**Capstone Project Submission**

**Instructions:**

1. Please fill in all the required information.
2. Avoid grammatical errors.

**Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**

**Introduction**

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars and luxury limousines), commuter vans, and paratransit vehicles. Over 200,000 TLC licensees complete approximately 1,000,000 trips each day.

**Problem statement**

Our task is to build a model that predicts the total ride duration of taxi trips in New York City. Our primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

**Approaches**

Initially we started understanding and analysing the data. After some initial analysing, we started data analysing. The data we have noticed that some columns like id and last review are not useful for us so we drop both the columns. Some columns and rows have null values which also need to removed or replaced. We also found outliners in some columns and rows, such as price column. We replaced those values with in a suitable price range. This whole process is called as Data Cleaning, Transformation and Analyses. After that we have applied different regression model on cleaned dataset and selected best model based on the R2 value.

**Conclusions**

* We can see that MSE and R2 and Adjusted R2 which are the metrics used to evaluate the performance of regression model of **Linear Regression, Lasso, Ridge.**
* The Linear models don't show good performance on our training and testing environment.
* From above table we can conclude that **Linear Regression** gives us **R2=91%.** Which is the best model as compare to the other models to predict the trip duration for a NYC taxi.

**Team Member’s Name, Email and Contribution:**

1. **Akshay Kumar Saini (**[**ak663348@gmail.com**](mailto:ak663348@gmail.com)**)**

1.1. Descriptive Analysis

1.1.1. Data frame description

1.1.2. Data frame shape

1.2. Data cleaning, Transformation and Analysis:

1.2.1. Drop unwanted columns

1.2.2. Checking of null values if exists

1.2.3. Identification and removal of outliers

1.3. Data Wrangling

1.3.1 Univariate Analysis of different features of dataset

1.3.2 Bivariate Analysis between different features of dataset

1.3.3 Creation of new features

1.3.4 Creation of dummy variable

1.4 Data Visualisation

1.4.1 Relationship between different features

1.4.2 Identified different trends among different features

2. **Kumkum Singh (**[**kumkum24091999@gmail.com**](kumkum24091999@gmail.com)**)**

1.5. Data Precreation

1.5.1. Identification of multicollinearity between independent features

1.5.2. Identification of dependent and independent feature

1.5.3. Splitting of dataset into train and test dataset

1.5.4. Standardization of dependent features

1.6. Data Modelling

1.6.1 Application of Linear Regression Model on dataset

1.6.2 Application of Lasso Regression Model on dataset

1.6.3 Application of Ridge Regression Model on dataset

1.7. Conclusion

1.7.1 Comparison of the result of different regression models

1.7.2 Selection of the best regression model

1.7.3 Final Conclusion of the project

Github Repository link:

Kumkum Singh: <https://github.com/Kumkum24/NYC-Taxin-Trip-Predection>