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.

    i. number_customer_service_calls
    ii. international_plan
    iii. total_day_minutes
    iv. voice_mail_plan

    I use xgboost feature_importances_ to find out feature importance.

3. ***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