Lyft Data Challenge Writeup

Jen Sheng Wong, Zhihao Guo

September 2019

The goal of this report is to recommend a driver's Lifetime Value (LTV), analyze drivers' behavior, and provide actionable insights for Lyft.

Conclusion

We adapt the following formula for LTV[1] and churn rate[2],

$$\text{LTV} = \frac{\text{Average Income Generated by Driver}}{\text{Churn Rate}} \times 365 \tag{1}$$

$$\begin{matrix}\text{Average Income} \\ \text{Generated by the Driver}\end{matrix} = \frac{\text{Total income generated by a driver}}{\text{Number of days the driver drove}} \tag{2}$$

$$\text{Churn Rate} = \frac{\text{Number of drivers who have stopped driving}}{\text{Total number of drivers}} \times 100\% \tag{3}$$

Our findings can be summarized as follows:

- LTV is affected by the number of days the driver drove, total revenue generated by the driver, which is related to miles traveled, minutes traveled, percentage increase in the fare, and the churn rate.
- The average projected lifetime of a driver is 3.03 years.
- Drivers can be segmented into 3 main clusters by applying KMeans clustering algorithm to features associated with LTV. We also study the relationship between LTV and other features that were not used to calculate LTV. Results show that there are several features that are significantly different based on each group. For instance, the number of rides completed on weekdays and time elapsed between a driver's drop-off and pick-up time are fairly distinct depending on clusters.
- Based on the features we created, we recommend that high value drivers should be rewarded with higher share of fare; mid value drivers should be nudged to encourage driving more consistently based on unusual inactivity; low value drivers should be incentivized via gamification strategies. Loyalty program can be introduced to all drivers to discourage "dual apping".

Assumptions and EDA

Interestingly, even we have 937 unique `driver_id`s for both `driver_ids` and `ride_ids`, only 854 overlap. For the 83 drivers that appear in `driver_ids` but do not have any record in `ride_ids`, we assumed that they registered but did not drive because they lack `ride_id` and thus do not exist in both `ride_ids` and `ride_timestamps`. For the 83 drivers that appear in `ride_ids` but not in `driver_ids`, we assumed they

[1] https://blog.hubspot.com/service/how-to-calculate-customer-lifetime-value
[2] https://blog.innertrends.com/customer-lifetime-value/

registered earlier than the start of the dataset. We also found that 94 drivers do not have any record in `ride_timestamps`. This poses a challenge to our analysis because a driver's lifetime value is dependent on the information in `ride_timestamps` table. Note that the 94 drivers altogether completed 8,684 rides but the corresponding entries in `ride_timestamps` are absent. Hence, we decided to impute all the null values of those features with the mean values of the respective features.

Methodology

LTV

First, we calculated the fare of each ride. We employed the assumptions on Lyft rate card given in the prompt. The formula we use is as follows:

$$\begin{aligned}\text{Fare} &= (\text{base fare} + \text{cost per mile} \times \text{miles traveled} + \text{cost per min} \times \text{mins traveled})(1 + \frac{\text{prime time}}{100}) \\ &\quad + \text{service fee} \\ &= (2 + 1.15 \times \text{miles traveled} + 0.22 \times \text{mins traveled})(1 + \frac{\text{prime time}}{100}) + 1.75\end{aligned} \tag{4}$$

Since the fare is limited by a lower bound of \$5 and an upper board \$400. The final fare for a ride is as follows:

$$\text{Fare} = \min\{400, \max\{5, \text{fare}\}\} \tag{5}$$

We first converted the original distance in metres to miles and the original duration of the ride in seconds to minutes. The fare is then summed based on drivers to determine the total income generated by a driver.

The figure below shows the distribution of LTV. We calculated the average lifetime value of a Lyft driver to be about \$90,142.

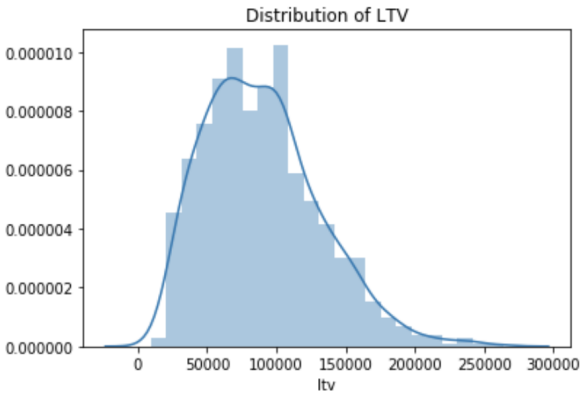

Figure 1: Distribution of LTV

We concluded that the main factors that affect a driver's lifetime value are:

- the number of days the driver worked, `unique_days`
- total revenue generated by the driver, `fare`
- the churn rate, `churn_rate`

There are other features we designed to reflect a driver's behavior, such as `is_weekday`, `is_late_ride` and others. Correlation of these factors are shown in the correlation heat map in Figure 2.

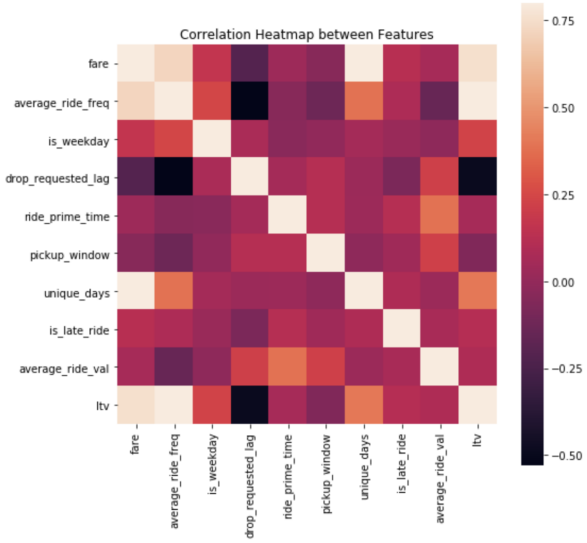

Figure 2: Correlation Heatmap

From the heatmap, we observe that there is no strong correlation between other features with LTV apart from the features that we used to calculate LTV.

Average projected lifetime of a driver

To calculate the churn rate, we treated drivers who have been inactive for more than 7 days from the last day of the dataset, which is 2016/06/27, as drivers who have stopped driving. In other words, drivers whose last day of activity ended before 2016/06/20 are drivers who will be churned. We find that there are 309 drivers have churned from a total of 937 drivers.

Using Equation(3), we calculated the average projected lifetime of a driver as:

$$\begin{aligned}\text{Churn Rate} &= \frac{309}{937} \times 100\% \\ &\approx 33.0\%\end{aligned}$$

$$\begin{aligned}\text{Average Projected Lifetime} &= \frac{1}{\text{churn rate}} \\ &= \frac{1}{0.33} \\ &\approx 3.03 \text{ years}\end{aligned}$$

Hence, the average projected lifetime of a driver is 3.03 years.

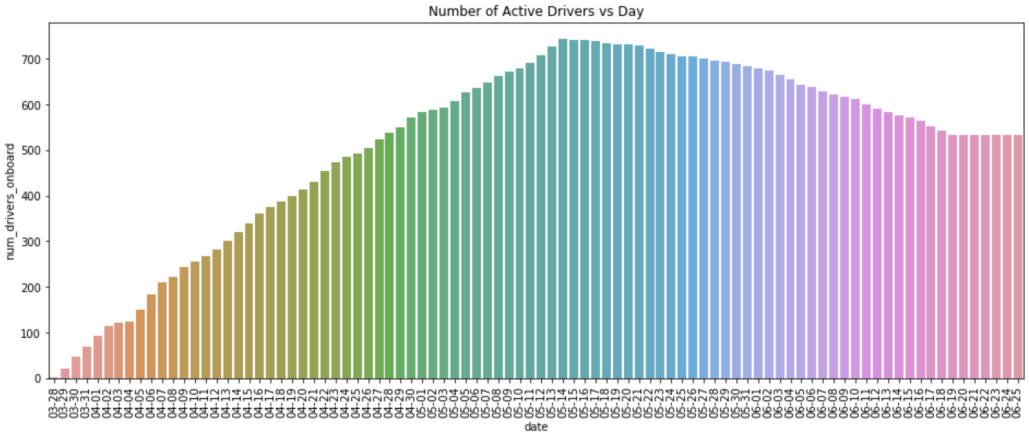

Figure 3: Active Drivers based on Days

From the bar graph, drivers start to onboard and the number of drivers peaked on 15th of May. As some began to drop out, the number of drivers started to fall and then flattened out because we assumed drivers who have activities during the last week have not dropped out, i.e., they might be taking a break.

Segment of drivers

Ideally, we would like to divide all Lyft drivers into segments of high, mid, and low value. Hence, we applied KMeans clustering algorithm to segment the drivers into 3 main clusters. To test how "predictive" the other features we created (i.e., features that are not used to calculate the LTV), we decided to fit our KMeans clustering algorithm to only the features used to calculate LTV. We call them `base_metrics` and they comprise of `ride_count`, `fare`, `unique_days`, and `ltv`. We also studied the other features that were not used to fit the model to see if there is any clear distinction between different clusters of drivers.

The radar chart below shows how each cluster performs based on all metrics.

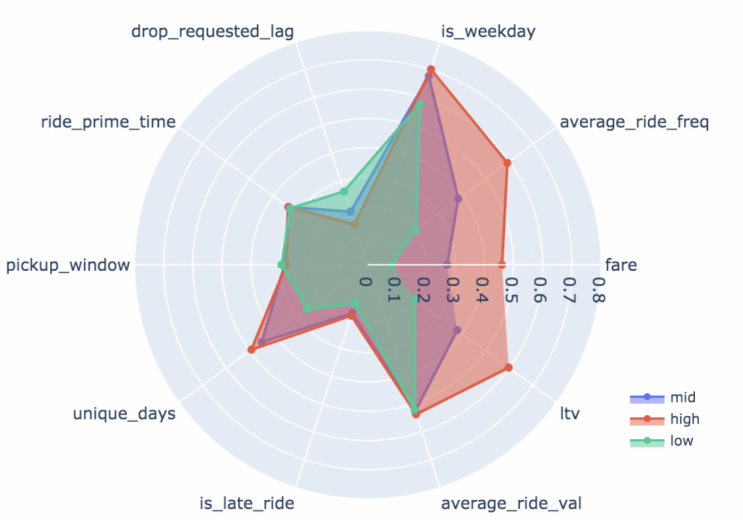

Figure 4: Radar Chart of Features

It can be seen that the red area has the highest LTV among all 3 areas. We conclude that this is the group of the drivers that generates the highest value for Lyft, followed by the blue area and green area.

For the features used to calculate LTV, such as `average_ride_freq`, `average_ride_val`, `fare` and `ltv` itself, we observed fairly distinct area coverage in the radar chart, indicating that they are well-separated.

Although not all of the features we created are helpful in segmenting the drivers, some features stand out. The most prominent feature is `is_weekday`. We can see that the red and blue clusters have high values of `is_weekday` while the green cluster has the lowest value of `is_weekday`. This allows us to conclude that drivers who drive on weekdays are most likely high value drivers.

Another feature is `drop_requested_lag`. Drivers with higher LTV tend to have lower `drop_requested_lag`, meaning they are more keen to pick up the next passenger's request after dropping off a passenger. This "characteristic" bodes well for Lyft as this means they are also generating higher value. A high `drop_requested_lag` value might indicate "dual apping", whereby drivers toggle back and forth between Lyft and Uber to decide which offers a more lucrative ride.

In conclusion, we are able to segment drivers into clusters that generate low and high value based on the features used to calculate their LTV.

The figure below shows 3 selected features and their distribution based on clusters.

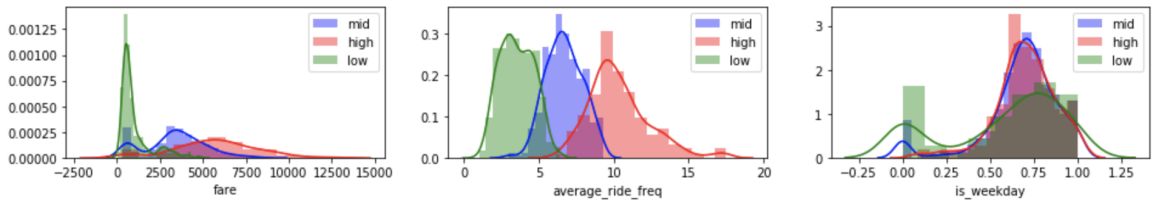

Figure 5: Distribution of Selected Features based on Clusters

Actionable recommendations

High value drivers are the "star" drivers. We recommend Lyft to focus on retaining them by rewarding them with an increase in the percentage share of fare the driver splits with Lyft and offering discounts on gas and car maintenance.

For mid value drivers, we should try to "nudge" them into becoming high value drivers. The best way is to offer a bonus based on "streaks", such as rewarding them if they drive more consistently. This can help minimize their gap of `average_ride_freq` with high value drivers (as seen in the radar plot). Notifications can be sent when their last activity is somewhat unusually long.

Low value drivers are more of a "marginal" driver. This category could possibly comprise of drivers who would like to experiment being a driver or drive temporarily while they are transitioning into different roles in their main career. It might be best to show them the tiers or milestones they can progress as a Lyft driver and the progress itself after every ride. Noticing the gap of `drop_requested_lag` between low value drivers and high drivers, we recommend introducing a feature such as "pick up your next customer in 5 minutes". This might be able to "hook" the drivers' interest as a form of gamification.

Another measure that can be implemented for all three parties is introducing loyalty program to discourage "dual apping". Drivers should be able to redeem the "miles" they have driven in exchange for some benefits.