Synopsis

Our objective was to explore and analyze the data in order to recommend a Driver's Lifetime Value along with finding the main factors that affect it, to identify subgroups of drivers that act differently from one another and what Lyft can do to address this to increase the Driver Lifetime Value across the board, and to determine average projected Driver Lifetime by utilizing linear regression.

Data Transformation

Timestamps and onboard date and time: (table ride_timestamps, driver_onboarddate2, four_hours)

Onboard dates and timestamps came as factors which are not particularly helpful. As a result we converted the timestamps to date objects, date-time objects, and time objects and we converted onboard dates to date objects.

Drivers Lifetime Value:

We calculated the price of every trip using the rate card provided. We decided to use the equation below:

$$Base\,fare\ +\ (Cost\ per\ mile \times ride\ distance)\ + (Cost\ per\ min \times Ride\ duration) \times (\ 1 + (prime\ time \div 100)) + Service\ fee$$

We then grouped the data by Driver ID and summed each ride fare to find each driver's individual revenue. It still cant be said that this is specifically revenue generated for the drivers or Lyft itself but even if it is just revenue generated for the drivers, it would still be directly proportional to Lyft's revenue. Every ride was assigned a "time of day" based on the timestamp (time objects) of drivers accepting ride requests. Rides that took place from 6am - 12pm were classified as morning rides, rides that took place from 12pm - 6pm were classified as afternoon rides, rides that took place from 6pm - 12am were classified as evening rides, and rides that took place from 12am - 6am were classified as night rides. Various driver attributes such as the number of rides completed, the number of rides completed with Prime Time rates applied, the number of days a driver has been with Lyft, and various metrics of average speed and average times were calculated through empirical counting methods.

Data Visualization

Figure 1. shows the ranked the drivers by the descending number of rides completed. The top 300+ drivers in terms of number of rides generate ~15% of all driver revenue while the lion's share of the revenue is generated by the lowest ~500 drivers. The decrease in the slope also suggests that past this point it's economically inefficient to encourage drivers to complete more rides under the timeframe of this dataset (~3 months).

Figure 2. shows that for this period of time Lyft retains a high retention rate, which means that the days someone has been with Lyft is a relatively accurate way to account for projected driver lifetime in a statistical model. However it also reveals that in terms of the scope of the data provided, a driver's lifetime will be heavily tied to their onboard date.

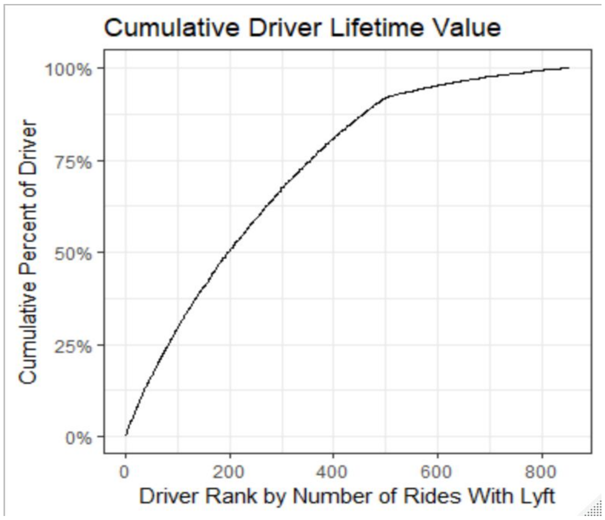


Figure 1

Figure 2

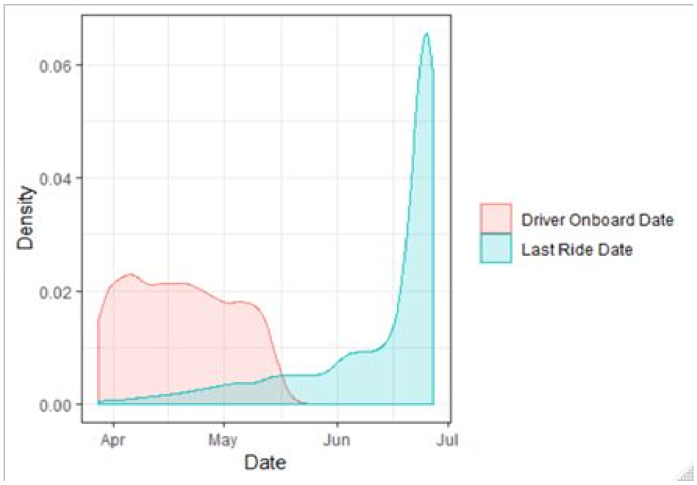


Figure 3

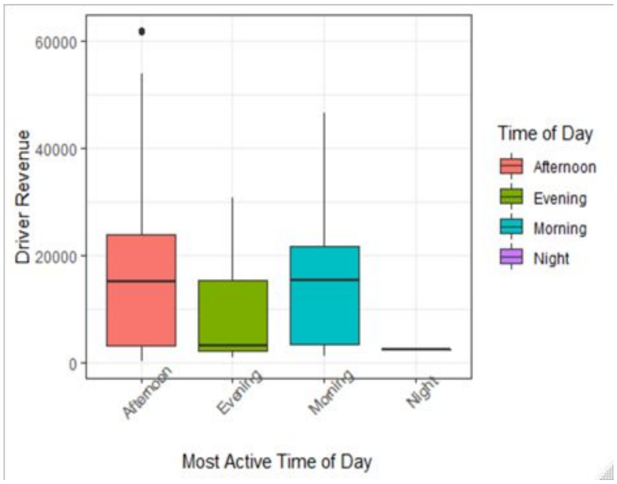


Figure 3 shows that drivers that are mostly active in the afternoon and drivers that are mostly active in the morning generate around the same median revenue. Drivers that are most active in the afternoon and drivers that are most active in the morning outperform drivers that are more active in the evening and drivers that are more active at night. Lyft should encourage drivers to be more active in the morning and afternoon.

Figure 4 shows that majority of the drivers are most active in the afternoon and not a lot of drivers are most active in the morning. This is interesting because they groups of drivers generate similar median driver revenues. Lyft should encourage more drivers to be more active in the morning especially since drivers mostly active in the morning generate around the same driver revenue.

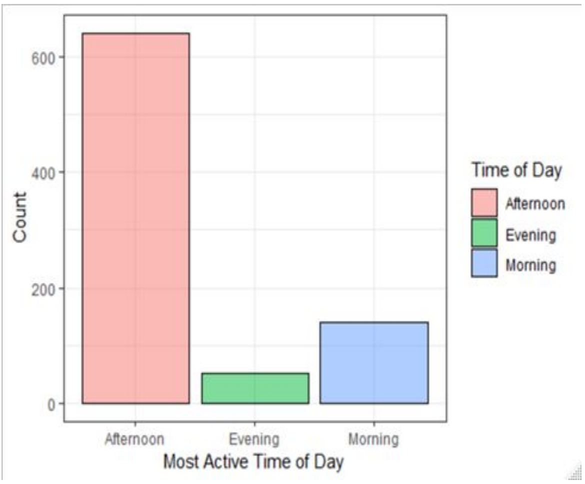
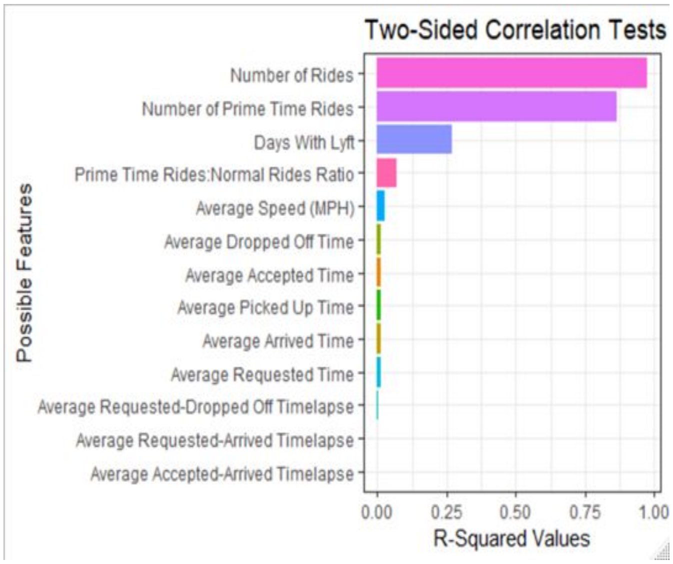


Figure 4 (above)

Figure 5 (below)



Two-sided correlation tests were carried out on possible factors of driver revenue to calculate Pearson's R. This was done to better understand the relationship of variables with driver revenue. The number of rides and number of prime-time rides are the biggest factors on the variation of driver revenue (97.4% and 86.2% of variation in driver revenue is explained by the number of rides and the number of prime time rides respectively). However, the ratio of prime-time to normal rides has little effect on the variation of driver revenue so Lyft probably cannot generate more revenue by artificially increasing the number of rides prime time is applied to. The number of days a driver has been with Lyft has a considerable effect on the variation of driver revenue (27.2 % of variation in driver revenue is explained by the number of days a driver has been with Lyft). All measures related to time and speed seem to have little or no effect on driver revenue. Every single one of these two-sided tests are extremely statistically significant.

Figure 6

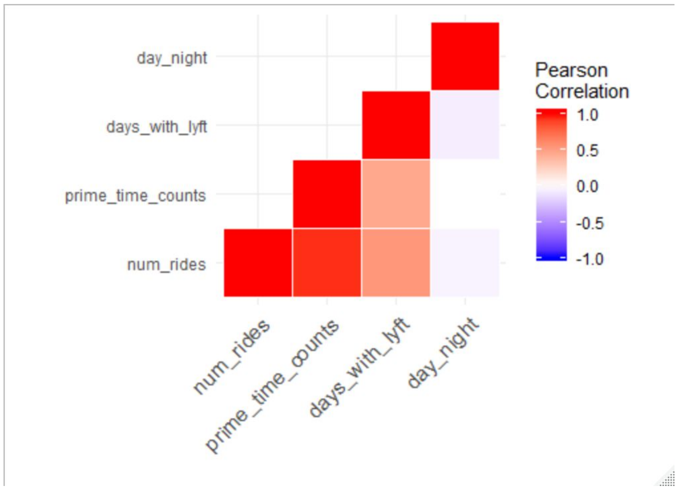


Figure 7

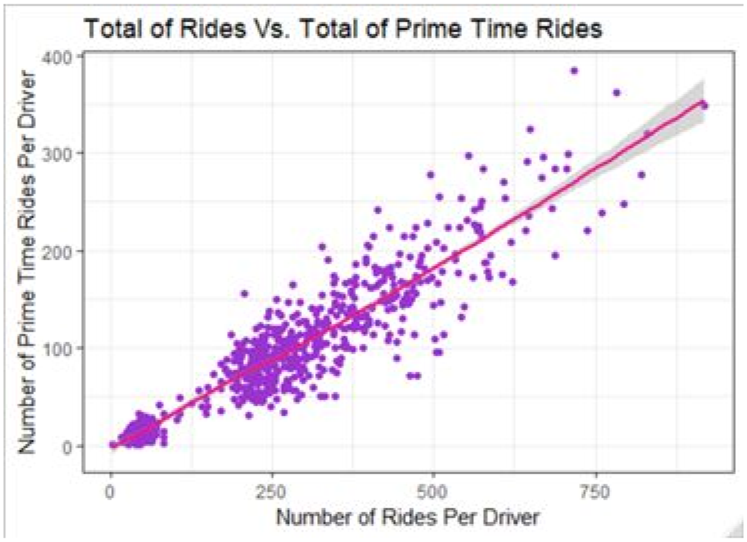


Figure 6 is a correlation matrix heatmap that was created to observe interdependence among factors. Every square in the heat map represents factor interaction and the extremely orange squares represent a Pearson's R close to 1.0 or strong positive correlation while the whitish squares represent a Pearson's R close to 0 or little to no correlation. There is heavy positive correlation between the number of rides a driver completes and the number of rides that have Prime Time rates applied to them as represented in Figure 7. Since the ratio of Prime Time rides to non-Prime Time rides is not a significant factor, we recommend that Lyft focuses on encouraging drivers to complete more rides to increase the amount of rides with Prime Time rates applied to them instead of focusing on increasing the number of rides Prime Time rates are applied to.

Statistical Modeling & Prediction

In order to determine average projected Driver Lifetime, a linear regression model was employed to predict how many days each driver in the dataset stays with Lyft over the timeframe of three months. Then, in order to calculate the average projected Driver Lifetime we applied 180 variable scenarios over the timeframe of the average days drivers stay with Lyft and attempted to account for the outcomes of the total amount of interactions over a span of 45 years (because its expected work duration of any individual from the age of 20 - 65 and because there are 540 months in 45 years).

The response variable/outcome that the linear regression model predicts is the days that a driver will stay with Lyft over the course of three months. In the first iteration of the model the following features were used:

Weekly driver revenue, mean number of rides per week, mean trip duration, mean trip distance, mean requested-arrived timelapse, mean accepted-arrived timelapse, mean requested-dropped off timelapse, time of day, the ratio of the number of Prime Time rides to number of total rides

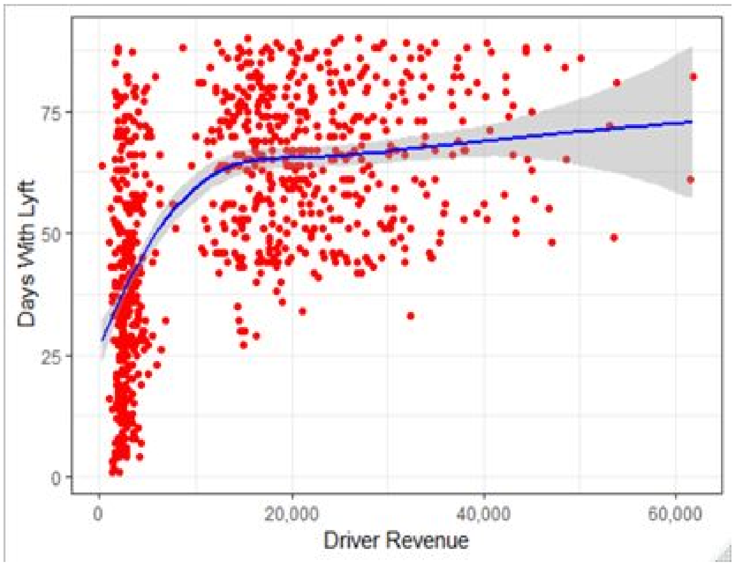
Also, the interactions between average trip duration and average trip distance and the interactions between mean requested-arrived timelapse and mean accepted-arrived timelapse were accounted for in the model.

For the next 5 iterations of the model, features were assessed for statistical significance (p-value < 0.05) and non-significant features were removed from the following iterations. With the exception of the average requested-arrived and accepted-arrived timelapses due to their interaction being highly statistically significant. The penultimate iteration of the model is as follows:

Days With Lyft=Weekly Driver Revenueβ_1*+Mean Rides Per Week*β_2*+Requested-Arrived Timelapse*β_3*+ Accepted-Arrived Timelapse*β_4 *+ Requested-Arrived Timelapse:Accepted-Arrived Timelapse*β_5

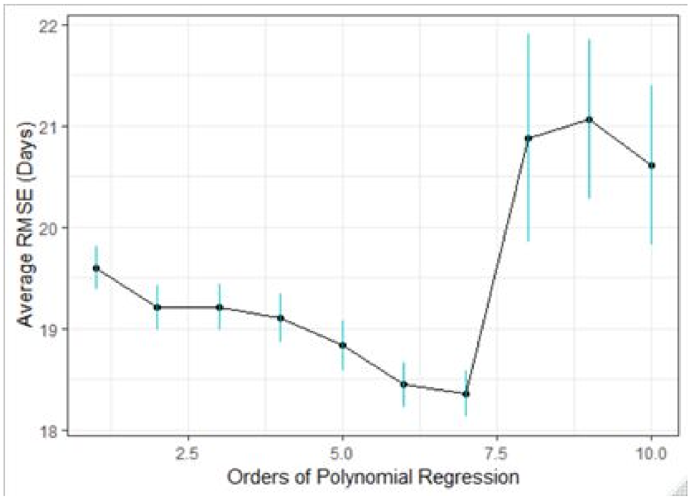
It was found that the most significant feature was weekly driver revenue. We advise that Lyft should take direct measures to increase driver revenue which would lead to a subsequent increase in the projected lifetime of their drivers. It was found that when driver revenue was graphed against the days drivers stay with Lyft that the locally weighted line of best fit had heavy curvature. This means that polynomial regression would allow the model to resemble the data more accurately.

Figure 8



In order to determine the order of the polynomial to be used for the regression model, k-fold validation was performed. For five loops, the data is randomly split into five sections. In each loop, four sections are used to train the model and one section is used to validate the model. The root means squared error, or RMSE, value is taken to assess model performance. The lower it is the more accurate the model is. At the end of the five loops the mean RMSE is recorded for polynomial variations of our model from order 1 to 10. The polynomial model with the lowest RMSE has the best performance in projecting the number of days a driver stays with Lyft.

Figure 9



The model with a polynomial order of 7 is revealed to have the lowest RMSE of around 18 days. The model is able to predict the number of days a driver stays with Lyft in the timeframe of ~3 months while being an average of 14 days off. The final model is as follows:

Days With Lyft=Weekly Driver Revenueβ_1^7*+ Mean Rides Per Week*β_2*+Requested-Arrived Timelapse*β_3 *+ Accepted-Arrived Timelapse*β_4 *+ Requested-Arrived Timelapse:Accepted-Arrived Timelapse*β_5

Using the outcomes of the model for every driver, the average predicted days a driver stayed with Lyft was determined to be ~55. In order to calculate the average Driver Lifetime, we applied a total of 180 variable scenarios over the timeframe of 55 days and attempted to account for the outcomes of the total amount of interactions over a span of 45 years. The average projected Driver Lifetime was determined to be ~9938 days or ~27 years.

Conclusion

After analyzing the data we have a few actionable business recommendations for Lyft to follow:

- The top 300+ earners only generate about 15 percent of the Lyft's revenue. Past a certain point its economically inefficient for Lyft to encourage drivers to ride more.
- Since drivers that drive mostly in the afternoon and drivers that drive mostly in the morning generate similar amounts of revenue and the number of afternoon rides greatly outnumber the number of morning rides, Lyft should encourage drivers to drive more in the mornings since rides in the mornings generate more revenue per ride.
- The number of rides and the number of prime-time rides are the biggest factors on the variation of driver revenue. However, the ratio of prime-time to normal rides has little effect on the variation of driver revenue so Lyft probably cannot generate more revenue by artificially increasing the number of rides prime time is applied to.
- Since the ratio of Prime Time rides to non-Prime Time rides is not a significant factor, we recommend that Lyft focuses on encouraging drivers to complete more rides to increase the amount of rides with Prime Time rates applied to them instead of focusing on increasing the number of rides Prime Time rates are applied to.
- Since the ratio of Prime Time rides to non-Prime Time rides is not a significant factor and the total number of rides and number of Prime Time rides are heavily correlated, we recommend that Lyft focuses on encouraging drivers to complete more rides to increase the amount of rides with Prime Time rates applied to them instead of focusing on arbitrarily increasing the number of rides Prime Time rates are applied to.
- Since weekly driver revenue played a huge part in predicting Average Driver Lifetime, we recommend that Lyft focuses on increasing the amount that drivers earn to incentivize them more in order to increase the amount of time they work for Lyft.