Data Intake Report

Name: G2M Insight for Cab Investment firm

Report date: 12 October 2022

Internship Batch: LISUM14

Version: 1.0

Data intake by: Kushal Samani

Data storage location: https://github.com/KushalSamani/Go-to-Market-Insight-for-Cab-Investment-firm

Tabular data details

Total number of observations	359392
Total number of files	4 CSV files were converted into 1 main CSV file.
Total number of features	18
Base format of the file	.csv
Size of the data	61.2 MB

Proposed Approach:

- 4 CSV files were downloaded from the link provided.
- The file, Cab_Data.csv, had a column named 'Date of Travel,' which needed correction in formatting and was done using excel.
- The files were then read in Jupyter Notebook using a pandas data frame, and four hypotheses were formed.
- After understanding the data in all four files using Jupyter Notebook, all the tables were joined into one main data frame using Inner Joins.
- The main data frame, after inner joins, consisted of 14 columns and 359392 rows. Four more columns were calculated and added to the main data frame for analysis. These columns are Profit, Profit/km, Year, and Price/km.
- All the columns have non-null values, and it is assumed that there is no repetition of transactions in the final data frame.

- No dates have been disregarded. The starting date is 02/01/2016, and the ending date is 31/01/2018.
- It is assumed that Income does not play a significant role in deciding which cab the user will book.
- After the Final Data Frame was ready, with 18 columns and 359392 rows, the data was visualized using Matplotlib and Seaborn for basic visualizations and Tableau for intermediate visualizations.
- Following the EDA, answers to the hypotheses and a final recommendation were provided.

Model Building:

- Building a model aimed to predict whether a person will choose Yellow Cab or Pink Cab when specific parameters are valid.
- It has been assumed that Company, Customer ID, Income, Transaction ID, Date of Travel, Cost of Trip, Payment Mode, Population, Users, Profit, Profit/km, Year, and Price/Km do not play a role in an individual's decision of booking a cab.
- The remaining non-numerical data was then converted to numerical data with the help of the 'dummies' method in Pandas.
- With the help of sklearn, a logistic regression model with an average accuracy of 79% was created.