Ultimate Take Home Challenge Report
by Scott Lew

Part 1 Exploratory Data Analysis

The data of login times of Ultimate users was read from a JSON file, converted into DateTime objects and the data was used to create a Pandas dataframe. Using a Pandas dataframe seemed the easiest way to analyze and create visualizations.

The month, day, hour, and minutes were extracted from each datetime object corresponding to each login time using date library functions in order to create columns in the dataframe. In addition, quarter of hours were studied by using a function to convert the minute of the login to a corresponding quarter of the hour, where each quarter consists of 15 minute intervals.

Analysis of logins per hour during the day revealed that 10 PM -2 AM, late evening and early morning hours, were the peak periods for rides. Using all the login data for the period of ~3.5 months, the logins per hour revealed 3 peak times. Another similar analysis of logins per hour for a single day in March also revealed similar 3 peak hours of activity by users.

As shown in Table 1 below, analysis of login data that is grouped by hour and quarter of hour revealed the busiest hour and quarter hours of the day in terms of number of logins are: the first, second and third quarter of the 22 hour (10 PM) and the first quarter of the 23rd hour (11PM).

TABLE 1: The hour and quarter hour of the day with the greatest number of logins

Hour	Quarter	Number of Logins
22	**1**	**1746**
22	**2**	**1677**
22	**3**	**1666**
23	**1**	**1677**

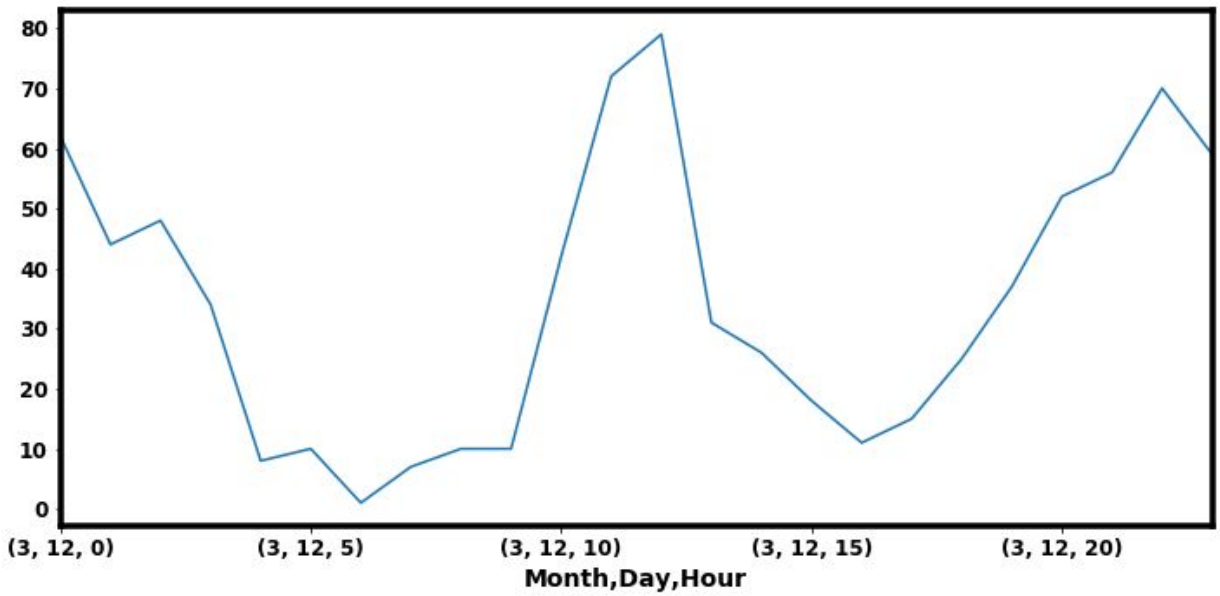

Figure 1. Ultimate ridership activity per hour of day for one day in March.

As shown in the analysis of the data in Table 1, the first, second, and third quarter of 10 PM had the most logins. Moreover, as shown in Figure 1 above which shows logins per hour for one day in March, ridership activity had three peak periods. Analysis of logins per hour of the day for the entire dataset revealed there are also three times of the day where ride demand is the greatest. As shown below in Figure 2, the three peak hours of Ultimate rider usage are: 1 AM, 11 AM and 10 PM.

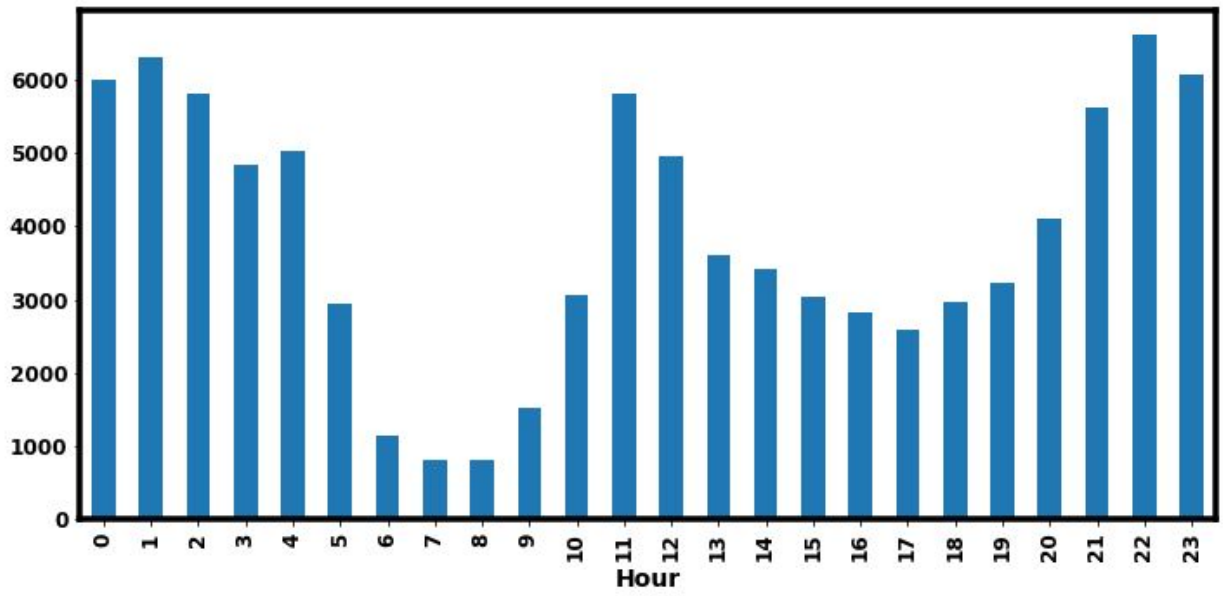

Figure 2. ridership demand per hour of the day for the entire dataset.

Interestingly, the first four months of the year are the busiest in terms of ridership demand, and March is the busiest month of the year.

Part 2 Experiment and Metrics Design

The managers of Ultimate propose an idea to encourage drivers from Gotham and Metropolis to provide service in both cities by offering to reimburse the bridge toll for each driver partner that crosses the bridge to work in another city. A metric is needed to see if Ultimate's idea for encouraging drivers from each city to service riders in another city actually works. It is assumed that the bridge connecting the two cities is the main method or only way of transportation between the two cities.

One proposed metric is the mean number of toll crossings for each direction of the bridge in a time interval such as an hour. Then, this metric would be studied and compared for two periods of time: one period would be before the toll reimbursement idea was proposed and the second would be after the reimbursement idea went into effect. In this case, hypothesis testing of the mean number of tolls crossing per hour for two groups, two time periods in this case. It would be especially useful to study the number of bridge crossings during the hours of greatest rider demand in both Gotham and Metropolis. If the reimbursement of toll policy works as intended, one would expect that bridge crossings as measured by number of bridge tolls would increase during those times when demand for rides in Gotham and Metropolis are the greatest

when compared to the normal number of bridge crossings before the reimbursement initiative was started. For instance, if 2-3 PM is a peak hour of ride service in Metropolis, a comparison of the average number of bridge crossings during that time period for one month before the reimbursement policy would be compared with the average for a month after the reimbursement policy was started. And, If there is no significant increase in bridge crossings during those peak hours, then the idea would seem to be ineffective in encouraging drivers to service both cities.

Part 3 Predictive Modeling

Ultimate is interested in predicting customer retention. To accomplish this goal, supervised machine learning models will be constructed to predict the retention status of Ultimate customers. In other words, Machine Learning methods will predict which users of Ultimate will use the service ~5-6 months later after signing up for the service, long-term users.

The data was supplied in the form of a JSON file. Features such as the city of the user, average distance of trips taken in the first 30 days of using Ultimate, and other relevant information are given.

A delta time metric which is calculated by subtracting the two datetime objects, signup_date and the last_trip_date, and calculating the difference in months and or days. This delta time metric was used to determine if a user was assigned a long-term or short-term status. Using a delta time metric one can achieve a 100% accuracy in classification since the metric is used to assign a class to each user. However, this approach is of little value in determining what other characteristics of a user can predict retention status. Such an approach is analogous to defining obesity by Body Mass Index (BMI) and than using BMI to predict whether a person is considered obese or not. In other words, there is no need for Machine Learning Classification methods when a simple function based on a single metric can achieve accurate results. So, for the purpose of training Machine Learning classification models, the time delta metric was not included. Numerical features and categorical features were used to train classification models. Categorical data was converted into numbers with label-encoding methods prior to being used for Machine Learning.

In this case, two classification labels are used: long-term users and short-term users. Different Machine Learning models were tried with mixed results: Logistic Regression failed to predict any long-term users. Decision Tree and Random Forest classifiers are

able to predict long-term users, but the XGBoost method had the best accuracy of 79% when categorical features were included in the model. The random search method was utilized for parameter tuning of the XGBoost model because random search can usually find a good combination of hyperparameters in fewer iterations than the grid search method. This random search optimization improved the F1 score for the long-term class.

Ultimate Users EDA

A customer is considered retained if he or she takes a trip in the first 30 days after signing up for the service. For this cohort of users, ~69% of customers are retained.

As shown in Figure 3 below, iPhone owners are more likely to be long-term users of Ultimate's ride service than Android phone owners.

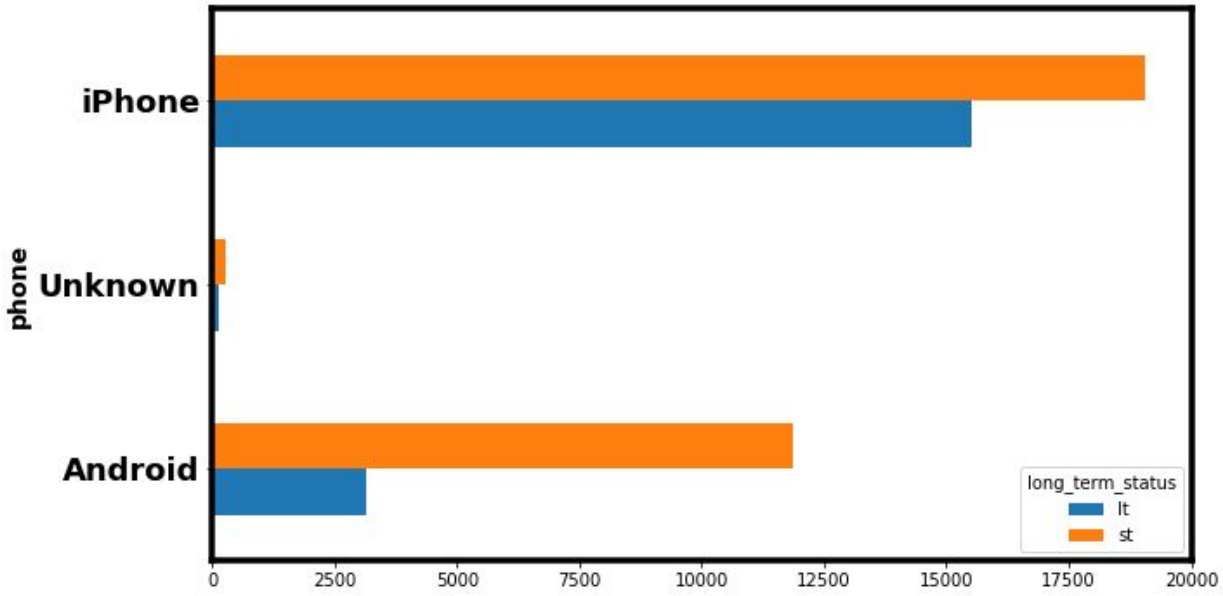

Figure 3. Customer Retention Status Grouped By Phone of User

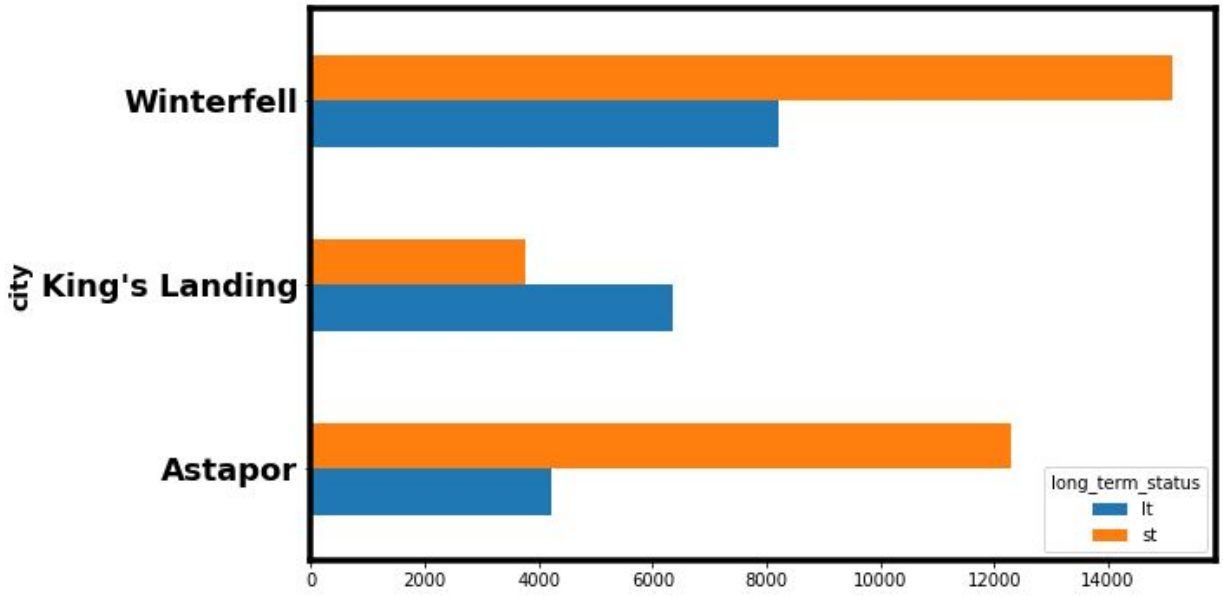

Figure 4. Customer Retention Status Grouped By City of User

An analysis of Ultimate user distribution for all three cities showed that the residents of the city of King's Landing have the highest ratio of long-term users to short-term users compared to Winterfell and Astapor. Although, Winterfell has the most long-term riders of all the three cities studied.

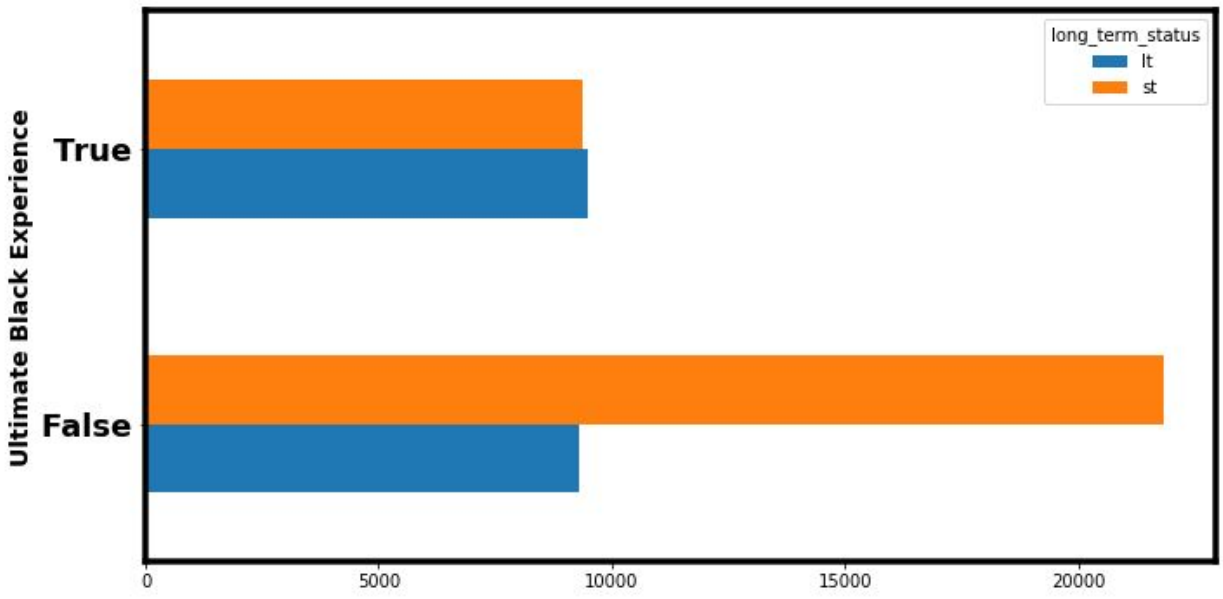

Figure 5. Customer Retention Status Grouped By Ultimate Black Experience

Another insight revealed by the exploratory data analysis is that riders who tried the Ultimate Black Experience are more likely to become long-term users of the service than those who never tried the experience.

TABLE 2: Average Weekday Percentage and Surge Percentage of long-term users vs. short-term users

Status	Weekday Percentage (Mean)	Surge Percentage (Mean)
Long-term	61.4	9.15
Short-term	60.6	8.67

As shown in Table 2, on average, long-term users used the ride service slightly more during the weekday than short-term users, and the long-term users were more willing to pay higher fares with the surge multiplier during busy hours than the short-term users.

Classification

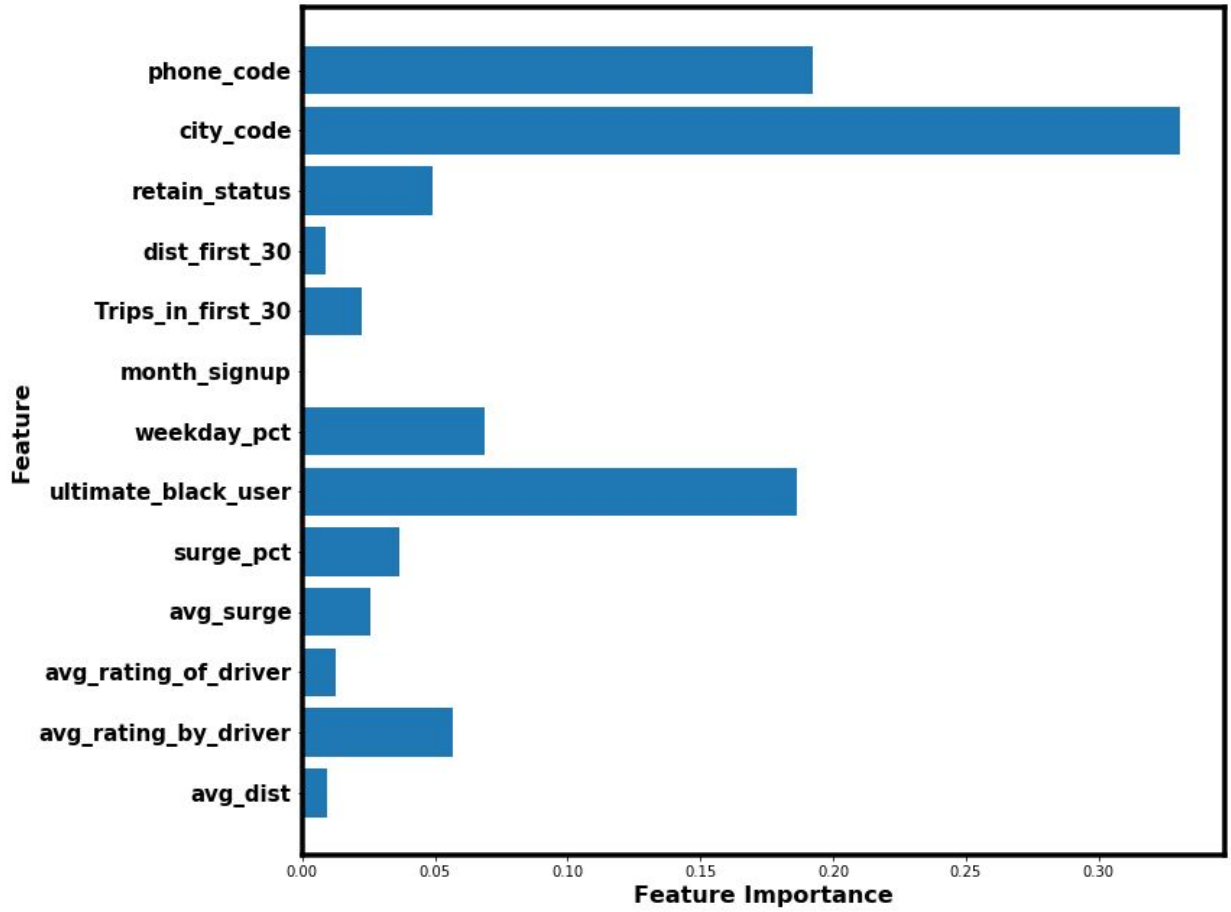

Figure 6. Feature Importance using XGBoost Classifier

Feature importance determined by the XGBoost classifier is shown in Figure 6. The phone and the city of the user are important factors in predicting the long-term retention of an Ultimate rider. As mentioned before, iPhone users were more likely to become long-term users than Android owners, and the residents of Winterfell and King's Landing were more likely to be long-term riders. In addition, the usage of the ultimate black experience by a user was also significant in predicting whether the customer would be a long-term user. Surge_pct and weekday_pct features are higher for long-term users than short-term users on average and significant in predicting long-term use. Another predictor of long-term usage is the average rating of the driver for the customer. Although it is not clear why this is a factor since the average rating by the driver for long-term customers and short-term customers is almost the same.

Classification Metrics

A confusion matrix for the XGBoost model is shown below in Figure 7. As shown, 5354 customers were correctly classified as short-term status and 2518 customers were correctly classified as long-term customers. And, 830 customers were incorrectly classified as long-term customers and 1298 customers were incorrectly classified as short-term customers.

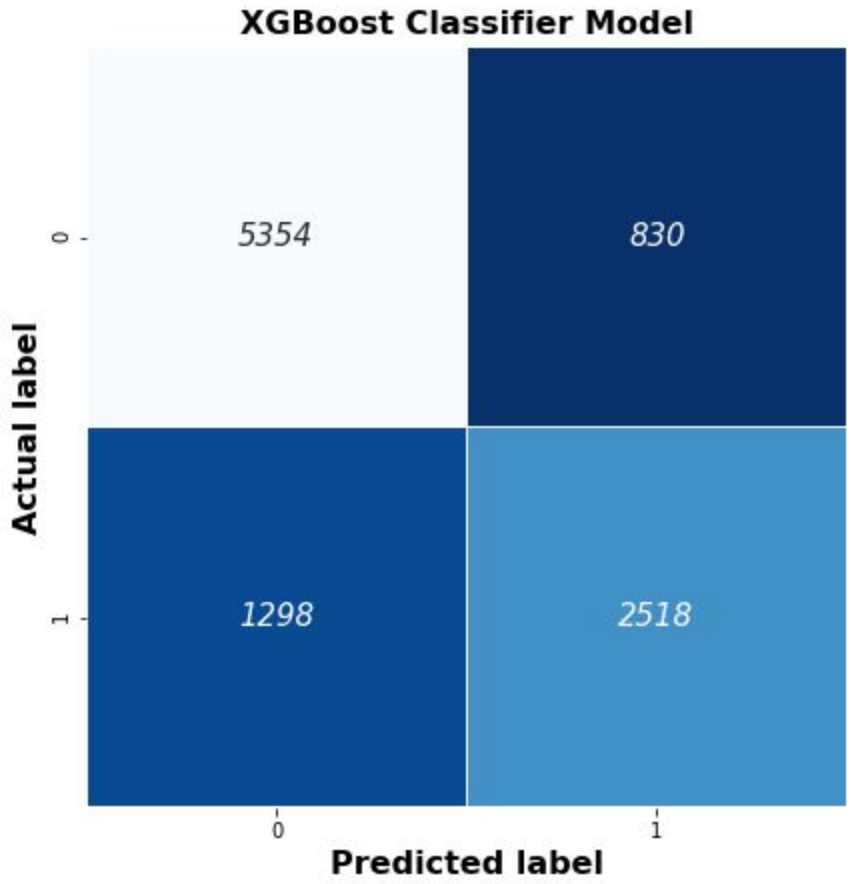

Figure 7. Confusion Matrix for XGBoost Classifier

TABLE 3: Classification Metrics for XGBoost Model

User Status	Precision	Recall	F1 Score	Support
Short Term	0.87	0.81	0.84	6732
Long Term	0.66	0.76	0.71	3268

The F1 scores for long-term and short-term riders are 0.71 and 0.84 respectively. The Area Under the Curve (AUC), a performance measurement for classification, for this XGBoost model is 0.76.

Summary and Recommendations

In summary, the type of phone owned and the city of the user are both useful for predicting the retention of a rider. Another important predictor of long-term retention is the use of the Ultimate Black Experience: users of the experience were more likely to use Ultimate in the future.

Therefore, one suggestion for improving customer retention is to offer the Black Experience at a discount for new Ultimate customers which would hopefully entice riders to stay with Ultimate.