Team #10 **DS4A Capstone Project Proposal** **September 20, 2020**

PROJECT PROPOSAL #1

Problem Statement: *Predict potential usefulness and positivity of product reviews*

Which questions do you want to explore? Why do you think this particular question is interesting?
Both consumers and e-commerce businesses rely on the ubiquity of user-generated product reviews to make better-informed decisions. Reducing the costs of identifying relevant information from vast amounts of product reviews aids potential consumers with their purchase decisions and informs businesses of key product development areas to focus on based on their importance to their consumers. For example, the latter may be achieved by grouping products that possess similar product characteristics deemed valuable to consumers within a particular product category in order to improve recommendation systems used. Furthermore, the lack of quality reviews may signal the necessity of paid reviewers for businesses. To cope with the inundation of product reviews available, such stakeholders may wish to focus on those that matter the most to consumers by assessing and ranking reviews by their usefulness and positivity. While websites like Amazon allow consumers to vote for reviews found to be helpful, newly published ones that contain relevant, up-to-date information may go unnoticed when buried under top reviews.

Through this project, we will focus on Amazon product reviews and create predictive models that output the helpfulness and positivity of newly published product reviews based on the review content. To do so, we need to answer the following: *as suggested by their votes (i.e. number of people that found them useful), what properties of product reviews signal their helpfulness?*

Which datasets do you plan to use? Why? Are there any data sources that you have failed to find?
We plan to start by exploring the provided Amazon reviews datasets for multiple product categories (https://nijianmo.github.io/amazon/index.html) and may select product categories to focus on based on availability of reviews with non-zero votes (see schema here for reference). Such datasets contain the summary, message, rating, and votes received for reviews, which may be used to develop the aforementioned predictive models.

Please describe the plan or methodology that you will use to answer your question:
We are planning on to:

- Perform exploratory data analysis and visualization to identify and understand patterns in past review data
- Implement predictive modeling to predict the usefulness and positivity of product reviews based on review content (i.e. title and message)
    - e.g. Sentiment analysis to predict whether a review is positive or negative
- Incorporate predictions to assess and rank product reviews in order to identify the top 5 most helpful reviews that are positive and top 5 helpful reviews that are negative

PROJECT PROPOSAL #2

Problem Statement: *Finding contributing factors to Lyft drivers' lifetime value*

Which questions do you want to explore? Why do you think this particular question is interesting?
Lyft's on-demand ridesharing services is directly dependent on its enormous transportation network consisting of drivers and riders. Due to the double-sided nature of the marketplace, it is imperative for Lyft to maintain an adequate level of supply — its drivers. To increase drivers' lifetime value and minimize their churn in order to ensure that a consistent, reliable supply chain is in place to accommodate consumer demand, this project aims to uncover contributing factors related to drivers' lifetime value and implement strategies to maximize their value by answering the following questions:

- What are the main factors that affect a driver's lifetime value?
- What is the average projected driver lifetime value?
- Do all drivers act alike? How do drivers across different segments behave differently in terms of the lifetime value that they generate?
- What are the commonalities that exist among drivers who churned? Are churning rates higher among a specific segment (in terms of lifetime value and other demographic characteristics)?
- What are some hypotheses that would explain why drivers churn?

Which datasets do you plan to use? Why? Are there any data sources that you have failed to find?
We plan to use the following three datasets that Lyft will provide for us:

1) driver_ids.csv
2) ride_ids.csv
3) ride_timestamps.csv

as well as datasets related to drivers' demographic information if Lyft grants us the permission to access.

We are using the above data sets because we believe that in order to examine the factors relevant to drivers' churn and calculate his/her lifetime value, we need data on the trips that they have completed (mileage, duration, time). The three datasets above would allows us to gain insights into:

1) drivers' time of working and how many hours/trips they complete in one day (which may be used to deduce the reason why a driver joined Lyft and whether they are doing it as a side gig or a full-time job)
2) drivers' average mileage in a trip (a proxy for drivers' comfort level with short or long trips)

Please describe the plan or methodology that you will use to answer your question:
We are planning on to:

1) Calculate lifetime value for each individual driver
2) Perform exploratory data analysis to examine churn rate across different cohorts (summary statistics, data visualization)
3) Implement predictive modeling to predict:
    a) to predict a driver's lifetime value (Linear Regression, Polynomial Regression with Regularization if needed)
    b) to predict whether a driver will churn or not (Decision Tree as baseline and explore a combination of other models to improve the baseline results)
4) Implement clustering algorithm to identify potential contributing factors of churning