Lyft Driver Dataset Challenge

By: Cassia M and Andrew

0. Summary of Findings

The main takeaways from our analysis is that: one, there are two major clusters of drivers; two, to promote a longer lifetime, consistency should be valued over daily productivity; and three, the productivity of a driver is mostly dependent on the amount of time they spend as drivers.

To begin, we found that the drivers indeed do not act alike and can be easily clustered into two major groups based on the frequency of their rides. We called these groups full-time and part-time drivers, and it seems the main divide comes from some people seeing Lyft as a main source of income, while others viewing it as a "side-hustle". Generally, we found the full-time drivers to be more valuable to Lyft.

Secondly, we found that amongst all drivers, the projected lifetime of a driver is longer amongst drivers that exhibit consistency over productivity; generally, this means drivers that tend to work more days, but may take less rides per day.

Lastly, we found that the revenue generated per day is associated with drivers that take on more rides and work longer. In simplified terms, people that work more often and for longer hours tend to have a higher rate of revenue generation.

1. Introduction

The goal of this analysis was to analyze the value lifetime value that a driver brings to Lyft. Though a measurement of value an employee brings to a company can be a holistic measurement that factors in the totality of their contributions, we feel that a holistic measurement is impossible to achieve with the limited nature of the dataset. For example, if possible, given the data, we would have integrated the quality of the driver through a proxy measurement, like the ratings and tips the driver receives. Also, attempting to creating this "holistic measurement" might inadvertently cause us to bias our analysis to our own prejudices. Therefore, we settled on using the estimated lifetime revenue as a proxy measurement of a driver's value.

We reduced the term of revenue to the two measurements: one, the amount produced per day of work, and two, the lifetime of the driver. We feel that these two measurements encode the two major things that a company would likely look for in an employee: productivity & longevity. Also, by analyzing these two aspects separately, we can see how they are related. For example, are more productive drivers retained by Lyft, or do they tend to experience "burn out" and stop driving. Also, this allows us to see the qualities of the driver that contribute to each aspect separately.

2. Analysis of Projected Lifetime

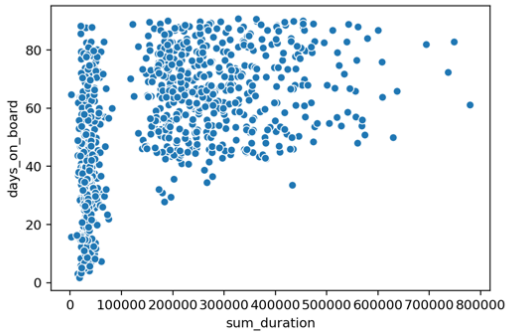
Though productivity through providing more rides may be the primary indicator of the productivity of a driver, equally important is the driver's retention. We view keeping a driver onboard as being equally as important, because maintaining a large number of experienced drivers will improve the service.

In our analysis, we found that driver retention is heavily clustered, and in order promote driver retention, Lyft should promote a higher consistency in driving rather than daily productivity. Also, we attempt to make an estimate of the average lifetime of a driver.

Also, as a note, we are limiting our definition of a driver's lifetime as being the temporal difference between day of their on-boarding and their most recent ride as an attribute we called days_on_board.

3.1 – "Full-Time" & "Part-Time" Drivers

To begin, we found that there are two major clusters of Lyft drivers, specifically based on their number of rides. This seemed to bias the analysis of any other factor, because the same clusters would form. There are two examples on the side.

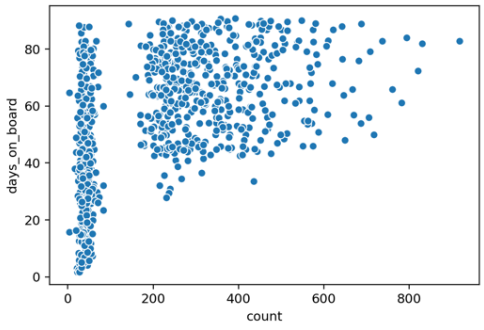


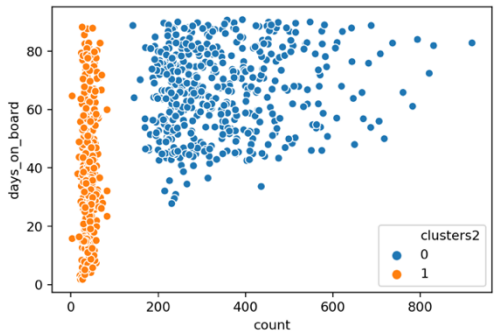
Therefore, we used clustering algorithm to separate an agglomerative clustering algorithm through sklearn to separate the two clusters.

We hypothesize that this clustering of drivers is analogous to how the split between part-time and full-time employees (or contractors).

There seem to be drivers that view Lyft as a large source of income, therefore performing many rides, while other drivers view Lyft as a "side-hustle". Making this distinction made our analysis much clearer, and we would recommend Lyft to also acknowledge the difference between these 2 types of drivers in future analyses and future business decisions.

As a side note, we also found that the average retention of "part-time" drivers is much lower compared to



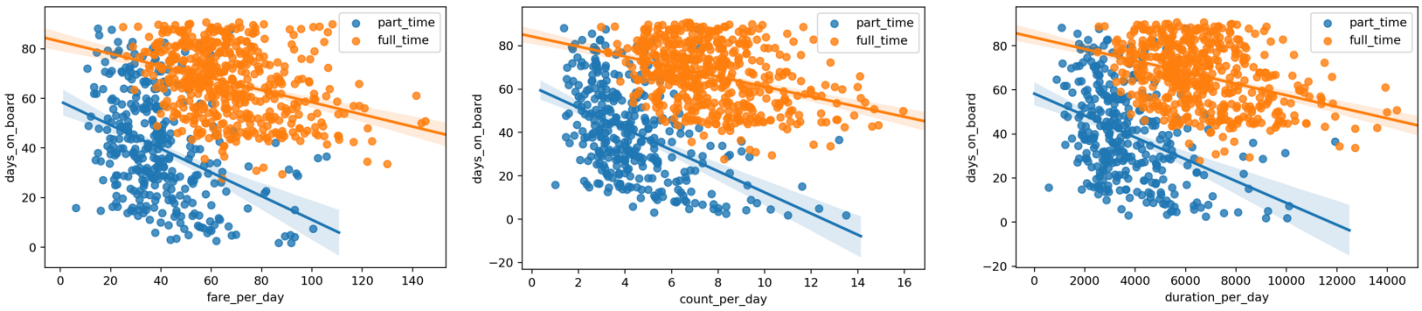


"full-time" drivers (66 days vs. 40 days), so an initiative should be for Lyft to incentivize drivers to see Lyft more as a major source of income rather than a "side-hustle".

3.2 – Consistency over Daily Productivity

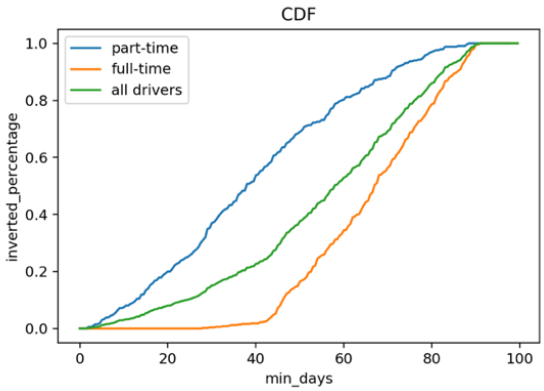
A common trend we found shared between the "part-time and "full-time" drivers is that retention tends to suffer when drivers attempt to drive heavily on few days. Clearly from the plots below, the longer a person drives per day, the lower their projected retention would be.

We recommend to Lyft that in order to incentive more frequent rides for drivers, such as promoting working more days per week and disincentive working heavily on only a few days. In summary, our recommendations for increasing driver retention is to promote seeing Lyft as a major source of income rather than a "side-hustle", and to promote higher consistency as a driver rather than raw productivity.



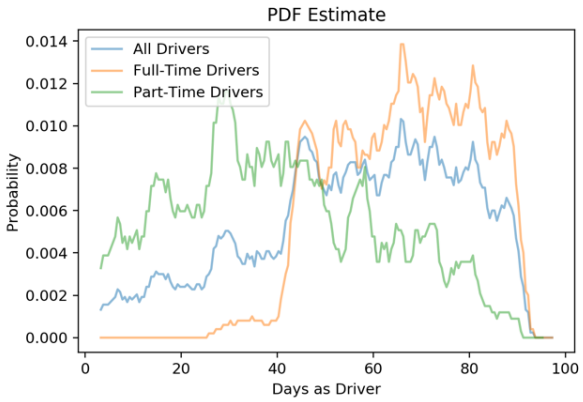
3.3 – Driver Lifetime Estimations

Lastly, we were asked to provide an estimate for lifetime of a driver. We feel that this is a tough ask given the relatively sparse amount of data present. For example, the largest time a person is a driver based on the data (difference between on-boarding to latest ride) is 90 days. However, we expect that there would be many drivers on the service that would drive for much longer than this. Therefore, we acknowledge that our estimate is likely a biased underestimate of the average lifetime of a driver.



Our method for creating an estimate is to define a random variable X that represents the probability of a driver spotting before X days. Given the data, we can build a cumulative

distribution function for X, which represents the percent of drivers leave before and including X days. Then, we can take an estimated derivative of this function to estimate the underlying probability distribution of X. At that points, finding E[X] is straight-forward.



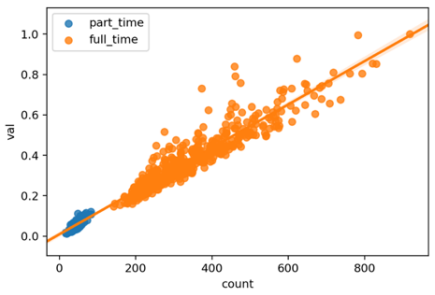
From our method, we estimated the lifetime of a driver to be 56.01 days. When looking at "full-time" and "part-time" drivers, we found the estimate is 66.63 and 40.22 days. From the estimated pdf, it is clear that is biased to underestimate, because it is clear that people work for more than 100 days. However, we would need more data to better understand the right tail of the distribution and produce a less biased estimate. Also, the distribution produced, even after smoothing has many irregular sharp peaks. Given more data, we would expect a more recognizable distribution to be produced.

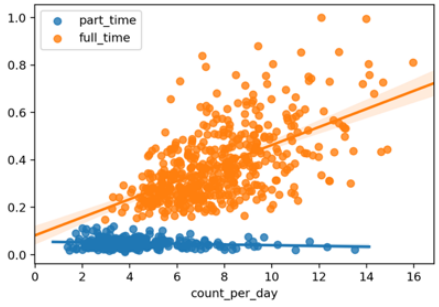
3. Analysis of factors that affect value

By creating linear regression plots of value against every other metric given, such as duration, distance, prime time etc. We were able to identify the factors which affects value most significantly. We want to see what correlates most to value and that will show the greatest attributing factor. Our initial hypothesis was that drivers who have been with Lyft the longest and have the highest average per day would be most valuable.

During our initial plotting, we found that when plotting value against metrics that were sums, such as the sum of duration, count and time, there was a very strong correlation, almost a score of 1, which basically means that the more that a driver has the higher his value, this is very obvious even without visualizing the data. This initially aligns with our hypothesis.

The interesting factor comes into play when plotting value against day to day metrics, this results in two very different clusters of data. From this fact, we decided that it was best to treat these two clusters differently so that we get a better understanding on what actually affects value. From these graphs, we can see that it's very hard to attribute any daily factors that affect value in part-time drivers. They do not get affected by any habit or characteristics like full time drivers. It is seen that there is a very big divide for part time vs full time drivers even within the summed metrics and it goes to show just how different these drivers act. An analysis of each factor can be found in the workbook.





We found that the value of a certain driver is very dependent on daily activity, especially in full time drivers. Part time drivers don't get affected by daily activity averages as their total value (0.047) is usually much lower compared to full time drivers (0.373). We found that conditions such as weekend percentages, and average pick up time did not affect value significantly. So, the most important factors for a driver's lifetime value would be:

1. Count, 2. Distance, and 3. Duration

All these three factors are very much interrelated and can be said as simply driver activity. In order to make this research more meaningful, we will attribute factors to be:

1. Driver status (Part time or full time)
2. Driver Activity (Ride Frequency) and Lifetime
3. Efficiency and Timing (Speed and busy hours (primetime))

In conclusion, our initial hypothesis was not entirely correct, although lifetime does play a part in value, it is not the most important thing. Rather the quantity of completed rides and activity of the driver is the defining factor for value. By grouping the drivers into part time and full time, we were able to get an even deeper understanding into the lifetime value of drivers where Lyft should value full time drivers much more than part time drivers because their average value is almost 10 times more.

In order for Lyft to effectively increase the value of drivers, Lyft needs to create more full-time drivers compared to part time drivers because they are 10x more valuable in terms of value. So, in order to do this, we should incentivize drivers earlier on to become full-time drivers, we can do this by increasing pay for the first month or two months in order to show that Lyft is viable as a full-time job and attract more drivers to become full time. This would also effectively create higher value drivers which would then generate more revenue for Lyft. Our recommendations for current drivers would simply be to increase driver activity and orders. Lyft should create new strategies that incentivize driving more often which could be a friendly competition or more bonuses when certain conditions are met. Further study is needed to create a more solid business plan.