A

PROJECT REPORT

ON

"UBER DATA ANALYSIS"

SUBMITTED TO

SAVITRIBAI PHULE PUNE UNIVERSITY

IN PARTIAL FULFILLMENT OF
MASTER OF BUSINESS ADMINISTRATION

SUBMITTED BY

Mr. KESHAV SHANKAR GHODAKE

UNDER THE GUIDENCE OF

Prof. KANIF SATAV

DEPARTMENT OF MBA
DHOLE PATIL COLLEGE OF ENGINEERING WAGHOLI, PUNE.
(2021-22)

DHOLE PATIL COLLEGE OF ENGINEERING, PUNE

CERTIFICATE

This is to certify that **Mr. KESHAV SHANKAR GHODAKE** is a student of Dhole Patil College of Engineering, Pune pursuing Masters Of Business Administration course of 2021-22, has successfully completed her Summer project titled “**UBER DATA ANALYSIS”** at **SIDDHANT D-LOGIC TECHNOLOGY.**

This project is accomplished adequately and submitted in partial fulfilment of **MBA BUSINESS ANALYTICS** curriculum as per the requirement of Savitribai Phule Pune University for batch 2021-2022.

Prof. Kanif Satav	**Prof. Shrikant Jagtap**	**Dr. Nihar Walimbe**	
Project Guide	**HOD, MBA Dept.**	**Principal**	**External**

ACKNOWLEDGEMENT

A summer project is a golden opportunity for learning and self-development. I consider myself very lucky and honored to have had so many wonderful people to lead me through in completion of this project.

I would like to extend my gratitude to the management of SIDDHANT D-LOGIC TECHNOLOGY for giving chance to work for this project in theirorganization.

I would like to thank entire team of SIDDHANT D-LOGIC TECHNOLOGY for their support and encouragement throughout the entire course of the project.

I would also like to thank my college Dhole Patil College of Engineering for giving me this opportunity to put in real time practice; the theoretical knowledge gained in the classroom with the practical reality in the external corporate world & I would like to thank Prof. Shrikant Jagtap (HOD of MBA Dept. DPCOE , Pune) & Prof. Kanif Satav (Project Guide of MBA Dept. DPCOE , Pune) who helped me in completing this project report.

KESHAV SHANKAR GHODAKE
M.B.A.(Business Analytics)
Dhole Patil College of Engineerning, Pune

DECLARATION

I hereby declare that this report is being submitted in fulfilment of the course curriculum of MBA at Dhole Patil College of Engineering.

The information and data given in the report is authentic and I further declare that this project report has not been submitted to any other university or institute for the award of any degree or diploma.

The summer internship gave me an opportunity to study, understand and have practical exposure on different aspects. This report is the result of the authentic work carried out by me, under the guidance of Prof. Kanif Satav (Project Guide, Dhole Patil College of Engineering, Pune) during the period from 1st Nov, 2021 till 30th Dec, 2021.

DATE: KESHAV SHANKAR GHODAKE

PLACE: M.B.A.(Business Analytics)

INDEX

LIST OF TABLE

LIST OF FIGURES

Chapter 1 - EXECUTIVE SUMMARY

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 overfitting. 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.

Chapter 2 - INTRODUCTION

2.1 Motivation and Overview

Uber Technologies, Inc., commonly known as Uber, was a ride-sharing company and offers vehicles for hire, food delivery (Uber Eats), package delivery, couriers, freight transportation, and, through a partnership with Lime, electric bicycle and 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.

Together, the pair developed the Uber app to help connect riders and local drivers. The service was initially launched in San Francisco 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 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.

2.2 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'.

2.3 Issues and Challenges

1. **Overfitting in Regression Problem:-** Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This problem occurs when the model is too complex. In regression analysis, overfitting can produce misleading R-squared values. When this occurs, the regression coefficients represent the noise rather than genuine relationships. However, there is another problem. Each sample has its unique quirks. Consequently, a regression model that becomes tailor-made to fit the random quirks of one sample is unlikely to fit the random quirks of another sample. Thus, overfitting a regression model reduces its generalizability outside the original dataset.

2. **Strip-plot and Scatter diagram:-** One problem with strip plots is how to display multiple points with the same value. If it uses the jitter option, a small amount of random noise is added to the vertical coordinate and if it goes with the stack option it increments the repeated values to the vertical coordinate which gives the strip plot a histogram-like appearance.

   Scatter plot does not show the relationship for more than two variables. Also, it is unable to give the exact extent of correlation**.**

3. **Label Encoding:-** It assigns a unique number(starting from 0) to each class of data which may lead to the generation of priority issues in the training of data sets. A label with high

value may be considered to have high priority than a label having lower value but actually, there is no such priority relation between the attributes of the same classes.

4. **Computational Time:-** Algorithms like support vector machine(SVM) don't scale well for larger datasets especially when the number of features are more than the number of samples. Also, it sometimes runs endlessly and never completes execution.

2.4 Contribution

Each team member is responsible and has willing participation in the group. The work within the group is equally done by each team member. First, the project work is divided like one has to be done the exploratory data analysis part, two members work on feature engineering, and the rest work of modeling and testing was equally divided among all four members. And the second part i.e. written work is done in pairs like two members work on report and the other two works on presentation.

2.5 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.

Chapter 3 - COMPANY PROFILE

Siddhant D Logic Technology Private Limited is an unlisted private company incorporated on 25 February, 2019. It is classified as a private limited company and is located in Pune, Maharashtra. It's authorized share capital is INR 1.00 lac and the total paid-up capital is INR 1.00 lac.

- Company name - Siddhant D Logic Technology Private Limited
- Established date - 25 February, 2019
- Physical address per location -
    - 645, Pundalic Laxman Path, Gondhalenagar, Satavwadi, Hadapsar,
      Pune, Maharashtra 411028
- Phone and fax numbers - 089753 17596
- Website URL - https://siddhant-d-logic-technology.business.site/

Siddhant D Logic Technology Private Limited has director & CEO - Naveenkumar Disale .

Chapter 4 – 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’.
- We use machine learning algorithms to predict the price of Uber, so that it is easy for the company to do analysis on price based on certain features.
- At Uber, this analysis is automated to drive the following results: Uber sends out weekly communications to drivers at real time. Weekly communications inform about high demand areas, with specific recommendations. Enabling driver-partners to make best decisions, increase earnings and lower ETAs.
- Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Chapter 5 – SCOPE

- We can use this data for training a model using ML and building a smart AI based predictive system. Model can automatically send the insights to the authorities or drivers related to areas having most trips and passenger count in certain areas. This big data can be used to study passenger's behavior.
- In the beginning, we saw that a successful ML in a big company like Uber needs more than just training good models – you need strong, awesome support throughout the workflow. We found that the same workflow applies to many different situations, including traditional ML and in-depth learning; surveillance, unsupervised, and under surveillance; online learning; batches, online, and mobile distribution; and time-series predictions. It does not mean that one tool provides everything (although this is how we did it) but it is important to have an integrated set of tools that can handle all the steps of the workflow.

1. Define

- Defining a business need is an important part of a business known as business analysis. This includes understanding and identifying the purpose of the organization while defining the direction used. In addition, you should take into account any relevant concerns regarding company success, problems, or challenges.

2. Prototype

- The users can train models from our web UI or from Python using our Data Science Workbench (DSW). At DSW, we support extensive deploying training of in-depth learning models in GPU clusters, tree models, and lines in CPU clusters, and in-level training on a wide variety of models using a wide range of Python tools available.

- Finding the right combination of data, algorithms, and hyperparameters is a process of testing and self-replication. Going through this process quickly and effectively requires the automation of all tests and results.

3. Production

- Once the working model has been trained, it is important that the model builder is able to move the model to the storage or production area. In Michelangelo, users can submit models through our web UI for convenience or through our integration API with external automation tools. Deployed model is used to make predictions.

4. Measure

- Models are trained and initially tested against historical data. This means that users may not know that the model would work well in the past. But once you have used the model and used it to make predictions on new data, it is often difficult to make sure it is still working properly. Models can degrade over time because the world is constantly changing.

- Data scientists, our use of tools makes it easier to create and produce on the side of building and shipping ML systems, enabling them to manage their work ultimately. For developers, Uber's ML tool simplifies data science (engineering aspect, modeling, testing, etc.) after these programs, making it easier for them to train high-quality models without the need for a data scientist. Finally, for the most experienced engineering teams forming special ML programs, we provide Michelangelo's ML infrastructure components for customization and workflow.
- Successfully measuring ML at a company like Uber requires much more than just the right technology – rather than the critical considerations of process planning and processing as well. In this section, we look at critical aspects of success across all three pillars: structure, process, and technology.

Chapter 6 - LITERATURE REVIEW

- As we are researching on Uber and found what different researchers had done. So, they do research on the Uber dataset but on different factors. The rise of Uber as the global alternative has attracted a lot of interest recently. Our work on Uber's predicting pricing strategy is still relatively new. In this research, "Uber Data Analysis" we aim to shed light on Uber's Price. We are predicting the price of different types of Uber based on different factors. Some of the other factors that we found in other researches are:
- Abel Brodeurand & Kerry Nield (2018) analyses the effect of rain on Uber rides in New York City after entering Uber rides in the market in May 2011, passengers and fare will decrease in all other rides such as taxi-ride. Also, dynamic pricing makes Uber drivers compete for rides when demand suddenly increases, i.e., during rainy hours. On increasing rain, the Uber rides are also increasing by 22% while the number of taxi rides per hour increases by only 5%. Taxis do not respond differently to increased demand in rainy hours than non-rainy hours since the entrance of Uber.
- Surge Pricing is an algorithmic technique that Uber uses when there is a demand-supply imbalance. It occurs when there is a downward shift in both the rider's demand and driver's availability. During such a time of the rise in demand for rides, fares tend to usually high. Surge pricing is essential in a way that it helps in matching the driver's efforts with the demand from consumers. (Junfeng Jiao, 2018) did an investigation of Uber on surge multiplier in Austin, Texas founds that during times of high usage, Uber will enhance their prices to reflect this demand via a surge multiplier. According to communications released by (Uber, 2015), this pricing is meant to attract more drivers into service at certain times, while also reducing demand on the part of riders. (Chen & Sheldon, 2016) While some research is mixed, in general, surge pricing does appear to control both supply and demand while keeping wait time consistently under 5 minutes.
- Anna Baj-Rogowska (2017) analyses the user's feedback from social networking sites such as Facebook in the period between July 2016 and July 2017. Uber is one of the

most dynamically growing companies representing the so-called sharing economy. It is also a basis for the ongoing evaluation of brand perception by the community and can be helpful in developing such a marketing strategy and activities, which will effectively improve the current rating and reduce possible losses. So, it can be concluded that feedback should be an important instrument to improve the market performance of Uber today.

- Anderson (2014) concluded from surveying San Francisco drivers that driver behavior and characteristics are likely determining the overall vehicle miles traveled (VMT). Full-time drivers are likely to increase overall VMT, while occasional drivers are more likely to reduce overall VMT. We also analyze the research on the driving behavior of the driver while driving on the road. The driver has been categorized based on ages and genders that focus on their driving reactions from how they braking, speeding, and steer handling. For gender differences, male driver practice higher-risk of driving while female drivers are lacks of pre-caution over obstacles and dangerous spot. More or less, adult drivers which regularly drive vehicles can manage the vehicle quite well as compared with young drivers with less experience. In conclusion, the driver's driving behavior is related to their age, gender, and driving experiences.
- Some papers take a comparison between the iconic yellow taxi and its modern competitor, Uber. (Vsevolod Salnikov, Renaud Lambiotte, Anastasios Noulas, and Cecilia Mascolo, 2014) identify situations when UberX, the cheapest version of the Uber taxi service, tends to be more expensive than yellow taxis for the same journey. Our observations show that it might be financially advantageous on average for travelers to choose either Yellow Cabs or Uber depending on the duration of their journey. However, the specific journey they are willing to take matters.

Chapter 7 - RESEARCH METHODOLOGY

7.1 What is Machine Learning?

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

7.2 Types of Learning Algorithms

The types of machine learning algorithms differ in their approach, the type of data they input, and the type of task or problem that they are intended to solve.

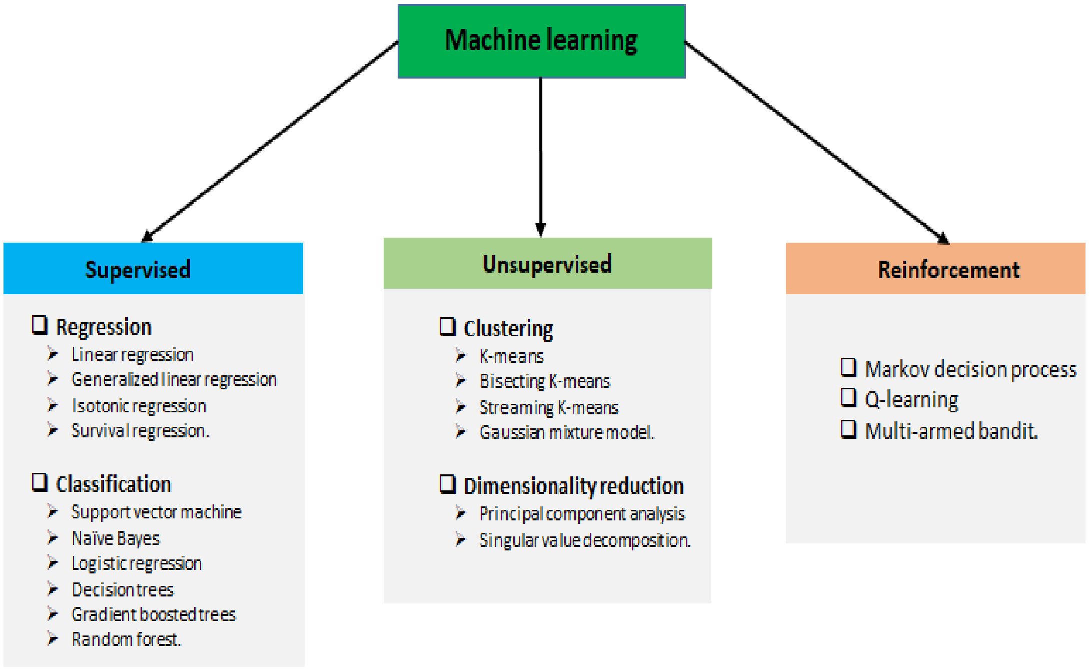


Fig. 7.1 Types of ML Courtesy of Packt-cdn.com

7.2.1 Supervised learning

Supervised learning is when the model is getting trained on a labelled dataset. The **labelled** dataset is one that has both input and output parameters. Supervised learning algorithms include classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range.

7.2.2 Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified, or categorized.

7.2.3 Reinforcement learning

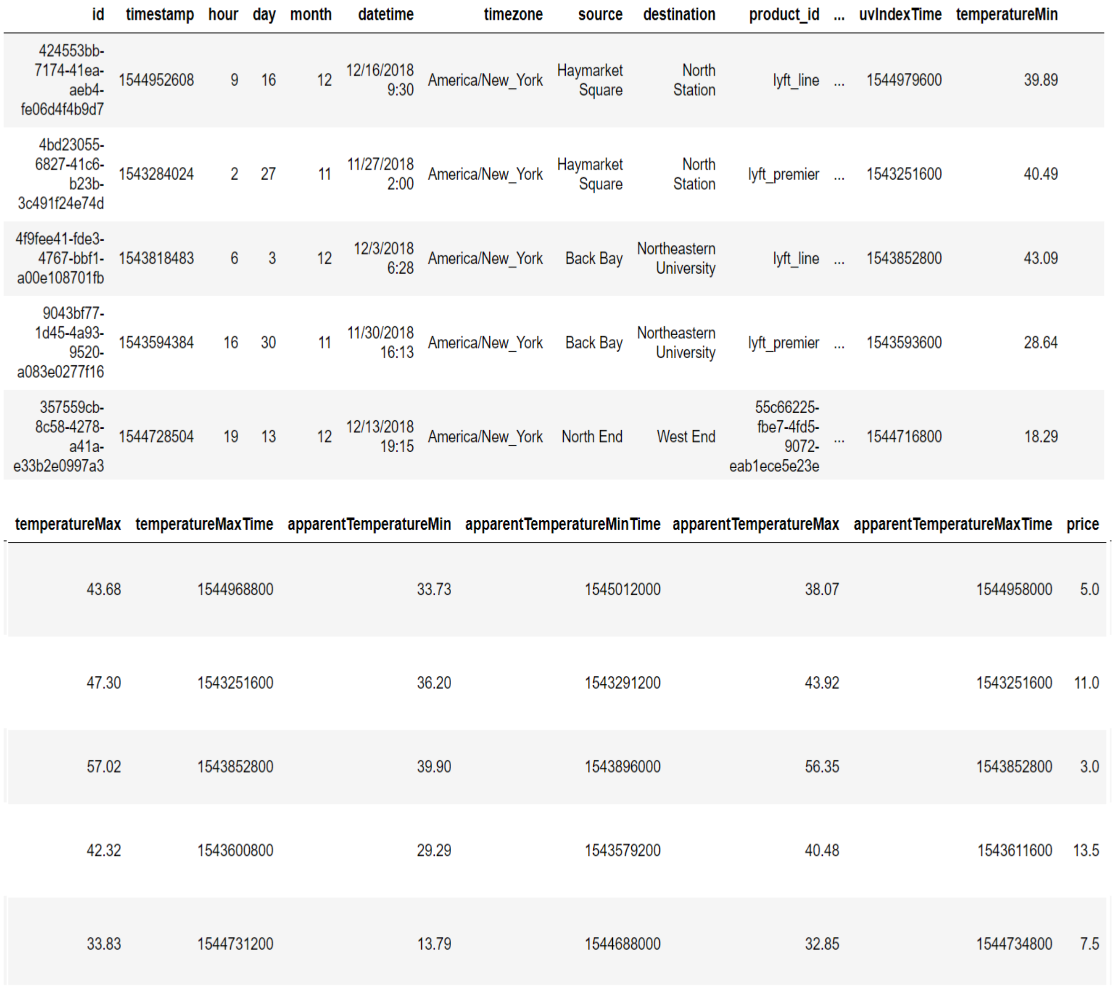
Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment to maximize some notion of cumulative reward. In this learning, system is provided feedback in terms of rewards and punishments as it navigates its problem space.

Chapter 8 - DATA ANALYSIS & INTERPRETAION

8.1 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. But for our analysis, we just need the Uber data so we filter out the data according to our purpose and got a new dataset that has 322844 rows and 56 columns. 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:- integer, float, and object. The dataset is not complete which means we have also null values in a column named price of around 55095.



id	timestamp	hour	day	month	datetime	timezone	source	destination	product_id	...	uvIndexTime	temperatureMin	temperatureMax	temperatureMaxTime	apparentTemperatureMin	apparentTemperatureMinTime	apparentTemperatureMax	apparentTemperatureMaxTime	price
424553bb-7174-41ea-aeb4-fe06d4f4b9d7	1544952608	9	16	12	12/16/2018 9:30	America/New_York	Haymarket Square	North Station	lyft_line	...	1544979600	39.89	43.68	1544968800	33.73	1545012000	38.07	1544958000	5.0
4bd23055-6827-41c6-b23b-3c491f24e74d	1543284024	2	27	11	11/27/2018 2:00	America/New_York	Haymarket Square	North Station	lyft_premier	...	1543251600	40.49	47.30	1543251600	36.20	1543291200	43.92	1543251600	11.0
4f9fee41-fde3-4767-bbf1-a00e108701fb	1543818483	6	3	12	12/3/2018 6:28	America/New_York	Back Bay	Northeastern University	lyft_line	...	1543852800	43.09	57.02	1543852800	39.90	1543896000	56.35	1543852800	3.0
9043bf77-1d45-4a93-9520-a083e0277f16	1543594384	16	30	11	11/30/2018 16:13	America/New_York	Back Bay	Northeastern University	lyft_premier	...	1543593600	28.64	42.32	1543600800	29.29	1543579200	40.48	1543611600	13.5
357559cb-8c58-4278-a41a-e33b2e0997a3	1544728504	19	13	12	12/13/2018 19:15	America/New_York	North End	West End	55c66225-fbe7-4fd5-9072-eab1ece5e23e	...	1544716800	18.29	33.83	1544731200	13.79	1544688000	32.85	1544734800	7.5

Fig. 8.1 Data Head

8.2 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.

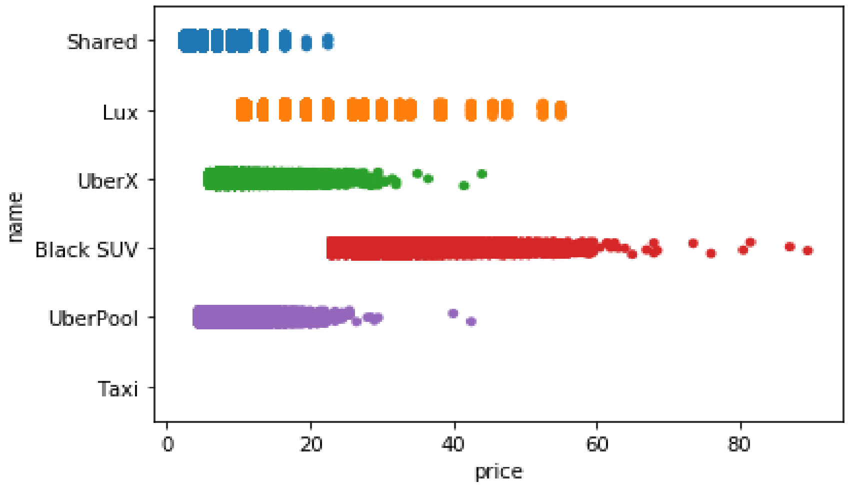


Fig. 8.2 Strip-plot between Name and Price

From the above chart, it was clear that Shared trip was cheapest among all and BlackSuv was most expensive. UberX and UberPool have almost same prices and Lux has moderate price. There is no graph for taxi which reveals that in the dataset there were no values of taxi was given.

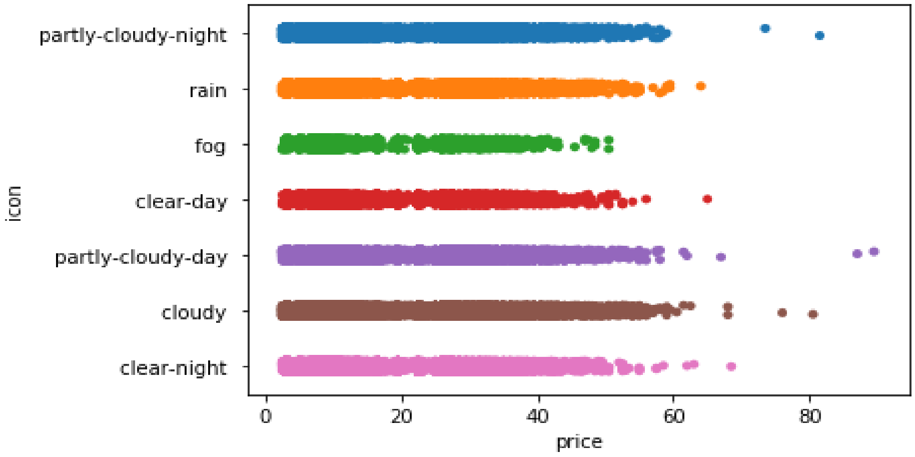


Fig. 8.3 Strip-plot between Icon and Price

From the above chart, it was clear there were some outliers in cloudy type weather, some data has an anonymously high price above 80 while the other was below 60. In this plot, we analyze that in cloudy-day weather price was the highest while in foggy weather price was minimum.

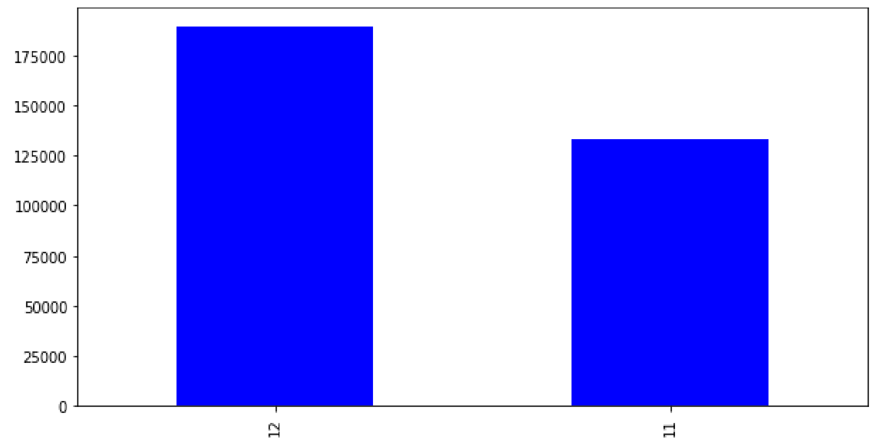


Fig. 8.4 Bar-Chart of Month

From the above bar chart, it was clear that the data consists of all the information of only two months that is November and December.

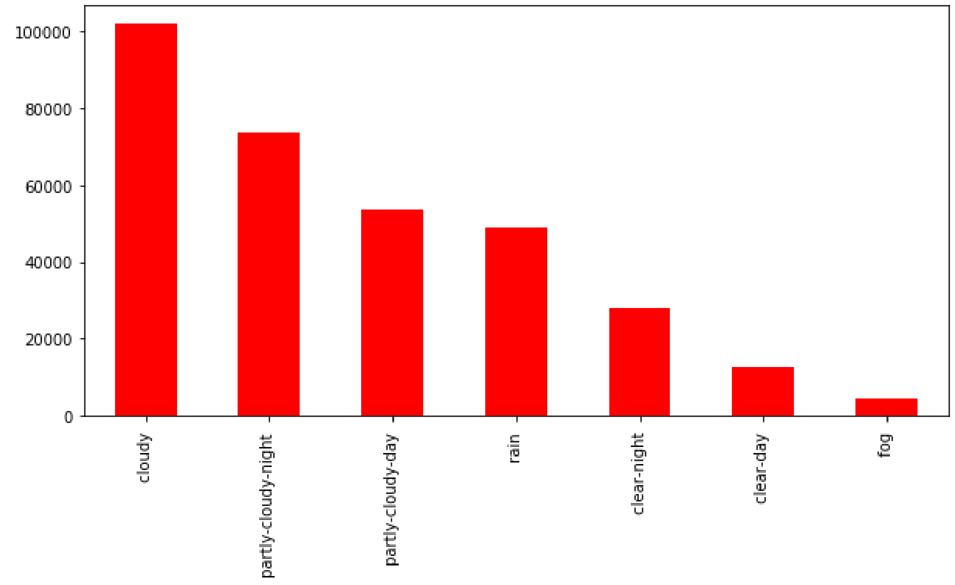


Fig. 8.5 Bar-Chart of Icon

The above bar chart represents the value count of the icon column and from the graph, it was clear that cloudy weather has the most data due to which we can say that may be in cloudy weather cab also opted most.

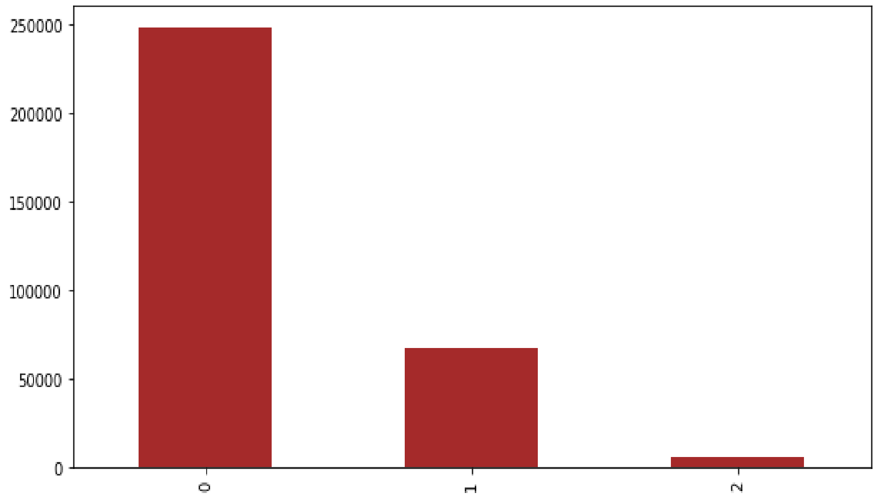


Fig.8.6 Bar-Chart of UV-Index

The above bar chart represents the value count of the UV-index column and from the graph, it was clear that when UV-index is 0, the dataset has the most data due to which we can say that when there is less UV-index cab was opted most.

8.3 Feature Engineering

Feature engineering is the most important part of the data analytics process. It deals with, selecting the features that are used in training and making predictions. All machine learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristics to work properly. A bad feature selection may lead to a less accurate or poor predictive model. To filters out all the unused or redundant features, the need for feature engineering arises. It has mainly two goals:

- Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- Improving the performance of machine learning models.

"According to a survey in Forbes, data scientists spend 80% of their time on data preparation."

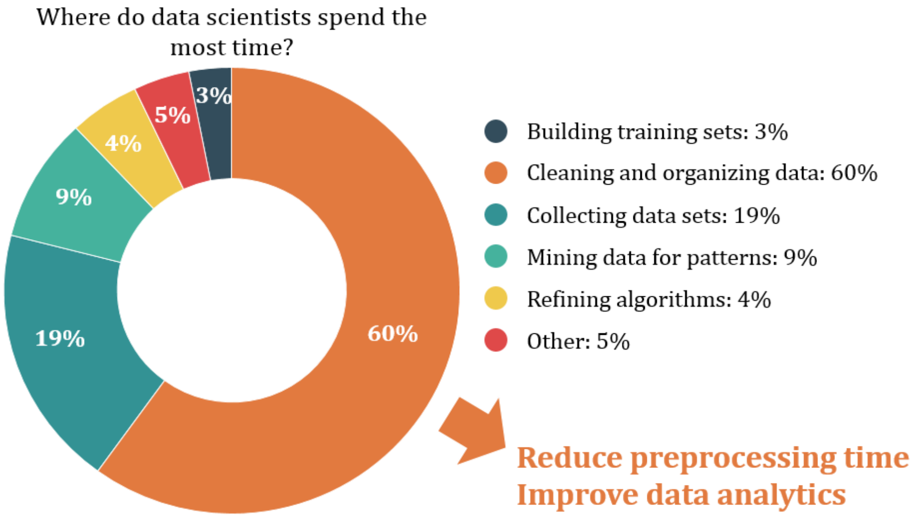


Fig. 4.7 Feature Engineering Courtesy of Digitalag

8.3.1 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.

8.3.2 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.

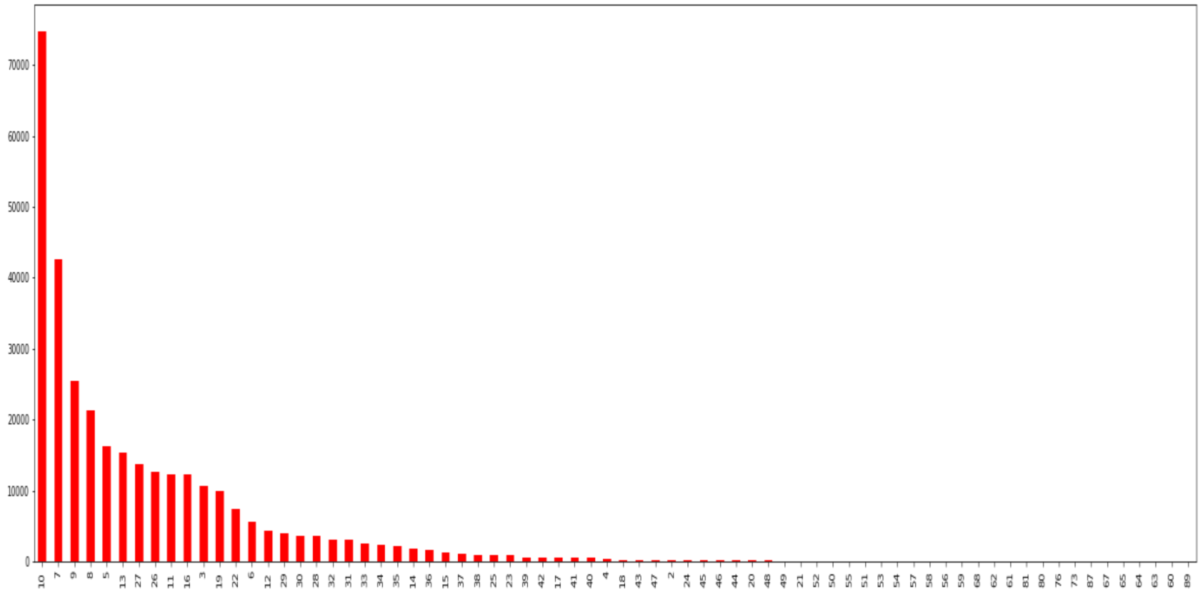


Fig. 4.8 Bar Chart of Price

4.3.3 RFE (Recursive Feature Elimination)

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.

There are two important configuration options when using RFE:

- The choice in the number of features to select (k value)
- The choice of the algorithm used to choose features.

RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remain. Hence RFE technique is effective at selecting those features (columns) in a training dataset that are most relevant in predicting the target variable.

We are implementing recursive feature elimination through scikit-learn via sklearn.feature_selection.RFE class.

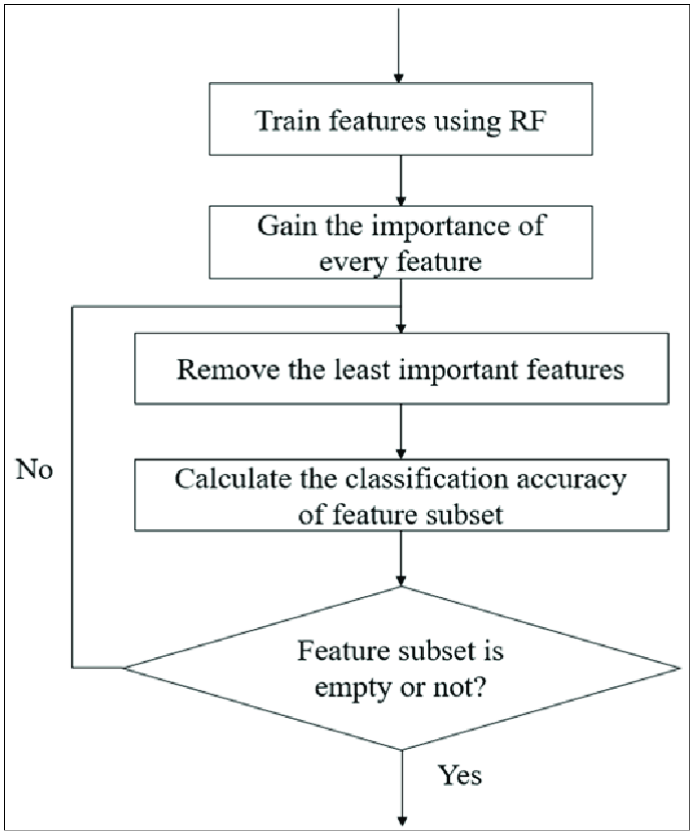

Train features using RF
Gain the importance of every feature
Remove the least important features
Calculate the classification accuracy of feature subset
Feature subset is empty or not?
No
Yes

Fig. 8.9 Recursive Feature Elimination Courtesy of Researchgate

On applying RFE in our dataset with Linear Regression model first we divide our dataset into dependent (features) and independent (target) variables then split it into train and test after that we found different accuracies in different number of features (k value) as follows:

Table 8.1: RFE Accuracy Table

Serial No.	No. of Feature (K)	Accuracy
1	56	0.8054834220
2	40	0.8050662132
3	25	0.8055355151
4	15	0.8050457819

Table 8.1 Continued

From the above table, it was clear that 25 features have the highest accuracy as compared to all other k values which mean these 25 features are the best features given by RFE. So, we only consider these 25 features for further working and rest we eliminate. Now our dataset reduces from 56 features to 25 features.

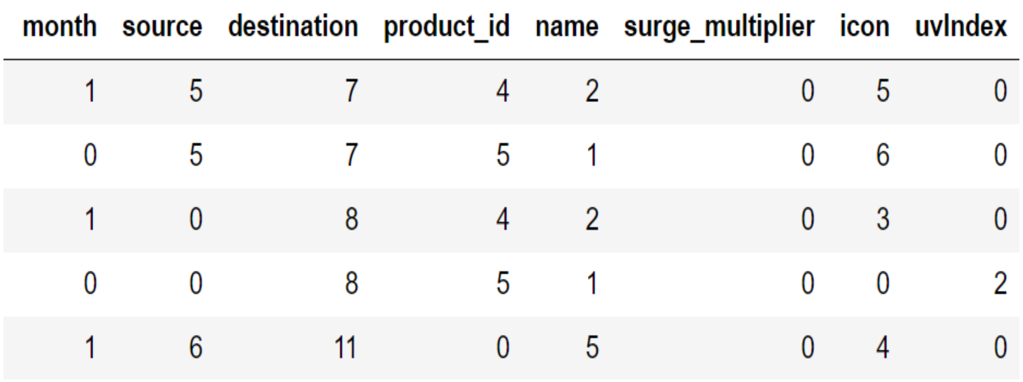
8.3.4 Drop Useless Columns

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.

8.3.5 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
1	5	7	4	2	0	5	0
0	5	7	5	1	0	6	0
1	0	8	4	2	0	3	0
0	0	8	5	1	0	0	2
1	6	11	0	5	0	4	0

Fig. 8.10 Final Dataset after Feature Engineering

8.4 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.

8.4.1. 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. The name linear regression was come due to its graphical representation.

There are two types of Linear Regression:-

- **Simple Linear Regression**- In a simple linear regression algorithm the model shows the linear relationship between a dependent and a single independent variable. In this, the dependent variable must be a continuous value while the independent variable can be any continuous or categorical value.
- **Multiple Linear Regression**- In a multiple linear regression algorithm the model shows the linear relationship between a single dependent and more than one independent variable.

8.4.2. 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. The order of the question as well as their content is being determined by the model. In addition, the questions asked are all in a True/False form. Here in our project, we are focusing on decision tree regression only. It is used for the continuous output problem. Continuous output means the output of the result is not discrete. It observes features of an object and trains a model in the structure of a tree to predict data that produce meaningful continuous output.

8.4.3. 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. A more accurate prediction requires more trees, which results in a slower model. In most real-world applications the random forest algorithm is fast enough, but there can certainly be situations where run-time performance is important and other approaches would be preferred. A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which **aggregates many decision trees**, with some helpful modifications. Random forest first splits the dataset into n number of samples and then apply decision tree on each sample individually. After that, the final result is that predicted accuracy whose majority is higher among all.

Random Forest depends on the concept of ensemble learning. An ensemble method is a technique that **combines the predictions from multiple machine learning algorithms** together to make more accurate predictions than any individual model. A model comprised of many models is called an **Ensemble model**.

Random forest is a bagging technique and **not a boosting** technique. The trees in **random forests** are run in parallel. There is no interaction between those trees while building random forest model.

8.4.4. Gradient Boosting

Gradient boosting is a technique which can be used for both classification and regression problem. This model combines the predictions from multiple decision trees to generate the final predictions. Also, each node in every other decision tree takes a different subset of features for selecting the best split. But there is a slight difference in gradient boosting in comparison to random forest that is gradient boosting builds one tree at a time and combines the results along the way. Also, it gives better performance than random forest. The idea of gradient boosting originated in the observation by Leo Breiman that boosting can be interpreted as an optimization algorithm on a suitable cost function. Gradient Boosting trains many models in a gradual, additive, and sequential manner.

The modeling is done in the following steps:-

- First, we split the dataset into a training set and a testing set.
- Then we train the model on the training set.
- And at last, we test the model on the testing set and evaluate how well our model performs.

So after applying these models we get the following accuracy:

Table 8.2: Model Accuracy Table

Serial No.	Models	Accuracy
1	Linear Regression	0.747545073
2	Decision Tree	0.961791729
3	Random Forest	0.962269474
4	Gradient Boosting Regressor	0.963187213

8.4.5 K-fold Cross Validation

We also apply cross validation using linear regression algorithm. It is a technique where the datasets are split into multiple subsets and learning models are trained and evaluated on these subset data. It is a resampling procedure used to evaluate machine learning models on a limited data sample. It is one of the most widely used technique. In this, the dataset is divided into k-subsets (folds) and are used for training and validation purpose for *k iteration* times. Each subsample will be used at least once as a validation dataset and the remaining (*k-1*) as the training dataset. Once all the iterations are completed, one can calculate the average prediction rate for each model. The error estimation is averaged over all k trials to get the total effectiveness of our model.

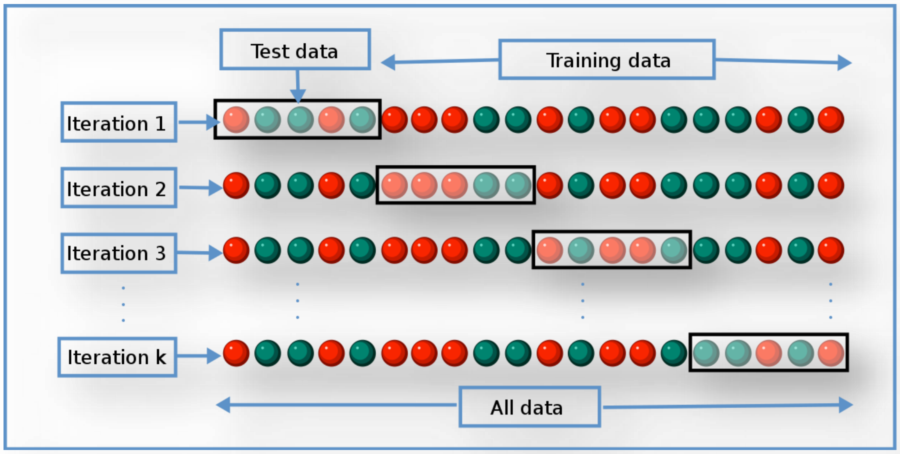


Fig. 8.11 Cross-Validation Courtesy of Wikimedia

8.5 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. For example, if 99% of emails aren't spammed, then classifying all emails as not spam gets 99% accuracy through chance. Therefore, you need to check your model for algorithmic correctness. Hence testing is required. Testing is a subset or part of the training dataset that is built to test all the possible combinations and also estimates how well the model trains. Based on the test data set results, the model was fine-tuned.

Mean Squared Error (MSE), **Mean Absolute Error** (MAE), and Root Mean Squared Error (RMSE) are used to evaluate the regression problem's accuracy. These can be implemented using **sklearn**'s **mean_absolute_error** method and **sklearn**'s mean_squared_error method.

8.5.1 Mean Absolute Error (MAE)

It is the mean of all absolute error. MAE (ranges from 0 to infinity, lower is better) is much like RMSE, but instead of squaring the difference of the residuals and taking the square root of the result, it just averages the absolute difference of the residuals. This produces positive numbers only and is less reactive to large errors. MAE takes the **average** of the error from every sample in a dataset and gives the output.

Hence, **MAE = True values – Predicted values**

8.5.2 Mean Squared Error (MSE)

It is the mean of square of all errors. It is the sum, overall the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points. MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.

8.5.3 Root Mean Squared Error (RMSE)

RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model. RMSE (ranges from 0 to infinity, lower is better), also called Root Mean Square Deviation (RMSD), is a quadratic-based rule to measure the absolute average magnitude of the error.

In our project, we perform testing on two models: Linear Regression and Random Forest.

Linear Regression Model Testing:

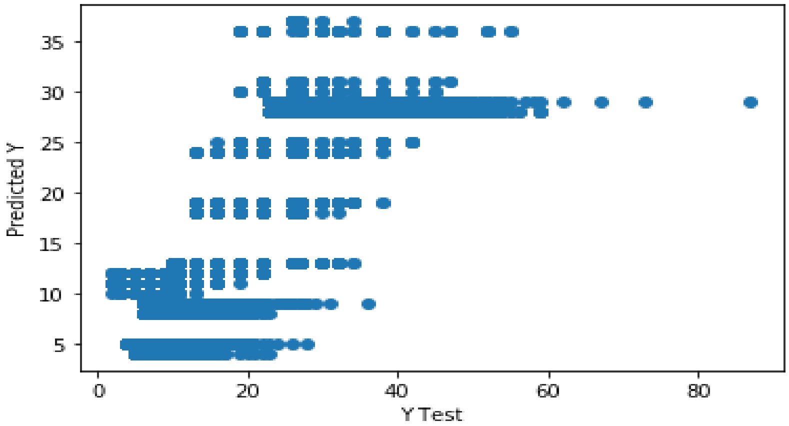


Fig. 8.12 Scatter Plot for Linear Regression

We draw a scatter plot between predicted and tested values and then find errors like MSE, MAE, and RMSE. After that, we also draw a distribution plot of the difference between actual and predicted values using the seaborn library. A **distplot** or distribution plot represents the overall distribution of continuous data variables.

Table 8.3: Error table for Linear Regression

Serial No.	Models	Accuracy
1	Mean Absolute Error	3.40607721
2	Mean Squared Error	20.0334370
3	Root Mean Absolute Error	4.47587277

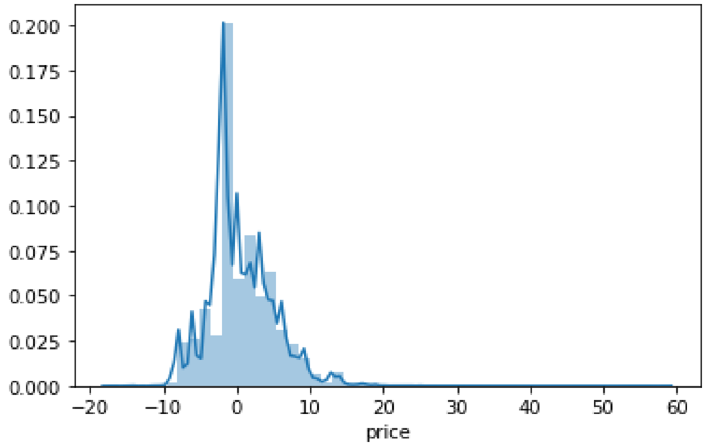


Fig. 8.13 Dist Plot for Linear Regression

Random Forest Model Testing:

Similarly, we draw scatter plot, dist plot, and find all three errors for random forest also.

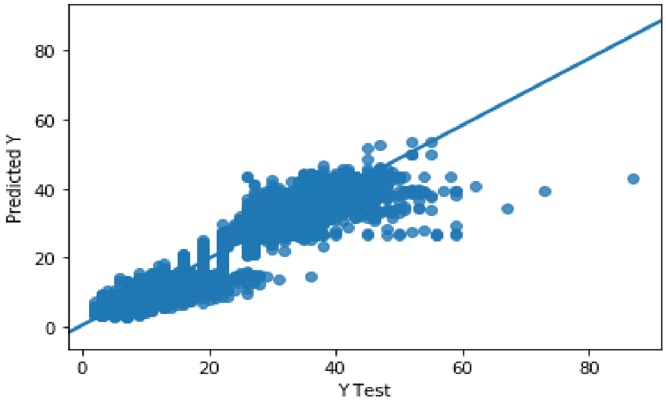


Fig. 8.14 Scatter Plot for Random Forest

Table 8.4 Error table for Random Forest

Serial No.	Models	Accuracy
1	Mean Absolute Error	0.99813700
2	Mean Squared Error	2.94465361
3	Root Mean Absolute Error	1.71599930

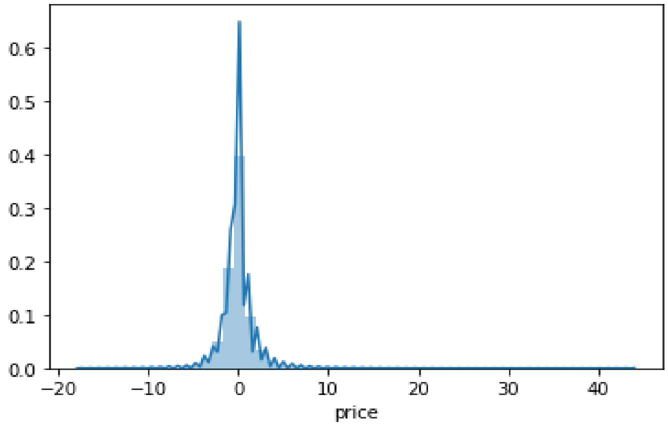


Fig. 8.15 Dist Plot for Random Forest

8.6 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 thelength 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.

Chapter 9 – LIMITATIONS

- Before you start managing and analyzing data, the first thing you should do is think about the PURPOSE. What it means is that you have to think about the reasons why you are going to do any analysis. If you are unsure about this, just start by asking questions about your story such as **Where? What? How? Who? Which?**

- Data visualization is certainly one of the most important stages in Data Science processes. While simple, it can be a powerful tool for prioritizing data and business context, as well as determining the right treatment before creating machine learning models.

- **Negative Impact of Price Competition**

- Price competition can be destructive for any industry. Increasingly, Uber, Lyft, and other e-hail services are engaged in an intense battle to provide the cheapest service. They are directly competing with each other, and with traditional taxi and car services for both customers and drivers.
- With competition from other ride-sharing services and the continuous hiring of new drivers, average earnings are being pushed downward. This means that drivers have to work longer hours to earn an income comparable to what they would have earned a year or two ago.

Chapter 10 – FINDINGS

❖ Insights from Data Exploration and Visualization.

- Early in 2017, the NYC Taxi and Limousine Commission (TLC) released a dataset about Uber's ridership between September 2014 and August 2015. This dataset contains features such as destination, trip distance, and duration that were not available in other sets released before and thoroughly analyzed by others.

- The combination of trip distance and duration allows for estimating Uber's revenue for each trip in NYC. In another hand, the pickup and drop-off locations were anonymized and grouped as taxi zones instead of geographic coordinates. This is a better attempt to preserve data privacy, but it precludes the positioning of such locations on a map.

- Before diving into the data, let me clarify what the term *"very large"* in the title means. The data comprises one complete year of trips, with a total of about 31 million entries. The uncompressed file itself is 1.4 GB, which is still fine to work on a laptop with 16 GB of RAM. However, some objects will be large enough to require better reasoning about how to efficiently apply transformations to them, from date-time parsing to arithmetic functions.

❖ Data Quality and Consistency.

- The NYC TLC requires that all taxi and for-hire-vehicles (FHV) companies operating in the city, which include Uber, Lyft, and others release their data periodically. An update is published twice a year. It's noteworthy that on their website the TLC warns about the non-audited nature of the data:

- "The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information."

- There were very few clearly erroneous entries in the dataset and a small proportion of suspicious cases or *anomalies* that warrant further internal analysis. These cases are, for example, those with very long distance traveled, but destination still recorded within New York City, or those with average speed slower than walking, but very long duration (beyond a reasonable assumption for the amount of time taken to get out of some really bad traffic gridlock, or the unlikely situation of a driver left waiting).

- In addition, there was a small proportion of cases with distance and duration equal to zero. Do they represent canceled trips? A small subset actually shows distinct origin and destination zones, indicating that some distance was driven but not recorded. In other cases, the recorded distance was zero, but the trip duration was more than that, even beyond 5 minutes in rarer cases. Are these system errors, fraud?

- The suspicious and anomalous data points were not changed, but the trips with a duration greater than 16 hours (123 cases out of nearly 31 million, mostly system errors) were removed from the dataset. In addition, the data was censored at exactly 365 days for convenience, which left only 1852 cases out.

- Finally, about 4% of the destination data were missing, and an extremely small number of cases had missing trip distance and destination. The imputation method chosen for the latter set was the mean distance and duration of their respective origin-destination pair. The entries with missing destination were left unchanged, although the information from the vast number of complete cases could potentially be used to determine the most probable destination.

❖ The Story from the Data: Uber's Growth in NYC.

- Uber launched in NYC in May of 2011, the first city outside of its San Francisco headquarters. NYC is probably the largest and most lucrative rideshare market in the world, with a total demand (for taxis and for-hire vehicles) in 2017 of more than 240 million trips per year.

- The number of Uber trips per day in NYC is still growing significantly. In 2017 so far, this number has often surpassed 200,000, but the plot below shows that by mid-2015 it was hovering around 120,000.

- Another interesting insight from the plot above is the effect of major events on the number of trips. For the period of time analyzed, negative impacts are related to Thanksgiving, Christmas, Memorial Day, and Independence Day. A lingering (two consecutive days) drop in activity is seen for all these holidays but Memorial Day. It turns out that the July 4th holiday was observed on Friday in 2015.

- In addition, an apparently odd and very significant drop in the number of trips is shown on January 27th. This was a result of a curfew imposed by the NYC's mayor in preparation for a blizzard.

- In the other hand, the plot also highlights which events have positively impacted the number of trips that year, with the International Marathon and the Gay Pride Week standing out as the strongest contributors. The latter attracted as many as 2 million people to the events in NYC and could be easily identified through an internet search, but figuring out the spike caused by the marathon required some “domain knowledge”, and having a friend who used to live in the city was definitely helpful.

❖ **Trends in the Demand for Rides in the City.**

- The data also allows us to visualize other interesting trends over time. In the bar charts below, we can see that the demand for Uber is higher from 4 PM until around midnight. Saturday has the highest demand. Interestingly, Sunday shows a level of demand similar to Wednesday, which is higher than Monday or Tuesday. When looking at the total demand per month along the period of time analyzed, seasonal effects are masked by the consistent month-to-month growth.

- Data analysis is the most crucial part of any research. It involves the interpretation of data gathered through the use of analytical and logical reasoning to determine patterns, relationships or trends
- Data Preparation is usually a stage that requires lots of work around data formatting, cleansing and manipulation, but making your data CONSISTENT is surely a success factor for your analysis and future modeling.

Chapter 11 – SUGGESIONS

"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

- **Data requirements**

  - The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis (or customers, who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

- **Data collection**

  - Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data; such as, Information Technology personnel within an organization. The data may also be collected from sensors in the environment, including traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

- **Data processing**

  - Data, when initially obtained, must be processed or organized for analysis.For instance, these may involve placing data into rows and columns in a table format (*known as* structured data) for further analysis, often through the use of spreadsheet or statistical software.

- **Data cleaning**

  - Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for *data cleaning* will arise from problems in the way that the datum are entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation.Such data problems can also be identified through a variety of analytical techniques. For example; with financial information, the totals for particular variables may be compared against separately published numbers that are believed to be reliable.Unusual amounts, above or below predetermined thresholds, may also be reviewed. There are several types of data cleaning, that are dependent upon the type of data in the set; this could be phone numbers, email addresses, employers, or other values. Quantitative data methods for outlier detection, can be used to get rid of data that appears to have a higher likelihood of being input incorrectly.Textual data spell checkers can be used to lessen the amount of mis-typed words. However, it is harder to tell if the words themselves are correct.

- **Exploratory data analysis**

- Once the datasets are cleaned, they can then be analyzed. Analysts may apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data.The process of data exploration may result in additional data cleaning or additional requests for data; thus, the initialization of the *iterative phases* mentioned in the lead paragraph of this section. Descriptive statistics, such as, the average or median, can be generated to aid in understanding the data. Data visualization is also a technique used, in which the analyst is able to examine the data in a graphical format in order to obtain additional insights, regarding the messages within the data.

- **Modelling and algorithms**

- **Mathematical formulas** or **models** (known as **algorithms**), may be applied to the data in order to identify relationships among the variables; for example, using correlation or causation. In general terms, models may be developed to evaluate a specific variable based on other variable(s) contained within the dataset, with some *residual error* depending on the implemented model's accuracy (*e.g.*, Data = Model + Error).

- Inferential statistics, includes utilizing techniques that measure the relationships between particular variables. For example, regression analysis may be used to model whether a change in advertising (*independent variable X*), provides an explanation for the variation in sales (*dependent variable Y*). In mathematical terms, *Y* (sales) is a function of *X* (advertising).It may be described as (*Y* = *aX* + *b* + error), where the model is designed such that (*a*) and (*b*) minimize the error when the model predicts *Y* for a given range of values of *X*. Analysts may also attempt to build models that are descriptive of the data, in an aim to simplify analysis and communicate results.

❖ **Data product**

- A **data product** is a computer application that takes *data inputs* and generates *outputs*, feeding them back into the environment. It may be based on a model or algorithm. For instance, an application that analyzes data about customer purchase history, and uses the results to recommend other purchases the customer might enjoy.

❖ **Communication**

- Once data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements.The users may have feedback, which results in additional analysis. As such, much of the analytical cycle is iterative.
- When determining how to communicate the results, the analyst may consider implementing a variety of data visualization techniques to help communicate the message more clearly and efficiently to the audience. Data visualization uses information displays (graphics such as, tables and charts) to help communicate key messages contained in the data.Tables are a valuable tool by enabling the ability of a user to query and focus on specific numbers; while charts (e.g., bar charts or line charts), may help explain the quantitative messages contained in the data

Chapter 12 – 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.

- In this report, I aimed to expose all the interesting insights that can be derived from a detailed analysis of the dataset, without even doing any machine learning. I particularly had fun investigating the "anomalies" in the plot of the total daily trips, which I have illustrated with icons to visually emphasize them.
- Having identified these change points will be very useful for the next step I intend to take on this project: **forecasting demand**. As the NYC TLC has published the aggregated total count of trips per week, I will have the data to compare the results of my experiments. Forecasting is an exercise of "science and art", but there are some more recent packages that make playing with Bayesian modeling more practical.
- Early in 2017, the NYC Taxi and Limousine Commission (TLC) released a dataset about Uber's ridership between September 2014 and August 2015. This dataset contains features such as destination, trip distance, and duration that were not available in other sets released before and thoroughly analyzed by others.

- The combination of trip distance and duration allows for **estimating Uber's revenue for each trip** in NYC. In another hand, the pickup and drop-off locations were anonymized and grouped as taxi zones instead of geographic coordinates. This is a better attempt to **preserve data privacy**, but it precludes the positioning of such locations on a map.

BIBLIOGRAPHY & RFERENCES

- Abel Brodeurand & Kerry Nield (2018) An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC
- Junfeng Jiao (2018) Investigating Uber price surges during a special event in Austin, TX
- Anna Baj-Rogowska (2017) Sentiment analysis of Facebook posts: The Uber Case
- Anastasios Noulas, Cecilia Mascolo, Renaud Lambiotte, and Vsevolod Salnikov (2014) OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs
- https://www.singlegrain.com/blog-posts/business/10-lessons-startups-can-learn-ubers-growth/
- https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston
- https://matplotlib.org/1.3.1/users/legend_guide.html
- https://www.sciencedirect.com/science/article/abs/pii/S0167268118301598
- https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S2210539517301165
- https://ieeexplore.ieee.org/abstract/document/8260068
- https://www.researchgate.net/publication/305524879_Dynamic_Pricing_in_a_Labor_Market_Surge_Pricing_and_Flexible_Work_on_the_Uber_Platform
- https://www.sciencedirect.com/science/article/abs/pii/S0167268118301598#:~:text=We%20look%20at%20the%20effect,rides%20in%20New%20York%20City.&text=The%20number%20of%20Uber%20(Lyft,higher%20when%20it%20is%20raining.&text=The%20number%20of%20taxi%20rides,higher%20when%20it%20is%20raining.&text=Taxi%20rides%2C%20passengers%20and%20fare,after%20Uber%20entered%20the%20market.
- https://arxiv.org/abs/1503.03021

- https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789808452/1/ch01lvl1sec19/label-encoding
- https://github.com/Ankush123456-code/house_price_prediction_end_to_end_project/blob/main/model/python.ipynb
- https://github.com/ankita1112/House-Prices-Advanced-Regression/blob/master/Housing_Prediction_full.ipynb
- https://scikit-learn.org/stable/modules/multiclass.html
- https://www.codegrepper.com/code-examples/python/confusion+matrix+python
- https://topepo.github.io/caret/recursive-feature-elimination.html
- https://arxiv.org/pdf/1503.03021.pdf
- https://www.sciencedirect.com/science/article/abs/pii/S0167268118301598
- https://www.kaggle.com/punit0811/machine-learning-project-basic-linear-regression
- https://gdcoder.com/decision-tree-regressor-explained-in-depth/
- https://medium.com/towards-artificial-intelligence/machine-learning-algorithms-for-beginners-with-python-code-examples-ml-19c6afd60daa
- https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
- https://blog.paperspace.com/implementing-gradient-boosting-regression-python/
- https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared
- https://statisticsbyjim.com/regression/overfitting-regression-models/
- https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/striplot.htm#:~:text=One%20problem%20with%20strip%20plots,increment%20to%20the%20vertical%20coordinate.

- https://www.toppr.com/ask/question/what-are-the-limitations-of-a-scatter-diagram/
- https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/#:~:text=Limitation%20of%20label%20Encoding,in%20training%20of%20data%20sets.
- https://digitalag.osu.edu/sites/digitag/files/imce/images/ag_sensing/Figure3.png
- https://www.researchgate.net/publication/325800934/figure/fig1/AS:638132596768768@1529154075317/The-main-procedure-of-the-recursive-feature-elimination-RFE-method.png
- https://upload.wikimedia.org/wikipedia/commons/thumb/b/b5/K-fold_cross_validation_EN.svg/1200px-K-fold_cross_validation_EN.svg.png
- https://static.packt-cdn.com/products/9781789345070/graphics/108f2a01-3e31-4907-a1a5-4baf441c3eed.png

ANNEXURE& QUESTIONERY

- How many times have I traveled in the past?
- How many trips were completed and canceled?
- Where did most of the layoffs take place?
- What type of product is most often selected?
- What a measure. fare, distance, amount, and time spent on the ride?
- Which days of the week have the highest fare?
- Which is the longest / shortest and most expensive / cheapest ride?
- What is the average lead time before requesting a trip?