**qwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmrtyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnmqwertyuiopasdfghjklzxcvbnm**

|  |
| --- |
| UBER DATA ANALYSIS  Price Prediction using Machine Learning  3/30/2022  R. Suresh ram |

MACHINE LEARNING PROJECT REPORT

**UBER-DATA-ANALYSIS**



As a project work for Course:

**MACHINE LEARNING FOUNDATION (INT 247)**

Submitted by:

Name : R. Suresh Ram (B54)

Registration Number : 11908772

Program : CSE B.TECH

Semester : Sixth

School : School of Computer Science and Engineering

Name of the University : Lovely Professional University

Date of submission : 27th MARCH 2022

DECLARATION

I, hereby declare that the work which is being presented in the B.tech. Project **“Uber Data Analysis”**, in partial fulfilment of the requirements for the award ofthe ***Bachelor of Technology*** in Computer Science and Engineering and submitted to the Department of Computer Engineering and Applications of Lovely Professional University, Punjab, is an authentic record of my own work carried under the supervision of **Dr. Sagar pandey.**

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Machine Learning Project undertaken during B. Tech. Third Year. This project in itself is an acknowledgement to the inspiration, drive and technical assistance contributed to it by many individuals. This project would never have seen the light of the day without the help and guidance that we have received.

We owe special debt of gratitude to Dr. Sagar pandey for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance.

Have been a constant source of inspiration for us. He has showered us with all his extensively experienced ideas and insightful comments at virtually all stages of the project & has also taught us about the latest industry-oriented technologies.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind guidance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

**R. Suresh ram**

ABSTRACT

Uber was founded just eleven years ago, and it was already one of the fastest-growing companies in the world. In Boston, UberX claims to charge 30% less than taxis – a great way to get customers' attention.

Nowadays, we see applications of Machine Learning and Artificial Intelligence in almost all the domains so we try to use the same for Uber cabs price prediction. In this project, we did experiment with a real-world dataset and explore how machine learning algorithms could be used to find the patterns in data.

We mainly discuss about the price prediction of different Uber cabs that is generated by the machine learning algorithm. Our problem belongs to the regression supervised learning category. We use different machine learning algorithms, for example, Linear Regression, Decision Tree, Random Forest Regressor, and Gradient Boosting Regressor but finally, choose the one that proves best for the price prediction.

We must choose the algorithm which improves the accuracy and reduces over fitting. We got many experiences while doing the data preparation of Uber Dataset of Boston of the year 2018. It was also very interesting to know how different factors affect the pricing of Uber cabs.

INTRODUCTION

Uber Technologies, Inc., commonly known as Uber, was a ride-sharing company and offers [vehicles for hire](https://en.wikipedia.org/wiki/Vehicles_for_hire), [food delivery](https://en.wikipedia.org/wiki/Food_delivery) ([Uber Eats](https://en.wikipedia.org/wiki/Uber_Eats)), [package delivery](https://en.wikipedia.org/wiki/Package_delivery), [couriers](https://en.wikipedia.org/wiki/Courier), [freight transportation](https://en.wikipedia.org/wiki/Freight_transport), and, through a partnership with [Lime](https://en.wikipedia.org/wiki/Lime_(transportation_company)), [electric bicycle](https://en.wikipedia.org/wiki/Electric_bicycle) and [motorized scooter](https://en.wikipedia.org/wiki/Motorized_scooter) rental. It was founded in 2009 by Travis Kalanick and Garrett Camp, a successful technology entrepreneur. After selling his first startup to eBay, Camp decided to create a new startup to [address San Francisco’s serious taxi problem](http://www.tc.umn.edu/~ssen/IDSC6050/Case15/Group15_index.html).

Together, the pair developed the Uber app to help connect riders and local drivers. The service was [initially launched in San Francisco](http://blog.uber.com/tag/history/) and eventually expanded to Chicago in April 2012, proving to be a highly convenient great alternative to taxis and poorly-funded public transportation systems. Over time, Uber has since expanded into smaller communities and has become popular throughout the world. In December 2013, [USA Today](https://en.wikipedia.org/wiki/USA_Today) named Uber its tech company of the year.

In Supervised learning, we have a training set and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. We applied machine learning algorithms to make a prediction of Price in the Uber Dataset of Boston. Several features will be selected from 55 columns. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data.

**Objective:**

The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on the dataset and to know the effect of each field on price with every other field of the dataset. Then we apply different machine learning models to complete the analysis. After this, the results of applied machine learning models were compared and analyzed on the basis of accuracy, and then the best performing model was suggested for further predictions of the label ‘Price’.

**Organization of the Project Report:**

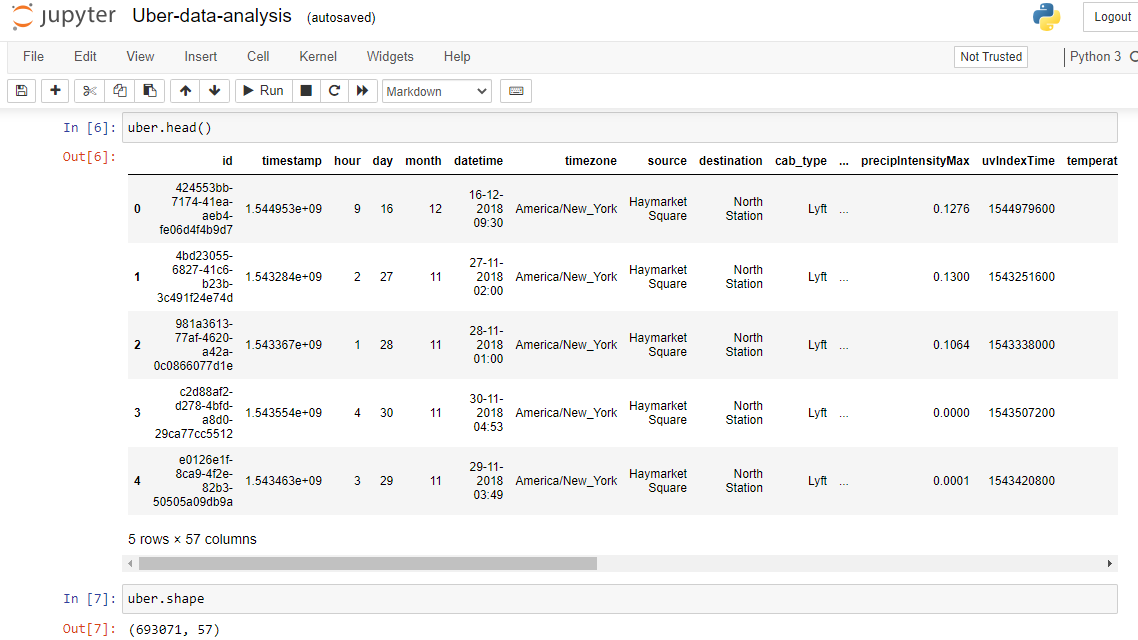
The first section of this paper presents the concept of exploratory data analysis which told general information about the dataset. Then from the next section feature engineering part was started in which we plot many charts and deal with columns to extract the features helpful for our predictions in many ways. In the last part, we did modeling and testing in which we apply different models to check the accuracy and for further price prediction.

**PROPOSED WORK & IMPLEMENTATION**

* 1. **Data Preparation:**

The data we used for our project was provided on the www.kaggle.com website. The original dataset contains 693071 rows and 57 columns which contain the data of both Uber and Lyft. The dataset has many fields that describe us about the time, geographic location, and climatic conditions when the different Uber cabs opted.

Data has 3 types of data-types which were as follows: - integers, float, and object. The dataset is not complete which means we have also null values in a column named price of around 55095.



**Fig: 1.1 Data head**

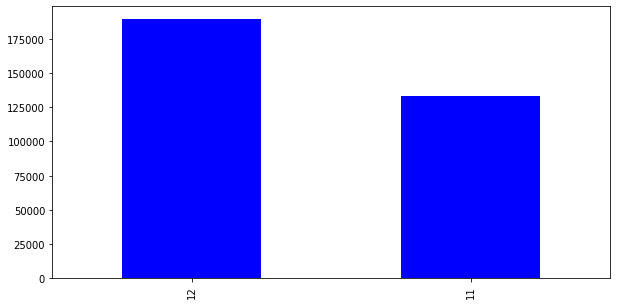
* 1. **Data Visualization:**

Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

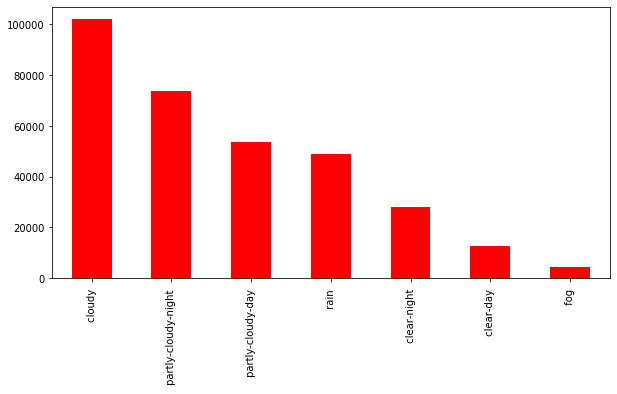
For the same purpose, we have to import matplotlib and seaborn library and plot different types of charts like strip plot, scatter plot, and bar chart.



**Fig: 1.2 Strip-plot between Icon and Price**



**Fig: 1.2** **Bar-Chart of Month**



**Fig: 1.2** **Bar-Chart of Icon**

* 1. **Label Encoding:**

Our data is a combination of both **Categorical variables**and**Continuous variables,** most of the machine learning algorithms will not understand, or not be able to deal with categorical variables. Meaning, machine learning algorithms will perform better when the**data is represented as a number** instead of categorical. Hence label encoding comes into existence. Label Encoding refers to converting the categorical values into the numeric form to make it machine-readable. So we did label encoding as well as class mapping to get to know which categorical value is encoded into which numeric value.

* 1. **Filling NAN Values:**

To check missing values in Pandas DataFrame, we use a function isnull(). So we find that the price column in our dataset consists of 55095 Nan values. Now to fill these null values we use the fillna() function. We fill missing values with the median of the remaining dataset values and convert them to integer because price cannot be given in float. Now for the visualization purpose, we make a bar chart of the value count of price.

* 1. **ML Model:**

Feature selection is an important task for any machine learning application. This is especially crucial when the data has many features. The optimal number of features also leads to improved model accuracy. So we use RFE for feature selection in our data.

RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is wrapped by RFE, and used to help select features. This is in contrast to filter-based feature selections that score each feature and select those features with the largest score

* + 1. **Drop Useless Columns:**

After applying RFE we get our 25 best features but still, there are many features which do not affect the price directly so we drop those features according to it. And eight features remained in our dataset. We use a method called drop() that removes rows or columns according to specific column names and corresponding axis.

* + 1. **Binning:**

Many times we use a method called data smoothing to make the data proper. During this process, we define a range also called bin and any data value within the range is made to fit into the bin. This is called the binning. Binning is used to smoothing the data or to handle noisy data.

So after dropping useless features, some features are not in range so to make all the features in the same range we apply binning and get our final dataset which is further used for modeling.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | month | source | destination | product\_id | name | surge multiplier | icon | uvIndex |
| 0 | **12** | **5** | **7** | **8** | **7** | **1.0** | **5** | **0** |
| 1 | **11** | **5** | **7** | **12** | **2** | **1.0** | **6** | **0** |
| 2 | **11** | **5** | **7** | **7** | **5** | **1.0** | **1** | **0** |
| 3 | **11** | **5** | **7** | **10** | **4** | **1.0** | **1** | **0** |
| 4 | **11** | **5** | **7** | **11** | **6** | **1.0** | **5** | **0** |

**Fig: 1.5.2 Final Dataset after Feature Engineering**

* 1. **Modeling:**

The process of modeling means training a machine-learning algorithm to predict the labels from the features, tuning it for the business needs, and validating it on holdout data. When you train an algorithm with data it will become a model. One important aspect of all machine learning models is to determine their accuracy. Now to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model.

In this project, we use Scikit-Learn to rapidly implement a few models such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting.

* + 1. **Linear Regression:**

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous in the range such as salary, age, price, etc. It is a statistical approach that models the relationship between input features and output. The input features are called the **independent variables**, and the output is called **a dependent variable**. Our goal here is to predict the value of the output based on the input features by multiplying it with its optimal coefficients.

* + 1. **Decision Tree:**

Decision tree is a supervised learning algorithm which can be used for both classification and regression problem. This model is very good at handling tabular data with numerical or categorical features. It uses a tree-like structure flow chart to solve the problem. A decision tree is arriving at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model gets confident enough to make a single prediction.

* + 1. **Random Forest:**

Random forest is a supervised learning algorithm which can be used for both classification and regression problem. It is a collection of Decision Trees. In general, Random Forest can be fast to train, but quite slow to create predictions once they are trained. This is due because it has to run predictions on each tree and then average their predictions to create the final prediction.

**Table 1.6: Model Accuracy Table**

|  |  |  |
| --- | --- | --- |
| **Serial No.​** | **Models** | **Accuracy​** |
| 1​ | Linear Regression​ | 0.384679328570529 |
| 2​ | Decision Tree​ | 0.956031845214901 |
| 3​ | Random Forest​ | 0.957199573274482 |

1.7 **Testing:**

In Machine Learning the main task is to model the data and predict the output using various algorithms. But since there are so many algorithms, it was really difficult to choose the one for predicting the final data. So we need to compare our models and choose the one with the highest accuracy.

Machine learning applications are not 100% accurate, and approx. never will be. There are some of the reasons why testers cannot ignore learning about machine learning. The fundamental reason is that these applications learning limited by data they have used to build algorithms.

**Mean Squared Error** (MSE) [**MAE = True values – Predicted values**],

**Mean Absolute Error** (MAE)[ **MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data**], and

Root Mean Squared Error (RMSE)[**RMSE is a quadratic-based rule to measure the absolute average magnitude of the error.**]are used to evaluate the regression problem's accuracy. These can be implemented using scikit’s mean absolute error method and scikit’s mean squared error method.

**Table 1.7: Error table for Linear Regression**

|  |  |  |
| --- | --- | --- |
| **Serial No.​** | **Models** | **Accuracy​** |
| 1​ | Mean Absolute Error | 5.312289434 |
| 2​ | Mean Squared Error | 49.740742343 |
| 3​ | Root Mean Absolute Error | 7.052711701 |

* 1. **Price Prediction Function:**

After finding the errors for both linear regression and random forest algorithm, we build a function name “predict\_price” whose purpose is to predict the price by taking 4 parameters as input. These four parameters are cab name, source, surge multiplier, and icon (weather). As the dataset train on the continuous values and not on categorical values, these values are also passed in the same manner i.e. in integer type. We create a manual for users which gives instructions about the input like what do you need to type for a specific thing and in which sequence.

We use random forest model in our function to predict the price. First, we search for all the desired rows which have the input cab name and extract their row number. After then we create an array x which is of the length of the new dataset and it’s initially all values are zero. After creating the blank array we assign the input values of source, surge multiplier, and icon to the respected indices. Following it we check the count of all desired rows if it was greater than zero or not. If the condition gets true, we assign the value 1 to the index of x array and return the price using the predict function with trained random forest algorithm.

It somehow works like a hypothesis space because it gives an output for any input from input the space.

**CONCLUSION**

Before working on features first we need to know about the data insights which we get to know by EDA. Apart from that, we visualize the data by drawing various plots, due to which we understand that we don’t have any data for taxi’s price, also the price variations of other cabs and different types of weather. Other value count plots show the type and amount of data the dataset has. After this, we convert all categorical values into continuous data type and fill price Nan by the median of other values. Then the most important part of feature selection came which was done with the help of recursive feature elimination. With the help of RFE, the top 25 features were selected. Among those 25 features still, there are some features which we think are not that important to predict the price so we drop them and left with 8 important columns.

We apply four different models on our remaining dataset among which Decision Tree, Random Forest, and Gradient Boosting Regressor prove best with 96%+ accuracy on training for our model. This means the predictive power of all these three algorithms in this dataset with the chosen features is very high but in the end, we go with random forest because it does not prone to overfitting and design a function with the help of the same model to predict the price.

**REFERENCES**

1. Abel Brodeurand & Kerry Nield (2018) An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC
2. Junfeng Jiao (2018) Investigating Uber price surges during a special event in Austin, TX
3. Anna Baj-Rogowska (2017) Sentiment analysis of Facebook posts: The Uber Case
4. Anastasios Noulas, Cecilia Mascolo, Renaud Lambiotte, and Vsevolod Salnikov (2014) OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs.