Problem statement and datasets

1. Problem Statement

Many users have expressed their dissatisfaction regarding the ride sharing apps' policy of charging for a ride cancellation. Even with implementation of grace periods where the user can cancel for free or not charge the users if the cancellation was carried out by the driver, a hidden cost is accrued in the form of wasted program runtime, bandwidth and loss of time.

With the sudden increase in heavy competition amongst various corporations in the ride sharing market, retaining users and addressing user churn is a very real priority for any ride sharing groups should they succeed against their competition.

This data science project attempts to create a machine learning model which will predict whether a ride will get cancelled, whether from the users' side or the driver's side. Although this project will not go beyond this, knowing the possible factors which cause cancellations is the first step to optimizing the app algorithm to prevent those user-driver matches in the first place.

Possible Users / Beneficiaries: Ride sharing app users, ride sharing app corporations.

Impact: Less customers being charged cancellation fees, improvement of customer experience and satisfaction, reduced customer churn, better user retention, preventing wastage of time, processing power and electricity as hidden cost, improvement of corporate profit margins.

2. Dataset Description

Datasets Involved:

https://www.kaggle.com/datasets/yashdevladdha/uber-ride-analytics-dashboard
Uber Data Analytics Dashboard - 148,770 records

This dataset contains 21 features excluding the index. They are:

Date, Time, Booking ID, Booking Status, Customer ID, Vehicle Type, Pickup Location, Drop Location, Avg Vehicle Time at Arrival (Driver to Customer), Avg Customer Time at Arrival (Customer to Destination), Cancelled Rides by Customer, Reason for cancelling by Customer, Cancelled Rides by Driver, Driver Cancellation Reason, Incomplete Rides, Incomplete Rides Reason, Booking Value, Ride Distance, Driver Ratings, Customer Rating, Payment Method

The project relies on a single dataset obtained from Kaggle.com. Unfortunately the team was unable to find any other datasets which could meet the quality and availability of appropriate information provided by the first dataset Uber Data Analytics Dashboard by Yash Dev Laddha.

Other Considered Disregarded Datasets:

https://www.kaggle.com/datasets/yasserh/uber-fares-dataset

https://github.com/Geo-y20/Uber-Rides-Data-Analysis/blob/main/UberDataset.csv

https://catalog.data.gov/dataset/2023-yellow-taxi-trip-data

https://developer.uber.com/docs/businesses/data-automation/data-download

https://www.kaggle.com/datasets/abhi231092/uber-rides-data-bw-city-and-airport

Here is the list of some other datasets that we considered but ultimately disregarded due to a variety of factors:

1. It is missing a feature which records whether the ride was cancelled or not. Most of these datasets only concerned rides where rides took place successfully, with the exception of the fifth dataset, however the fifth dataset only concerned rides to and from the airport which is incredibly limiting.
2. The second most fatal flaw with the information provided by all of the above dataset concerns that it only contains the pickup and the destination coordinates with no information regarding the time it took the driver to pick up the rider which is another important factor that we require for modelling in our project and is provided by Yash Dev's data set.

The variables in the dataset are divided into two categories:

- **Predictor / Feature - Independent Variables (X)**: Pick up location, Drop location, Average Vehicle Time at Arrival, Booking Value, and Ride Distance.
- **Target / Label - Dependent Variable (Y)**: Booking Status (Completed, Cancelled by Customer, Cancelled by Driver, No Driver Found, Cancelled by Customer, Incomplete)

3. Justification

The team's justification for using this dataset as well as it being the only dataset is based on several key points:

- **Sufficient Data Volume**: The dataset provided by Mr Yash Dev Laddha roughly contains around 150,000 entries, which we believe is sufficient enough to train our models effectively and satisfactorily.
- **Contains Relevant Label Variables:** The dataset contains appropriate variable data which the team believes is both relevant and necessary to build our model. Most of the other datasets were rejected on this premise of not containing data about cancelled rides.
  The last dataset was rejected for lacking pick up time, as well as lacking destination diversity.
- **Feasibility of Preparation**: Although the dataset contains excessive variable data such as data regarding incomplete rides and random cancellations due to random

circumstances such as inputting the wrong destination, changing their minds and the driver not moving.
We have a clear plan to handle this by dropping the irrelevant rows by first considering their relevancy to the problem at question.

We will also discuss with Professor Dr Chantri regarding whether it is viable and within the bounds to establish a project with only one dataset or whether we should consider changing our project.

Team Members:
Biki Nath Newa - st126199
Hein Min Htet - st126459