Ultimate Technologies Challenge

Part 1: Exploratory Data Analysis for User Logins

Key Observations:

- Clear daily cycles were observed based on the counts of user logins at different time intervals of the day, shown in Figure 1.
- Users were **highly active** during the following time periods of the day:
    - At noon 11:00 to 12:30 and late night time 20:45 to 05:15
- Users were **active** for:
    - Afternoon and night time 12:30 to 20:45
- Users were **not active** for:
    - Early morning 05:15 to 11:00
- The time period with the most user logins was 22:30 to 22:45.
- The time period with the least user logins was to 19:45 to 20:00.

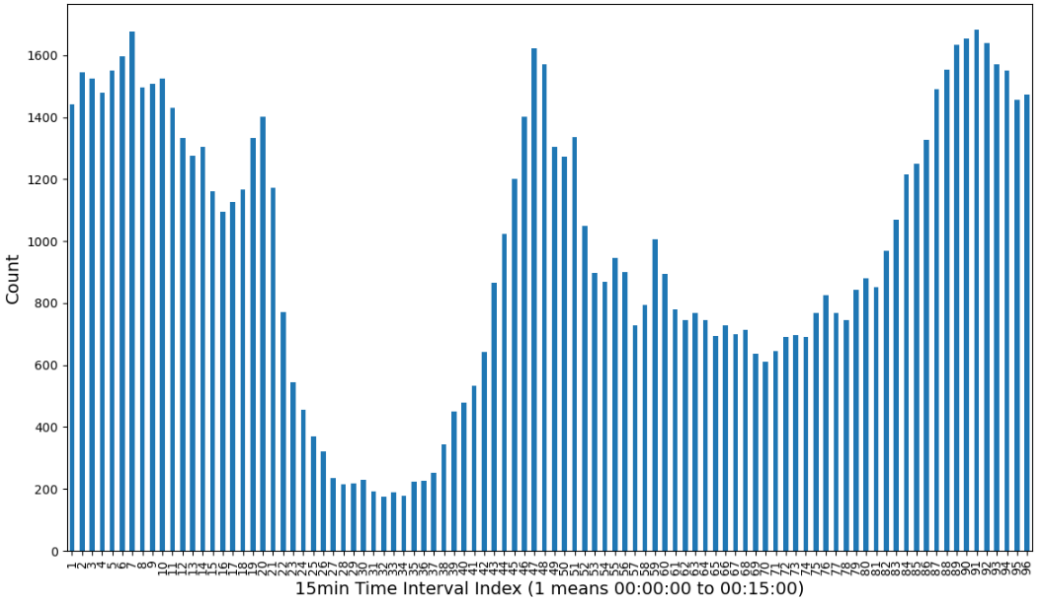


Figure 1: User Logins of Different Time Intervals of a Day

Data Issue:

The logins data started on 1970-01-01 which doesn't look right. It should not be that old. This could be related to mistakes when converting the raw timestamps to the current ones. Usually the raw timestamps are stored as time intervals from a reference date, '1970-01-01 00:00:00' if Unix timestamps are used.

There were issues on the data. The daily cycles should remain true. But the exact time period of “highly active”, "active" and "not active" can be questionable.

Part 2: Experiment and Metric Design for A New Policy

1. The key measure of success of the experiment

I will choose the **number of trips that travel between the two cities of each driver**. Because this is the straightforward metric of interest. The managers proposed this new policy to reimburse all toll costs for drivers in order to encourage the driver partners to be available in both cities. Plus, the daily number of trips that travel between two cities of each driver is an easily measurable property.

Besides the key measurement metric above, **the average customer rating of each driver** can be considered as the secondary metric because it plays an important role in determining if the new reimbursement policy should be implemented or not.

2. Experiment Design and Recommendations

Experimental Design and Evaluation

Two groups, control and treatment, are needed for this experiment. Each group should have the **same number of drivers randomly selected** from both cities.

- **Control group**: No toll reimbursement offered
- **Treatment group**: Toll reimbursement offered

The number of drivers should be large enough to minimize both type 1 and type 2 errors. It can be estimated by using legacy data, 5% as type 1 error and 0.8 as power.

The experiment needs to run for **at least 2 months** to collect enough data to evaluate the key measurements for the two groups.

After collecting enough data, **statistical tests** are required to evaluate the difference of the two groups. **Bootstrapping** techniques can be used to evaluate if there is any significant difference between distributions of the key metrics in the two groups. If the distributions are close to normal distributions, **T-test** can be used to compare the mean or median number of the key metrics.

Recommendations

If there is a **statistically significant increase in the number of trips traveling between two cities in the treatment group**, then I will recommend this new imbursement policy which should be able to stimulate driver partners available in both cities. If there is no significant increase, then I will not recommend this new policy.

If there is significant increase in both the cross-city trips and the average customer ratings in the treatment group, then I will confidently recommend this new policy, which might not only increase driver availability but also customer satisfaction.

However, there are some caveats about this experiment and the recommendations. For this experiment, **external factors** such as holidays, weather are not taken into consideration. The experiment should run in the same time period for the two groups and for a longer time period to minimize such external impact. Or take care of them in the analysis.

The experiment **lacks long-term sustainability**. The experiment should be continuously monitored if possible to ensure the increase in key metrics is sustainable.

There is **no cost evaluation** in this experiment. The cost of the imbursement and the added revenue should be analyzed to further justify the implementation of this new policy.

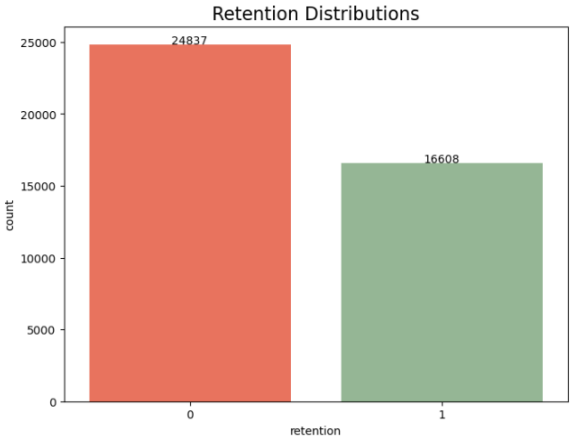
Part 3: Predict Retention

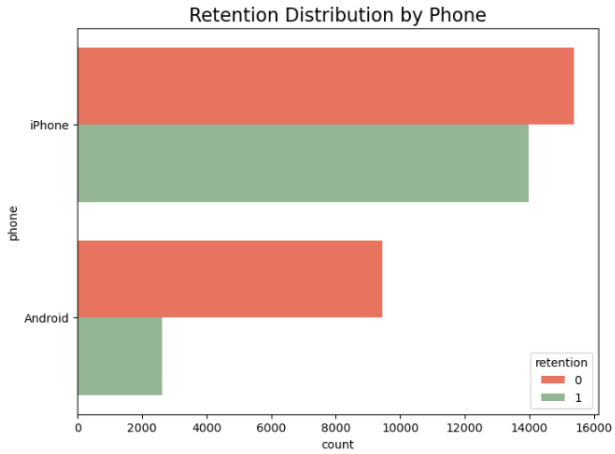
1. Data Wrangling and EDA

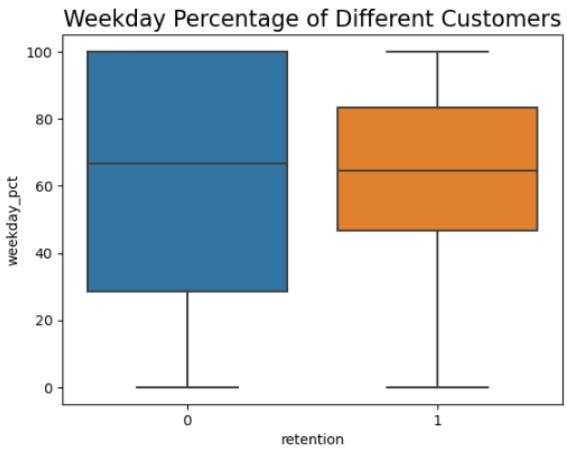
Missing data were dropped since there was sufficient data for modeling. For **numerical features** with skewed distributions, logarithm was applied. For **categorical features**, one-hot encoding was applied with the drop first option.

A user is considered as retained if he/she was “active” (i.e. took a trip) in the preceding 30 days. **A binary retention label was created** based on if the user’s last trip date is within the 30 days period before the latest last trip date. The latest last trip date was assumed to be the reference date to determine retention.

Below are plots generated during EDA. The first plot showed the distribution of retention in the whole data. The second and third plots showed how retention distribution changed with different categorical features. The last plot showed how numerical features differed between retained customers and non-retained customers.



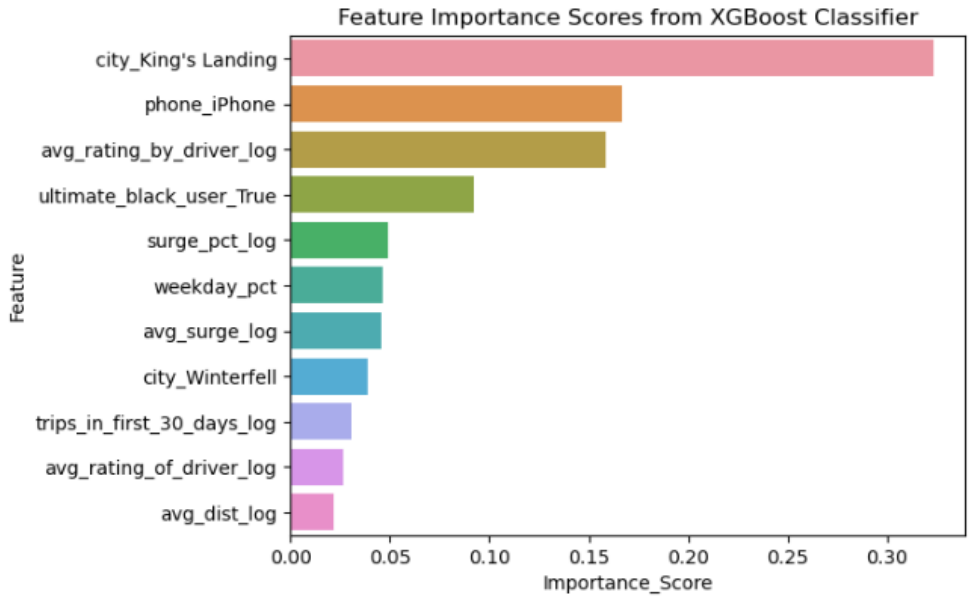




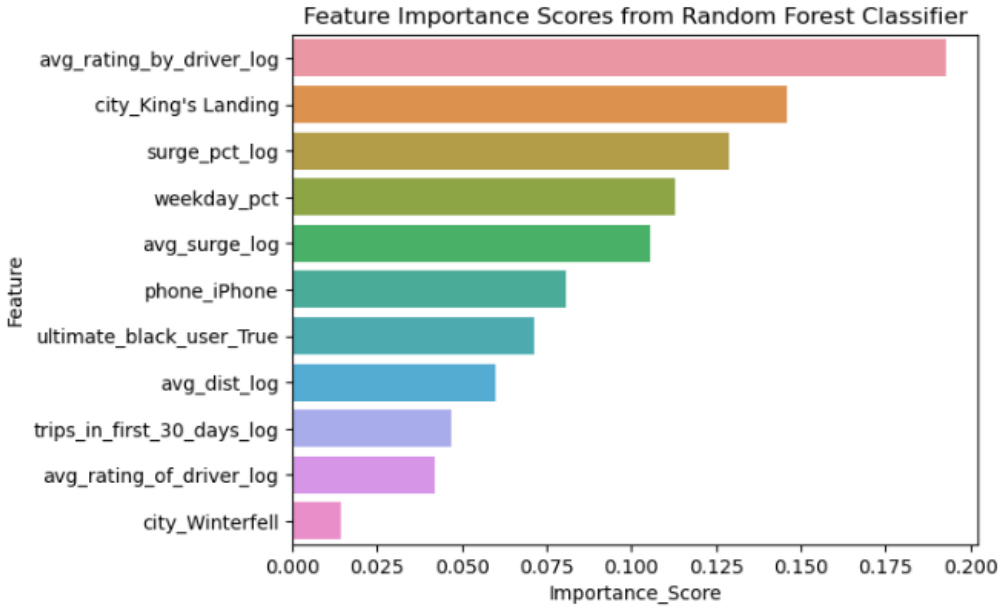
2. Model Building and Evaluation

Both **Random Forest** and **Gradient Boosting Machine** techniques were used for modeling. Two classification models were built with similar accuracy, ~0.78. Random Forest and XGBoost were chosen due to their capabilities to output feature importance scores, which could guide business recommendations. And they don't require additional normalization on the data since no need for similarity or distance calculations. They both perform well on classification tasks due to the ensemble nature of Random Forest and boosting technique in XGBoost.

Feature importance scores from the two models were provided in the figures below.



These features showed high scores from both models: **city, phone, ultimate_black_user, avg_rating_by_driver, surge_pct, avg_surge, and weekday_pct**. How they impact the retention were further investigated by EDA to guide business decisions to improve customer retention.



3. Business Recommendations

Some recommendations can be made to improve long-term customer retention for Ultimate:

- Consider **increasing the marketing or promotions in King's Landing** to attract more customers since there are better chances they will be retained.
- Research why King's Landing behaves differently from other cities for potential strategies to improve retention rates in other cities.
- Investigate the difference between the **app design** on iPhone and on Android then improve app design on Android to increase Android user retention.
- **Promote ultimate black to new users** can help increase user retention.
- Consider **offering weekday discounts** to improve customer's commuting use and to increase customer retention.
- **Implement dynamic pricing strategy more often** since the retained customers showed high acceptance on surging rates.
- Consider increasing marketing on benefits of surging rates, like more ride options and less waiting time, to attract more customers willing to take rides with surging rates.