Lyft data challenge – Report

CONCLUSIONS:

1. A driver's lifetime value (over the period of 90 days) is 3593.14$.
2. The main factors that affect lifetime value are as follows:
    a. Average rides per day
    b. Average earnings per day
    c. Number of days using the app.

    It's fairly obvious that if a driver has more rides per day and uses it for a long time, they will bring in more value to Lyft.
3. On an average (over the 90 days period) a driver will be with lyft for around **84** days after he is onboarded.
4. The driver that brings more value to lyft, typically has more rides per day, arrives to the destination faster, uses the platform more regularly, makes shorter trips, more trips in the weekend (Friday nights to Sunday night) and during odd hours and night than the average driver.
5. The actionable recommendations to the business addressing average/ less than average drivers are as follows:
    a. More notifications to the drivers who are riding less or not using the app giving them a sense as to what they are missing on.
    b. Notification to the rider at the end of the trip (if they arrived faster or did statistically better than the other riders) while the rider gives a tip. Hopefully, if we can re-iterate with numbers as to how well the trip went, the rider may get swayed to tip or tip more.
    c. Incentives to ride in the odd hours and night (if the demands aren't met)

Driver Lifetime value:

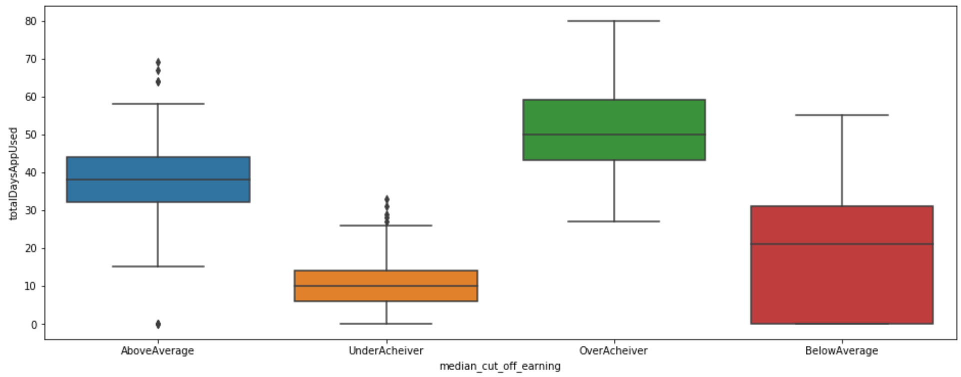
The driver lifetime value is calculated as follows.

1. Calculate the average earnings for a day. (Take average of every driver's average daily trip cost- 38.77$)
2. Calculate the average trips for a day. (Take average of every driver's average trips per day- 3.2)
3. Calculate the average value per day by multiplying average earnings and average trips- 126$
4. Now calculate the number of days the driver used the app. (If a driver takes at least one ride a day, he has used the app to bring in value). Calculate the average number of days the app was used among all drivers in the dataset – 28.51 days
5. Calculate the driver lifetime value by multiplying the average value per day and average number of days the app was used (in the dataset). - 3593.14$

Segments of drivers generating more value:

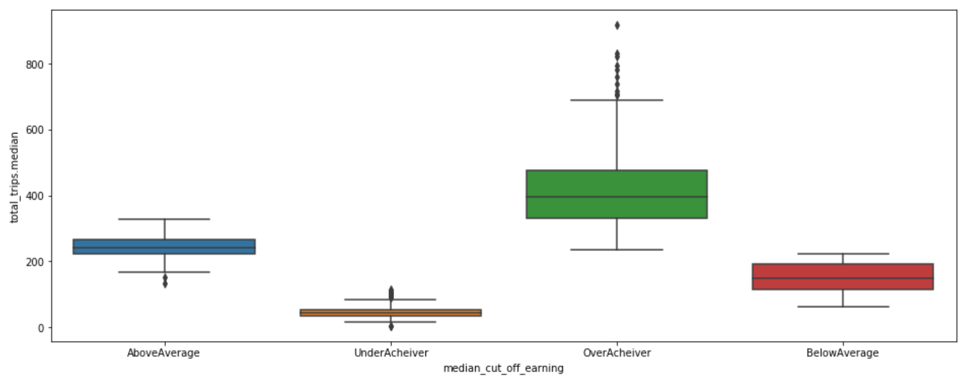
The following are the segments where there was an appreciable difference between the 4 classification of drivers. **Overachievers** - (greater than 1.5 * Median Total Earning), **Above Average** - (greater than the

Median total earning and less than 1.5 * Median total earning), **Below Average** (Less than Median total earning but greater than Median Total Earning – 0.5* Median Total Earning) and **Underachievers**-(Less than Median Total Earning – 0.5 * Median total earning). Before analyzing, the segments let's see how the different riders use the app, how many riders on an average and so on



We could see that the underachievers and the below average used the app far less than the others.

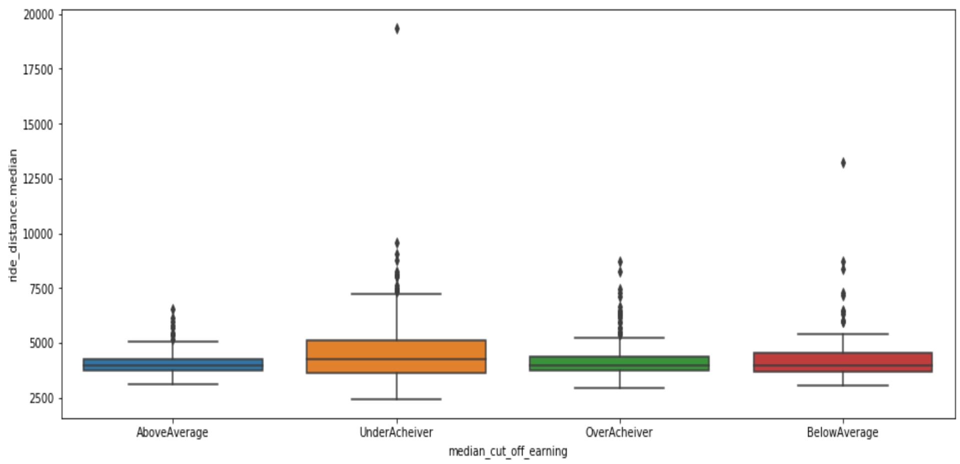
Now, let's see how many trips they have driven.



Again, overachievers and above average have more rides.

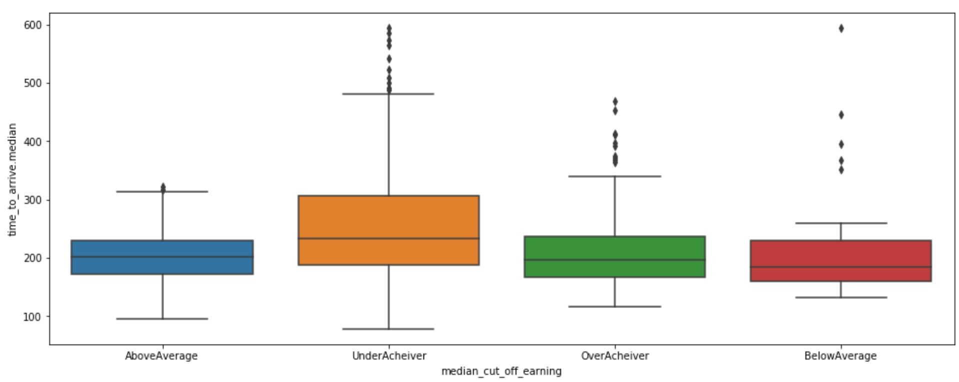
More days on the app -> more rides -> more money

1. **Shorter trips across rides**



The underachievers have a higher ride distance median even though the number of trips were far less among the other three types. This indicates that these types of drivers know that they are in for the short term and hence they take longer rides to maximize their profit at that time.

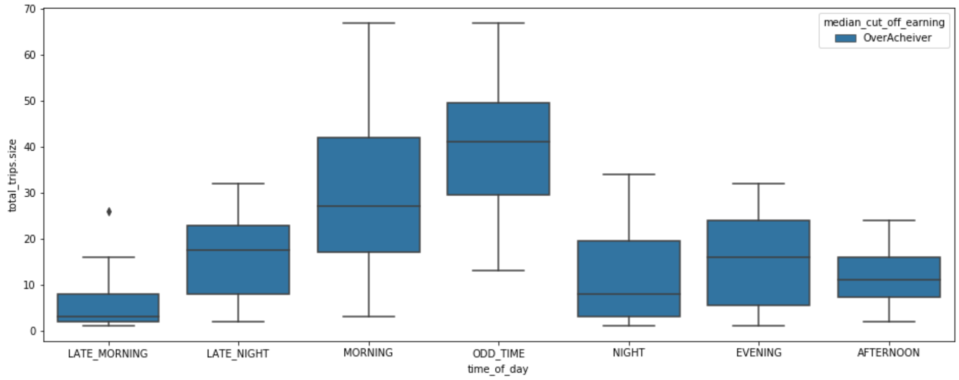
2. **Arriving to the destination faster**



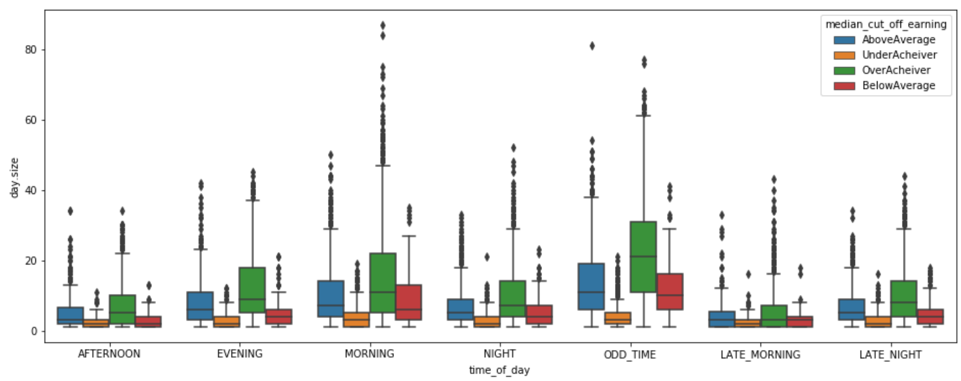
Here, apart from the under achievers everyone else has a lower time to arrive median. Even though the overachievers took many trips their median arrival time was less, as they want to maximize the number of rides each day

3. **More preference to odd hours**

The overachievers prefer odd hours to get more rides where there is a demand that many don't satisfy.



This is the pattern of ride taking when only the top 5 over achievers were analyzed.



The above again reiterates the preference of overachievers for odd hours.

Lifetime of a driver:

Time at lyft – Difference between the last ride and the date onboarded.

This was the trickiest part of the challenge as there were many challenges as to how to frame the question. There was no clear definition of lifetime here and adding to it there were data discrepancies like, 83 drivers with no onboard date, riders with no rides.

If we just simply calculate the median of each individual's duration in this dataset (date of the last ride – date onboarded), it results in **69** days. This analysis would be crude because of two reasons.

1. Not everyone started at the same time
2. Since, there is no clear definition of "death" or no case of driver being thrown off the platform you simply just can't take the median alone.

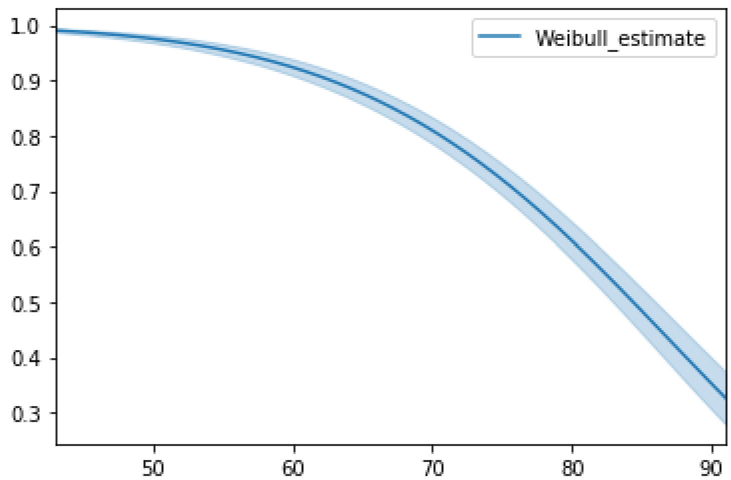
Solution: Use **Survival Analysis technique** to get the unconditional survival curve and correspondingly the median lifetime.

Definition of a lifetime – Iteration 1:

Take the riders with no rides as a failed case (they weren't mapped to any rides in the dataset. Assuming it isn't a data error). This assumption is okay as lyft as a platform wants to benefit riders as well as the passengers, if someone isn't using it, the purpose of the platform is put to waste. Since, the proportion of failed cases in this case is very low, the survival curve at the last timestep, didn't reach below 0.5 and hence it was impossible to calculate the median lifetime.

Definition of a lifetime – Iteration 2:

To increase the number of "failed" cases, taking a much more stringent definition. Riders with less than 20% presence in the app and who have onboarded for more than 45 days would be considered as failed cases. Presence in the app is calculated by counting the number of days the rider has taken at least one ride and dividing by the time at lyft. With this approach the number of "failed" cases increased to more than 200. Using a Weibull fitter, the unconditional survival median/ lifetime was calculated to be **84** days (in a 90-day period). The lifetime definition cannot be taken out of the context of this dataset as the behavior of the driver outside of the dataset is unknown and not considered in the model.



Ideal approach if there was some time to work on it: (Technical)

- Fit a parametric survival regression model (tried cox model with 0.89 concordance but the linear hazard model assumptions were violated. The covariates that I used didn't vary linearly in the hazard domain) and get the unconditional survival curve
- Get the conditional survival probability for all the censored observations. With this way, we can get the conditional survival probability for the riders with no onboard date as well (The conditional after would be 0 days in this case as opposed to other drivers whose time at lyft value is known)
- Get the remaining useful life from conditional survival curve. (Draw n samples from a uniform distribution, 0-1 and get the duration. Calculate the mean of the duration of n samples)
- Add it to the time at lyft value for each case. Calculate the mean for all cases.

Advantage of this approach is that, if the covariates changes for an observation in a bad way indicating he is not using the platform or suggesting cases of failure then the survival probability becomes less and hence the remaining time would also be less.