Project Proposal

Team Members: Gagan Arora, Shagufta Pathan, Deepak Khirey (Group1)

Title: New York City Taxi Fare Prediction (https://www.kaggle.com/c/new-york-city-taxi-fare-prediction#description)

Abstract:
Most of us have taken taxi rides before and in today's world with the advent of technology-based cab services, it is become so convenient to book a taxi ride at our finger-tips. But knowing the taxi fare and trip duration even before you can take the ride, has proven to be extremely useful and a time-saving process. So, here our goal is to accurately **predict a rider's taxi fare** using only factors that could be known when booking the ride, such as pickup and drop off location and so on. A good prediction mechanism can be instrumental for drivers in optimizing their returns, while also saving the customers from the uncertainties attached to a trip. We plan to achieve this by using **The New York City Taxi Fare Prediction dataset** which is a challenge hosted by **Kaggle** in partnership with Google Cloud and Coursera. This dataset uses a selection from the massive New York City (NYC) Taxi and Limousine Commission (TLC) Yellow Cab dataset that is also publicly available on Big Query.

This is an open competition and is accepting late submissions. The dataset is divided into train and test in the form of csv files. The training set consists of 55M rows and the test set consists of 10K rows.

The training sets consists of the 6 following **input features**:

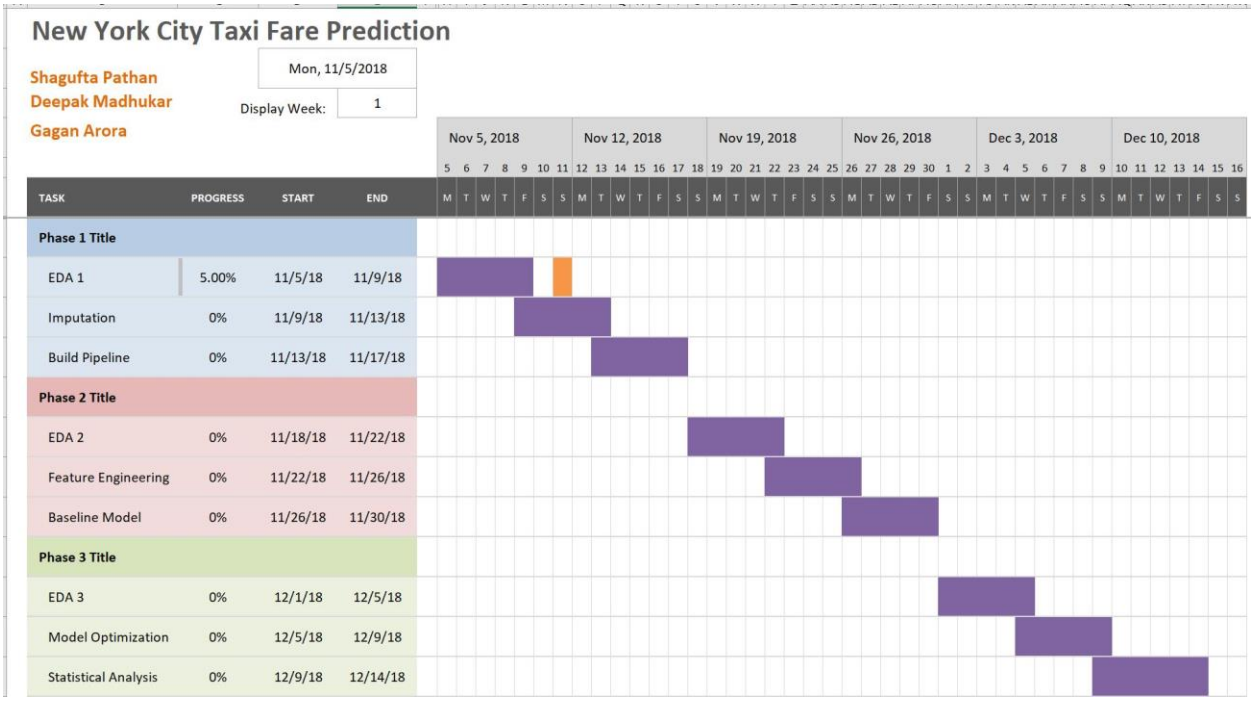
Feature	Data Type	Description
pickup_datetime	timestamp	value indicating when the taxi ride started.
pickup_longitude	float	longitude coordinate of where the taxi ride started.
pickup_latitude	float	latitude coordinate of where the taxi ride started.
dopoff_longitude	float	longitude coordinate of where the taxi ride ended.
dropoff_latitude	float	latitude coordinate of where the taxi ride ended.
passenger_count	integer	indicating the number of passengers in the taxi ride.

The **target variable** is the **fare_amount** (float) dollar amount of the cost of the taxi ride which is only available in the train.csv file.

Since we have the target variable (continuous) to determine, this is a **Supervised Regression** machine learning task and we aim to tackle this using some of the regression algorithms that we have learned so far like Linear Regression, Tree based models like Random Forest/ Decision Trees using the cross-validation and hyper-parameter tuning techniques.

We plan to use Root Mean Squared Error (RMSE) as the evaluation metric for this challenge which is the same used by Kaggle as well. This way we can benchmark against others along the way. RMSE measures the difference between the predictions of a model, and the corresponding ground truth. One nice property of RMSE is that the error is given in the units being measured, so we can get directly accuracy of the model on unseen data.

Gantt Diagram of Machine Learning Project With Timeline:



Any machine learning project will involve the following keys steps:

1. Understand the problem and data
2. Data exploration / data preparation
3. Feature engineering / feature selection
4. Model evaluation and selection
5. Model optimization/ Hyper-parameter tuning
6. Statistical analysis / significance test
7. Interpretation of results and predictions

Description of the pipeline steps:

Exploratory Data Analysis –
We are dealing with a dataset of large number of data points and few features. Most of the data is in numerical data type. Hence, we will check for any data inconsistencies and imputation requirements in our pipeline. This data also contains time-series feature and geo co-ordinate features which can be leveraged for new feature creation.

Feature Engineering –
Number of passengers is a numerical feature but we will try to explore if it is useful to categorize it. Also, timeseries feature will help to derive time of day, season etc. which might improve prediction accuracy. Geo co-ordinates will be useful to derive categorization of pick-up and drop-off location. Once we have categories defined we will use OneHotEncoding to convert them into new features.

Model evaluation –
Kaggle mentions that basic estimate is around $5-$8 RMSE, we will try to achieve better accuracy in our predictions by evaluation of different models. We are planning to evaluate supervised regression models like Linear Regression, Support Vector Machine, Decision Tree/ Random Forest algorithms. We will choose best model based on least RMSE value.

Model Optimization –
We will fine tune selected model with feature selection, cross validation and hyper-parameter tuning. We will use RandomizedCV and GridSearchCV for experimenting different parameter values for our model. Best estimator will be used in pipeline for final predictions.

Significance test –
We will compare outputs of best 2 models for statistical significance. We will carry out t-test for arriving at a conclusion based on p value.

Interpretation of results –
We will conclude our study with our predictions by best model and what it means in business terms for both taxi operator and passenger.

Responsibilities:

Gagan	Deepak	Shagufta	Team Effort
• Data Cleaning including imputation of missing values, removing outliers etc. • Building SkLearn pipelines	• Basic EDA (1) • Model Optimization using hyperparameter tuning techniques	• Baseline Models • Feature Engineering/ Feature Selection • Basic EDA (2)	• Integrating SkLearn pipelines • Discussions, results and conclusions • Creating slides and presentation.