Lyft Data Science Assignment

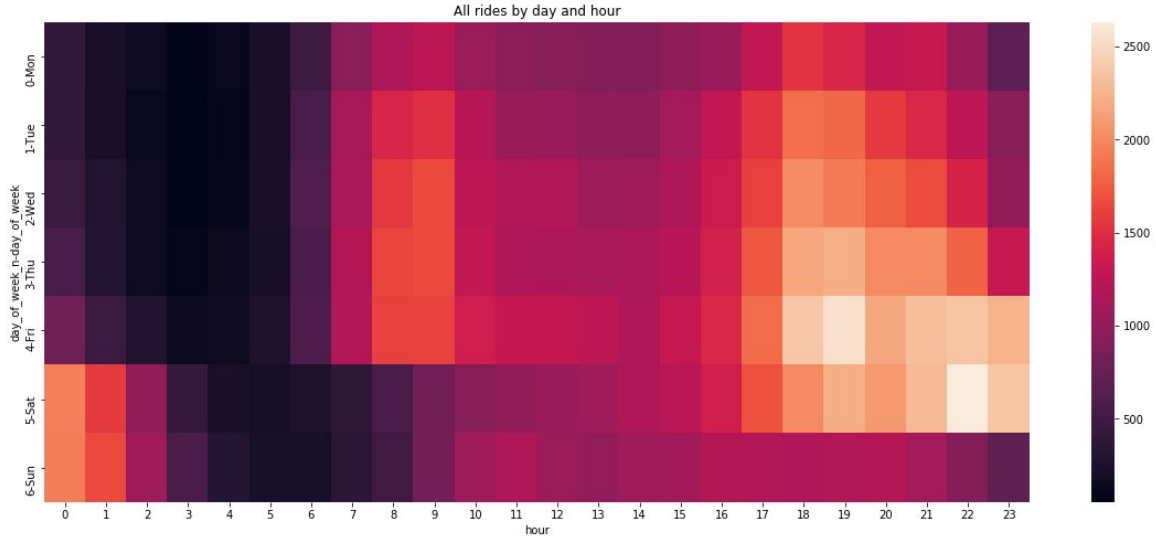
Thanks for evaluating my assignment and opportunity to interview!

Assignment

Lyft is a two-sided marketplace with drivers and passengers. Every day new drivers join the platform and existing drivers either drive or they do not. Suppose you are working as a Data Scientist on the Driver Retention team whose primary goal is to reduce the rate of churn of activated drivers (a driver becomes 'activated' once they complete their first ride).

Exploratory Data Analysis: Spending some deep diving into Rides Data.

Hourly Rides volume per hour heatmap Lighter color shows higher volume.



- All rides_heatmap - Friday night and Saturday Night hours dominate rides_per_hour- could be riders are taking rides to restaurants/bars/events that are are open late night and public transit/driving is inconvenient.

- So far It looks like Visually Sunday 5pm - Friday 5pm is (More of weekday pattern) and (Friday 5pm- Sunday 5pm) are more night heavy segment of Socializers

- From this data Looks like a Bustling Downtown city with Lots of night events / Restaurants

More Charts in ipython Notebook

a) The team would like to understand churn better. Explore the data to:

1) Define and justify the criteria for a driver to be considered churned:

My Strategy to understand Churn

1) understand the trend using Linear regression as baseline model on all the days_since_onbaord_date

2) Use Kmeans clustering to understand the clusters that have some special characteristics

3) Need to explore some vector model where all the drivers can be clustered based of the activity and other features vector -Future Work

Churn Criterion:

'''Define CHURN

There seems like 2 main clusters drivers (mainly based of rides volume per week) with low rides volume, (last 6weeks) avg_rides_per_week < 5 and avg_rides_per_week > 5

The retention strategy is going to be different for each cluster.

Low ride volume drivers are probably only looking at ways to get more rides - Ideally to increase retention low rides driver need to be converted into 5+ rides_per_week.

Defining Power Drivers and Low volume Drivers

Our Last Ride pickup data is 2016-06-27 and Last onboard date is 2016-05-15 - so pretty much have 6 weeks range since last onboarded drivers ride data.

Since we have Six weeks of data for all the onboarded Drivers - Using moving average of weekly rides on Last 6 weeks to determine if the Drivers are low volume or power drivers

driver_stats_all['last6_weeks_avg_rides_ma'] =
driver_stats_all['last6_weeks_avg_rides_ma'] = (driver_stats_all['12'] + driver_stats_all['11'] + driver_stats_all['10'] + driver_stats_all['9'] + driver_stats_all['8'] + driver_stats_all['7'])/6

Power drivers retention is critical and needs a different strategy and keeping conservative cut off for 1 week (7 Days) to) last Drive to determine the drivers who have high probability to churn.

Non Power Drivers (<5 Rides in last 6 weeks) - are given more relaxed timeframe of 4 weeks of inactivity to determine Churn

For low_ride drivers avg_rides_per_week <5 and no ride pickedup since last 4 weeks
For Power drivers avg_rides_per_week>5 and no ride pickedup since last 1 week

Also would be good to limit false positives by letting driver mark as in preference for days/months to drive, vacation.
so incase driver went on vacation during that time or has seasonal patterns in work and drives only certain months would be good to set up in the model as a different category of drivers

Ideally I should have set up way to test the criterion by separating out June Data for Test - something to do in future

'''

2) Are there specific segments of drivers (based on activity / patterns of driving /contribution to the marketplace / etc.) that churn earlier or later than average

Power users (avg_weekly_rides>5) are different than Low volume users (avg_weekly_rides<5)

Refer to Jupyter Notebook on more details: here are my top variables summary:

Main Top Variables to help lower the churn rate

Power_Driver (Mean here is Churn Rate) : Power Drivers have Lower Churn rate than Low Volume Drivers

:

power_driver	size	sum	mean
0	461	153	0.331887
1	376	32	0.085106

#prime_time_pct:
Churn rate drops down for low power drivers to get the Prime time rides vs Non prime drivers where prime_time_pct > 50%
#Probably the Power drivers that are Driving >50% of Rides as Prime time rides end up competing/ waiting for prime time rides hence wasting the ride opportunity

 # Miles_per_hour Analysis:

Churn rate Drops with Higher Miles_per_hour as that will mean higher Profit/Earnings per hour for Both Power Driver and non Power Driver

#**Day1_rides:**
#Most of the Drivers get first ride on Day1 of onboard 804/837 = 96.4% binned by 5 rides
I see higher churn rate with increase in Day1_rides for both groups of drivers
#-Probably Drivers who take much more rides get frustrated on the day1- as they are still learning the ropes - so end up not doing well - So better training will probably help Driver retention

#Week1_rides:

#bins here are 0:0, 30:1-30, 60:31-60, 90:61-90

As we see as driver gets more rides in week1 motivates or is a signal for driver who is intrinsically committed to drive more - hence lower churn rate is for drivers who drove 90+ rides in first week.

Adding week1 Rides has much more and different impact for Power_Drivers [higher churn for 30-60 bin vs 0-30, 60-90] than Low Volume Drivers

#So opportunity sizing Depends which bin we are adding more Drivers to.

week1 Rides has little impact for non Power_Drivers
#and for power drivers somehow the 30-60 bin is not ideal (see an increase there) but it drops again for 90-120 week1 rides set

#n_active_days_pct
More active days for low power driver will improve Churn rate

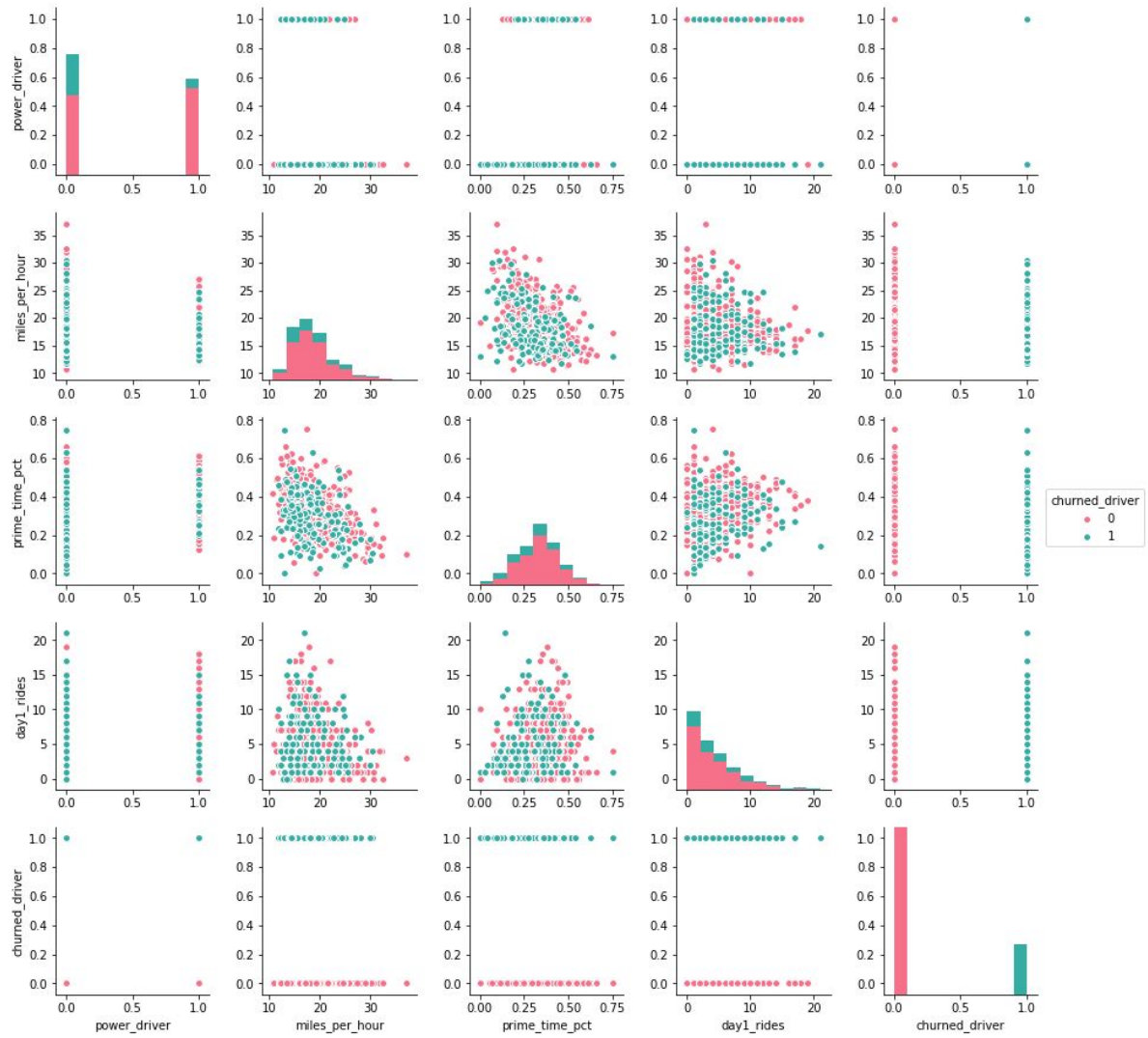
#mins_per_ride:
Churn Rate Does drop especially for non Power users to get the longer duration rides and probably longer duration should mean more money especially in this region

#Miles_per_ride:
Churn Rate Does drop especially for non Power users to get the longer distance rides (miles_per_ride)

#Weekday_pct:
#Higher Weekday_pct rides are Adding to higher Churn Rate for both power and non power drivers

#FS_nighters_pct:
#For non Power Driver Churn rate is Higher for drivers who dont do Friday/Saturday Nights FS_nighters_pct=0,and it drop down quite a bit if Drivers are doing the Friday/Saturday Night Drives.

#For power user Doing more than 30% of Rides as Friday/Saturday night are showing bit higher churn



3) What are relevant business uses of accurate prediction of driver churn?

Prediction churn Accuracy impacts:

- Right timing of intervention and developing products to retain drivers
- Build Virtuous cycle: Happy drivers=>[More referred Happy Drivers] => Happy passengers => Happy LYFT :)
- Saving Costs in Promoting False Positives

b) Next, the team would like to opportunity size reducing the rate of churn using interventions at different stages of a driver's life-cycle in order to prioritize their roadmap: Estimate the impact of:

1. Doubling the number of rides in an activated driver's first week.

Increasing rides for drivers in the low rides volume driver can impact the churn rate in meaningful way.

2. Having each driver give their first ride on their onboarding date.

Too much variability in Day1 rides data to make useful recommendation and 96.4% had day1 ride and almost 100% had week1_ride

What are possible product features or driver interventions that might accomplish

1 or 2?

1) Promotion [$30-$50] for drivers who make at least 30 rides in week1: Promotion as in if they drive at least 30 rides in 1st week- they can claim additional bonus of ($30 - $50) on the rides they did. For lyft it will mean assuming avg_cost is $2 per mile and median miles_drove is 2.5 miles, total_revenue_per_driver = 2.5*2*30 = $150 - hence a promotion of $30-$50 will be good intervention.

2) Product: Training sessions in person/online by experienced drivers in onboarding to help improve the earning potential that fits to different driver needs. Build online Community or events for drivers to share their experiences.

3) Product: Increase frequency of In app Driver Tips notifications for new drivers - to learn the platform faster and prioritize the help tickets for them.

4) Product: [Driver location favorites]: Driver chooses the favorite locations as work /home and time_range associated with it to help them pick rides when they are commuting to/from the work/home. Pretty Much anything that can drive the rides pickup will help drivers to stay on the platform.

Which hypothesis from b) should the team work on first? Why?

Promotion [$30-$50] is usually the most effective and easier form of intervention to deliver the results. All other product features can be tested but it takes more time and plan to develop / improve the onboarding program or improve ride recommendation algorithm: more of longer term effort.

c) Finally, suppose the team wants to design an experiment to test the following hypothesis: "*eliminating the Prime Time feature will decrease driver churn*".

Eliminating The Prime Time feature:

1)Control it for Power Drivers and Low volume Riders, As there could be the reason primetime is causing the Drivers to Drive more - More they drive - Better for Retention.

2 Hypothesis:

1 : Eliminating Prime time does not impact the Driver Ride Volume

2: Eliminating Prime time will positively impact the Driver_Retention_rate

Experiment Setup:

Setup a T-Test experiment to test the impact of No prime_time in Treatment_City versus Control_City- Choose city groups and Random Sample Drivers to get a balanced sample. Randomly pick the drivers in both groups to ensure randomization and balanced samples. It will be inaccurate to do testing in one city area by randomly assigning drivers - as Prime time Drivers can impact positively or negatively non prime time drivers to ride less or more during peak hours hence biasing the data.

1. What will be your control and treatment populations? Why?

   Control: City A group with Prime time ON

   Treatment: City B - Similar to City A with Prime time OFF, Would be great if we could store data on rides which are expected to be Prime time rides , so we can compare on across The Prime time / Non Prime Time metrics delta as well on the Control and treatment.

2. Which primary metric and secondary metrics will you track? Based on these metrics, what is your criteria for rolling out Prime Time feature elimination?

Main metric:
Rides_per_driver
Driver_Retention_rate/ Driver Churn Rate

Other metrics:

Revenue Metrics: Avg/ Median Earnings_per_mile, Avg_Earnings_per_hour
Funnel Metrics: Ride Completion Rate as in Ride_complete/Ride_request
Quality metrics: Avg Driver rating, Customer Support Tickets_total

Mainly if we see impact of Rides_per_driver is not getting affected by Prime Time elimination And Prime Time Elimination improves the Driver Retention

Prime Time Elimination is a big deal as 35% of Rides use Prime Time.
So need to balance the impact of Eliminating Prime time on Rides Volume vs impact of Driver Churn Rate.

3. Based on the dataset, how many weeks will you need to run the experiment to reach statistical significance?

Power analysis:

a) Statistical significance level - 95%, p-value =.05
b) Power (1-Beta) = 80%
c) Effect Size: 35% Rides are Prime Time Rides essentially 65102 rides are prime time in 12 weeks timeframe . To get detect + 5% change in Rides_per_driver and -10% change in Driver_churn_rate

d) Time frame: Based of sample size plan the Test for 2 weeks

Total rides in 12 weeks: 184209
Prime time total rides in 12 weeks: 65102
Avg Rides_per_Driver 220
Standard Deviation Rides_per_Driver 178

Sample Size for each group to detect 5% delta on avg_rider_per_driver: 4120

('Detect Churn Rate Delta between baseline and effect size lowering churn rate by 10% ', 0.221, 0.199)
Sample Size for each group to get 10% Delta on churn Rate : 7131

Assuming We are able to record for treatment population Rides which would from Prime time algo be Prime time ride then we can limit our data to only compare Prime Time rides on Control and treatment

('Weeks it will take to get the Sample size setup for Prime Time Ride comparison for Avg_rides_per_driver: ', 1.0)
('Weeks it will take to get the Sample size setup for Prime Time Ride comparison for Churn rate:', 2.0)