Ultimate Take Home Challenge

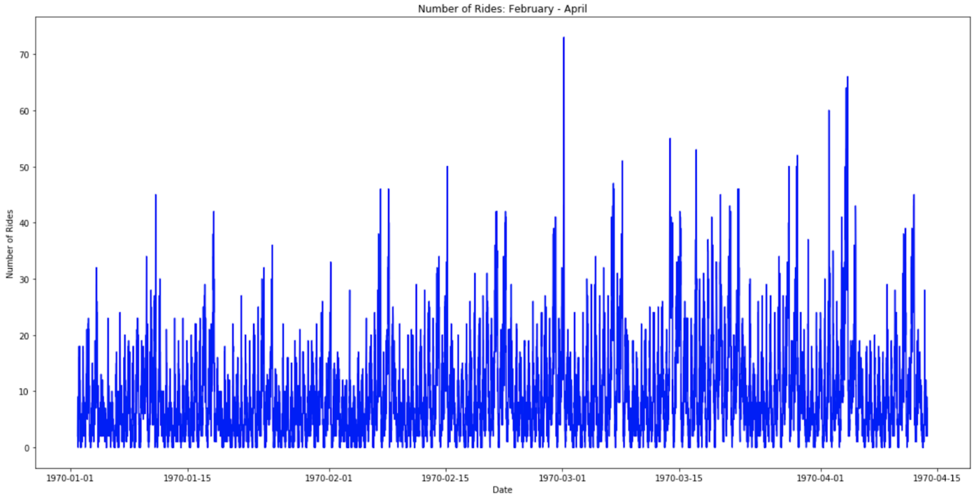
by: Lauren Broussard | Springboard DSC

All code can be found in /notebooks/01.ultimate_data_science_challenge_solution

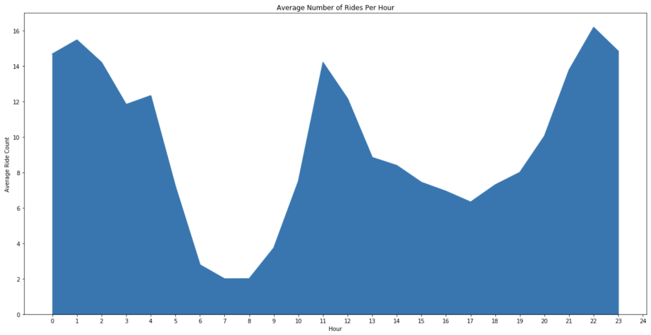
Part 1: Exploratory data analysis

Instructions: *The attached logins.json file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15-minute time intervals, and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand. Please report/illustrate important features of the demand, such as daily cycles. If there are data quality issues, please report them.*

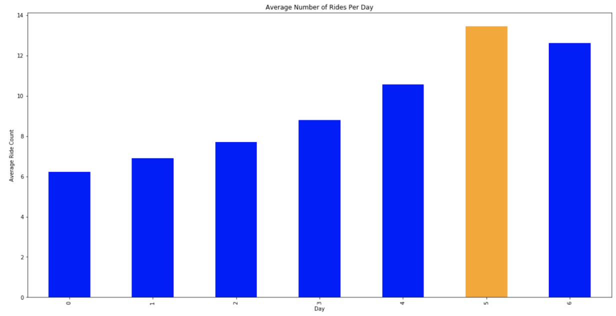
Initially plotting the entire dataset as a time series, we can see a definite daily cyclical trend in the data. The pattern looks fairly similar every day, with high and low points in certain points of each day.



Rides Per Hour. If we look further at the average number of rides per hour, we can see the trend more clearly. While one might expect demand to peak at 8am and 5pm - to coincide with rush hour in most cities - the demand is actually vastly different for Ultimate. Demand peaks a bit before noon, and again near midnight - with an average of 16 rides at this time. It is at its lowest point early in the morning, between the hours of 6pm and 8pm.



Rides Per Day. The most common day for rides is Friday, with an average number of rides of 13.5, followed by Saturday, with an average of 12.6. The demand simply seems to steadily increase as the days get closer to the weekend, and then dips dramatically on Sunday.



Ultimate Take Home Challenge
by: Lauren Broussard | Springboard DSC

Possible Data Quality Issues. As a note, there are 877 duplicated timestamps in the dataset. Since the timestamp information is granular to the second, it would be important to ask additional questions to determine if these are true duplicates that should be removed. For instance, are multiple people interacting with the product at once? Is it possible for more than one person to use the service simultaneously?

Part 2: Experiment and Metrics Design

Instructions: *The neighboring cities of Gotham and Metropolis have complementary circadian rhythms: on weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day. On weekends, there is reasonable activity in both cities. However, a toll bridge, with a twoway toll, between the two cities causes driver partners to tend to be exclusive to each city. The Ultimate managers of city operations for the two cities have proposed an experiment to encourage driver partners to be available in both cities, by reimbursing all toll costs. 1. What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities, and why would you choose this metric? 2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success. Please provide details on: a. how you will implement the experiment b. what statistical test(s) you will conduct to verify the significance of the observation c. how you would interpret the results and provide recommendations to the city operations team along with any caveats.*

1. **KPI:** To evaluate the effect of the toll incentive, an **important metric for success would be the percentage of drivers picking up rides outside of their usual city**. For instance, the percentage of Gotham drivers picking up Metropolis rides, and vice versa. If the implementation is a success, we should see an increase in drivers originating in one city and picking up/dropping off in another.

   In the future, drivers between the two cities should be indistinguishable, so at another point it would be important to look at the average number of rides between the two cities.

2. **Experiment Design:**
   a. **Design & Implementation:** We could set up an A/B test, where some portion of drivers receive a toll reimbursement and another portion does not. The group should be chosen completely at random. We could run it for a two-month period to account for anomalies on certain days (i.e. big events, big holiday weekends, etc.). We'd need to note which city the driver normally drives in, and collect data on all of their pick-ups and drop-offs to note the increase (or not) in cross-city pick-ups.

   b. **Statistical Test(s):** We will declare our null hypothesis that there is no difference in the percentage of cross-city pick-ups between groups that receive a toll reimbursement and those that do not receive a toll reimbursement. From there, we'd do a bootstrap test with our dataset. This would simulate rerunning the experiment x amount of times. We could do 10,000 simulations, for example. After that, we would find our p-value, to determine the probability that our results would be this extreme or more extreme. Depending on those results, we could then accept or reject our null hypothesis.

c. **Results and Recommendations:** Depending on our results - i.e. if we found that there was in fact a difference in percentage of cross-city pick-ups between groups - we could roll out the toll reimbursement to the rest of the drivers in the cities. Another caveat would be to consider who is getting chosen in this A/B test to get their toll reimbursed. It would be worth considering the fairness of giving a financial incentive to some drivers but not all. Therefore, it may be possible to run the experiment not as an A/B test, but as a trial run for all drivers, using data from the previous year as a comparison.

Part 3: Predictive Modeling
Rider Retention

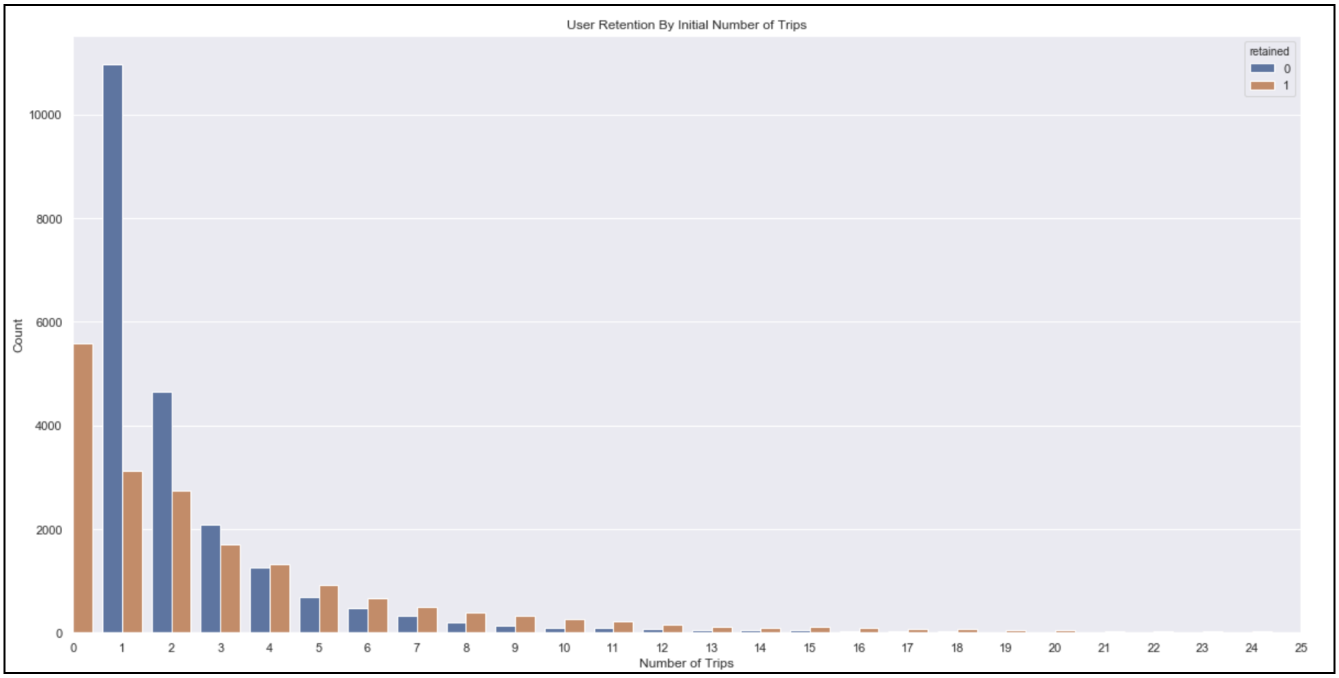
Instructions: *Ultimate is interested in predicting rider retention. To help explore this question, we have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. The data was pulled several months later; we consider a user retained if they were "active" (i.e. took a trip) in the preceding 30 days. We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate.*

Cleaning and EDA: The dataset consisted of 50,000 user transactions. The first step in cleaning involved labeling users who were retained and those that were not. **From this, we found that 18,804 users or 37.61% were labeled as 'retained'.** Other cleaning steps involved filling duplicate values. For instance, fields for both *avg_rating_by_driver* and *avg_rating_of_driver* were filled with their mean values in the case of missing values. Extreme outliers were also removed. Initial analysis of the data suggested a few features that may be indicative of user retention:

- **City:** "King's Landing" had higher retention than some other cities.
- **Phone:** iPhone users had a higher percentage of retention than Android users.
- **Initial Number of Trips:** Looking at trips taken in the first 30 days, at four trips or more, the retention rates outpace the non-retention rate.

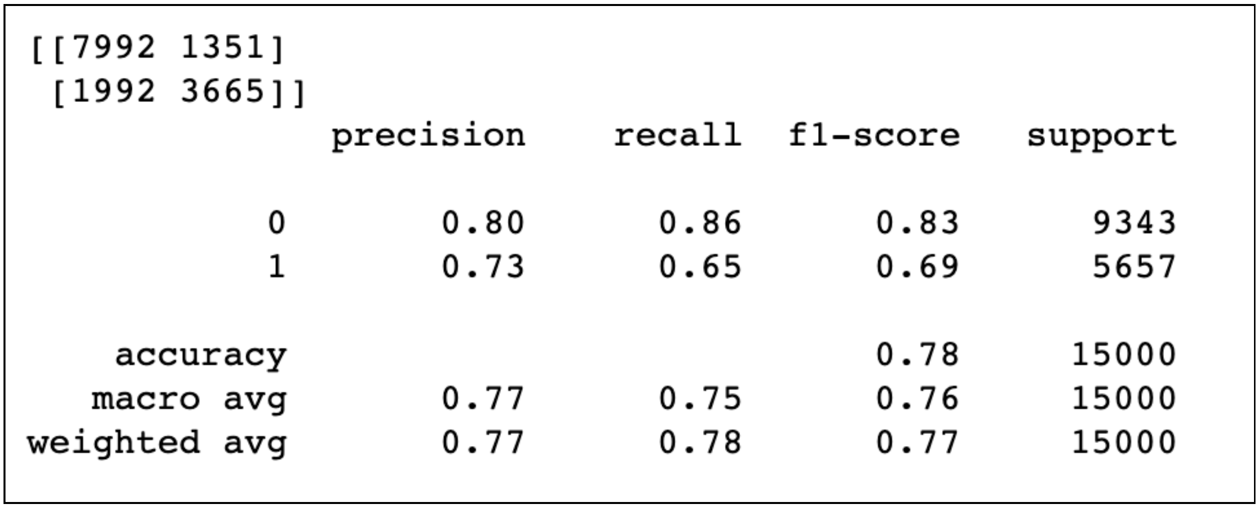
Ultimate Take Home Challenge

by: Lauren Broussard | Springboard DSC



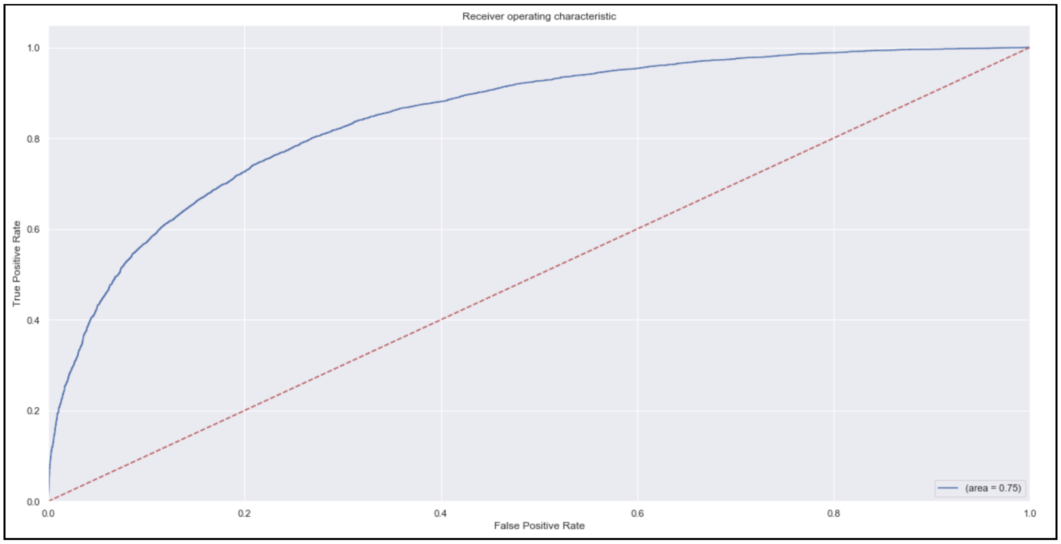
Predictive Model: A Random Forest Classifier was used to predict user retention. This type of algorithm is appropriate for a supervised classification problem like this one, and can also handle outliers and a large amount of features/dimensions. Another alternative considered for this problem was a Logistic Regression.

The Random Forest Classifier achieved an accuracy of 0.77 in predicting user retention after 6 months. However, since there was an imbalance of classes (38% retained vs 62% not retained), it is more appropriate to look at other measures of success in addition. The following classification report shows the precision, recall, and F1-score for the 15000 predictions done on testing data points.



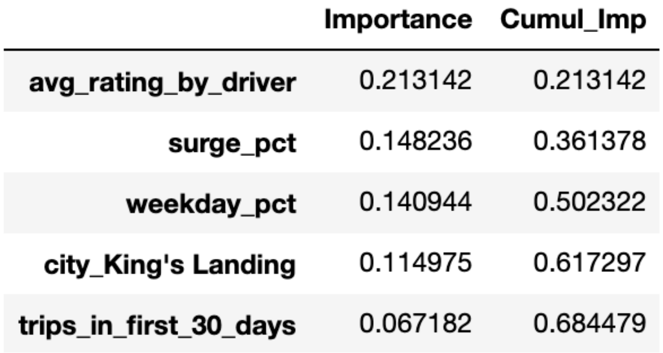
```
[[7992 1351]
 [1992 3665]]
              precision    recall  f1-score   support

           0       0.80      0.86      0.83      9343
           1       0.73      0.65      0.69      5657

    accuracy                           0.78     15000
   macro avg       0.77      0.75      0.76     15000
weighted avg       0.77      0.78      0.77     15000
```

Finally, looking at the ROC curve, we can see that the model predicts much better than chance, in fact at 75%.



Results & Recommendations: From our early analysis and from the model, a few of the most important features in predicting user retention are:

- **Average Rating By Driver:** This accounted for 21% of the model's prediction. Also interesting was that the average rating given *by* the driver and the average rating given *of* the driver were only slightly positively correlated - meaning that drivers and riders are not always on the same page about their experience.
- **Surge and Weekday Percentage:** Higher surge pricing and higher weekday percentages may be indicative of the reasons some riders are using the service. For instance, riders may be using the service for their weekly commute as an alternative to public transportation. It will be important to look further into this group of patrons.
- **Kings Landing:** This city was a predictor for higher retention. It will be important to look further at what makes riders in this city come back - signage, better drivers, a lack of other nearby transportation options, etc.



	Importance	Cumul_Imp
avg_rating_by_driver	0.213142	0.213142
surge_pct	0.148236	0.361378
weekday_pct	0.140944	0.502322
city_King's Landing	0.114975	0.617297
trips_in_first_30_days	0.067182	0.684479

It will also be important to look at other models in the future, add additional data points to the dataset, and work for a better class balance to improve the predictive power of the model.