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 the drivers' behavior, and provide actionable insights for Lyft.

Summary

We propose the following formula of LTV,

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

Our findings can be summarized as follows:

- LTV is affected by the number of days the driver worked, total revenue generated by the driver, and the churn rate.
- The average lifetime value of drivers is 2.8 years.
- Drivers can be segmented into 3 main clusters by applying KMeans clustering algorithm to features used to calculate LTV and validated on other features, such as the drivers' tendency to drive on weekdays versus weekends.
- 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".

Methodology

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:

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

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:

$$\text{Fare} = \min\{400, \max\{5, \text{fare}\}\} \tag{3}$$

We 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.

To calculate the base LTV, we referred to the formula from https://blog.hubspot.com/service/how-to-calculate-customer-lifetime-value.

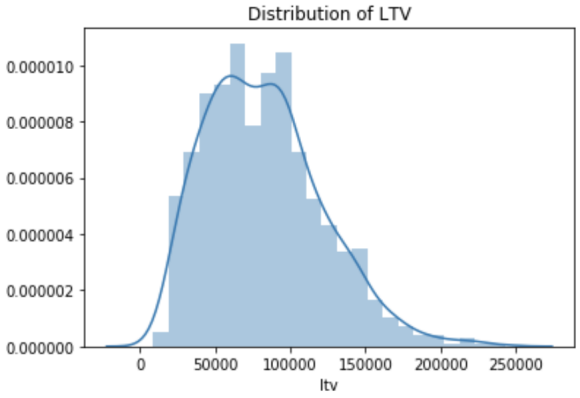
$$\text{LTV} = \frac{\text{Average Income Generated by Driver}}{\text{Churn Rate}} \times 365 \tag{4}$$

$$\text{Average Income Generated by a Driver} = \frac{\text{Total income generated by a driver}}{\text{Number of days the driver drove}} \tag{5}$$

$$\text{Churn Rate} = \frac{\text{Number of drivers who have stopped driving}}{\text{Total number of drivers}} \times 100\% \tag{6}$$

We define drivers who have stopped driving as drivers who have been inactive for more than 7 days from the last day of the dataset, which is 2016/06/27. In other words, drivers whose last day of activity ended before 2016/06/20 are drivers who have churned.

The figure below shows the distribution of LTV.



Challenges

Interestingly, even we have 937 unique `driver_id`s for both `driver_ids` and `ride_ids`, only 854 overlap. That means, 83 drivers that appear in `driver_ids` do not have any record in `ride_ids` while 93 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_ids` table. Seeing that 9,262 rides do not have the corresponding `driver_id`, we speculated that those rides might possibly belong to the drivers that do not have any `ride_ids`. Hence, we decided to impute all the null values of the features for the 83 drivers with the mean values of the respective features.

Questions

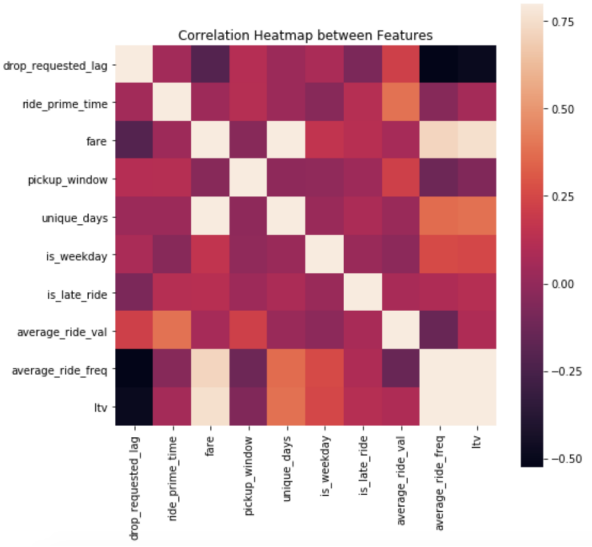
Main factors that affect a driver's lifetime value

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`

A secondary feature that correlates with LTV but not included in the calculation is `is_weekday`. Correlation of some factors can be seen in the correlation heat map in Figure 2.



Average projected lifetime of a driver

We have already defined churn rate in Equation (5) and average projected lifetime is in fact, the inverse of churn rate.

$$\begin{aligned} \text{Churn Rate} &= \frac{305}{864} \times 100\% \\ &\approx 35.7\% \end{aligned} \tag{7}$$

$$\begin{aligned} \text{Average Projected Lifetime} &= \frac{1}{\text{churn rate}} \\ &= \frac{1}{0.357} \\ &\approx 2.8 \text{ years} \end{aligned} \tag{8}$$

We consider drivers who have stopped driving as drivers who have been inactive for more than 7 days from the last day of the dataset, which is 2016/06/27. In others, drivers whose last day of activity ended before 2016/06/20 are drivers who have churned.

Note that there are 305 drivers whose last activity was more than 7 days more the end day of the dataset. Given the 83 drivers who have no corresponding `ride_timestamps` entry, we assume they have already churned.

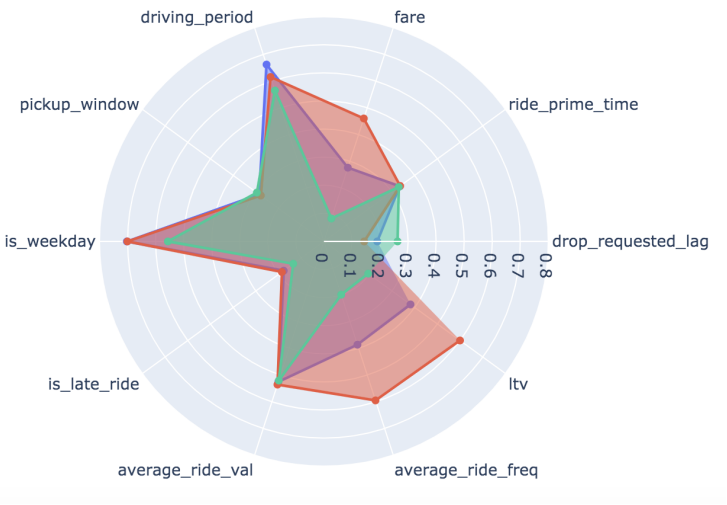
Hence, the average projected lifetime is 2.8 years.

Segment of drivers

We applied KMeans clustering algorithm to segment the drivers into 3 main clusters. To test how "predictive" the other features we created 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 this `base_metrics` and it comprises of `ride_count`, `fare`, `unique_days`, and `ltv`.

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



Red color represents cluster 2, green color represents cluster 1 and purple color represents cluster 0.

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 green area and purple area.

We can tell that features used to calculate LTV, such as `average_ride_freq`, `average_ride_val`, `fare` and `ltv` itself have fairly distinct area coverage in the radar chart.

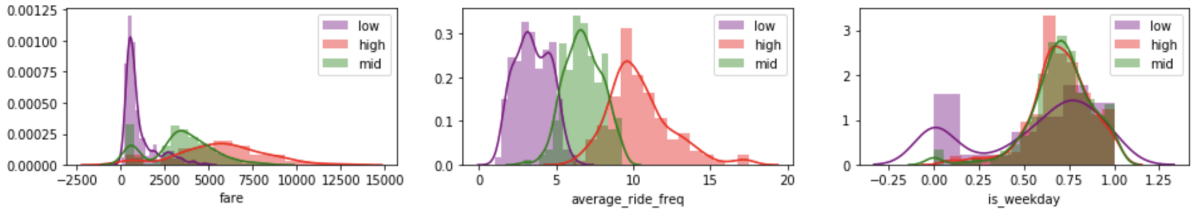
We think using the features above to segment the driver is not as interesting as using features that have been excluded in constructing the LTV. Hence, we included the excluded features to see if we can gain some insights into them.

Although not all of the features we created are helpful in segmenting the drivers, some clear features stand out. The most prominent feature is `is_weekday`. We can see that cluster 2 has the highest `is_weekday` while cluster 0 has the lowest `is_weekday`. This allows us to conclude that full-time drivers mostly drive on weekdays and part-time drivers mostly drive on weekends as they might be driving Lyft to supplement their incomes. Hence, we conclude that drivers who have higher `is_weekday` on average are full-time drivers and they generate higher value for Lyft.

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. In addition to that, we also showed full-time drivers generate higher value.

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



We proceeded to fit KMeans on the 9 features that were present in the radar chart above. The diagram below shows how does each feature correspond to the 3 clusters respectively. We mapped the cluster values 0,1, and 2 such that they roughly correspond to the old clusters.

all_feats_clusters	0	1	2
drop_requested_lag	-0.39549	1	-0.562289
ride_prime_time	1	-0.0109757	-0.246306
fare	-0.787238	-1	0.862204
pickup_window	0.156404	1	-0.702237
unique_days	-0.302311	-1	0.739236
is_weekday	-1	0.196024	0.123698
is_late_ride	-1	-0.100609	0.32024
average_ride_val	0.212839	1	-0.716548
average_ride_freq	-1	-0.676026	0.701498
ltv	-1	-0.694586	0.713795

It seems like the values here do not quite align with our hypotheses. For instance, the LTV for cluster 1 leans towards the negative region but we expect it to be neutral.

Actionable recommendations

We conclude that cluster 2 (red) consists of high value drivers, cluster 1 (green) consists of mid value drivers and cluster 0 (purple) consists of low value drivers.

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. Part-time drivers are more likely to fall into this category. 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 they can progress as a Lyft driver and the progress itself after every ride. Milestones, such as "fetch your 1,000th customers" can also be introduced to help drivers achieve targets. This might be a feature 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.