Ultimate Data Science Challenge

# By: Nicholas Roller

## Part 1

EDA for our logins.json dataset netted several key insights. After aggregating the login counts per 15 minutes and plotting the 96 timeframe rolling average (daily login counts) we could visually see that our data was mostly stationary and exhibited seasonality.

Shape

Description automatically generated with medium confidence

We then ran an adfuller test to confirm that the time series was indeed stationary. We got a p-value of 2.73e^-18 which was indeed larger than our 5% p-value coefficient value of -2.86.

We then plotted the data against it’s -1 time frame lag using lag\_plot and could see that it has a strong correlation with its lag.

Chart, scatter chart

Description automatically generated

We then wanted to look more closely at the time periods where seasonality might be occurring, so we first plotted the average hourly login counts throughout the day. We were able to determine that there were two major times during the day in which logins occurred, around mid-day and around midnight. Chart, line chart

Description automatically generated

We did the same for daily logins throughout the course of a week, and were able to see that they trend upwards and peak on Friday’s.

Chart, line chart

Description automatically generated

Again, we did the same for days throughout the entire month. This did not provide such an obvious conclusion as the others, aside from higher peaks and lower troughs towards the end of the month.

Chart, line chart

Description automatically generated

## Part 2

1. A certain percent increase in completed rides during each city’s respective down-period.
2. A practical experiment would be to collect drive partner app data on completed rides and if, for example, we saw a 20% increase in complete rides during the low-activity periods in each respective city, we could potentially determine the experiment a success.
   1. I would analyze the weekday and weekend data from before and after the initiation of the toll reimbursement change and calculate whether there was a statistically significant increase in completed rides during the weekdays after the change.
   2. I would use a logistic regression statistical test to verify the hypothesis that traffic increased into each city during their respective low-activity periods since we are seeking a binary outcome variable using continuous predictor variables. with the two cities traffic into one another as two population groups and use the weekend traffic as a control.
   3. We would be looking for a p-value below 0.05 to determine if our null-hypothesis that the number of rides did not increase is nullified.

## Part 3

1. Three variables had missing data, avg\_rating\_of\_driver, phone and avg\_rating\_by\_driver, so we imputed these missing data points with the median for each respective variable. We then converted the two date columns from objects to datetime formats and took their difference to create a new feature called ‘days active’. This is the feature we used to create our independent variable, ‘retained’, if they were active for longer than 5 months after signing up, to meet the criterion of being active during their 6th month. From this we created a binary column where 1s represent retained, and 0 if they were not retained. These are the labels we used in our predictive model. 12,714 users out of 50,000 were retained, or roughly 25%.

Chart, bar chart

Description automatically generated

1. Because we had 50,000 samples in our dataset and were looking to make a categorical prediction, we narrowed down our estimators to linear SVC, KNN or some ensemble method. I first ran a simple logistic regression model which had a roughly 58% accuracy is user retention prediction. Then we setup a SVM classifier, which improved this estimation to about 63%.
2. With our support vector machine (SVM) classifier, we achieved a 63% accuracy score in estimating user retention, and 76% precision in estimating negative retention. King’s landing had higher retention than other cities, Winterfell was rather neutral and Astapor had the worst retention. Ultimate black users were more likely to be retained. The user’s activity in their first 30 days was also a positive indication as to their retention.