Carta Data Science Challenge

1. ***Did you preprocess any of the features? If so, why? If not, why not?*** 
   * Yes, first of all, *'total\_day\_charge', 'total\_eve\_calls'* contain NULL/NaN values so I checked data distribution in both columns and based on that I impute both columns with their respective mean/average values.
   * After that, I found [*'state', 'area\_code', 'international\_plan', 'voice\_mail\_plan' ]* contain categorical values. So, I converted state and area code into one-hot vectors and the other two into Boolean since there are only yes/no values in them.
   * At last, I found the highly correlated features and remove them. Also, I normalize the data (standard scaling)
2. ***Which features are the most relevant for predicting the output? How did you measure feature importance?***

Here are few of top most important feature for predicting outcome.

1. number\_customer\_service\_calls
2. international\_plan
3. total\_day\_minutes
4. voice\_mail\_plan

I use xgboost feature\_importances\_ to find out feature importance.

1. ***What metrics did you use to measure the performance of your model? How did you determine how well your model generalizes?*** 
   * Confusion matrix
   * F-1 score
   * AUC