**Matthew DeGiacomo**

**DAT- 690**

**Capstone Final- Customer Churn**

**1/5/2020**

**Introduction**

The business problem we are dealing with at GE is that their customers are churning at high rate within their Health based subscription app and GE doesn’t know why, nor do they have the information to anticipate which customer will churn next. The purpose of the research is to create a predictive analytical model which shows GE which customers are more likely to churn than others and at what stage, this way GE can focus their efforts on those customers who need more attention. Informing the stakeholders and their needs for data analytics starts with what’s important to them, for example, wanting to know why that business unit is losing money, how we are going to fix it, and what future business impact with data analytics have within GE. The needs can be assessed by explaining the churn situation and how using data analytics would help cut down on churn, improve efficiencies across multiple business units like retention and sales within the healthcare app team, and help us gain customer knowledge by having more meaningful phone calls. My central concerns for this project will be within the dataset, dealing with missing data, and trimming down the dataset to get us the most accurate model possible.

**Pilot Evaluation**

Throughout the pilot plan I ran into many challenges and some of them I successfully overcame throughout the class with essential feedback from my classmates as well as my professor. My biggest challenges in the pilot run came from the data set its self and learning that I did not do enough data cleansing, data preparation, and making constant model adjustments in the beginning with my data and when I went to model the first time my categorical data had to many unique variables to run the random forest in R. The dataset had a lot of missing data as well as unneeded variables which created unnecessary noise and in turn made it difficult to model at first. I would say my success from the initial pilot run in DAT 650 came from being able to adjust when a problem arrived, when I couldn’t fit the data into a Random Forest, I read in a case study that used a Support Vector Machine that scored well, so I ran that. I ran a correlation analysis to see which the most important variables to the outcome were to trim down the data. Looking back at my initial pilot results from my SVM model’s error matrix (\*1), I never even handled the missing data as the matrix shows the NA in it. This model was made without much cleansing, the error rate may look nice, but it was setup that way and is not a stable or reproducible model. For the full implementation results I chose the route of the random forest and through feedback, cleansing, PCA, variable selection and many trials I got to a model that we can use moving forward (\*2). With full implantation the ROI can be broken down as so, if GE loses a customer due to churn it’s a loss of $1,000 spent over the lifetime of that customer, if we have 41 customers churning per the confusion matrix that equals $41,000 GE loses per year. If GE didn’t have this model and tried to replace all the lost customers with new ones, it takes 3 times more work and resources to get a new customer, so we can say that trying to replace that 41,000 with 41 new customers will cost GE $123,000. This model predicts churn at a 78% accuracy for churned customers (32/41 true positives), If we can flip 25% of customers to not churn that’s about $30,750 GE get back per year and with a steady increase of 5% or two customers GE got to stay per year, in 5 years we get back 50% of customers and $61,500 GE saves.

**Plan Modification**

Throughout data analytics you are always looking to make modifications to your plan or model to keep making it the best model or project it can be. I chose my original target variable CHURDEP that I used in DAT 650 after testing out both that and CHURN as my target. I have made many adjustments since my first data analytic plan and created two flow charts to visualize my revised plan for the data, flow chart 1 (\*3) shows the prepping of the data before the modeling starts while flow chart two (\*4) shows modeling phase, the constant adjustments, and how we evaluate the model. You can state flow chart 1 as phase 1 of the data analytic plan for GE, flow chart two as phase 2 of the data analytic plan, with phase 3 being full implementation of the results. Each block of the flow chart represents a task that needs to get done then breaks out the smaller tasks that are needed to complete overall task. I project each task on the flow chart to take about 4-6 weeks, some tasks get done quicker than others and some tasks take longer than other, with phase 3 taking about another 4 weeks. Each block in the flow charts is projected to be about 1 week and that’s what we will relay to stakeholders as this is based on rolling this project out in this class, in which we are using a 20-week period to complete the project. Strategies I want to implement for professional and effective collaboration throughout this process is first to have bi weekly meetings with the stakeholders to report on progress, where we are on the timeline, delays and reason for them, and what phase is next. This will create a line of communication where the stakeholders have an idea of the state of the project, where we are heading next, and gives them a chance to ask questions that they have.

**Plan Implementation and Results**

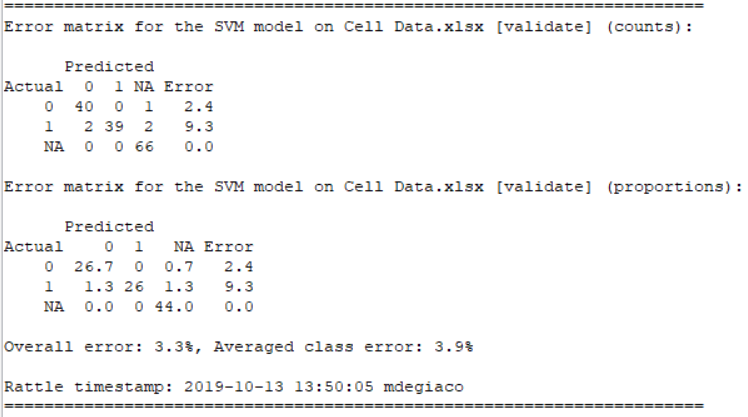
After many trial runs, modifications, and constant adjustments, I have gotten to a model I deem accurate in helping us figure out our business problem of not being able to predict customer churn. I have chosen a random forest model for my final model, after running my data though a decision tree, boost model, and regression model, the Random Forest became the most accurate model out of the bunch. I ended up excluding the following variables from my final model: DROPVICE, CALLFWDV, CHRUN, CSA, OCCPROF, OCCSTUD, OCCHMKR, OCCRET, OCCSELF, AND SETPR from my PCA and Variables importance tab from constantly running my random forest models. One of main adjustments I made for my model is upping the variables included in my Random Forest model, the default has us use 8 and I bumped it up from 8 🡪 14, then kept going to find my sweet spot of 18. My confusion matrix (\*2) shows my model is 62.8% accurate in predicting both people who will churn and who wont churn but more importantly 76% accurate true positive rate meaning people who did churn. Out of 41 people who churned in we predicted 31/41 and for people not churning it was at the 50% mark with a 20/20 split between the true negative rate and false negative rate. Since the goal was to predict churn, modeling at 76% is much better than random selection at 50%. My AUC sits at a .63 (\*5), not a perfect angle above the 50% mark but hitting 13% above random selection for the model will already give GE an advantage to what they had prior but looking at my OOB error rate, which is an unbiased estimate of error that when the resulting model is applied to new observations my error rate is 21.71% meaning 78.29% accuracy. Evaluating my model from a risk chart perspective (\*6), my strike rate is 51% meaning that 51% of the customers are of interest to churn, which is a bit concerning for GE but good info to have. When looking at the green performance line, this is used to achieve performance when using the model to prioritize churn, at 80% of the caseload we get 90% performance. This is useful because the retention team can’t call everyone so we are now narrowing down who is most likely to churn, we know that now we don’t have too call everybody to get the best performance which will help GE become more efficient. Now heading to the lift chart (\*7), we have a massive spike at the 20% mark, what that means is we can cover a little over 1.4 times the number of attritors (churn) by selecting only 20% of customers based on the model vs 20% in random selection. This proves just one more way this model will help GE work smarter, data driven, and more efficient. **(include scores)**

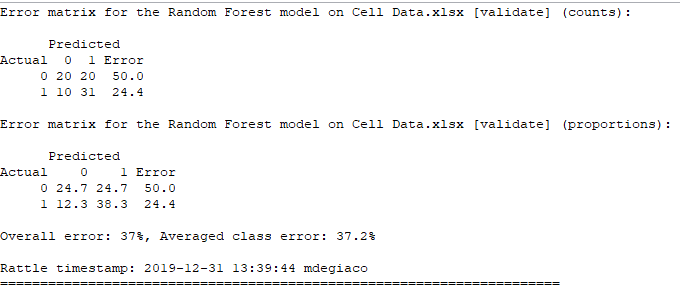
**Conclusions and Implications**

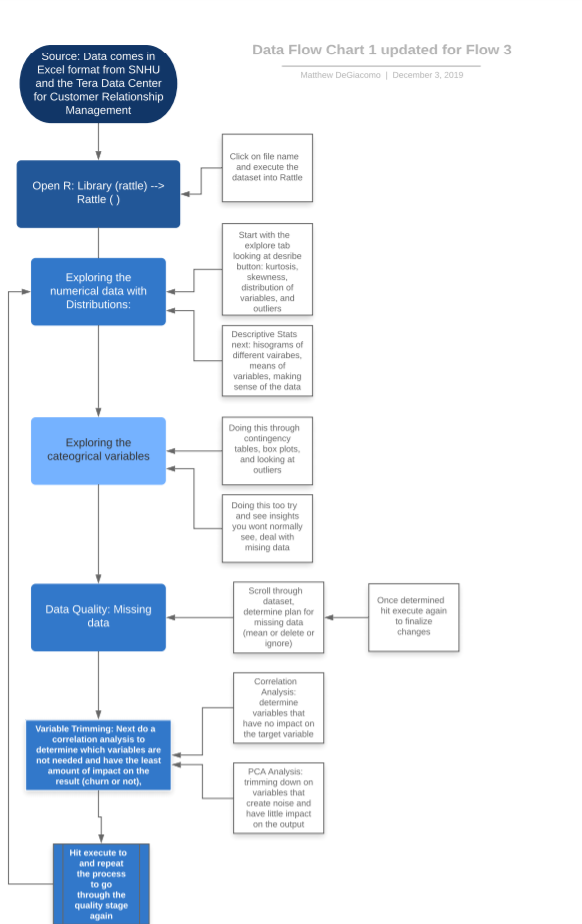
Summarizing the conclusions from my full implementation of the model, we have now provided GE with a model that predicts churn at a 76% accuracy for true positives, they now have the ammunition to be able to get a clear more data driven idea of when a customer will churn, who is next, and a threshold for max performance for more efficient retention calls when dealing with the high volume of calls. GE now has a competitive advantage with the ability to learn more about their customers faster and the ability to turn that knowledge into action. That is a massive impact for GE where none of this was possible beforehand. Another major impact comes from the cost of the problem, it’s been stated that it takes 3-4 times more effort and resources to get a new customer to replace on that has churned. When customers were churning at a high rate for GE and they were attempting to get new ones I predicted it would cost them over $120,000 per 40 churned customers where on the flipside a 2% increase in customer retention is equivalent to a 10% reduction in costs. This can impact GE as they make smarter more efficient calls to the right customers, the more customer knowledge they have on why a customer churned the improvements can be made, they can set indicators based off the model so the retention team can get alerts when a new customer is in jeopardy of churning. The impact this project’s results have on future projects in GE as this will set the baseline of what a data analytic solution can for a company or department which will motivate the rest of the business units to follow suit because customer churn happens to every business that has customers and having a successful model to help prevent churn with success compounding over time will save the company upwards of 20% of costs if GE hits my 5% projection of increased customer retention. One method for future prevention as well is to have tier indicators of points in time of the customer life cycle where churn happens most, score each customer in terms of probability of churn, and have the retention focus on the high probability customers first.

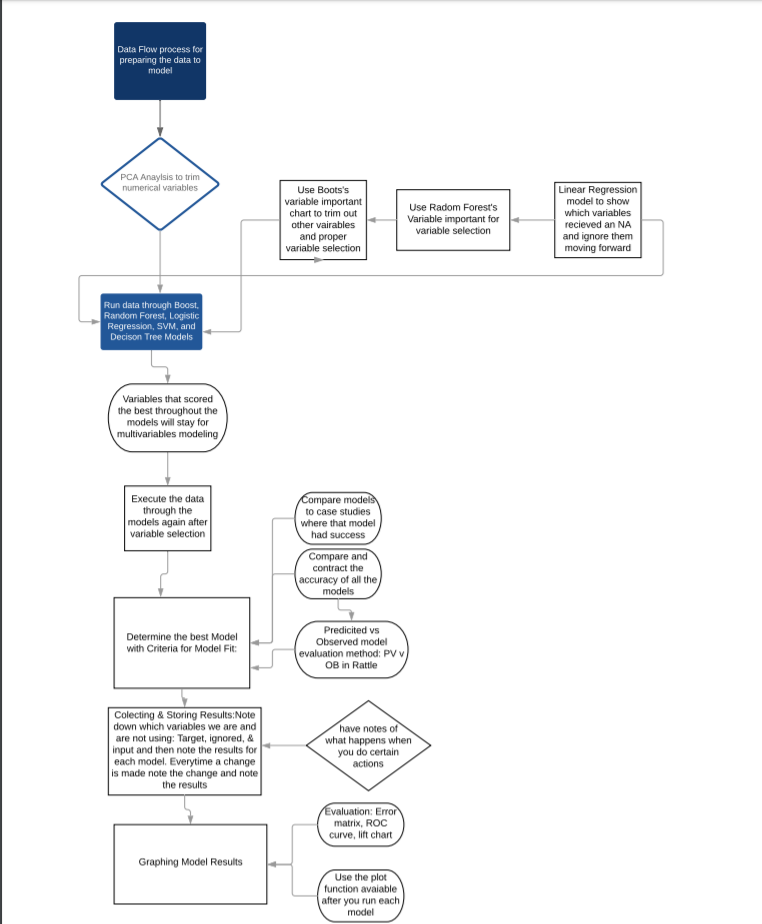
**Recommendations**

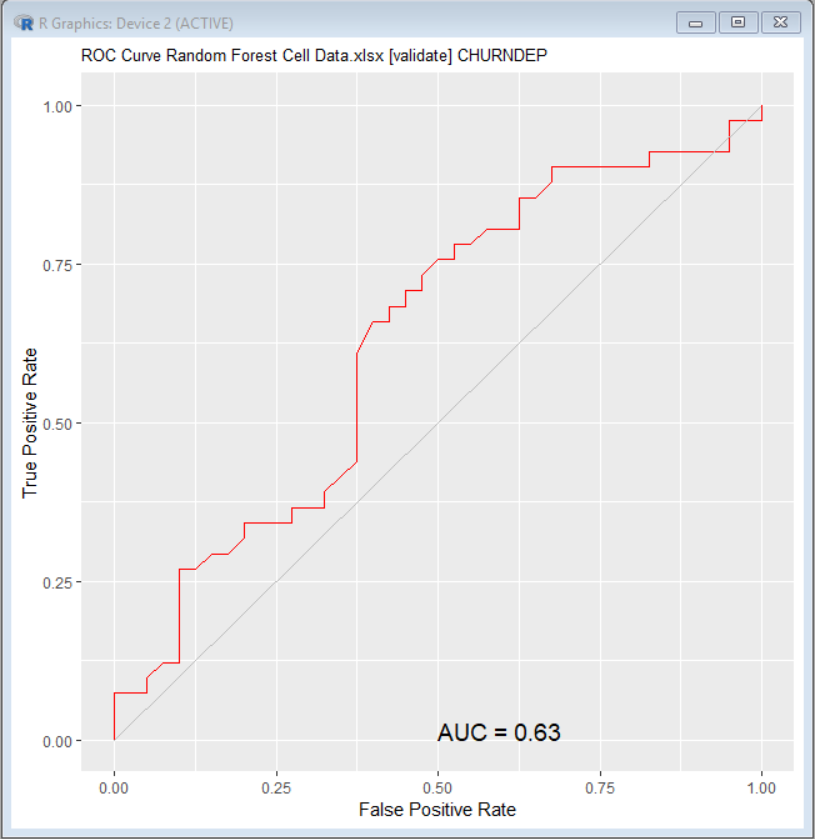
Moving forward we can use this model for a couple different alternative uses, for example for flipping our binary options, we were looking for customers who churned in this project where can also do the opposite and we can use that model with slightly different adjustments in the variable selection part to find out which customers are far from churning to gather information on why they’re satisfied with the app. GE can use this model to their advantage and apply the same principals to an employee churn model to help cut down on GE employee’s churning. This is valid use of this project because just as customers leave so do employee’s, so being able to retain employee’s is just as valuable and the alterations that would need to be made start with variable selection. The same variables that I chose to keep/ignore in my current model won’t be the same, but the process would be very similar in my steps to how I got there but I would need different employee data but overall each data set is different and will provides different problems for the analyst.

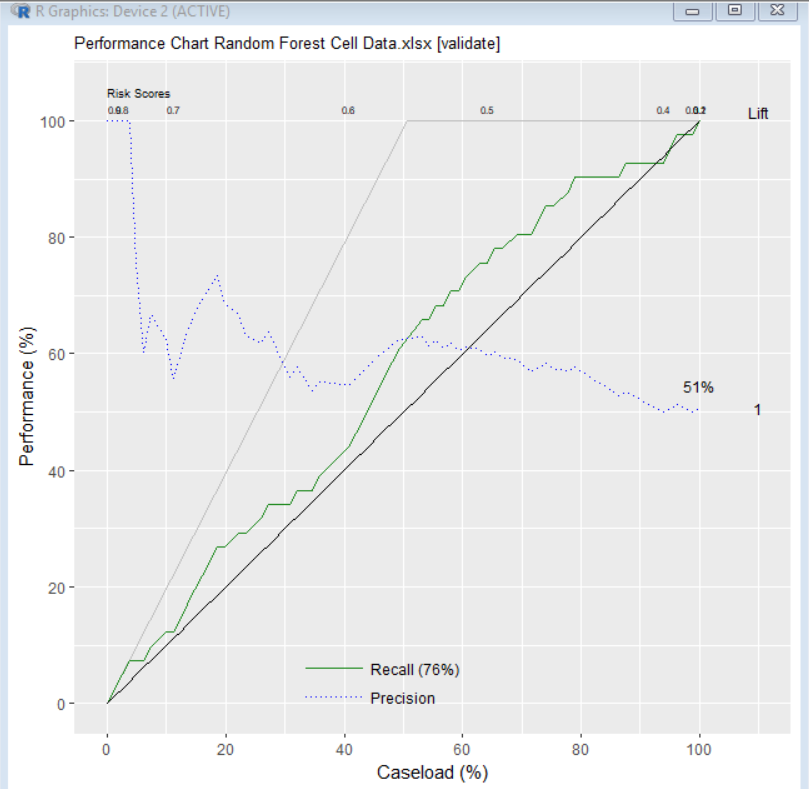
**(\*1) **

**(\*2)** 

**(3\*)** 

**(\*4)** 

**(\*5)** 

(\*6) 

(\*7) 