**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| |  |  |  | | --- | --- | --- | | Names | Email ID | Contribution | | Ankush Kumar  (Team Leader) | dsankushkumar@gmail.com | Code:  Different Variate Analysis and correlation.  Modeling  PCA Implementation  Conclusion and Summary  Power Point Presentation Preparation | | Nayan Kumar Jha | nayan8625@gmail.com | Code:  Data Insights  Data Visualization  Feature Transformation  Power Point Presentation Preparation | | Pinky Thakur | thakurpinky1896@gmail.com | Code:  Data Information  Data Wrangling  Data Preprocessing  Power Point Presentation Preparation | |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - <https://github.com/dsankush/NYC_Taxi_Trip_Time_Prediction_Capstone_Project> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**  There are several methods to move from one location in a city to another, but the taxi ride serves more purposes than any other means of urban transit. When provided the essential set of criteria that determine journey time, analysing and forecasting travel duration between two sites in the city becomes critical. The project is an ideal way to study New York City's traffic system in order to deliver a good taxi service and integrate it with the current transportation system. Predictions are made using variables such as pick-up latitude, pick-up longitude, drop-off latitude, drop-off longitude, and so forth.  The overall travel duration is calculated utilising these geographic areas, as well as other important variables such as the pick-up date and time. The primary goal of this study is to conduct an in-depth examination of the elements involved in a taxi journey in New York City.  The dataset was constructed using data from 2016 NYC Yellow Cab trip records made accessible in Big Query on the Google Cloud Platform. The information was initially given by the NYC Taxi and Limousine Commission (TLC). The data was cleansed and sampled for the purposes of this study. Based on the unique trip features, predict the length of each trip in the test set.  the practise set: Data on NYC Taxis (contains 1458644 trip records).  We have performed different types of visualization on the basis of Univariate, Bivariate and multivariate analysis and some leading insights are:  Vendor 2 there are a greater number of bookings which is of 54 %. When a taxi ride is booked by only a single person there are a greater number of bookings is high as compare to multiple people booking the taxi ride. Weekends 4 - Friday | 5 - Saturday there are high booking rate for taxi as compare to other days. This indicates that people use to go out for their celebrations | parties | or may be for other personnel works on weekends. Morning after 10 O’clock people use to book taxi because they want to go out to their work places. And at in the evening after 6 O’clock the taxi demand tends to in peak.  both pickup and drop-off day time count plot booking count is maximum in the EVENING Day time. **Least trip duration occurred in January and highest started occurring after the month of march.** Trip duration start increases from 3rd month before that it is quite constant. **'dropoff\_day', 'dropoff\_hour', 'dropoff\_month', 'dropoff\_weekday'** are highly corelated we can drop these features.We can see that we applied several models using various methodologies, and that, with the exception of Decision Tree and Random Forest Regression, XGBOOST performed flawlessly in each case.But after applying PCA to our features and successfully reducing the feature dimension, we noticed that Decision Tree and Random Forest were performing just as well as they had before transformation, if not even better.Additionally, XGBOOST is also performing better after PCA transformation, and it also displays marginally better results in terms of RMSE | MSE, which are lower than those of other models, as well as R2 | Adj R2 scores, which are also higher than those of other models.As a result, we can use the XGBOOST Regressor to get good prediction rates and less error prone results. It can also be further optimized by using more tuned hyperparameters. |
|  |