**Team # 42 - DS4A: Capstone Project Proposal**

**Proposal 1**

**Lyft: Driver Lifetime Value & Churn**

**Problem Statement: (please describe the project in 1-2 sentences)**

Lyft being one of the most popular ride sharing applications, offering vehicles for hire, scooters, bicycle-sharing, and much more, has acquired a significant number of customers/passengers. To retain the high market on demand side, Lyft has to keep up with the supply side of drivers. Thus, in order to ensure the organization’s success, it becomes very important for the company to understand the value of the drivers and to ensure the retention of profitable drivers in the company.

Our team would like to analyse the data to estimate the Driver’s Lifetime Value (LTV) and propose strategies to reduce churn. Recognizing the most profitable drivers based on LTV can help Lyft not only reduce the churn and maintain a steady pipeline of drivers but adjust the marketing and service costs to focus only on the most profitable segments.

**Which question(s) do you want to explore? Why do you think this particular question is**

**interesting?: (1-2 paragraphs elaborating on the project’s relevance)**

1. What are the main factors that affect a driver's lifetime value?

Driver’s LTV can be approximated as the lifetime driver revenue minus lifetime driver cost that lyft had to pay to keep the driver onboard:

* Lifetime driver revenue ~ Lifetime in a time-unit \* approximate revenue per time-unit

(timeunit : week/month)

Estimated lifetime of driver may be influenced by:

* Competing salaries in the market
* Number of Rides available
* Economy

Revenue per month =

* Frequency of rides (availability)
* Duration of rides
* Distance covered during the drive (Shortest path)
* Rides during Peak time
* Number of rides at Prime location
* Lifetime driver cost - marketing and service cost to keep the driver’s positive and happy, thereby ensuring their retention

1. What is the average projected lifetime of a driver? That is, once a driver is onboarded, how long do they typically continue driving with Lyft?

Understanding what is the lifetime of the driver and what are the factors leading to their retirement can help us divide the drivers into segments. These segments can help us determine how to increase the projected lifetime of a driver with minimum extra cost.

1. Do all drivers act alike? Are there specific segments of drivers that generate more value for Lyft than the average driver?

Estimating the lifetime value can help us figure out which drivers are more profitable, if there are any. It answers the question if more numbers of drivers are more profitable or few but smart drivers (with higher LTV) can bring the same revenue. Sorting the drivers based on their values and needs can give us more insight on how to upgrade Lyft’s driver retention policies.

1. How to onboard new drivers?

This question allows us to combine all of our findings above to solve the problem of maintaining the driver’s supply and further bring more revenue to Lyft with minumum cost.

**To answer the above questions, following are the few hypotheses we would like to test:**

1. A driver is more valuable to Lyft only if they drive more frequently or for more extended periods of time

2. Competing salaries in the market can determine the lifetime of the driver

3. A driver is more valuable to Lyft during peak hours vs driving more frequently

4. More rides for drivers based on location

5. Wait time vs rating:

* wait time1:time taken to accept the request after the customer has sent the request (accepted\_at-requested\_at)
* Wait time2: time taken by the driver to arrive after he has accepted the requested (arrived\_at-accepted\_at)

**Which datasets do you plan to use? Are there any data sources that you have**

**failed to find? (List relevant datasets)**

Apart from the three datasets given we plan to use the following datasets if available:

1. Ratings of the driver
2. Location of the ride(picked up,dropped at)
3. Some dataset where we can get how many rides were cancelled by the driver/customer
4. Competing data from other ride sharing apps to see correlations in booking
5. Existing strategies and the expenses associated to reduce churn

**Please describe the plan or methodology that you will use to answer your question (1-2**

**sentence description of statistical analysis techniques)**

1. Cleaning and pre-processing the datasets. Merging the datasets based on primary key.
2. Calculate each driver’s lifetime value by calculating every fair corresponding to each trip.
3. Add some new variables like, driver’s total days in lyft, percentage that each driver works during prime time, total distance travelled, total number of rides etc.
4. Exploratory data analysis: See correlations between variables and see what variables have a high impact on the lifetime of drivers. Perform descriptive statistics to get a sense of each variable and it’s distribution(univariate analysis). Perform bivariate analysis to understand the relationship between two variables. These will be accompanied by plots to enhance the readability.
5. Feature engineering: To derive more variables from existing variables. For example, adding binary variables like peak\_hour based on the time of the day of the ride, might also be possible that there are correlated variables so we’ll either have to remove one of them or combine to make a new variable.
6. Clustering (kmeans etc) to segment drivers in different groups to understand driver behavior. Variables like total distance travelled, total number of rides, percentage that each driver works during prime time, total days at lyft etc. can be used for clustering to understand driver behavior and form clusters based on that.
7. Modeling: Linear regression, random forest, xgboost etc. to estimate lifetime value of driver. Using this model we can predict the lifetime value of a new driver at lyft.
8. Feature importance: Based on the model, we can rank the variables in the order of the importance. Thus we can propose what variables affect the lifetime value of the driver the most. This will help us propose methods to reduce churn.

**Why would we like to work on this project and why are we a good fit for this project?**

This project will help us apply our data science skills on a real world problem. We would like to discuss with Lyft the possibility of making recommendations to help them get more (retain and attract) drivers, and make sure that our work adds more value and has a positive impact on Lyft as a company. Working on this project will give us an amazing opportunity to learn from and work with the mentors from Lyft. Their subject matter expertise will be very valuable as well as helpful for us in proceeding in the right direction.

All of us are well versed with Machine Learning & Deep Learning models (supervised, semi-supervised and unsupervised). We have done several projects that have involved applying knowledge of a data science pipeline by performing exploratory data analysis, data pre-processing, feature engineering, model selection along with hyperparameter tuning and presenting our findings. Some of us have upto 2 years of work experience and internship experience as data scientists. We also have experience in working with Churn Analysis, predicting customer churn and helping retain customers.

**Technical skills:** Python (Pandas, NumPy, Matplotlib, Sklearn, Pytorch, Keras, Tensorflow), R (ggplot2, R dataframes).

**Proposal 2**

**London Electricity Consumption**

**Problem Statement: (please describe the project in 1-2 sentences)**

The British government is trying to upgrade energy supply and tackle climate change by better understanding the energy consumption by installing smart meters in every home in England, Wales and Scotland.

**Goal**: Explore the data from smart meters and infer the energy consumption patterns to come up with strategies to upgrade energy supply and tackle climate change

**Which question(s) do you want to explore? Why do you think this particular question is**

**interesting?: (1-2 paragraphs elaborating on the project’s relevance)**

1. What factors would influence electricity consumption? What are the daily patterns of electricity consumption?

The potential factors that we think can influence the electricity consumption are Household income, house type, children in household, occupation, weekend/weekdays, public holidays, season, geography, ethnicity, age, social media activity

1. We will then forecast the energy consumption of a household-based on different factors we will explore in q1.

**Which datasets do you plan to use? Are there any data sources that you have**

**failed to find? (List relevant datasets)**

Smart meter data from London area: <https://www.kaggle.com/jeanmidev/smart-meters-in-london>

Following are the datasets we are planning to use

* Acorn\_details.csv: Acorn refers to a specific group of people with certain characteristics and then for each acorn there are demographic, geographic and other details.
* Informations\_households.csv: Information about the households, (the acorn group and tariff)
* uk\_bank\_holidays.csv: This dataset contains the data about bank holidays in UK
* weather\_hourly\_darksky.csv: Hourly visibility, temperature, precipitation, weather icons and other weather-related data

**Please describe the plan or methodology that you will use to answer your question (1-2**

**sentence description of statistical analysis techniques)**

1. Combine all blocks into a single dataframe- keeping on relevant columns.
2. Use day-level energy consumption data per household to normalize data for the inconsistent household count.
3. Explore relationships between weather conditions and energy consumptions using EDA and regressions. Create clusters for the weather data- using which we can add weather identifiers to day-level data.
4. Add UK holidays data to the day level data as an indicator.
5. Check correlations between variables and see what variables have a high impact electricity consumption.
6. Clustering (kmeans etc.) to identify groups which have similar energy consumption usage.
7. Fit multivariate time series forecast model to forecast daily energy consumption.
8. Fit neural network model (lstm etc.) to forecast daily energy consumption .