Michael Walker
October 9th, 2024
Ultimate Technologies Inc. Take-Home Challenge

Part 2 - Experiment and metrics design

Question 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:The key measure of success for this experiment should be ***the increase in the percentage of driver partners serving both cities*** (Gotham and Metropolis) before and after the toll reimbursement policy is implemented.

Why I chose this metric:

1. **Direct Alignment with the Experiment's Goal:** The objective of the experiment is to encourage drivers to be active in both cities. Measuring how many drivers start serving both cities would directly quantify whether the toll reimbursement incentive is working.

2. **Behavior Change Indicator:** Tracking the increase in cross-city driving behavior will highlight the effectiveness of removing the toll cost barrier. A rise in the percentage of drivers serving both cities would suggest that the toll cost was a deterrent, and the reimbursement strategy is motivating drivers to expand their service area.

3. **Impact of Incentive:** By comparing the data before and after toll reimbursement, the metric will clearly show the impact of the incentive. It captures the behavioral change of drivers who previously avoided crossing the toll bridge but now do so due to the financial incentive.

Question 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.

Phase 1: Set the Context for the Experiment

Purpose: The goal of the experiment is to **determine whether reimbursing toll costs encourages driver partners to serve both cities** (Gotham and Metropolis) more frequently. We hypothesize that by removing the toll cost barrier, drivers who previously worked exclusively in one city will expand their service to both cities.
Treatment: The treatment is the implementation of **toll cost reimbursement** for drivers who serve both cities.
Hypothesis: Drivers who are reimbursed for toll costs will increase their service across both cities compared to those who are not reimbursed.

Primary success metric: The percentage increase in drivers serving both cities (Key Metric). This directly ties to the experiment's goal of encouraging cross-city driving.

Guardrail Metrics:
Driver satisfaction scores: To ensure that the toll reimbursement does not negatively impact driver satisfaction (for example by adding inconvenience).
Average trip duration: To ensure that trips between cities don't lead to overly long drives, which might decrease driver retention or increase customer wait times.

Tracking Metrics:
Number of cross-city trips: Helps track movement and whether drivers are actively using the toll bridge to serve both cities.
Revenue per driver: To see if cross-city driving results in higher earnings, incentivizing further cross-city engagement.

Phase 2: Design the Experiment

Randomization:
Unit of Randomization: Randomize at the driver level, dividing drivers into four groups:

1. Control Group in Gotham: Drivers in Gotham who do not receive toll reimbursement.
2. Treatment Group in Gotham: Drivers in Gotham who receive toll reimbursement for cross-city trips into Metropolis.
3. Control Group in Metropolis: Drivers in Metropolis who do not receive toll reimbursement.
4. Treatment Group in Metropolis: Drivers in Metropolis who receive toll reimbursement for cross-city trips into Gotham.

Duration of the Experiment: The experiment will run for at least 4 weeks, covering both weekdays and weekends to capture different traffic patterns in both cities.

Geography-based Assignment: Both cities, Gotham and Metropolis, should be included in the treatment group to ensure that drivers from both cities can benefit from toll reimbursement.

Power Analysis: Conduct a power analysis to determine the sample size of drivers needed to detect a meaningful difference in cross-city driving. We would use a standard significance level ($\alpha = 0.05$) and a power of 80%.

Tracking Driver Activity: Use GPS data to track driver movements between Gotham and Metropolis, focusing on trip origin, destination, and whether they crossed the toll bridge.

b) Statistical Tests

1. **Hypothesis Test: Conduct a two-sample t-test** to compare the **mean percentage of drivers serving both cities between the control and treatment groups**. This tests whether there is a statistically significant increase in cross-city driving due to toll reimbursement.

2. **Alternative Test: If there are pre-existing biases** (e.g., certain drivers were already more likely to drive between cities), consider using a **difference-in-differences (DiD) approach**. This will help control for these biases by measuring the difference in cross-city driving before and after the toll reimbursement for both the control and treatment groups.

3. **Guardrail Metrics Analysis:** Run additional t-tests on the guardrail metrics (e.g., driver satisfaction scores) to ensure that no significant harm is done to these secondary factors.

4. **Multiple Metrics:** Use the False Discovery Rate (FDR) correction to account for multiple hypothesis testing since we will be measuring several metrics.

Phase 3: Post-Experiment Analysis

c) Interpretation of Results

Positive Outcome:

If the t-test shows a statistically significant increase in the percentage of drivers serving both cities ($p < 0.05$), and guardrail metrics (e.g., driver satisfaction) are unaffected or improved, the toll reimbursement program can be considered effective. The recommendation to city operations would be to roll out the toll reimbursement policy.

Secondary Insights:

Look at segment analysis to see if certain groups of drivers (e.g., drivers who normally work at night) benefit more from the toll reimbursement than others, which could lead to a more targeted toll reimbursement program.

Non-Significant Results:

If the p-value is >0.05, the increase in cross-city driving is not statistically significant. At this point, recommend extending the experiment (with larger sample size) or adjusting the incentive structure (e.g., provide additional bonuses for cross-city trips).

Negative Impact on Guardrail Metrics:

If guardrail metrics like driver satisfaction or average trip duration show significant negative effects, consider revising the implementation of the toll reimbursement program (e.g., offer higher reimbursements to drivers to offset any dissatisfaction).

Caveats:

Novelty Effect:

There may be an initial spike in cross-city driving due to the novelty of the reimbursement program. It's important to run the experiment long enough to capture sustained behavior changes, not just short-term excitement.

External Factors:

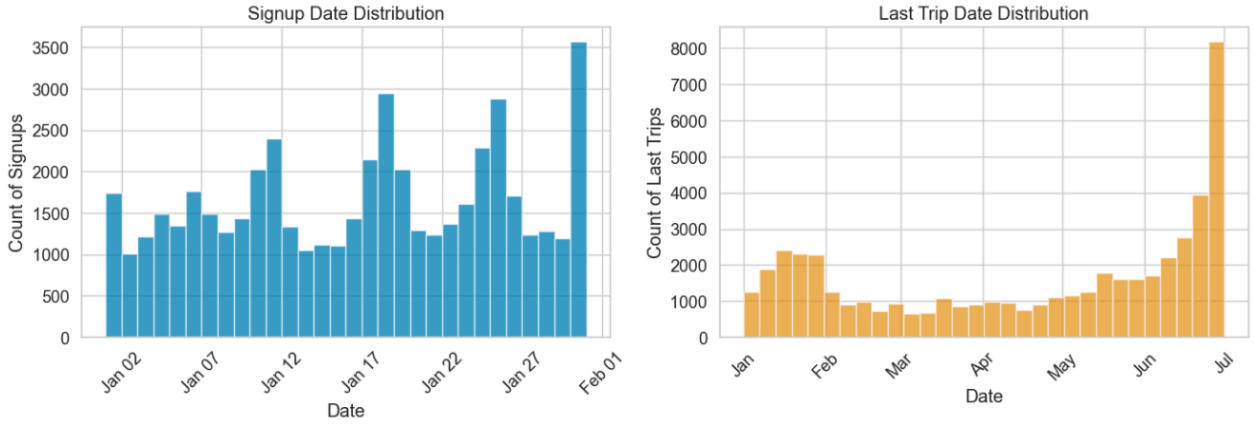
External variables, like weather or economic conditions, could impact driver availability or demand in one city, potentially skewing results. We can include a pre-experiment control period to account for these factors if we do not have the present data.

Part 3 - Predictive modeling

1. Perform any cleaning, exploratory analysis, and/or 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?

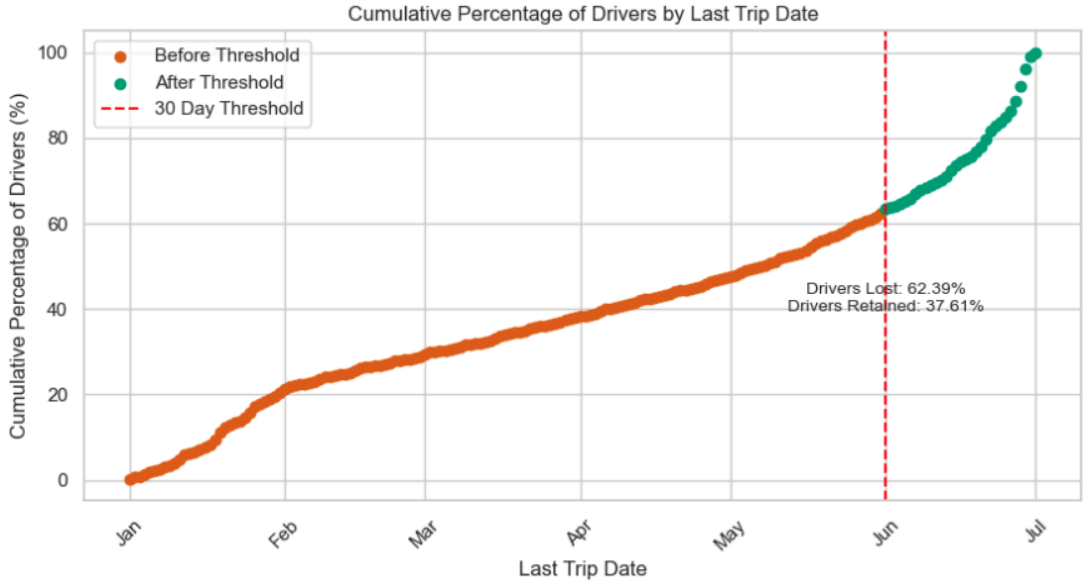
Signup Date + Last Trip Date Distribution:

- Signup Date Distribution: Signup spikes during the weekend.
- Last Trip Date Distribution: Many drivers have their last trip in January, then a steady loss of drivers from months 2-4 (Feb through April). Months 5 (May) and 6 Last Trip Dates start to rise.



Driver Retention:

- 37.61% of drivers who signed up in January were still active going into their 6th month on the system.
- Nearly 20% of drivers had their last trip date in January, not even lasting a month.
- From Feb to about June the dropout rate appears steady.

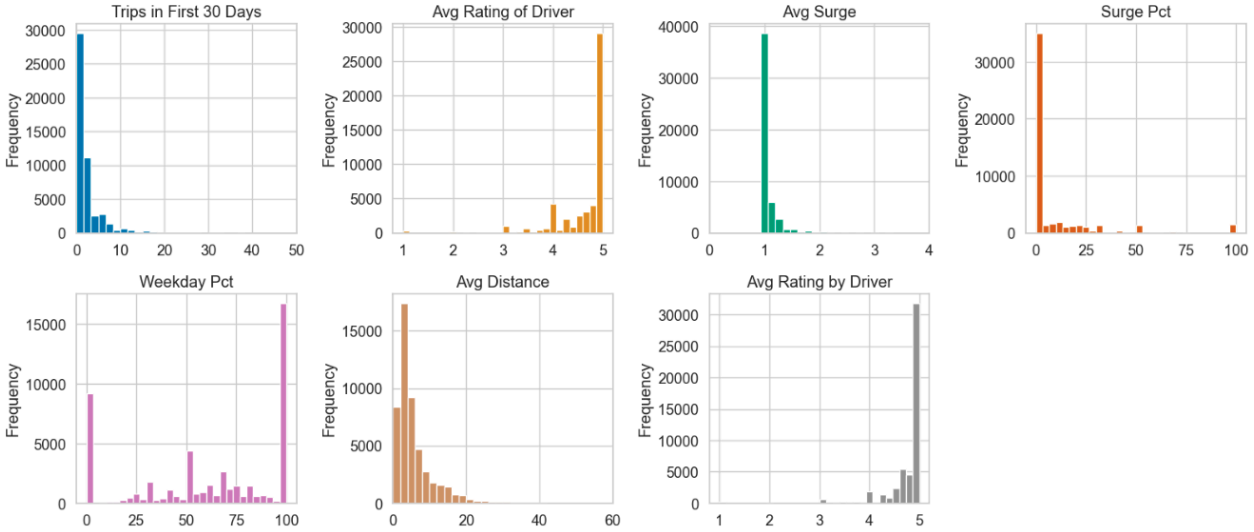


Summary of Driver Behavior and Performance Distributions

- Trips in the First 30 Days: Most drivers have under 5 trips, a few have more than 10.
- Avg Rating of Driver: Most drivers receive a 5 star rating by far, then a huge majority are 4+, a few lower.
- Avg Surge: Above 80% of drivers have an Avg. surge of 1.0 so are not taking advantage.
- Surge Pct: Majority of drivers have a Surge Pct of 0%.

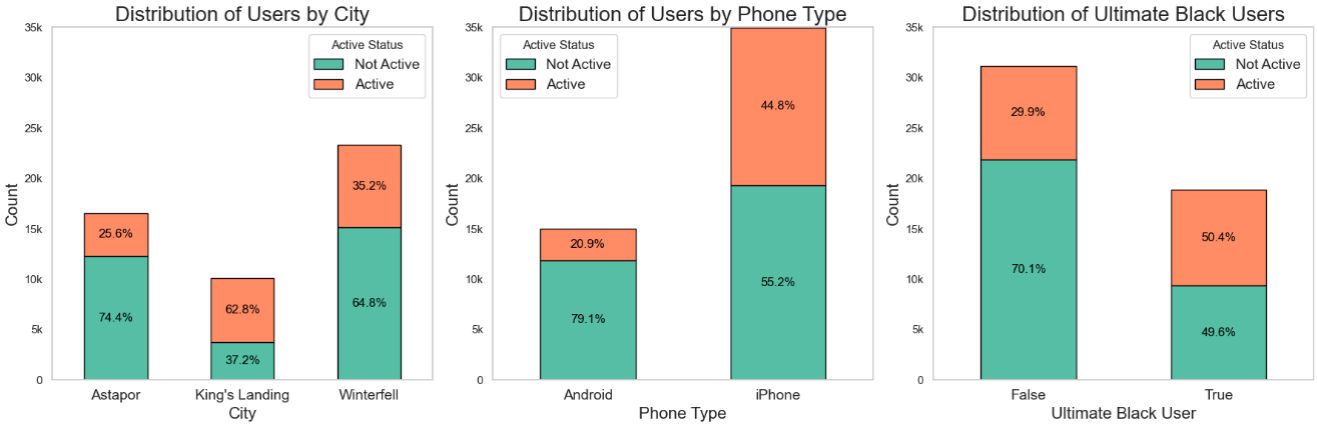
- Weekday Pct: Drivers are most likely to have either 0% or 100% of Weekday driving.
- Avg Distance: Most trips are just a few miles. Huge majority under 20 miles.
- Avg Rating by Driver: Drivers give a huge majority of 5 star ratings of their trips.

Key Takeaways: Most features exhibit non-normal distributions, with a significant concentration of values at the lower ends, such as under 5 trips in the first 30 days and average distances predominantly under 20 miles.

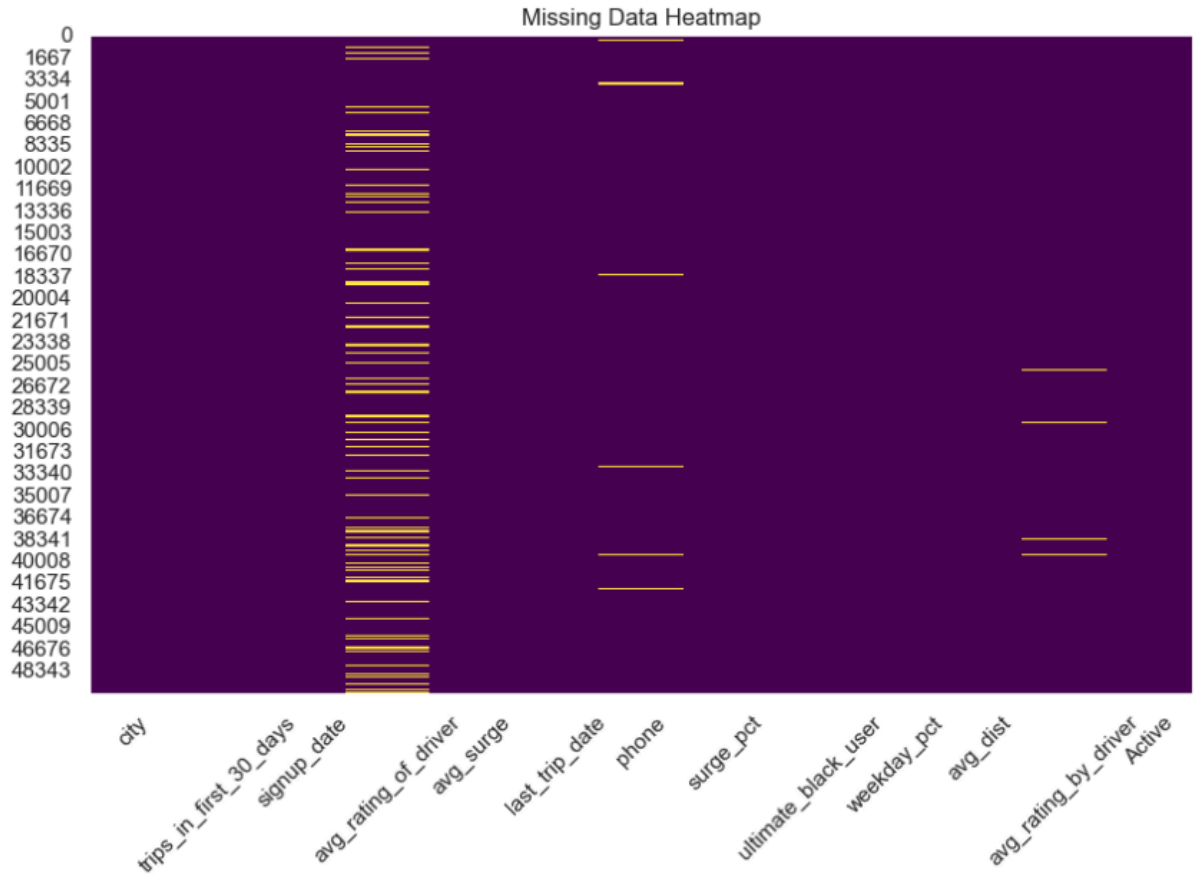


Summary of City, Phone Type, and Ultimate Black User Distributions

- Winterfell has most drivers, then Astapor, then King's Landing. Winterfell driver's have the highest retention.
- iPhone dominates with 35k out of 50k drivers. Also higher retention (44.8% vs. 20.9%) retention.
- Majority of users are not Ultimate Black Users, however Ultimate Black has much higher retention rate.



Missing Values Analysis:

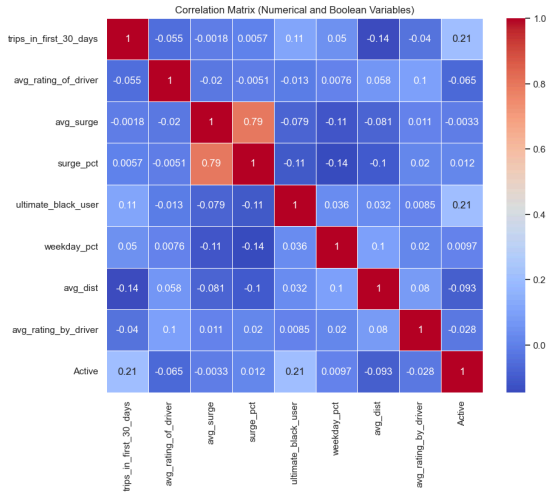


Feature	Missing Percentage (%)	Mean (Active=True)	Median (Active=True)	Mean (Active=False)	Median (Active=False)
trips_in_first_30_days	0.0	3.3	2.0	1.7	1.0
avg_rating_of_driver	16.244	4.6	4.8	4.6	5.0
avg_surge	0.0	1.1	1.0	1.1	1.0
surge_pct	0.0	9.2	0.0	8.7	0.0
weekday_pct	0.0	61.4	64.3	60.6	69.6
avg_dist	0.0	5.1	3.7	6.2	4.0
avg_rating_by_driver	0.402	4.8	4.8	4.8	5.0

Choose to fill in missing values with the median of their respective Active or Non-Active status due to distributions not being normal. For phones I fill missing values in the phone column by using the most common phone type (mode) for drivers based on their Active status (either active or inactive)

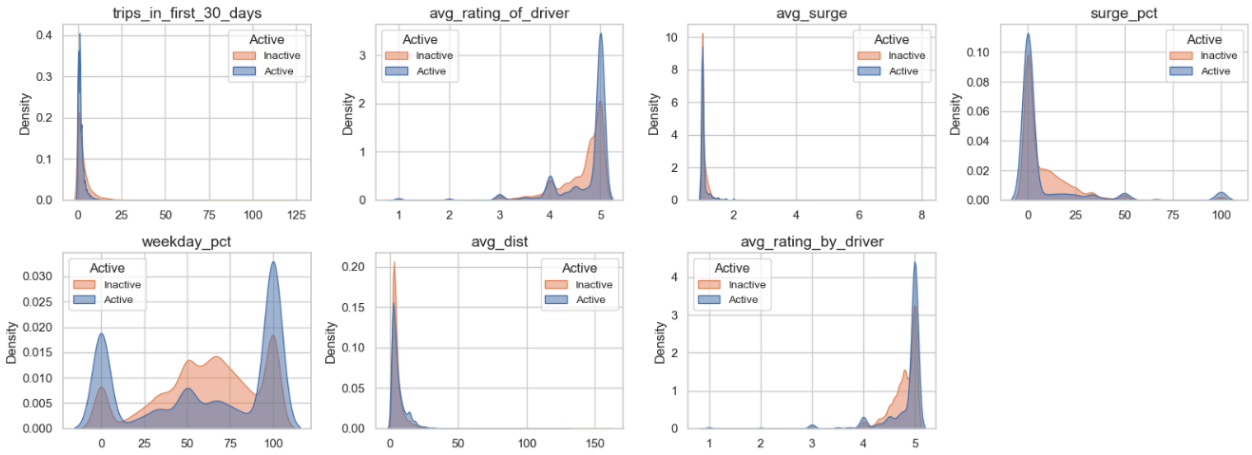
Correlation Matrix:

- 'Active' feature is the feature created to identify drivers who have driven in the past 30 days.
- Trips in the first 30 days and Ultimate Black User were the most highly correlated variables with respect to our target feature.



Active vs. Inactive Drivers KDE Plots:

- Trips in the first 30 days - Active and Inactive Drivers look mostly the same.
- Avg Rating of Driver - Active drivers tend to have a higher rating, more centered on 5.
- Avg Surge - Active and Inactive Drivers look mostly the same.
- Surge Pct - Surprisingly, Inactive Drivers had a slightly higher surge pct it looks like.
- Weekday Pct - Active drivers tend to work 0% or 100% of weekdays, fewer mixed.
- Avg Distance - Mostly the same.
- Avg Rating by Driver - Active tend to have more 5 star ratings by driver.
- **Key Takeaways:** Some of these KDE Plots should be useful for feature engineering and modeling, building stronger features to predict our target feature.



==========

Part 3 - Predictive modeling

2. Build a predictive model to help Ultimate 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.

Approach:

1. **Data Preparation**
    - **a. Feature Engineering:** Create relevant features that might influence model activity.
    - **b. Target Variable:** Binary variable indicating whether a user is active in their 6th month.
    - **c. Scaling of Features:** Standard scaler (mean of each feature is 0, and stdev is 1.0)
    - **d. Handling Class Imbalance:** Apply class weights for class imbalance

2. **Model Selection**: Looking to model complex interactions with interpretable results
    - **Baseline Model:** Mean
    - **Machine Learning Model Types**:
        - **Logistic Regression**: A good baseline for binary classification problems.
        - **Random Forest**: An ensemble method that reduces overfitting and improves accuracy.
        - **XGBoost**: An efficient and effective implementation of gradient boosting, often yielding high performance in classification tasks.
    - **Alternatives but not chosen**: SVM, Neural Networks, Naive Bayes
    - **Final Choice**: Based on preliminary performance, select the best-performing model.

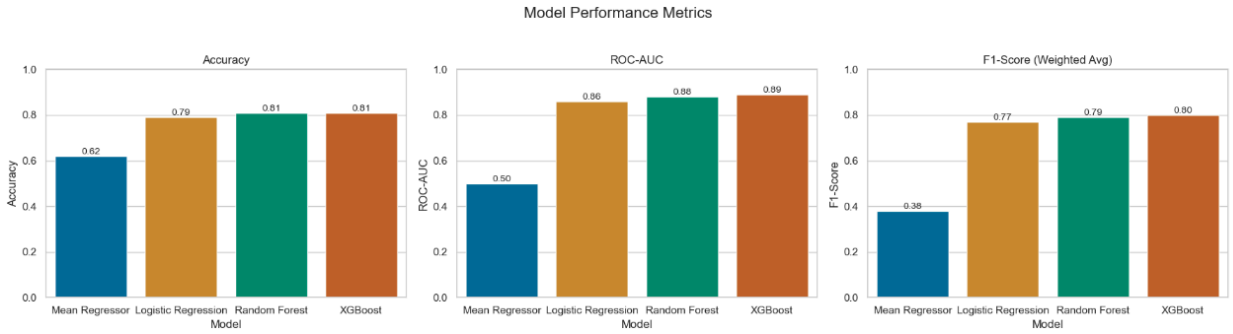
3. **Model Evaluation**:
    - **Cross-Validation**: 5 k-fold cross validation using a training/test split of 80/20%.
    - **Metrics**:
        - **Accuracy**: Overall performance measure.
        - **Precision and Recall**: To understand the balance between false positives and false negatives.
        - **F1 Score**: The harmonic mean of precision and recall, especially important for imbalanced classes.
        - **ROC-AUC Score**: To assess the model's ability to distinguish between classes.

4. **Hyperparameter Tuning**: Randomized Grid Search

5. **Feature Importance**:
    - **SHAP Values**: Use SHAP (SHapley Additive exPlanations) to interpret model predictions and understand the impact of different features on the outcome.

6. **Concerns:**
    - **Imbalanced Classes**: If the number of active vs. inactive users is significantly skewed, this may lead to biased model predictions. Techniques such as oversampling, undersampling, or using class weights can help.
    - **Overfitting**: Ensuring the model generalizes well to unseen data is crucial. Techniques like cross-validation and regularization can mitigate this risk.

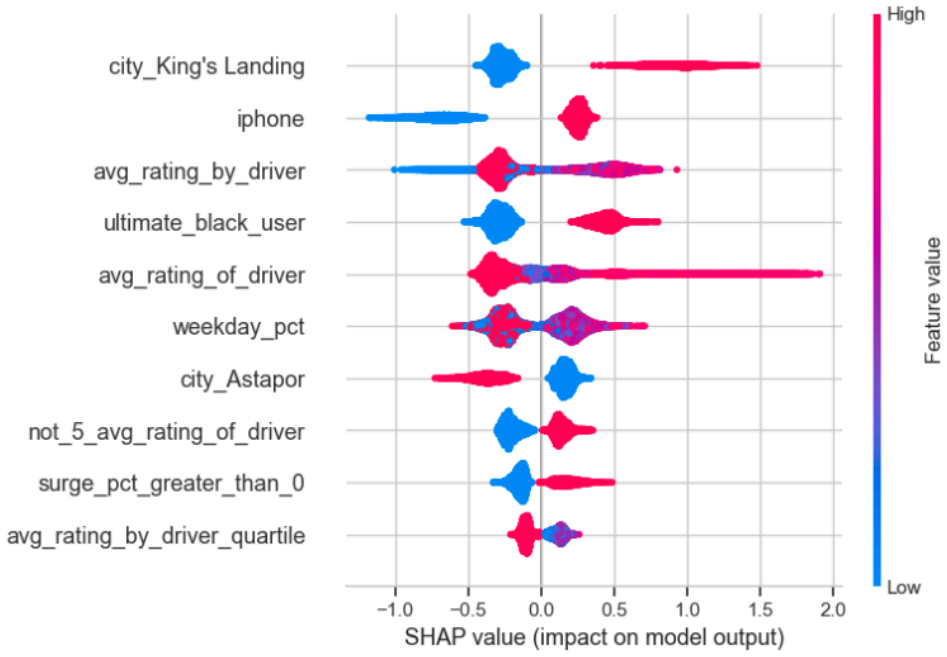
Modeling Results:



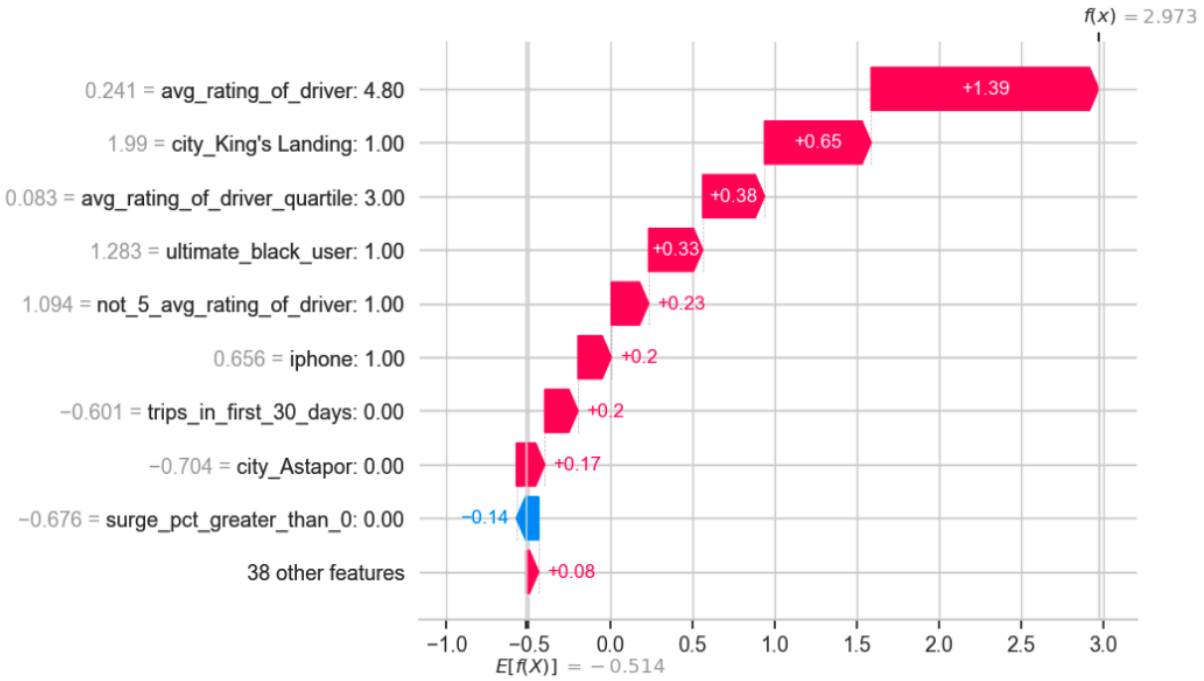
Model	Accuracy	ROC-AUC	F1-Score (Weighted Avg)
Mean Regressor	0.62	0.50	0.38
Logistic Regression	0.79	0.86	0.77
Random Forest	0.81	0.88	0.79
XGBoost	0.81	0.89	0.80

XGBoost was chosen as the best model in terms of accuracy, ROC, and F1 score.

SHAP Values for top Features:



Waterfall Plot for a specific instance:




```
Probability of positive class for instance 0: 0.9513
Expected value (baseline prediction): -0.5138
SHAP values sum for instance 0: 3.4868
Final prediction for instance 0: 2.9730
```

==========

Part 3 - Predictive modeling

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).

Based on the SHAP values among the top features, the model suggests the following to improve it's long term rider retention:

1. **King's Landing:** Focus marketing efforts in King's Landing, drivers show a higher likelihood of retention in this city.
2. **iPhone Users:** Tailor promotions and loyalty programs specifically for iPhone users to enhance their retention rates.
3. **Ultimate Black Users:** Implement targeted engagement strategies to encourage users to join Ultimate Black users, who have shown a positive retention trend.
4. **Astapor:** Investigate the factors leading to lower retention in Astapor and develop strategies to improve driver satisfaction in this area.
5. **Surge Percentage:** Optimize surge pricing strategies to encourage more rides during high-demand periods, which can positively impact driver retention.