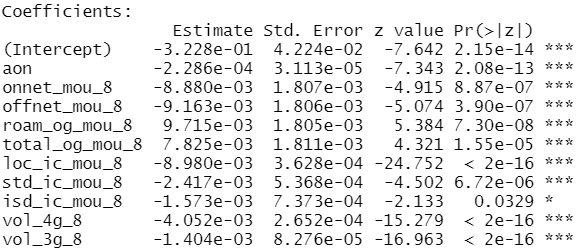


Exhibit 3:

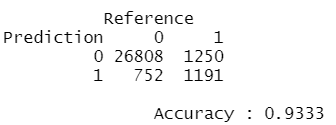
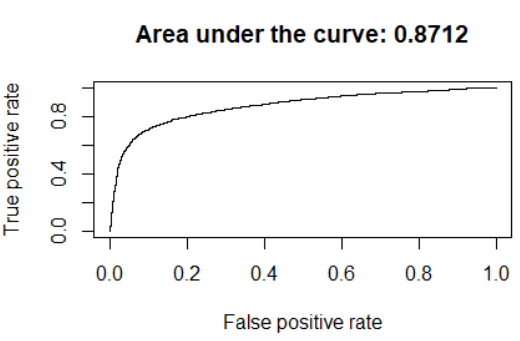


Exhibit 4:

