Data Analysis Challenge

ULTIMATE TECHNOLOGIES INC.

YUKA ABE 3/30/2019

Background

Ultimate Technologies Inc. is an American worldwide online transportation network company that has disrupted the taxi and logistics industry and is a prestigious companies to work at. This challenge has been adapted from an actual Ultimate Inc. data science challenge.

Data Analysis Challenge: Part One

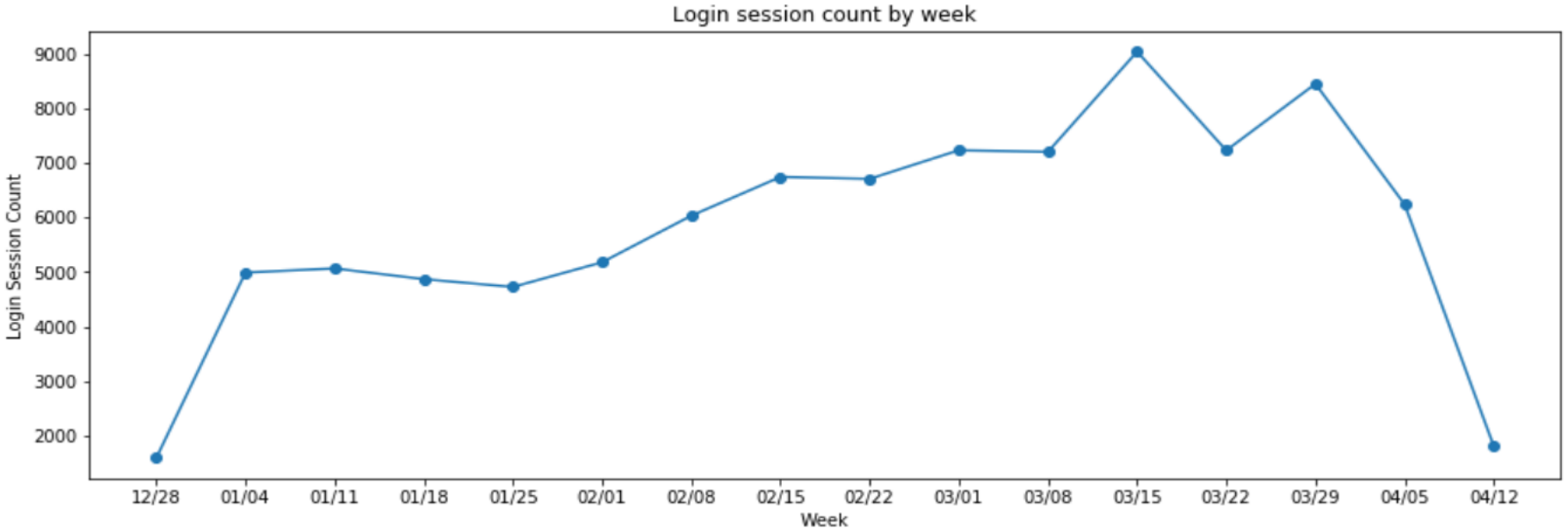
EXPLORATORY DATA ANALYSIS ON USER LOGIN TIMESTAMP

Part One Challenge Background

- Download (simulated) timestamps of user logins data 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.

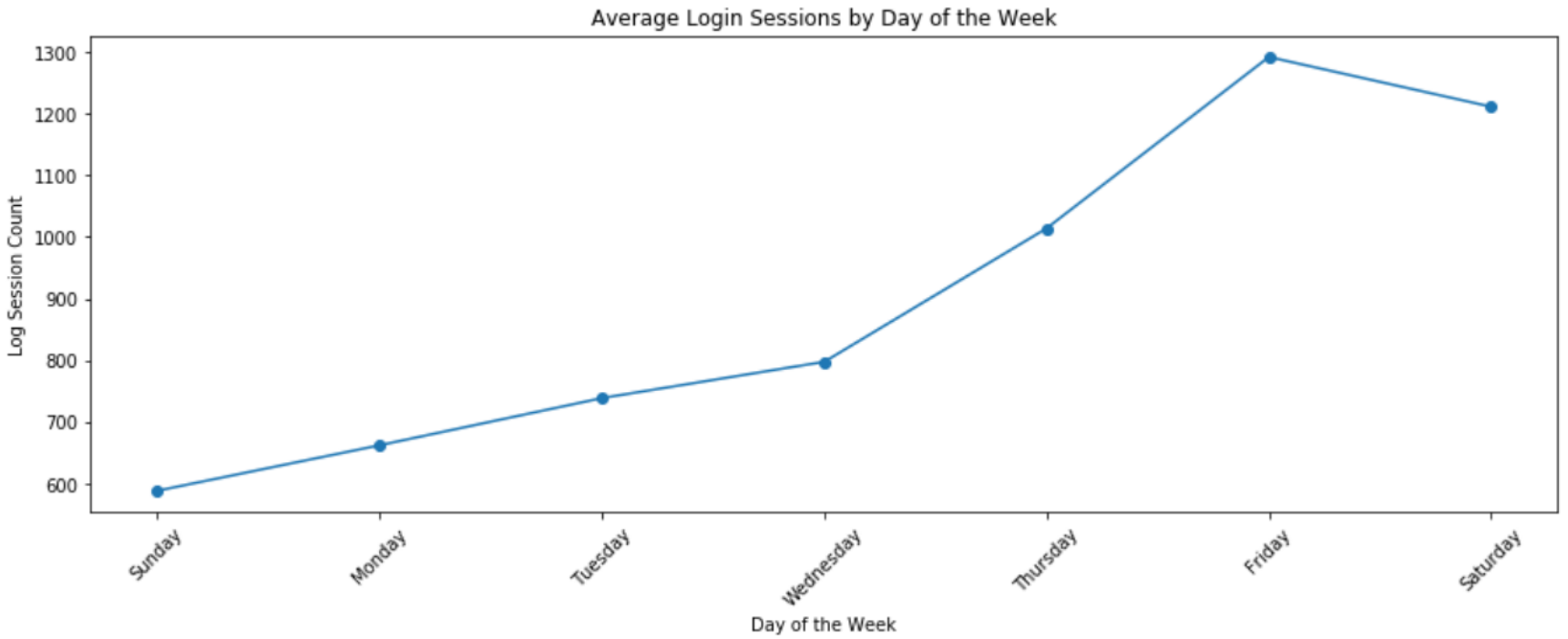
Solutions: Login Sessions By Week

- Usage of app in that specific location has an upward trend from January till end of March.
- However the usage is going down during the latest weeks in April.
- Need to look into systems to see if the dip in April is caused by any data/system issue.



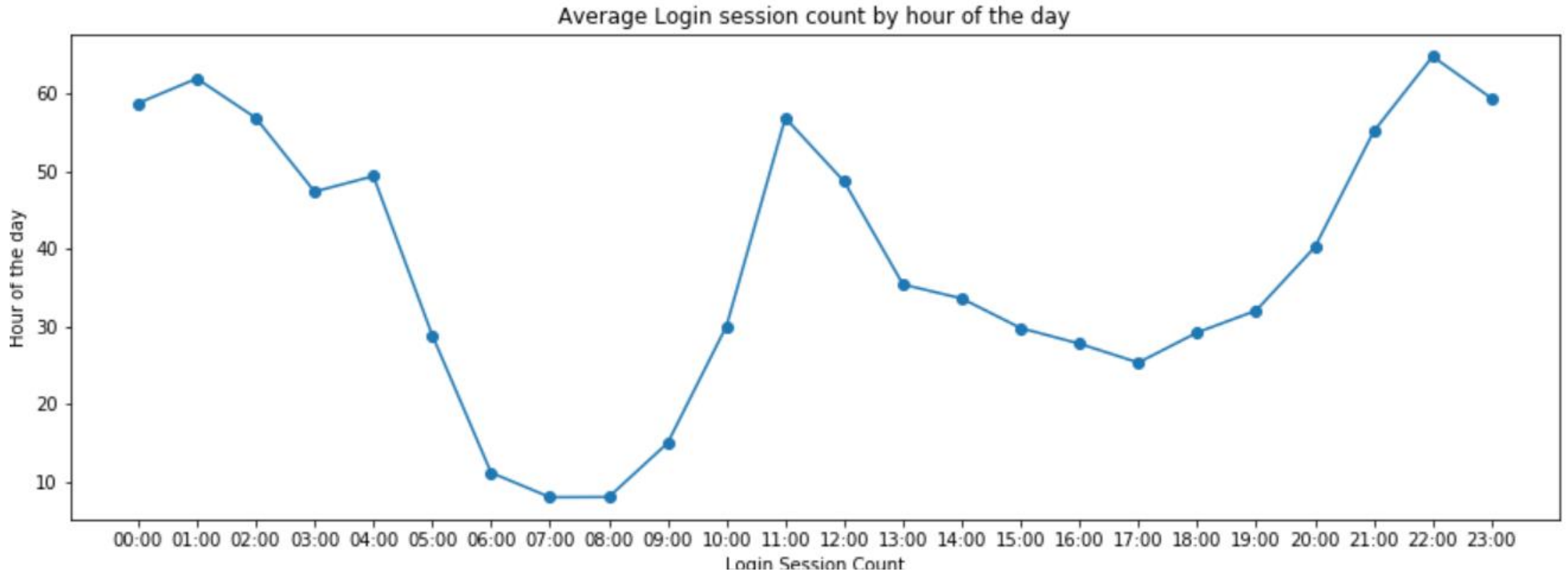
Solutions : Login Sessions by Day of the Week

- The closer towards the weekends, the more demands for taxis in that location.



Solutions: Login Sessions By Hour of the Day

- The peak hours for the app: 10 pm - 1 am
- off-peak hours: 6 am - 9 am



What Geo Location could the data be collected?

- The location might be a dynamic neighborhood full of entertainment, restaurants and nightlife (example: East Village, Lower East Side in NYC)

Data Analysis Challenge: Part Two

EXPERIMENT AND METRICS DESIGN

Part Two Challenge Background

- **Background information**

  The neighboring cities of Gotham and Metropolis have complementary circadian rhythms. However, a toll bridge, with a two way 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.

Part Two Challenge Background

- **Questions**:
  1. What key measure of success will you choose 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.

Solution: Key Metric

- **Key Metric:** Trip fulfillment rate in the other city by the driver partners working in these two cities
- Trip fulfillment Rate = # Trips fulfilled by driver partners in the other city / total # trips fulfilled by driver partners
- **Hypothesis:** After the new toll reimbursement policy gets rolled out, the probability of the driver partners to fulfill the trips in the other city will increase.

Solution: Test Design

- **A/B testing:** Pick two groups of driver partners working mostly in one of the cities. (one as control, the other as experiment)
- **The testing groups should meet with the following requirements:**
  - Usually active during peak hours of both cities
  - Have similar weekly working hours and driving miles
  - Each group should have a representative proportion of driver partners from both cities
- Control group will not receive reimbursement for toll fee. Experiment group will get notification on the new toll reimbursement policy.
- **A/A test:** Check if the key metric is robust enough before doing the test.
- **Test duration:** at least one month for driver partners to get adapted to the new policy.

Solution: Test Read & Interpretation

- **Statistical test:**
  - Use Z test to compare fulfillment rate given the testing population is large enough
  - Use One-tail test since we only care about whether the new change has positive effect
- **Test result interpretation:**
  - Stats significant result: The policy is actually working. Keep ramping up the test
  - No stats significant result: Post-test study/survey to get drivers' feedback
- **Additional recommendations:**
  - Keep track of the actual demand fulfillment rate when ramping up the test to understand whether customers actually need extra supply for drivers outside of the city
    - Demand fulfillment rate = # driver supply/ # requests made by customers in both cities.
  - Track revenue driven by the change as opposed to the increase in cost to ensure the profitability of the new policy

Data Analysis Challenge: Part Three

PREDICTIVE MODELING

Part Three Challenge Background

- **Data provided:** Sample data of a cohort of users who signed up for an Ultimate account in January
- **Objective:**
  - Predict rider retention 6 months after users signed up
  - Understand what factors are best predictors for retention
  - Offer suggestions to operationalize the insights to help Ultimate
- **Definition of active user:** consider a user retained if they were "active" (i.e. took a trip) in the preceding 30 days
- **Deliverables:**
  - Build predictive model to determine whether a user will be active in their 6^{th} month on the system
  - Discuss why you choose you approach and how valid your model is
  - Leverage insights gained from model to improve its long-term rider retention

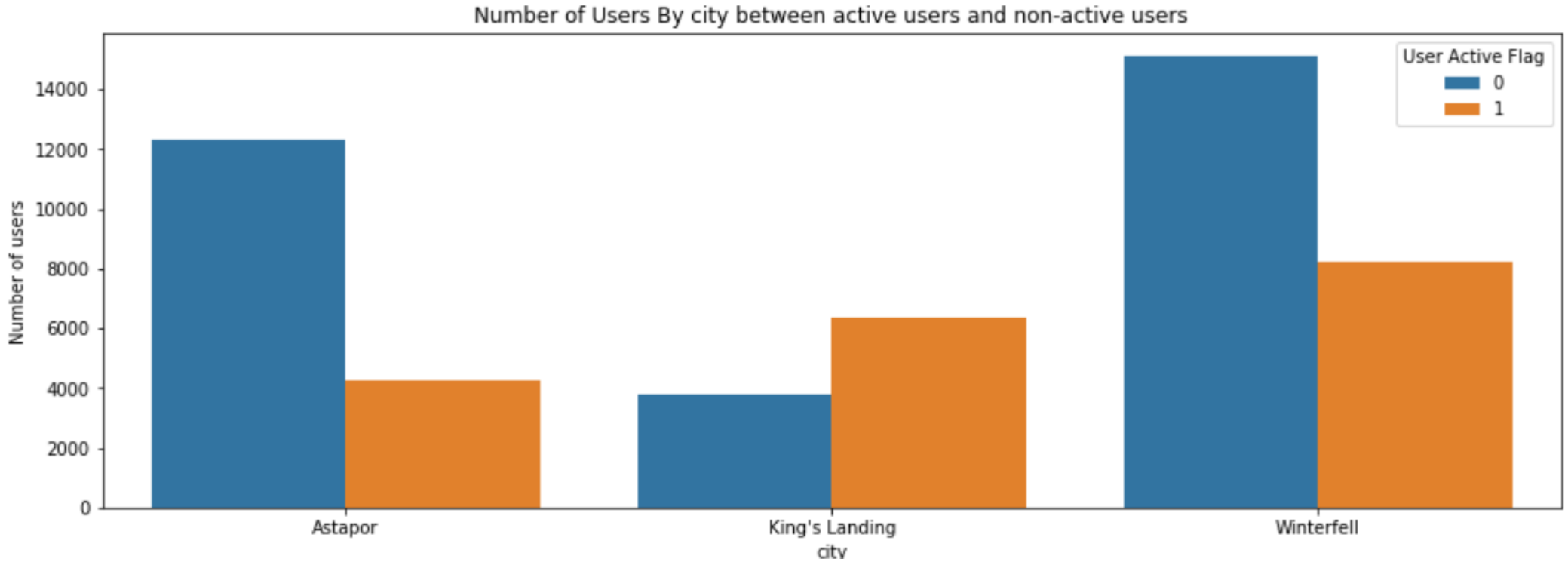
Part Three Challenge Background

- **Data Description**

  - **city:** city this user signed up in
  - **phone:** primary device for this user
  - **signup_date:** date of account registration; in the form 'YYYYMMDD'
  - **last_trip_date:** the last time this user completed a trip; in the form 'YYYYMMDD'
  - **avg_dist:** the average distance in miles per trip taken in the first 30 days after signup
  - **avg_rating_by_driver:** the rider's average rating over all of their trips
  - **avg_rating_of_driver:** the rider's average rating of their drivers over all of their trips
  - **surge_pct:** the percent of trips taken with surge multiplier > 1
  - **avg_surge:** The average surge multiplier over all of this user's trips
  - **trips_in_first_30_days:** the number of trips this user took in the first 30 days after signing up
  - **ultimate_black_user:** TRUE if the user took an Ultimate Black in their first 30 days; FALSE otherwise
  - **weekday_pct:** the percent of the user's trips occurring during a weekday

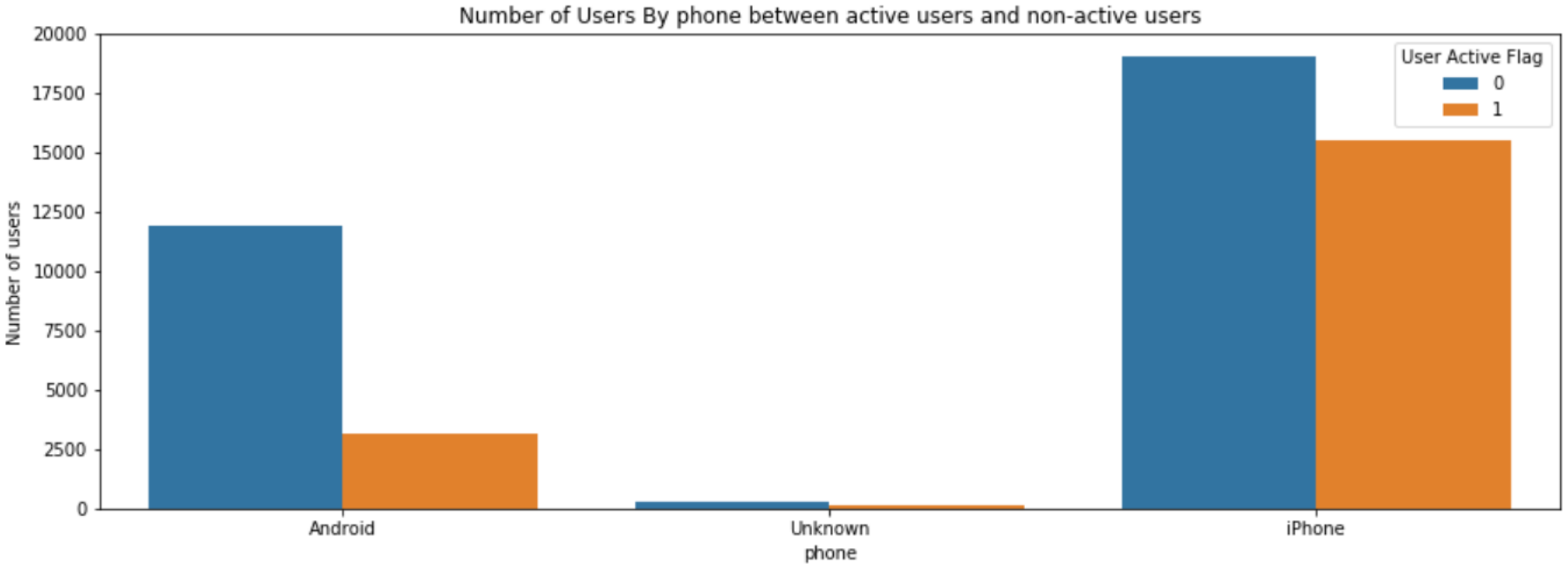
Solution: Exploratory Data Analysis

- Users who signed up in King's landing are more likely to be retained.



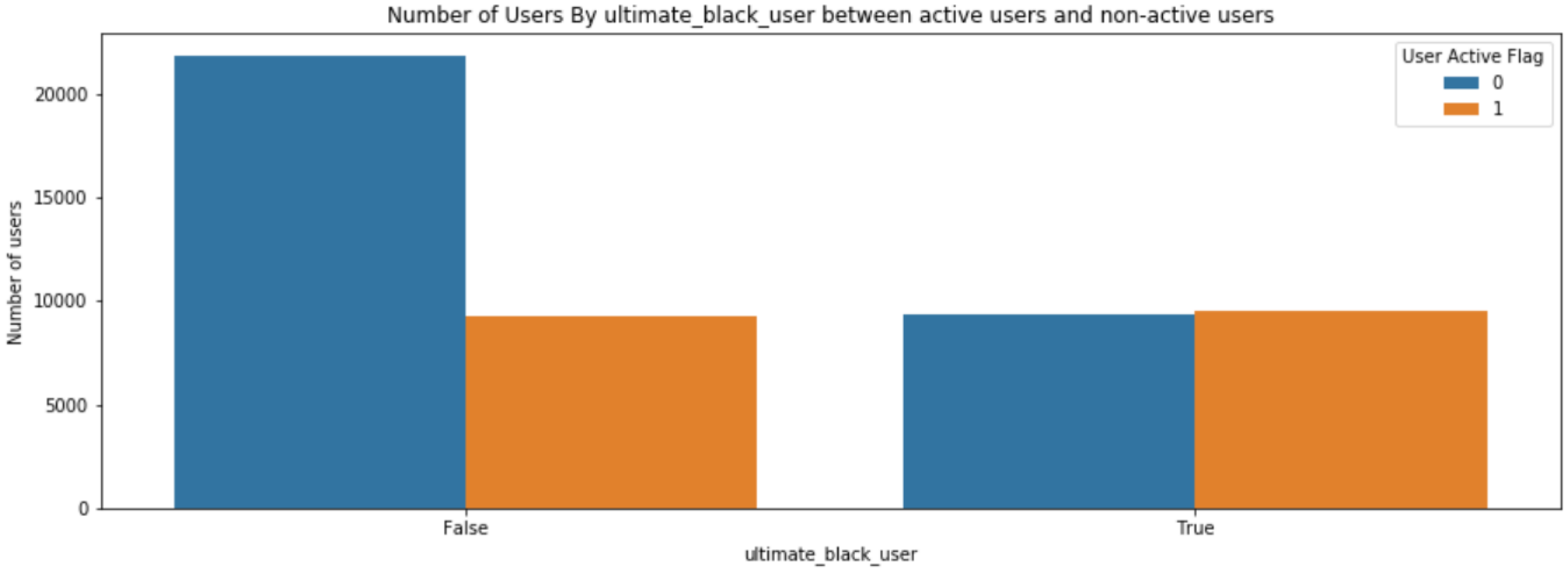
Solution: Exploratory Data Analysis

- Android users are less likely to be active users after 6 months.



Solution: Exploratory Data Analysis

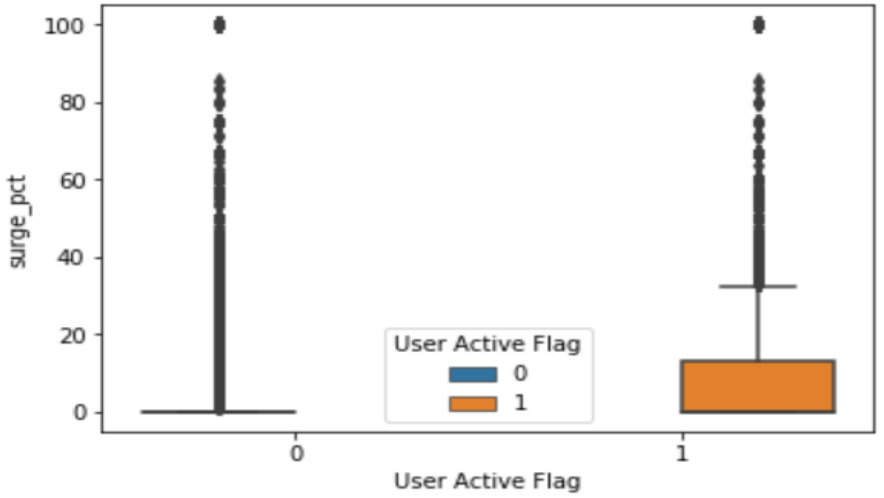
- Users who did not take an Ultimate Black in their first 30 days are less likely to be retained users.



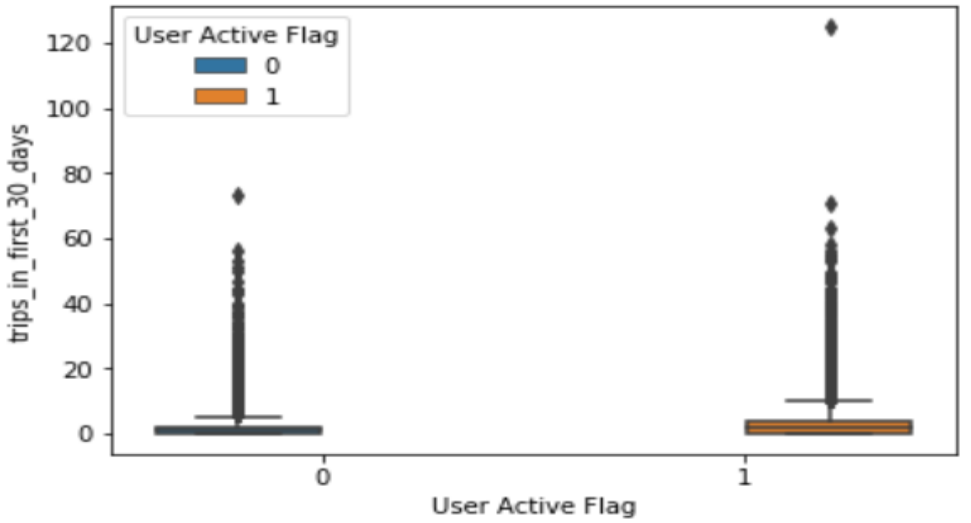
Solution: Exploratory Data Analysis

- Active users took slightly more trips in their first 30 days.
- Active users have taken more of the trips with surge multiplier compared with inactive users.

Percentage of trips with surge multiplier



Trips in the first 30 days



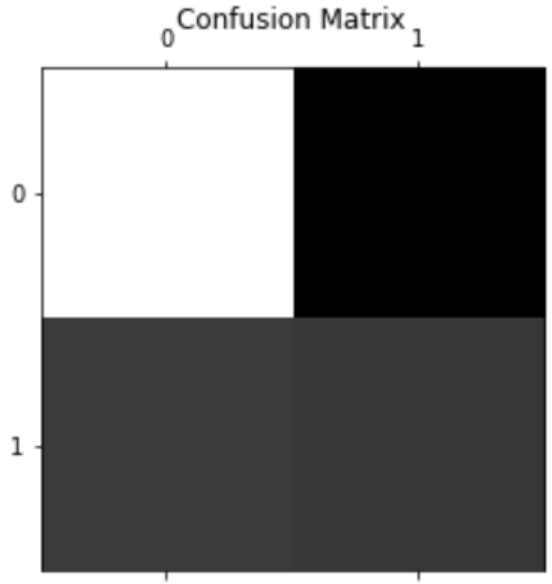
Solution: Classification Model

- **Label:** Active User Flag (Y/N)
- **Features:** User attributes in the dataset such as city, average rating on driver, etc.
- **Models used:** *Logistic Regression & Random Forest Classifier*
  - Logistic Regression is more interpretable as it gives insights on the weight of each feature on the user retention probability.
    - ✓**Techniques used:**
      1. Log transformation on highly skewed variables
      2. One hot encoding on categorical variables
      3. Adjust coefficients through regularization (Ridge) to avoid overfitting
  - Random Forest Classifier is less interpretable but it gives more accurate predictions, avoids overfitting and also provides insights on feature importance.
    - ✓ **Techniques used:** Randomized Search and Grid Search for hyper-parameter tuning

Solution: Logistic Regression Model

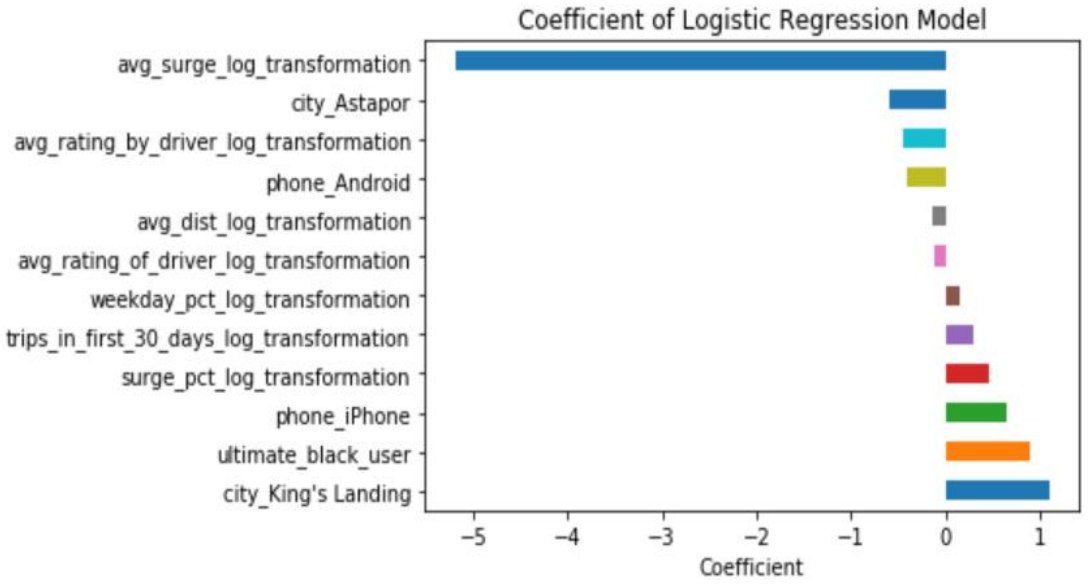
- **Model Performance**
  - The model has a good accuracy score and ROC AUC score since that data has unbalanced classes.
  - The recall score looks low, which means the model will have problem capturing real active users that are likely to be retained.

```
Below is the performance from the testing set:
Accuracy Score:  0.73
Precision Score:  0.6799733865602129
Recall Score:  0.5404547858276044
F1 Score:  0.6022392457277548
Roc Auc Score:  0.6928713298710232
```



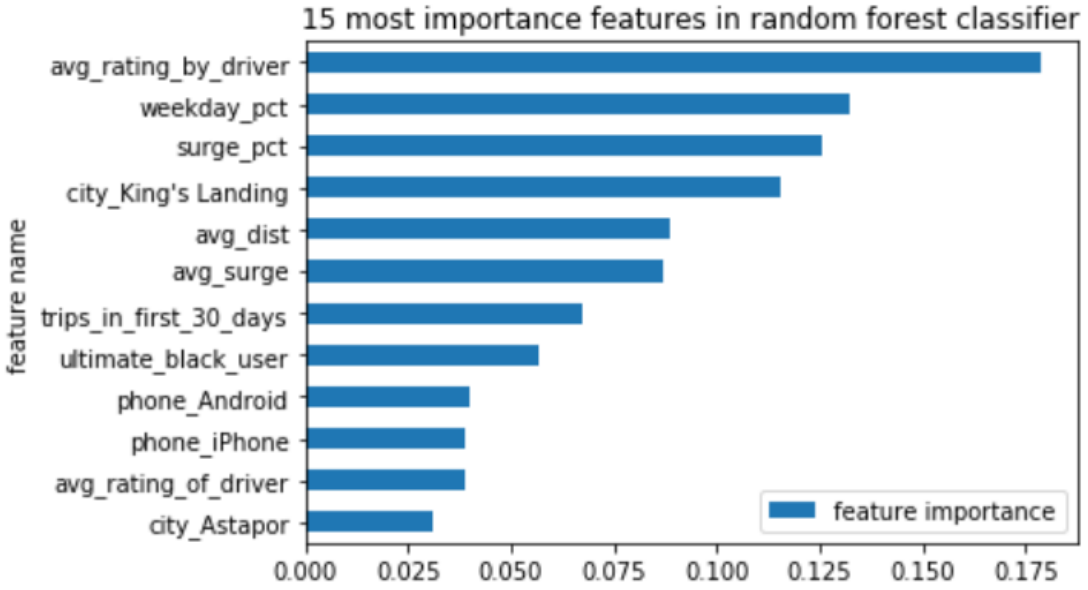
Solution: Logistic Regression Model

- **Top positive features:** “King’s landing” signup place, “Ultimate Black” usage and iPhone user device
- **Top negative features:** Average surge multiplier usage over all of users’ trips



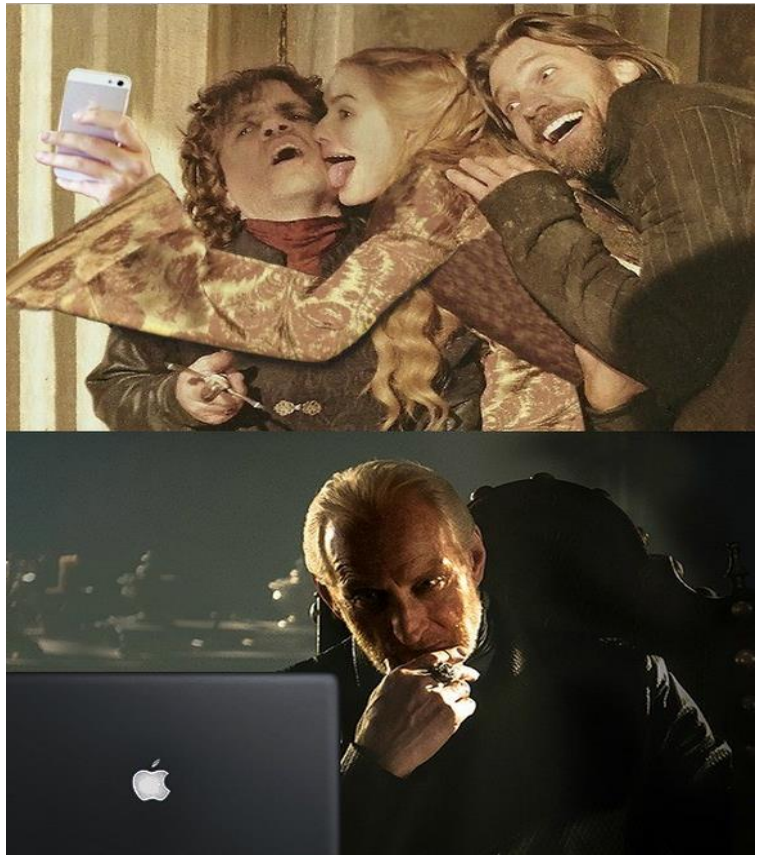
Solution: Random Forest Classifier

- **Top positive features:**
  Average rating by driver, Percentage of weekday trips over all trips, Percentage of trips using surge multiplier over all trips



Solution: Business Insights & Recommendation

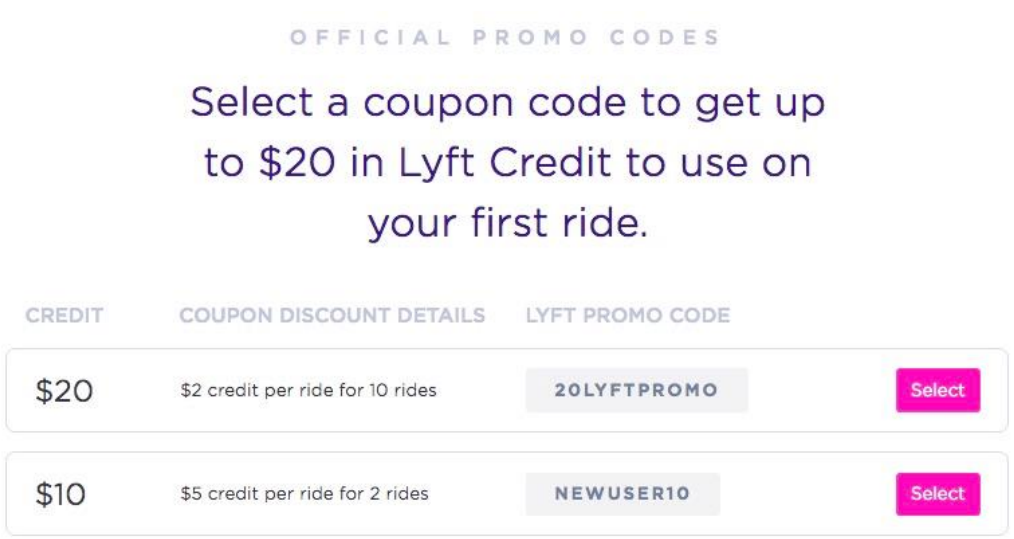
- Use the retention probability to score the new users: promote referral program towards high-quality users & reduce initial acquisition cost on low-quality users
- Spend more marketing budget for targeting at iPhone users at King's Landing for marketing acquisition

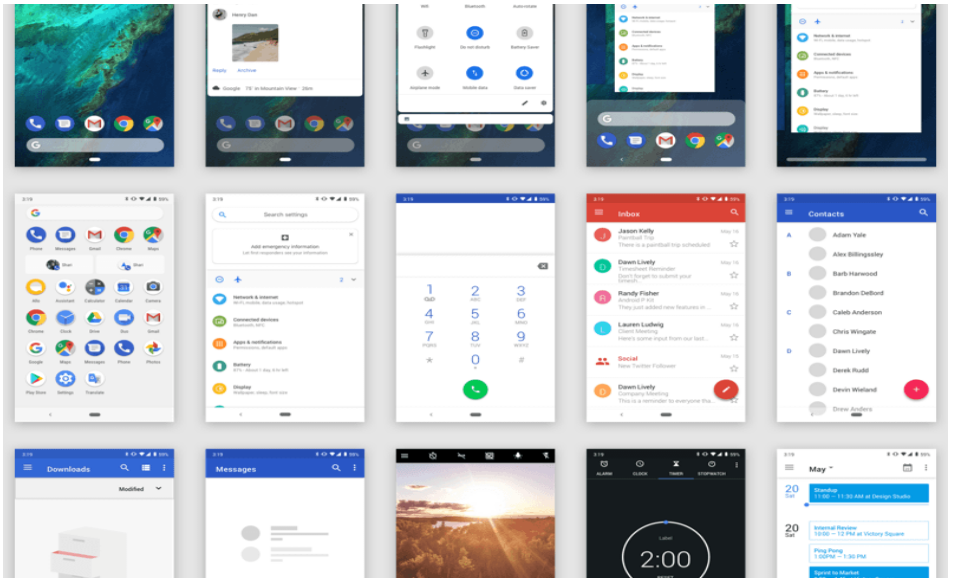


Details on Github: https://github.com/yukaabe/Data-Science-Projects-Portfolio-Repo/tree/master/Data%20Analysis%20Take%20Home%20Challenge%20Ultimate%20Technologies

Solution: Business Insights & Recommendation

- Improve UI experience of the Android app
- Experiment on promotions to motivate users to use more of the app during their first 30 day trial
- Adjust the pricing and user experience of surge multiplier function for long-term user retention





Details on Github: https://github.com/yukaabe/Data-Science-Projects-Portfolio-Repo/tree/master/Data%20Analysis%20Take%20Home%20Challenge%20Ultimate%20Technologies

THANK YOU

MORE DETAILS ON GITHUB:

HTTPS://GITHUB.COM/YUKAABE/DATA-SCIENCE-PROJECTS-PORTFOLIO-REPO/TREE/MASTER/DATA%20ANALYSIS%20TAKE%20HOME%20CHALLENGE%20ULTIMATE%20TECHNOLOGIES