Proposal for Lyft Data Challenge 2019

Team: GirlsWhoCode
Xin Hao Erin Liu
Sep 15, 2019

In this Lyft Data Challenge, we explored and analyzed the given dataset of drivers, rides and event timestamps. We've built our statistical models, tested given data, analyzed based on visualization and given our business recommendation.

PART ONE: Introduction

After exploring the three given datasets, we had the following findings:

As we define a driver's lifetime value as related to his career length, average rides and profit, we find that average distance, responding time, arrival time, standard deviation of earnings and proportions of night-time rides and weekday rides greatly contribute to a driver's lifetime value.

Our assumptions:

We are making the following assumptions regarding this data challenge:

1. We set the time of the most recent ride in the dataset we're given (2016-06-26 23:57:45) as the "current time".
2. Any Lyft driver who hasn't finished a ride within the 10 days from now is considered inactive. We know a 10-day break may be possible and even common for drivers, but considering the limited data we have and our need for training data to predict drivers' lifetime with Lyft, we're making this somehow bold assumption.
3. We don't know how much profit a ride provides to Lyft . So we are generalizing the fares of rides to represent both the profit drivers made and the profit Lyft got. We believe this assumption is fair because the actual values of drivers' and Lyft's earnings are proportional to drive's fares.
4. With base fare and service fee, each ride is priced directly proportional to its distance and duration. We priced ride sections during prime time 25% higher.

Data Cleaning

We began the data challenge by regular data cleaning, ignoring rides that have missing information (eg. driver's on-board date or drive's accepted time) and wrong information (eg. ride accepted time is before ride requested time). We then group the three files together into a drive-centered dataset (data_all_combined.py), containing each drive's driver id, event types information, drive distance (meter), and drive duration (second). This file contains information of 148639 drives & 837 drivers and is our essential exploring target.

PART TWO:

Main Factors of Drivers' Lifetime Value (Response to 2.a)

We have proposed, explored, and analyzed the following factors (main factors are colored in blue);

1) Basic Info of drivers:
    a) average speed ;
    b) average distance;
    c) Total ride durations;
2) Profit & Earnings:
    a) standard deviation of profits per day
3) Duration for rides:
    a) average duration per ride;
    b) average responding time per ride;
    c) average arrival time per ride;
    d) average waiting time per ride;
4) Time slot of main rides:
    a) proportions: 9 am to 5 pm rides;
    b) proportions: 10 pm to 6 am;
    c) proportions: 8 pm to 10 pm;
    d) proportions: weekdays;

1.a Average Speed

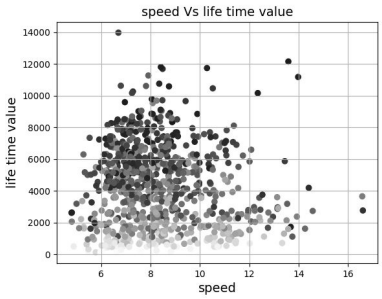
The average speed of drivers are distributed more sparsely than other factors, ranging from 4 m/s to 15 m/s. This leads to our assumption that Lyft drivers tend to drive in local and downtown area instead of suburban district. Yet the average speed does not affect a driver's career length too much overall.

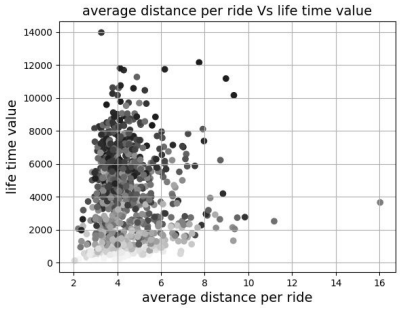
1.b Average Distance

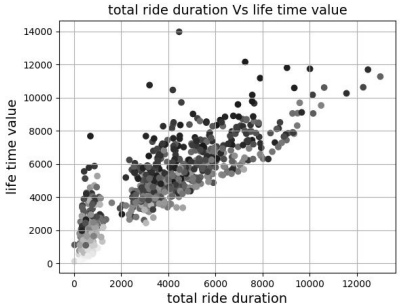
The average distance centers at 3 to 6 miles per ride. The average distance of drivers with short career length take clusters between 2 to 5. Generally, the drivers with average distance/ride over 8 miles do not have a consistently higher lifetime value, which is consistent with the pricing of Lyft and our definition of lifetime value. Based on our visualization and analysis of data, drivers with higher lifetime value prefer to accept drives with less than 5 miles. Due to pricing with the service fee and base fare, this corresponds with the drivers' general preferences.

1.c Total Ride Duration

There exists an approximately linear relationship between total ride duration and a drivers' lifetime value, which is expected since our lifetime value is based on a driver's career length, which is closely related with his total ride duration.

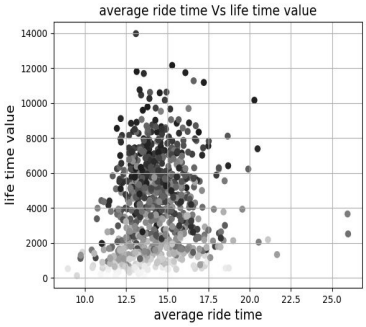






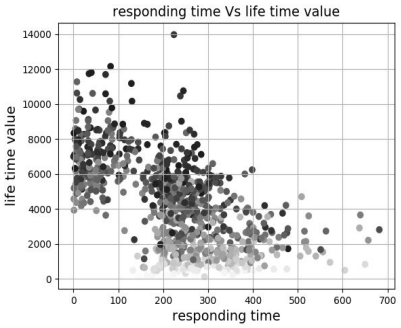
2.a. Standard Deviation of Profits per day

The average standard deviation of a drivers' earning per drive is approximately 9, higher than our expectation. We found an exponential growth of a driver's standard deviation of his earnings per day when it's less than 10. This result lead us to conclude that a driver with higher lifetime value would ride with various distances, time span and duration.



3.a Average Duration per ride

The average ride duration mainly clusters between 11 mins - 18 mins. We found that the drivers whose lifetime value is higher than 6000 mainly have drive length between 12 mins and 17 mins. Overall, the drivers' lifetime value increases linearly with the increase in their average duration per ride.



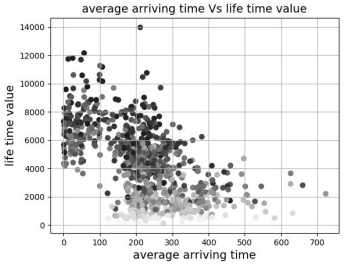
3.b Average Responding time

The average responding time is the time difference between the time a passenger requests the drive and the time the driver responds. This variable indicates the responsiveness and the activitiveness of a Lyft driver. We found that the drivers with higher lifetime value tends to respond in shorter time. There exists an inverse relationship between the responding time and lifetime value. Drivers whose average responding time less than 100s have no less than 4000 lifetime value. They also tend to have longer career lengths, indicating that the longer Lyft drivers have

worked, the more sophisticated they are, and the faster they will respond to nearby requests. Interestingly, drivers whose responding time is over 500s have approximately less than 4000 lifetime value.

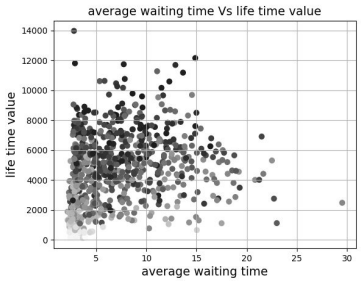
3.c Average arrival time per ride

The average arrival time is the time that a driver takes to arrive at appointed location after accepting the request. It's highly correlated to the quality of a Lyft driver. Similar to responding time, there's an inverse linear relationship between a driver's lifetime value and its arrival time. Based on our calculation, drivers with arrival time less than 150s have an average lifetime value higher than 7500 and career length longer than 40 days. This proves that shortening drivers' arrival time can potentially increase their lifetime value.



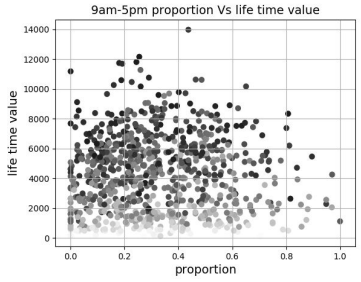
3.d Average waiting time

The average waiting time indicates how long it takes a Lyft driver to pick up his passenger after his arrival. As expected, this value depends a lot on the passager's performance. The average waiting time varies from 3 min to 9 min. After careful analysis, we don't think that this is a useful indication of the main factors that affect a driver's lifetime value.

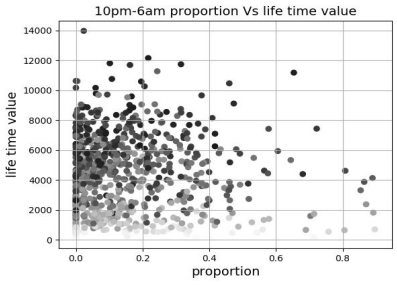


4.a Proportion of 9 pm to 5pm rides

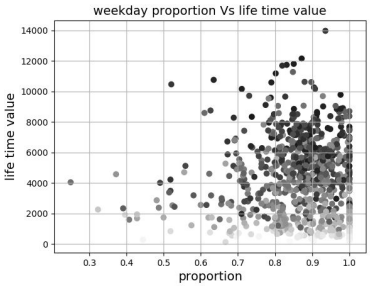
This is the proportions of the number of a driver's rides between 9AM to 5 PM to his all rides, which is their daytime rides. We first found that the average of this proportion clusters at 25% to 50%, which is less than our initial expectation. There does not exist an obvious contribution between the driver's work time slot and their lifetime value. And we are trying to narrow down the time range in the future.



4.b Proportion of 10 pm to 6 am rides

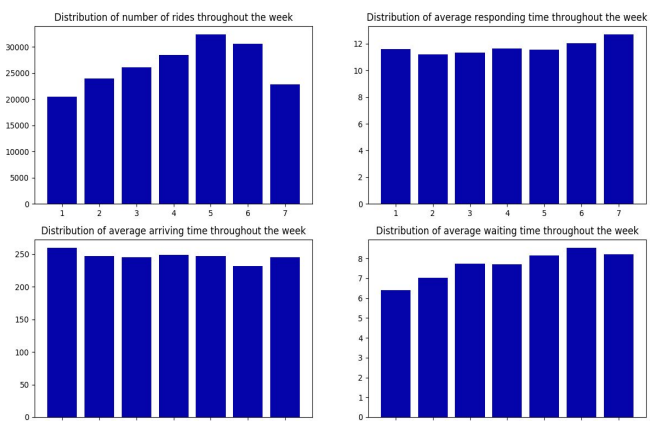


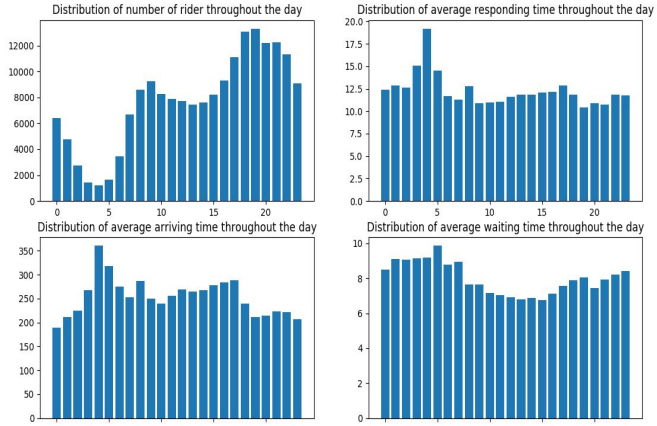
This is the proportion of the number of a driver's rides between 10PM to 6 PM among all his rides. Expectedly, most drivers' night time rides are less than 20%. However, drivers with over 20% night time rides have higher lifetime value and others. We believe this is related to the variable prime time, as the demand of night-time rides are generally higher than the supply.



4.c Proportions of rides on weekdays

This denotes the proportion of a driver's rides on weekdays. Over 80% of Lyft drivers have over 75% of rides on weekdays, many of whom also have high lifetime value. The graph also indicates that having more rides on weekends won't enhance a driver's lifetime value; and only a small fraction of Lyft drivers mainly work during weekends.





Driver's Lifetime Prediction (Response to 2.b)

Find Inactive Drivers & Build Predicting Models

Under our assumption, we found in total 287 drivers that have already left Lyft. Their average working length at Lyft is ~28 days. We first investigated the relationship between their working length with the following factors by calculating their pearson coefficient: the average number of rides they finish per day; the average profit they earn per ride; their average response time to a ride request; their average arrival time to an accepted request; their average waiting time for the passengers (pick-up time minus arrival time); their average speed during all rides; the proportion of their rides from 10:00pm to 6:00am & 9:00am to 5:00pm & 20:00pm to 22:00pm & on weekend.

To start off, we tried multivariate linear regression as it's in general a good place to start. We used 80% of these drivers' information regarding the factors in interest to build the model and the remaining 20% to validate and tune the hyperparameters.

	Unnamed: (	driver_id	career_len	les_per_da	profit	esponding	arrival	waiting	speed	al_duration	age ride tin	e dist per ri	10_to_6	9_to_5	prime_time	Unnamed: 0.1
Unnamed: 0	1		-0.668347	0.14359208	0.1022117	-0.0110789	-0.0279711	-0.0202703	-0.0905781	-0.2941214	-0.1364425	-0.1524584	0.01768435	-0.0263159	-0.2951555	-1
driver_id																
career_len	-0.668347		1	-0.1713066	-0.140512	-0.1435092	-0.1368484	-0.0323747	-0.0091387	0.40396922	0.16279468	0.08591701	-0.149219	0.04494391	0.0162874	0.66834703
rides_per_day	0.14359208		-0.1713066	1	0.96112936	-0.3488624	-0.3182916	-0.0451889	-0.2184623	0.49341587	-0.0207029	-0.2002274	-0.0332605	0.03036825	0.06878542	-0.1435921
profit	0.1022117		-0.140512	0.96112936	1	-0.2490817	-0.2034812	-0.0011099	-0.0501916	0.5172465	0.16931008	0.04060705	-0.0276478	0.06561443	0.03549239	-0.1022117
responding	-0.0110789		-0.1435092	-0.3488624	-0.2490817	1	0.96846668	0.1928549	0.47676678	-0.3861103	0.18736099	0.50113823	-0.0958656	0.21623266	-0.1900697	0.01107893
arrival	-0.0279711		-0.1368484	-0.3182916	-0.2034812	0.96846668	1	0.24268486	0.48934885	-0.3148061	0.22148413	0.52870587	-0.0932459	0.21369252	-0.191247	0.02797107
waiting	-0.0202703		-0.0323747	-0.0451889	-0.0011099	0.1928549	0.24268486	1	0.12296445	-0.0912168	0.07953441	0.15043778	-0.0724012	0.01961969	0.01852705	0.02027028
speed	-0.0905781		-0.0091387	-0.2184623	-0.0501916	0.47676678	0.48934885	0.12296445	1	-0.0590538	0.02582001	0.844797	0.37020147	-0.140644	-0.10534	0.09057812
otal_duration	-0.2941214		0.40396922	0.49341587	0.5172465	-0.3861103	-0.3148061	-0.0912168	-0.0590538	1	0.13511876	0.02321982	-0.0579492	-0.007715	0.14315926	0.29412137
erage ride time	0.1364425		0.16279468	-0.0207029	0.16931008	0.18736099	0.22148413	0.07953441	0.02582001	0.13511876	1	0.54380378	-0.3127734	0.34066512	-0.1218593	0.13644245
age dist per ride	0.1524584		0.08591701	-0.2002274	0.04060705	0.50113823	0.52870587	0.15043778	0.844797	0.02321982	0.54380378	1	0.12860981	0.06084135	-0.1558795	0.15245843
10_to_6	0.01768435		-0.149219	-0.0332605	-0.0276478	-0.0958656	-0.0932459	-0.0724012	0.37020147	-0.0579492	-0.3127734	0.12860981	1	-0.5398625	0.19709518	-0.0176843
9_to_5	-0.0263159		0.04494391	0.03036825	0.06561443	0.21623266	0.21369252	0.01961969	-0.140644	-0.007715	0.34066512	0.06084135	-0.5398625	1	-0.2283627	0.02631586
prime_time	-0.2951555		0.0162874	0.06878542	0.03549239	-0.1900697	-0.191247	0.01852705	-0.10534	0.14315926	-0.1218593	-0.1558795	0.19709518	-0.2283627	1	0.29515549
Unnamed: 0.1	-1		0.66834703	-0.1435921	-0.1022117	0.01107893	0.02797107	0.02027028	0.09057812	0.29412137	0.13644245	0.15245843	-0.0176843	0.02631586	0.29515549	1

The result looks only decent, so we then proceeded in trying logistic regression, the random forest model, and finally the decision tree model. Results are summarized as follows:

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Linear Regression	8.564	110.180	10.497
Logistic Regression	9.045	148.318	12.179
Random Forest	11.072	179.834	13.410
Decision Tree	12.837	252.511	15.891

As shown from the results, the simple linear regression models remains the best. Even though the error is still significant, we decided it's the best fit for our limited given dataset.

Predict All Drivers' Career Length with Lyft

After we've trained the model, we fed the information of all 837 drivers into the model and got their projected career length with Lyft.

	Unnamed: 0	driver_id	career_len	les_per_da	profit	responding	arrival	waiting	speed	al_duration	age ride tim	e dist per ri	10_to_6	9_to_5	prime_time
Unnamed: 0	1		0.75828758	-0.0902626	-0.0501718	-0.0311487	-0.0173322	-0.0542798	0.07357191	0.36307685	0.17956852	0.15812868	-0.0710536	0.06737181	0.14722136
driver_id															
career_len	0.75828758		1	-0.0999197	-0.0738407	-0.1500628	-0.143351	-0.1173922	0.00938039	0.4374134	0.14367976	0.08760404	-0.1366227	0.09335762	-0.0376096
des_per_day	0.0902626		-0.0999197	1	0.95882863	-0.3938864	-0.356633	-0.0588678	-0.2867587	0.53892948	-0.0199927	-0.2506042	-0.0310351	0.04485172	0.1253455
profit	-0.0501718		-0.0738407	0.95882863	1	-0.2926491	-0.2433184	-0.0130232	-0.1041476	0.57492001	0.17510763	-0.0052636	-0.0343667	0.07564679	0.07240296
responding	-0.0311487		-0.1500628	-0.3938864	-0.2926491	1	0.96139929	0.22256272	0.47818459	-0.4418569	0.22136181	0.50647886	-0.1425239	0.18967521	-0.2476655
arrival	-0.0173322		-0.143351	-0.356633	-0.2433184	0.96139929	1	0.26417578	0.48792056	-0.3815012	0.25007977	0.52762493	-0.1471366	0.19279496	-0.257758
waiting	-0.0542798		-0.1173922	-0.0588678	-0.0130232	0.22256272	0.26417578	1	0.15147304	-0.130865	0.08087198	0.17517595	-0.0701694	-0.0075658	-0.0153765
speed	0.07357191		0.00938039	-0.2867587	-0.1041476	0.47818459	0.48792056	0.15147304	1	-0.0828705	0.09983216	0.85707948	0.30912155	-0.1322106	-0.2059259
al_duration	0.36307685		0.4374134	0.53892948	0.57492001	-0.4418569	-0.3815012	-0.130865	-0.0828705	1	0.15325993	0.00729167	-0.030842	0.00790086	0.15503077
age ride tim	0.17956852		0.14367976	-0.0199927	0.17510763	0.22136181	0.25007977	0.08087198	0.09983216	0.15325993	1	0.58464688	-0.3461669	0.30300373	-0.168157
e dist per ri	0.15812868		0.08760404	-0.2506042	-0.0052636	0.50647886	0.52762493	0.17517595	0.85707948	0.00729167	0.58464688	1	0.0571848	0.04331434	-0.2561576
10_to_6	-0.0710536		-0.1366227	-0.0310351	-0.0343667	-0.1425239	-0.1471366	-0.0701694	0.30912155	-0.030842	-0.3461669	0.0571848	1	-0.5269551	0.20888725
9_to_5	0.06737181		0.09335762	0.04485172	0.07564679	0.18967521	0.19279496	-0.0075658	-0.1322106	0.00790086	0.30300373	0.04331434	-0.5269551	1	-0.2905311
prime_time	0.14722136		-0.0376096	0.1253455	0.07240296	-0.2476655	-0.257758	-0.0153765	-0.2059259	0.15503077	-0.168157	-0.2561576	0.20888725	-0.2905311	1

PART THREE: Driver's Lifetime Value (Response to 1 & 2.a & 2.c)

Here we propose a formula to calculate a driver's lifetime value to Lyft:

Lifetime Value = Average Rides / Day * Average Profit / Ride * Projected Career Length

This is essentially calculating the total amount of revenue a driver can make for Lyft throughout his/her entire driving time with Lyft.

With the definition, we easily calculated the lifetime value of all the 837 drivers as well as their pearson coefficient with several factors we thought might be influential.

Except for the three values we used in lifetime value calculation, we also identified responding time, arrival time, profits, proportions of night-time and weekday rides as important factors that influence drivers' lifetime value due to their strong correlation. To more vividly demonstrate the relationship between these factors, we once again resorted to the table above.

PART FOUR: Business Side Analysis (Response to 2.d)

Based on our analysis of factors above, we summarize our business proposal as following:

1. **Improve responding time & Arrival time**
   As discussed in Part 2, we found that drivers with smaller responding time and arrival time will generally produce greater lifetime value. We recommend Lyft to improve their responding time with some tips and trainings provided by Lyft. For business proposal, responding time and arrival time is also critical to the passenger's experience and feelings.
2. **Increase the number of drivers for night rides**
   Since most of the drivers prefer to work during the daytime, night-time drivers generally have higher profits and earnings. We propose that for Lyft, it is more user friendly to explore more night-time drivers, which will increase overall lifetime value for the drivers.
3. **Focus on the improvement of short-time rides and local rides**
   Based on drivers' average riding duration and distance, we conclude that short-time, low-speed rides and rides in local and downtown area are the major rides for Lyft drivers. For Lyft, it is significant to improve the drivers' driving experiences for short-time drives, in order for a higher lifetime value, longer career length for drivers, and a better riding experiences for passengers.

PART FIVE: Summary

We have listed all our essential findings above after exploring the limited given dataset. We believe in the thoroughness of our analysis as we've explored a great variety of factors that may affect Lyft drivers' career length and lifetime value. We've also found a good way to demonstrate their correlation by multiple scatter plots and grid of pearson coefficients. Admittedly, our machine learning predictive accuracy isn't very good, yet it still gave us usable prediction results. If given a bigger dataset covering more information regarding drivers' background, our model will certainly yield higher precision.