MIS-64036: Business Analytics
Assignment II

Date Graded 8 Dec 2019

Final Grade: 98/100

Group 4, Sumanth Avunuri, Krystn Hood, Chujun Huang, Alfred Rajan , Sushmita Singh

Thanks for your submission. Please find below my feedback with regard to your assignment submission.

## Data Exploration & Preparation

+ The group has rightly identified highly correlated variables (r=1) and removed them from the modeling.

+ The group has done a fantastic job in the exploratory data analysis, showing the breakdown of churn by usage and plan features (e.g., vmail and int plan)

+ Transforming skewed variables using log transformation

+ Creating new variables: It was interesting to see that the group created a new variable (total charge). However, please note that for a linear model (i.e., the logistic regression), adding a new variable that is simply the sum of other variables may not increase the accuracy of the model. However, it can be very useful for interpreting the model and what is happening.

+/- It was interesting that the group decided to re-label the state, which is a categorical variable with many levels. While getting rid of meaningless levels is a good practice, the group should note that just because the target variable (i.e., churn) respond similarly for few levels (i.e., high churn for a number of states) it does not justify to combine them as one level since you may lose some predictive power. For example, imagine if the churn was high both in California and Texas. However, the higher churn was more strongly associated with voice mail plan in CA and more strongly with international plan in TX. If you label both states the same, a model such as a decision tree can not distinguish between the two. For example, a decision tree rule could be if State is CA and voicemail plan is false then …… This is not, however, a significant problem for logistic regression.

## Modeling Strategy and Logic

+ Excellent justification of the modeling framework.

## Model Accuracy

+ the accuracy off your model is pretty good

- Your report should explicitly state whether the accuracy measures have been constructed based on the training data or based on out of sample records

- Your code could have been organized a little bit better. On line 87, I could see a model is trained and after that some code related to exploratory data analysis. Try to follow a logical flow. The use of extensive comments is appreciated though.

Perhaps the group could also include the confusion matrix and discuss a potentially suitable cut-off threshold for the given problem

## Recommendations and Insights

+ This part was very well organized and was very comprehensive.  A very big well-done to the team!

## Presentation

+ Very good presentation. Excellent use of visualizations and right to the point!