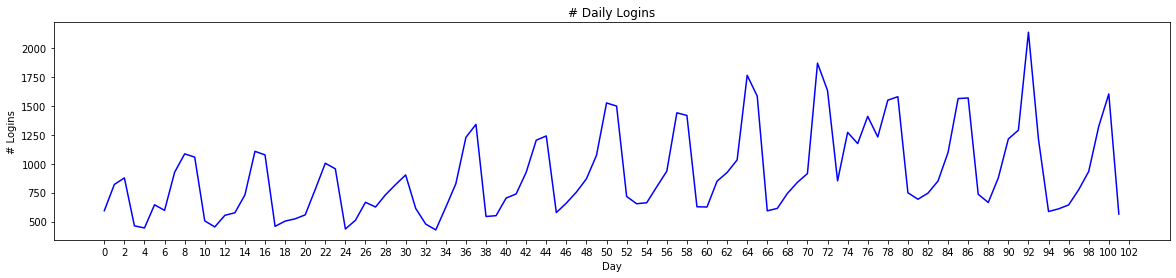
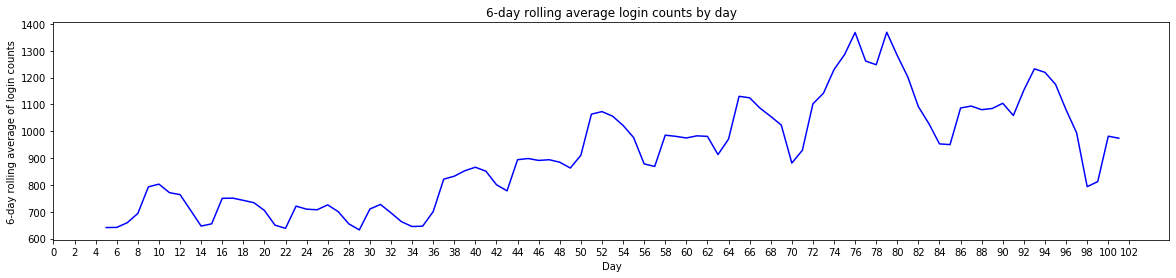
Ultimate Data Science Challenge

# Part 1 ‑ Exploratory data analysis





The second plot above shows the 6-day rolling average for the login counts observed in 102 days based on login time stamps. It clearly shows that there is a spike in the # login counts or demand between 70th and 80th days and then it stars going down again.

# Part 2 – Experiment and metrics design

Objective

1. Increase footfall/revenue/profit in a week
2. Improve awareness
3. Improve employment opportunities for drivers and other staff

## Success Metric

Number of cross city visitors per driver during the weekdays (Monday – Friday). The reason behind choosing this metric is that it can be directly linked to the drivers, which is the population being treated or reimbursed. But at the same time, it ties up with the ultimate business objective of increasing the number of cross city visitors.

## Experiment design

We can perform test and control analysis where the drivers can be divided into two groups:

1. Test group – Contains the group of drivers for which the toll is reimbursed
2. Control group – Contains the group of drivers for which the toll is not reimbursed

### Sample Size

The following factors help determine the sample size needed to run the experiment. Without the right sample size, results cannot be statistically significant.

1. Baseline for Success metric = Number of cross city visitors per driver during the weekdays (Monday – Friday)
2. Incremental % - How much is % incremental expected in the success metric based on pre and post results
3. Confidence level = 70%
4. Significance level = Alpha = 30%
5. 1 – Alpha/2 = 85%
6. Z at 1-Alpha/2 = 1.04
7. Power = 65% = It is the likelihood that the test is correctly rejecting the null hypothesis (i.e. “proving” your alternate hypothesis i.e. proving that the test will have an effect in the success metric). Power is your probability of not making a type II error.
8. Z at Power (65%) = 0.39
9. Minimum sample size for control group = 2\*std^2\*(Z at 1-Alpha/2 + Z at Power)^2/(Baseline success metric^2\*Incremental %^2)

After we determine the sample size for control and test groups, we need to ensure these groups should be similar in terms of the following parameters based on the historical data:

1. Average number of visitors per driver per week (Monday to Friday) in the last year
2. Average number of cross city visitors per driver per week (Monday to Friday) in the last year
3. Average # of hours per driver per week (Monday to Friday) for which the driver is available in the last year
4. Average revenue earned per driver per week
5. Other demographics-based distribution such as neighborhood they live in, age, gender etc.

We can perform pre-alignment t-tests to determine if the test was different from control.

### Experiment Results

Experiment duration:

The experiment can run for 4-6 weeks and then the impact can be measured at different windows such as 3-month, 6-month or 12-month time periods.

Therefore, for a particular time period, let’s say, 12 months:

Lift = Delta change in success metric in test - Delta change in success metric in control

Delta change in number of cross city visitors in test = Number of cross city visitors per driver in test (post period) - Number of cross city visitors per driver in test (pre period)

Delta change in number of cross city visitors in test = Number of cross city visitors per driver in control (post period) - Number of cross city visitors per driver in control (pre period)

Pre period = 12-months prior to when the experiment was run

Post period = 12-month post the experiment was run

We can again perform t-tests if the lift is statistically significant.

* Pooled std deviation = sqrt [std of delta spend (control)^2\*(n of control -1) + std of delta spend (test)^2\*(n of test -1)]
* T-statistic = $ Lift/ [Pooled std deviation \* sqrt (1/n of test + 1/n of control)]

### Recommendations

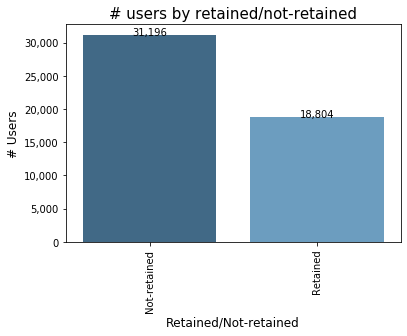
There could be four potential scenarios in terms of experiment results:

1. If the lift > 0 and statistically significant then we can say that they should go ahead and implement this policy on the entire population as we would see similar results on the overall population as well. And they should be able to see an increase in the # of cross city visitors per driver
2. If the lift <= 0 and statistically significant then we can say that they should not go ahead and implement this policy on the entire population as we would see similar results on the overall population as well. And they will not see an increase in the # of cross city visitors per driver
3. If the lift is not statistically significant, then we can’t prove that the same effect will be observed on the overall population. Also the duration of the experiment and sample size have a great impact on how statistically significant the results are

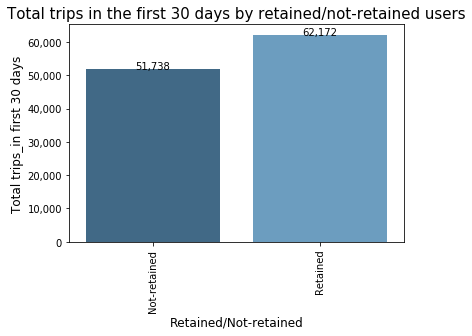
# Part 3 – Predictive Modeling

## Exploratory Data Analysis

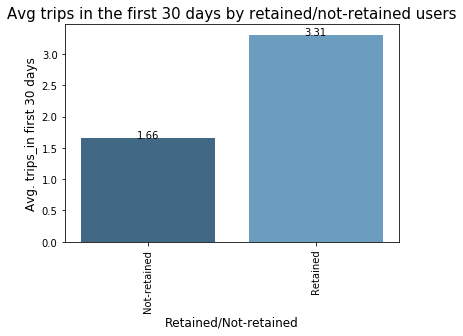
1. 37.6% of the users were retained i.e. 37.6% of the users took a trip in the last 30 days preceding July 2014



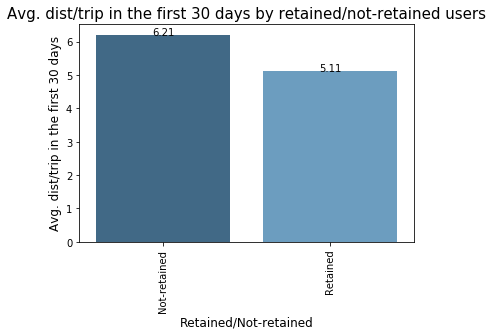
1. Higher total number of trips in the first 30 days by retained users

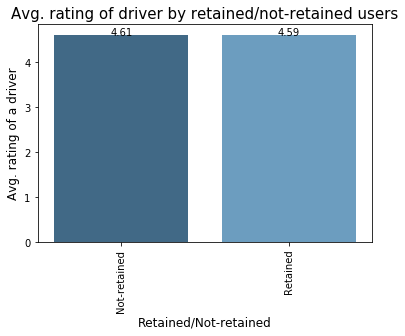


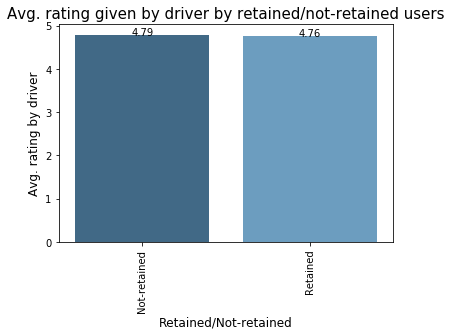
1. Higher average number of trips in the first 30 days by retained users



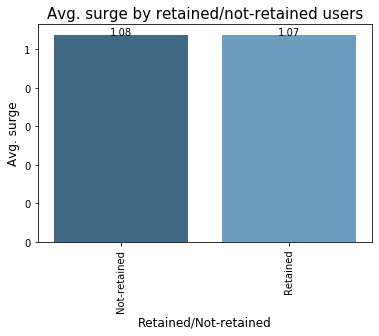
1. Avg. distance per trip in the first 30 days is higher for users which were not retained



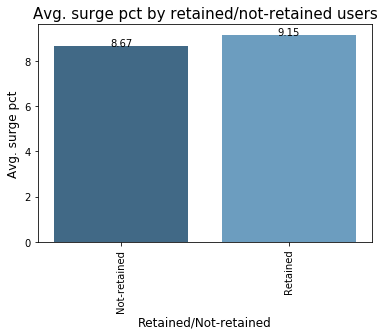
1. Avg. rating of driver is slightly higher for users which were not retained6
2. Avg. rating by driver is slightly higher for users which were not retained



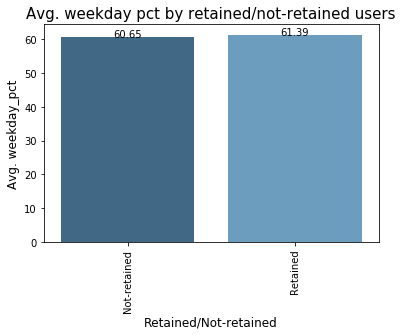
1. Avg. surge multiplier is slightly higher for users which were not retained



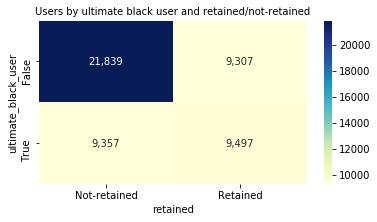
1. The average % of trips taken with surge multiplier > 1 is higher for users which were retained



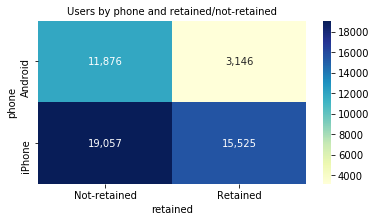
1. The avg. percent of the user’s trips occurring during a weekday is higher for users which were retained



1. 70% of users which were not retained didn't take Black Ultimate in their first 30 days



1. 83% of the users who were retained used iPhone



## Prediction model to predict retention

**Prediction model used:** XGBoost Classifier

Note: Also implemented Logistic Regression model but it has a significant lower accuracy in the OOS data than XGBoost classifier

### Hyperparameters

Best hyperparameters identified based on hyperparameter tuning:

1. learning\_rate = 0.1
2. max\_depth = 6
3. n\_estimators = 100
4. n\_jobs = 10

### Model performance

**Accuracy**

|  |  |
| --- | --- |
| Accuracy score for training data | 79.2% |
| Accuracy score for OOS (test data) | 78.9% |

**Precision and Recall**

|  |  |  |
| --- | --- | --- |
| **Confusion matrix** | |  |
| TN | 8,044 |  |
| FN | 1,831 |  |
| FP | 1,335 |  |
| TP | 3,790 |  |
| **Total** | **15,000** |  |
|  |  |  |
| Precision | 74.0% | % correct positive predictions out of all positive predictions |
| Recall | 67.4% | % correct positive predictions out of all positive predictions that could have been made |

### Variable Importance

|  |  |
| --- | --- |
| **Variable** | **Importance** |
| city | 29.0% |
| avg\_rating\_by\_driver | 20.6% |
| phone | 14.3% |
| surge\_pct | 10.7% |
| ultimate\_black\_user | 9.5% |
| weekday\_pct | 6.1% |
| trips\_in\_first\_30\_days | 3.1% |
| avg\_rating\_of\_driver | 2.7% |
| avg\_surge | 2.3% |
| avg\_dist | 1.7% |