Lyft Driver Dataset Challenge

By Michael Kusnadi & Udaikaran Singh

**Summarize your conclusions at the beginning of your writeup.**

1. **Introduction**

In order to accurately answer the question concerning value, we decided to define it through the company’s perspective. Through the dataset given, the least biased way that we can define value is simply through the revenue that they have generated throughout their lifetime. But when we look at it further, drivers that have a longer lifetime does not simply make them more valuable.

In terms of value for the company, we are defining value through revenue, efficiency and activity. Of course, value is more than just the amount that the driver earns, but in this case considering that the dataset does not provide a holistic review of the driver, we are going to define value as lifetime revenue generated. Since we are only looking at the information of drivers objectively and not holistically such as rating and satisfaction, the definition of value in this case would be the average amount of revenue they are able to generate multiplied by their lifetime which is then normalized. We are able to make a metric by getting the average revenue made. Since the average fare that they get takes into account the time and distance that they have traveled, we will simply need to calculate it against their lifetime in order to gain an understanding of how much revenue they have generated.

1. **Analysis of factors that affect value**

By creating 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 will be using linear regressions for most finding out which factors affect value because 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.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

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. (Udai’s explanation on clustering below)

From these graphs, we can see that its 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.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

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, speed 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 or count per day which translates to orders
  2. Distance
  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 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.

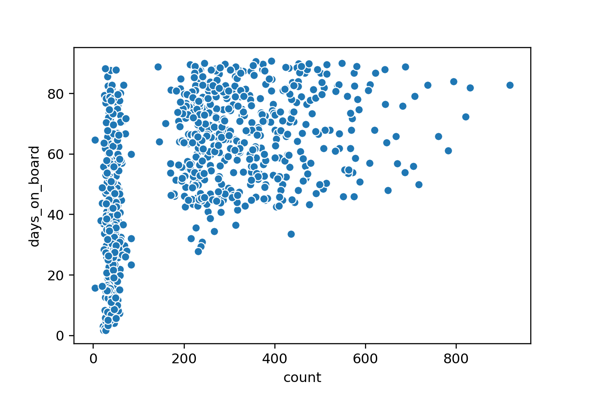
**3. Analysis of Projected Lifetime**

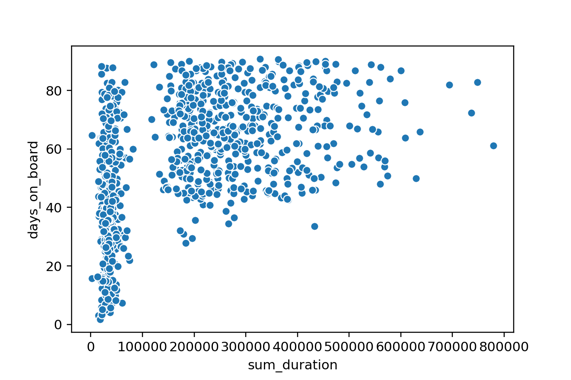
Though productivity through providing more rides may be the primary indicator of the productivity of a driver, equally important is the driver’s retention. We view keeping a driver onboard as being equally as important, because maintaining a large number of experienced drivers will improve the service.

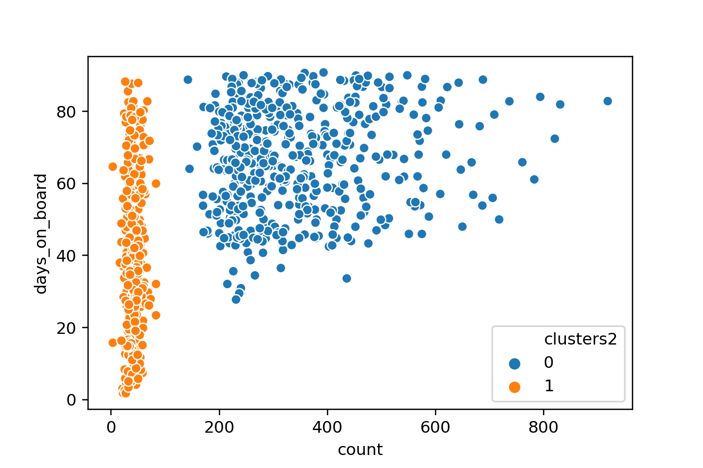
In our analysis, we found that driver retention is heavily clustered, and in order promote driver retention, Lyft should promote a higher consistency in driving rather than daily productivity. Also, we attempt to make an estimate of the average lifetime of a driver.

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 3 examples of this below.



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

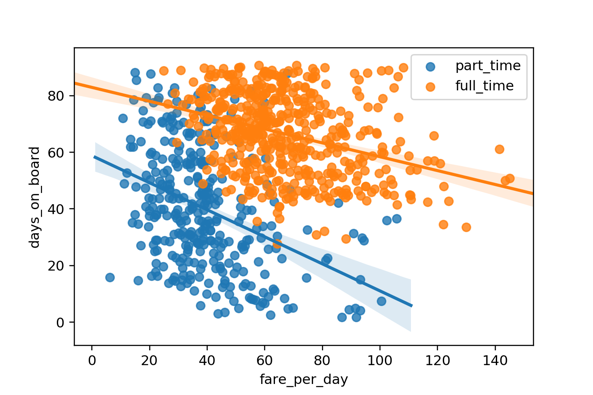
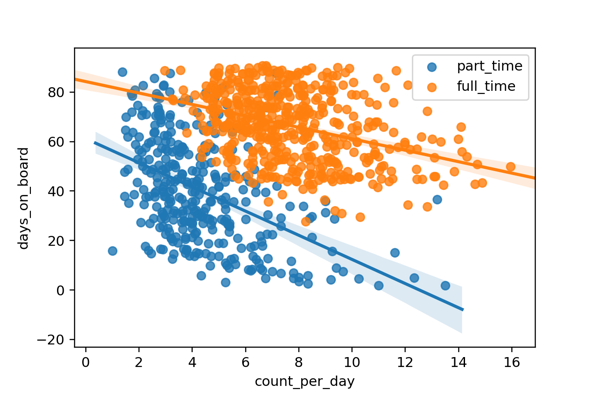
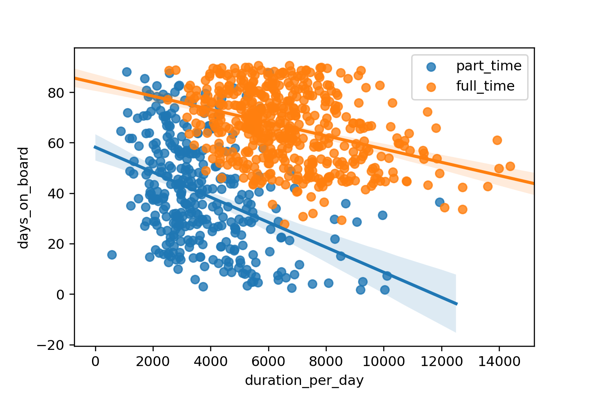
We hypothesize that this clustering of drivers is analogous to how the split between part-time and full-time employees (or contractors).

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”.

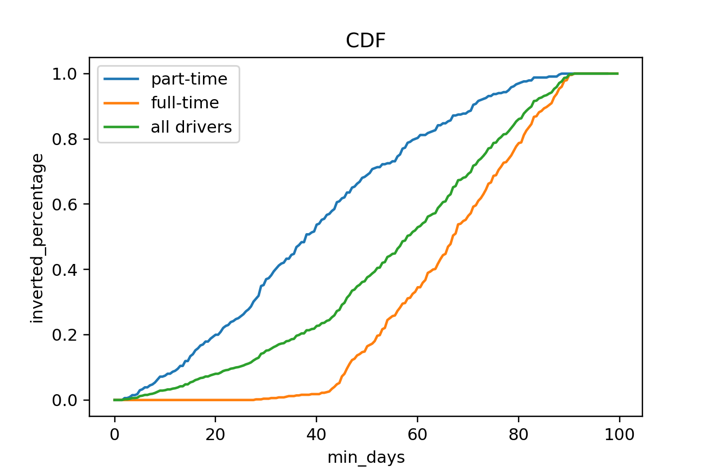
3.2 – Consistency over Daily Productivity

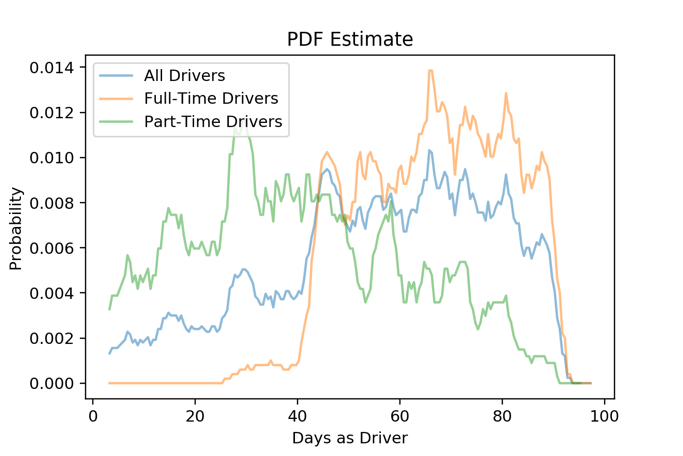
A common trend we found shared between the “part-time and “full-time” drivers is that retention tends to suffer when drivers attempt to drive heavily on few days. Clearly from the plots below, the longer a person drives per day, the lower their projected retention would be.

****

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 create a less biased estimate.

**4. Conclusion**

…