# Project Proposal: **Uber Ride Review Analysis**

Uber ride review analysis is a Supervised Learning Classification problem where we try to classify the ride reviews given by the people as Positive (1) or Negative (0) based on the ride rating.

Uber ride reviews is an integral part of the Uber System which puts the entire transaction in check leading to a significantly more pleasant & safer service. This classification of ride reviews will help Uber to better tackle the public problems related to uber such as uber not being on time, driver not good, quality of the car is bad & multiple uber drive cancellation issues.

A review-based system is very important for any business to increase its profit and market capitalization. It is through this review-based system the company can directly communicate with the customer & get a better sense of its credibility & reliability among its users. This is reason that motivated us to approach this problem & also as team we were very keen to work on a text classification problem. Therefore, we thought "Uber Ride Review Analysis" would be an interesting dataset to work on.

We got the dataset from Kaggle, the size of the data is around 15-20 mb which consist of 1300 rows & has 3 features ride review (text), ride rating (numerical), ride sentiment(numerical). Ride sentiment is our target feature having 2 classes 1 & 0. We will be using ride review & ride rating feature to classify the ride sentiment as positive or negative

The dataset is sufficiently complex to implement the machine learning algorithms that we have learned in class, the only thing that we would have to be slightly cautious about is the data formatting & feature scaling part where we will have to convert text into numerical data.

We plan to use all the classification model discussed in our class such as Logistic Regression, SVM, Decision Trees for our project. Quality of the model will be accessed using various performance metrics for classification such as Confusion matrix, AUC ROC Curve, Recall, Accuracy & Precision. Whichever model gives us the best result for all these metrics we will use that model as a final model.

Generalization is very important as it is a way of telling us how good our model predicted the results for an unseen data & we plan to do this in our project by creating some percentage of hold out data (test data) to use for our evaluation. We will be using K fold cross validation technique for our hold out data to measure its accuracy. K fold cross validation is an essential technique which uses every point for validation without letting the model learn about it & This process would really help us to evaluate if our model is doing well for an unseen data

Complications for this project is mostly related to conversion of text data to numeric data & use of special characters in text, also consistency & completeness of the data value is what we would have to work on

As this dataset was taken from Kaggle therefore the data is already in a tabular format & we don't have to web scrape it from any other websites using API's. We just need to focus on data conversion, data cleaning & data processing.