**Capstone Project Submission**

**NYC Taxi Trip Time Prediction**

|  |
| --- |
| **Summary** |
| A typical taxi company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. One of main issue is determining the duration of the current trip so it can predict when the cab will be free for the next trip.  Machine learning has been of significant help as it has helped businesses in abundant ways. We will use Machine Learning to efficiently build a model with a real-world dataset of Yellow Taxi  Service of NYC which will predict the estimated time duration of a tax trip for a given Pick up location, Drop location, Date, and Time.    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.  Now I started cleaning our data. So, we first identified the null values and we replaced this null value according to their data type. After dealing with null values I moved on to those columns which we don’t need. So I removed last review column. Then I replaced few data which don’t make sense with other values. Here I replaced zero price of property with mean price according to their room type and neighborhood type. Finally, our data is ready for EDA. Monthly trips analysis, Pickup hour analysis, Weekday trip analysis, Trip duration analysis were carried out during EDA. Total 5 machine learning regression model were implemented. At first linear regression had poor results while other tree-based models like Decision Trees, Random Forest, Extra Trees Regressor did very well on this dataset. Model based on boosting technique called XGBoost was also implemented and did show good results. After performing each model, Extra Trees Regressor was called as an optimal model for training.  This was all about my analysis and training that I did, and based on result and my inferences I am making following conclusions:   * Mostly 1 or 2 passengers avail the cab. The instance of large group of people travelling together is rare. * Most trips were taken on Friday and Monday being the least. * Fridays and Saturdays are those days in a week when peoples prefer to roam in the city. * The highest average time taken to complete a trip are for trips started in between 2 pm to 5 pm and the least are the ones taken between 5 am to 7 am. * Linear Regression doesn’t work well on this data. * The optimal model is Extra Trees Regressor. |
| **Contributor Role** |
| Ritik Vaidande ([vr171k@gmail.com](mailto:vr171k@gmail.com))   * Understanding Data   Understanding different column  Having overview of data   * Exploratory Data Analysis * Target feature analysis * Monthly trips analysis * Pickup hour analysis * Weekday trips analysis * Trip duration analysis * Visualizations * Bar graphs and Boxplot * Modelling * Feature Preparation * Linear Regression * Decision Tree * Random Forest * Extra Trees Regressor * XGBoost * Conclusions |
| **Github Repo and GDrive Link**  Github Link:-  <https://github.com/vaidande/NYC-Taxi-Trip-Time-Prediction>  Drive Link:-  https://drive.google.com/drive/folders/1BcksloqfAFh\_ONzo5EOSA9s7Tp-6CEL2?usp=sharing |
|  |