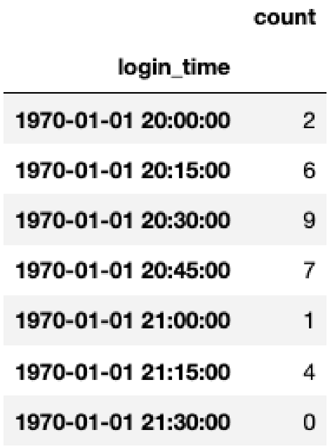
SUBMISSION: Ultimate - Data Analysis Interview Challenge

John Baselj

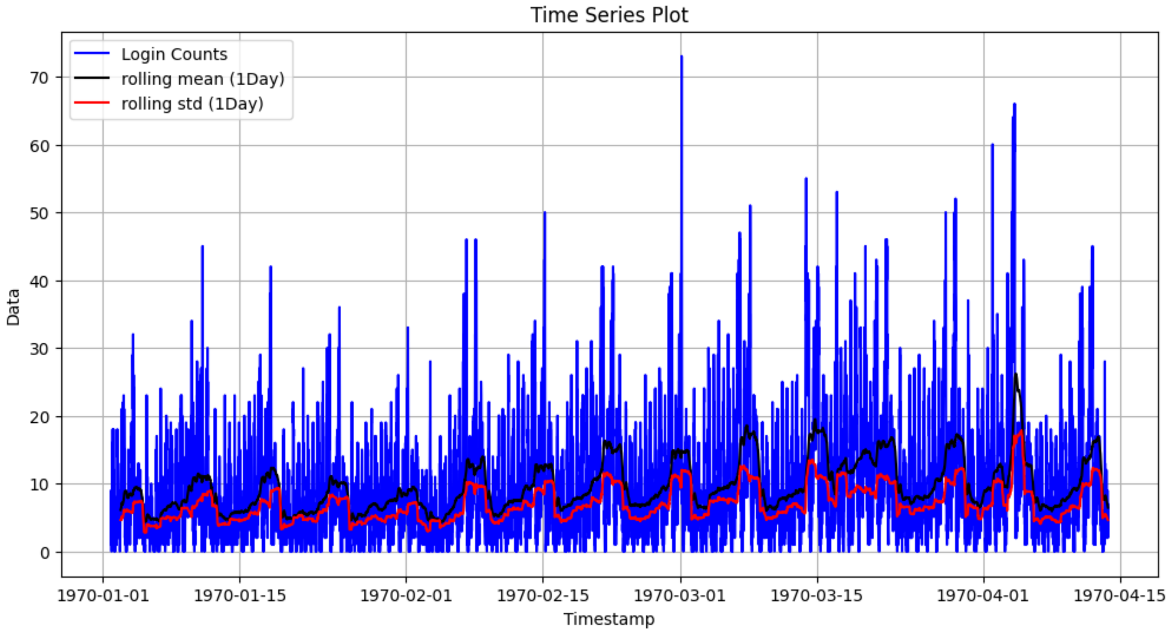
Part 1 - EDA

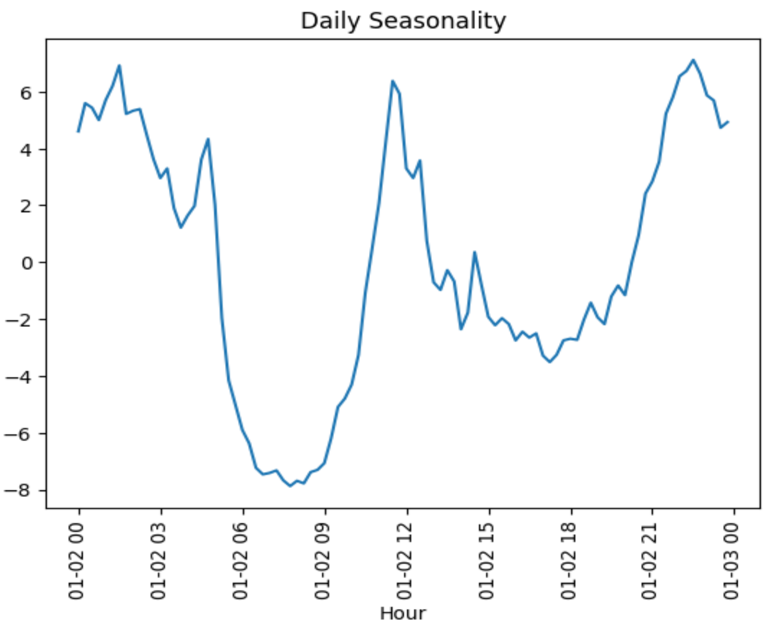
- Aggregate login counts based on 15min time intervals

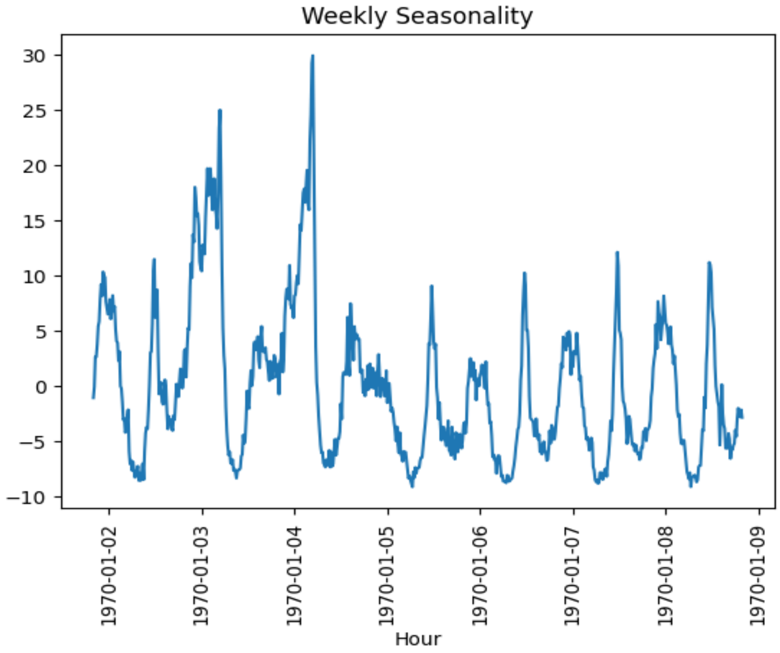


	count
login_time	
1970-01-01 20:00:00	2
1970-01-01 20:15:00	6
1970-01-01 20:30:00	9
1970-01-01 20:45:00	7
1970-01-01 21:00:00	1
1970-01-01 21:15:00	4
1970-01-01 21:30:00	0

- visualize remaining time series and characterize underlying patterns







- report/illustrate features of demand like daily cycles
    - Within each day, there is a cycle that has lowpoints around 6am-9am and 1pm-8pm, and highs around 11am and 9pm-4am
    - Within each week, there is a slight upward trend in volume from Monday through Friday, and then a sharp increase for Saturday and Sunday.

Part 2 - Experiment and Metrics Design

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?*

Answer: Metric: Count of rider pickups by drivers who live on the other side of the bridge from the pickup location (recorded per time hour/day/week etc.).
This metric is numeric and easily trackable for the company. It also represents a volume of productive use of this cross-bridge travel by counting the number of pickups by a driver while across the bridge.

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.

Answer A: Randomly select a group of drivers “A” from the full set of drivers in the two cities. Offer toll bridge reimbursement to all drivers in group A. The remaining drivers from the two cities will be defined group B. Group be serves as a control group for reimbursement. Over a predetermined time span, such as 1 week. Track the target metric of the two groups, total number of rides given on “the other” side of the bridge from each driver’s home.

B: Perform statistical t-test to determine whether or not the values of the target metric between groups A and B is statistically significant. If group A shows a statistically significant increase in number of rides on the other sode of the bridge compared to group B, this would indicate that the incentive is effective in encouraging cross-bridge driving.

C: If this test were to indicate that the new incentive is effective, it must still be modeled whether or not this incentive is financially beneficial for the stakeholders overall. i.e., the city must model or monitor whether the lost money from reimbursing tolls is outweighed by the benefits of increasing cross-bridge drivership.

Part 3 - Predictive Modeling

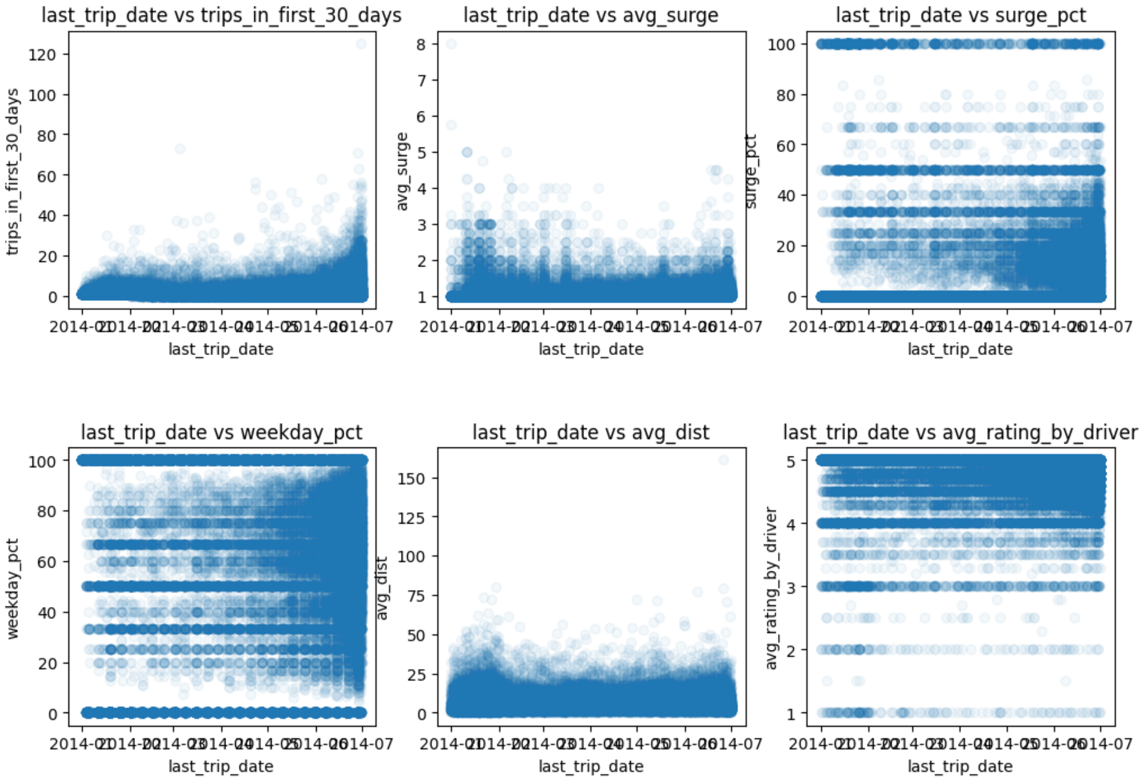
Predicting rider retention: sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. data from months later; user “retained” if they took a trip in the preceding 30 days.

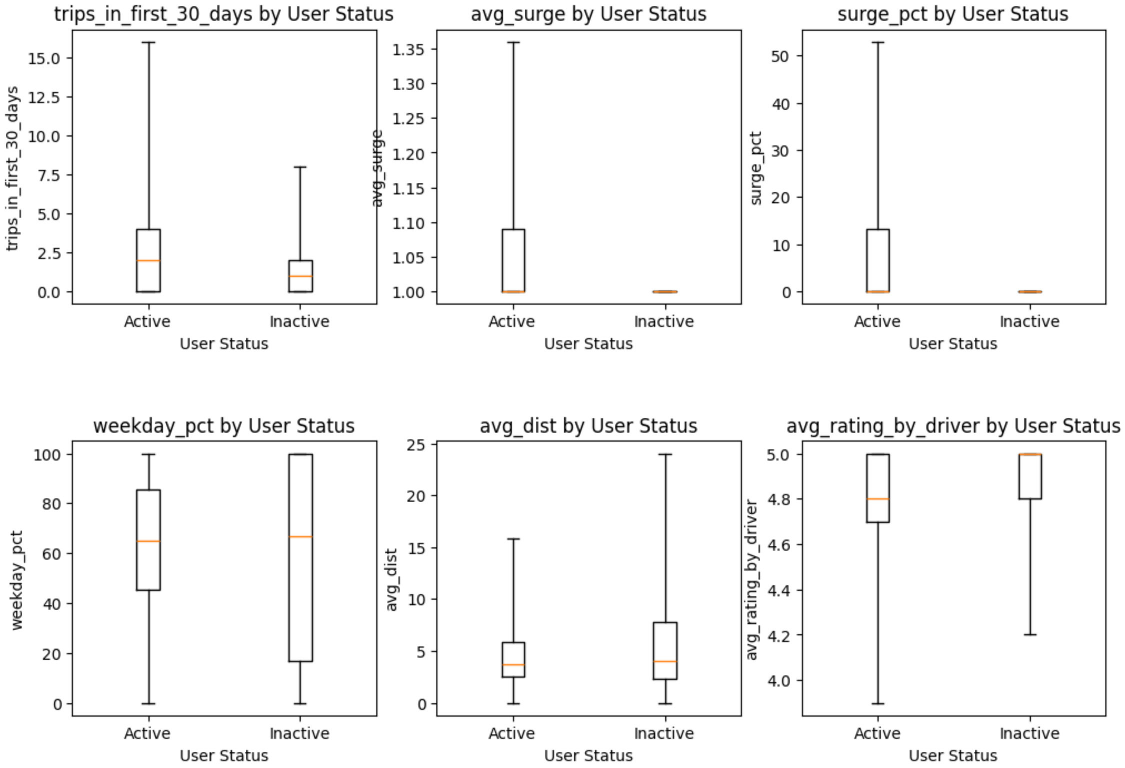
What factors are the best predictors for retention? Offer suggestions to operationalize those insights.

1. Cleaning, EDA, visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?

   About 36% of users were retained.

   Scatterplot each numerical field against last_trip_date to spot obvious trends. Notably, more trips in the first 30 days trends positively with more recent ridership, and weekday percentage may trend weakly positively with more recent ridership.

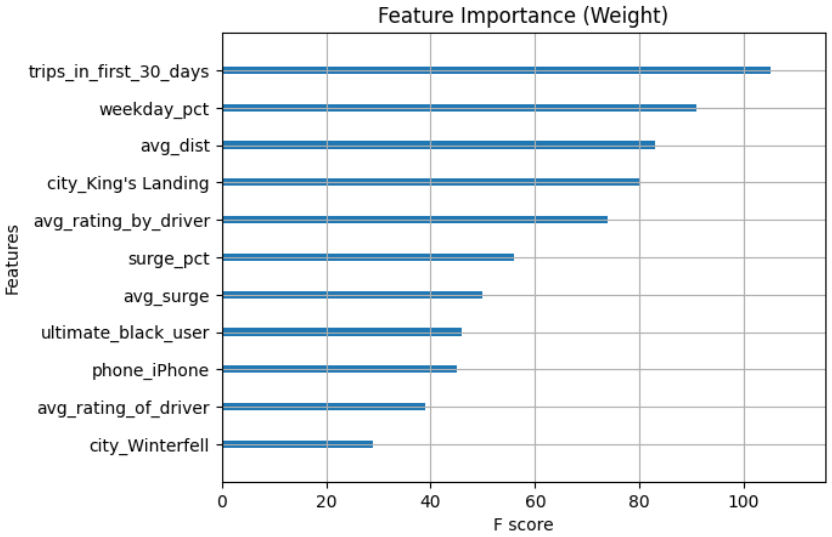



trips_in_first_30_days by User Status
trips_in_first_30_days
0.0
2.5
5.0
7.5
10.0
12.5
15.0
Active
Inactive
User Status
avg_surge by User Status
avg_surge
1.00
1.05
1.10
1.15
1.20
1.25
1.30
1.35
Active
Inactive
User Status
surge_pct by User Status
surge_pct
0
10
20
30
40
50
Active
Inactive
User Status
weekday_pct by User Status
weekday_pct
0
20
40
60
80
100
Active
Inactive
User Status
avg_dist by User Status
avg_dist
0
5
10
15
20
25
Active
Inactive
User Status
avg_rating_by_driver by User Status
avg_rating_by_driver
4.0
4.2
4.4
4.6
4.8
5.0
Active
Inactive
User Status

2. Build a predictive model to help determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.

   Answer: An XGBoost model was trainedbecuase of its ability ot have strong predictive performance while maintaining interpretability fo features. This is important so that the business can apply the modeling results to make decisions on how to improve rider retention.

   Test set accuracy: 79.52%
   Precision: 0.75
   Recall: 0.67
   F1-score: 0.71
   ROC AUC: 0.77



3. Briefly discuss how Ultimate might leverage the insights gained from the model to improve its long term rider retention (again, a few sentences will suffice).
   Answer: Based on feature importances that correlate with the company's target outcome, Ultimate could:
   - try to increase riders' average number of trips in the first 30 days. They could attempt to do this, for example, by having introductory offers/rates for new riders that are discounted for about 30 days
   - Offer discounted rides on weekdays. Many consumers weekdays are more routine than their weekends, so Ultimate can strive to penetrate as many weekday riders' routines as possible

In General, this modeling drew a picture for which consumer features have an influence on liklihood of retention, so the business can leverage this insight to target those features.

Code in Jupyter Notebooks in this folder -

Part1_ultimate_challenge | **Part3_ultimate_challenge**