Project Proposal

Predict Uber/Lyft prices based on weather conditions.

This would be a supervised regression problem, trying to predict the target variable: price of an Uber and Lyft ride**,** based on the predictors provided by both ride and weather datasets. This information would be valuable to clients such as Uber and Lyft as the purpose of this project is to understand what factors drive the demand and supply of rides due to differing conditions such as time, distance, and weather. This provides a variety of predictors in order to observe a pattern in price changes for companies of Uber and Lyft. Other ride-sharing companies or cab companies may use this information to rival Uber and Lyft by providing discounted prices for their rides to increase ridership.

Real time ride data, collected using Uber and Lyft API queries, containing 10 features including: ride application used (Uber, Lyft), distance between pickup and dropoff, timestamp of ride, pickup location of ride, destination of ride, price of ride, surge multiplier of ride (over much price was increased), specific type of ride (Uber Black, Lyft Lux XL), and corresponding id is included. There are over 690000 ride data instances recorded for this dataset. Weather data, containing shared features with the ride dataset such as the location of the ride pickup, at timestamps ranging between November - December, contains unique features such as rain, clouds, humidity, wind, and pressure measurements. Other datasets or factors, such as events (e.g. Red Sox game at Fenway) could come into consideration by playing a factor in the pricing decisions by Uber and Lyft. These datasets will be acquired from Boston archived databases and websites, as the time range for the main datasets range from November 2017 to December 2017. All datasets acquired will be combined with the ride and weather data into a csv file. Ride and weather data is provided from: https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices and is relatively clean.

Because the target variable is a real and continuous value, a variety of supervised regression algorithms will be implemented to determine the best regression model resulting in the most accurate prediction for the target variable, ride cost. An algorithm such as multivariate regression, due to the likelihood of there being multiple predictor variables with the added weather predictors or other predictors, will be implemented and optimized to reduce the loss function, mean square error, which will indicate the accuracy of the prediction. Visualized regression graphs will display the relationship between the predictors and the target variable and report the results of the regression model.

Deliverables will include python code of the implemented algorithms for the data wrangling, data storytelling (identifying unique trends between the predictors and the target

variable), regression modeling and visualization processes. A paper documenting the methodology, displaying in depth analysis of the results and visuals from each process, and discussion of the results will be available. A powerpoint slide deck presentation will also display the results and any unique trends in the data from the data storytelling process, as well as, a succinct discussion of the results of the project. All files and documents will be stored within a Github repository.