TABLE OF CONTENTS

Definitions, Acronyms and Abbreviations

1.0 Introduction 2

1.1 Overview 2

1.2 Scope 2

1.3 Objective

2.0 Literature Survey 3

3.0 Methodology 7

3.1 Proposed Approach 7

4.0 Environment Requirements 8

4.1 Hardware Requirements 8

4.2 Software Requirements 8

6.0 Results 13

7.0 Conclusions 14

8.0 Future Work 15

9.0 References 15

Introduction

Overview

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable.

Scope

The price of a flight ticket depends upon various factors, often it can be quite challenging to correctly determine the price of flight ticket. Our team tackles this problem by using Machine Learning to address this problem.

Objective

Now a Machine Learning model is as good as the data it uses. So we have taken the dataset from Machinehack.com, where different hackathons are listed.

The dataset uses features like the name of airline, date\_of\_journey, Source and Destination Airport, Route taken, Departure\_time, duration of the flight, total stops and any additional info. Taking these features into account we can train our model and then predict the price. Our training set consists of **10683** records.



# Literature Survey

One of the pioneers on optimal ticket purchase timing prediction is probably the work done by (Etzioni et al, 2003). The authors proposed a model that advises the user whether to buy a ticket or to wait at a particular point of time. For each query day, the model generates a buy or wait signal based on historical price information. The model uses various data mining techniques such as Rule learning (Ripper), Reinforcement learning (Q-learning), time series

Methods and combinations of these to achieve various accuracy levels.

The authors in (T.Wohlfarth et al., 2011) proposed an optimal ticket purchase time optimizing model based on a special preprocessing step known as marked point processes (MPP), data mining techniques (clustering and classification) and statistical analysis techniques. The MPP pre-processing technique was suggested to convert heterogeneous price series data such as international, national, long and short flights, different providers (low cost and regular) into an interpolated price series trajectory that can be fed to an unsupervised clustering algorithm. Once the MPP step is completed, the model applies clustering followed by classification and statistical processing techniques on historical price data to develop price decrease event predictive rules. First, the price series trajectory is clustered into groups based on similar pricing behavior. Next, a price evolution model that estimates price change patterns up to departure date is defined for each cluster. For a new test dataset, a tree-based classification algorithm is used to select the best matching cluster and then the corresponding price evolution model defined for that cluster is used to predict the price decreasing event. The dataset used by this research is obtained

From Liligo.com’s historical price data collected for 28 days. It covers data for 6 routes from 9 airlines. Unlike others, this paper also considers round-trips for 3, 7 and 14 days. The set of features in the analysis include: departure station, arrival station, departure date, return date, provider, day of week, day of month, day of year and demand. The authors claim that the model achieved 55% performance However, no details of performance evaluation steps were presented

Y.Chen et al., (2015) proposed a model that predicts the lowest price available for a given itinerary (a specific flight on a given route for a particular departure date). To be more precise, given the current day, d1, and a specific itinerary (r,d,n) identified by route r and departure date dn, the model predicts

the lowest prices available for consecutive days d2,d3...dn-1,dn where d1<d2<d3<dn-1<dn. However, the model considers only non-stop flights. Moreover, it is not possible to predict the price of a single flight as it works at the route level. An ensemble-based learning algorithm Learn++.NSE, is modified and trained to incrementally learn from past patterns of the price changes and to forecast future prices. A recursive strategy is used to estimate multiple future prices iteratively i.e. the price from previous predictions is used to predict the next price in multiple steps. Features such as prices of the same

itinerary, prices of recent itineraries before the target day, prices of

itineraries with the same day of the week and price of itineraries with the same day of the month are used for the model. The model is tested on a daily price dataset extracted from an OTA company in China for 5 different international routes. The collection was made for more than 3 months (Feb 11 to Jun 01, 2015, for 110 days in total). For each day, the lowest prices of itineraries leaving in the next 60 days were recorded for every route, resulting in 110

60 total observations. Experimental results reveal that the model performs relatively better on diverse routes i.e. routes in which pricing behavior of different flights is completely independent from each other with different price level and variation magnitude having no universal pattern.

When to Book: Predicting Flight Pricing

In the paper published by Qiqi Ren at Stanford University the project

was “When is the best time to purchase a flight?”. The project would

use machine learning classification to predict the flight price.

He focused on the properties of the customers and predict the binary

class that whether the price will increase or not i.e. a person should buy

it or not.He collected the dataset from Expedia, a major travel website. He looked at five different major locations: Boston (BOS), Chicago (CHI),Portland (PDX), Los Angeles (LAX), and New York (LGA). The features he used were current day of the week, the day of the week of the flight, the current time of day(categorized into four 6-hour time periods), the flights time of day, the number of hours until the flight, number of stops, the duration in minutes, and the price of the flight. It led to a total of 75000 examples. He split this data randomly with a 70-30 split for training versus test examples.

The methods used were:

1) Logistic Regression was used as the simple base model on the

dataset

2) SVM is a supervised learning model.

3) K-Nearest Neighbors was used for training set with k=3.

4) Random Forest

Result and Conclusion:

The initial base model Logistic regression that gave an accuracy 61%.

The classification models I tried (SVM, k-Nearest-Neighbors, Random

Forest) performed well. However, the fact that all of the test set

accuracies are a few percentage points less than their respective

training set accuracies show that these models are overfitting the data.

The performance indicator for any transportation system is its delay. Flight delays have negative impact, mainly economic, they also jeopardize the airlines marketing strategies. The main problems

related to ﬂight delay predictions are the scopes, models, and ways of handling ﬂight delay prediction problem. It considers ﬂight domain features, such as problem and scope, and Data Science perspectives, such as data and methods. There are three major concerns regarding the ﬂight delay prediction problem: delay propagation, root delay and cancellation. Root delay is caused due to weather conditions, acts of God, aircraft problems, may lead airlines to cancel ﬂights. Besides,

airlines may directly cancel a ﬂight, when factors like seat occupancy or cost savings are taking into consideration. In delay propagation, we try to understand how delay propagates through airlines and airports based on the assumption that an initial delay has already occurred in the transportation system. The data model used for flight delay prediction uses features, spatial, planning, temporal, weather, operations and the state of the system. Since the data generated has been increasing a lot

since the past years, data warehouse is used to store data. Data pre-processing includes data cleaning(outlier removal), feature selection, data transformation(normalization and discretization)and clustering. There are several methods to predict flight delays. Statistical Analysis

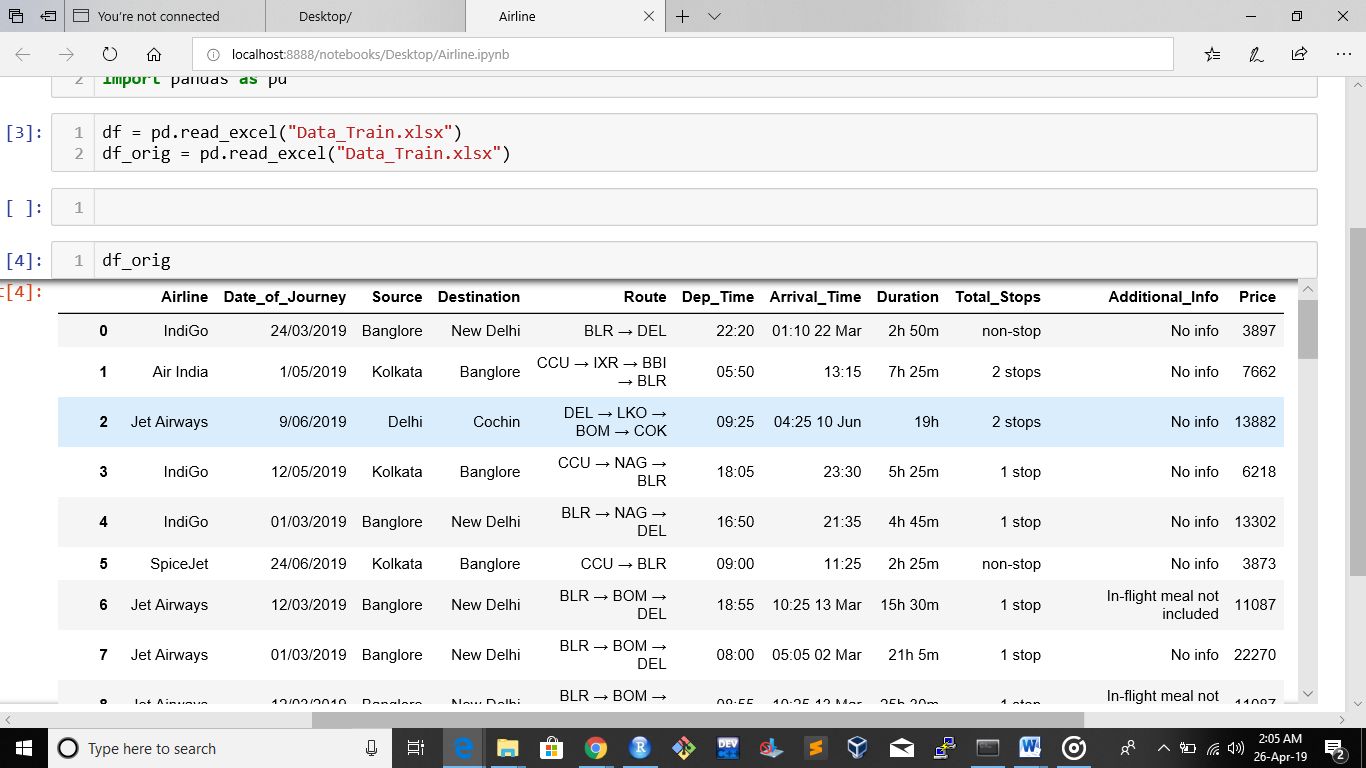
uses regression model in which both delay multiplier and recursive models can help airlines to understand delay propagation eﬀects through the network and to estimate the costs of delays. Machine Learning includes k-Nearest Neighbour, neural networks, SVM, fuzzy logic, and random forests. They are mainly used for classiﬁcation and prediction. Prediction of flight delays can improve marketing decisions for the airlines. Thus, flight delays are an important subject due to their economical and environmental impacts. They may increase costs to customers and operational costs to airlines.

Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system.

# Methodology

## Proposed Approach

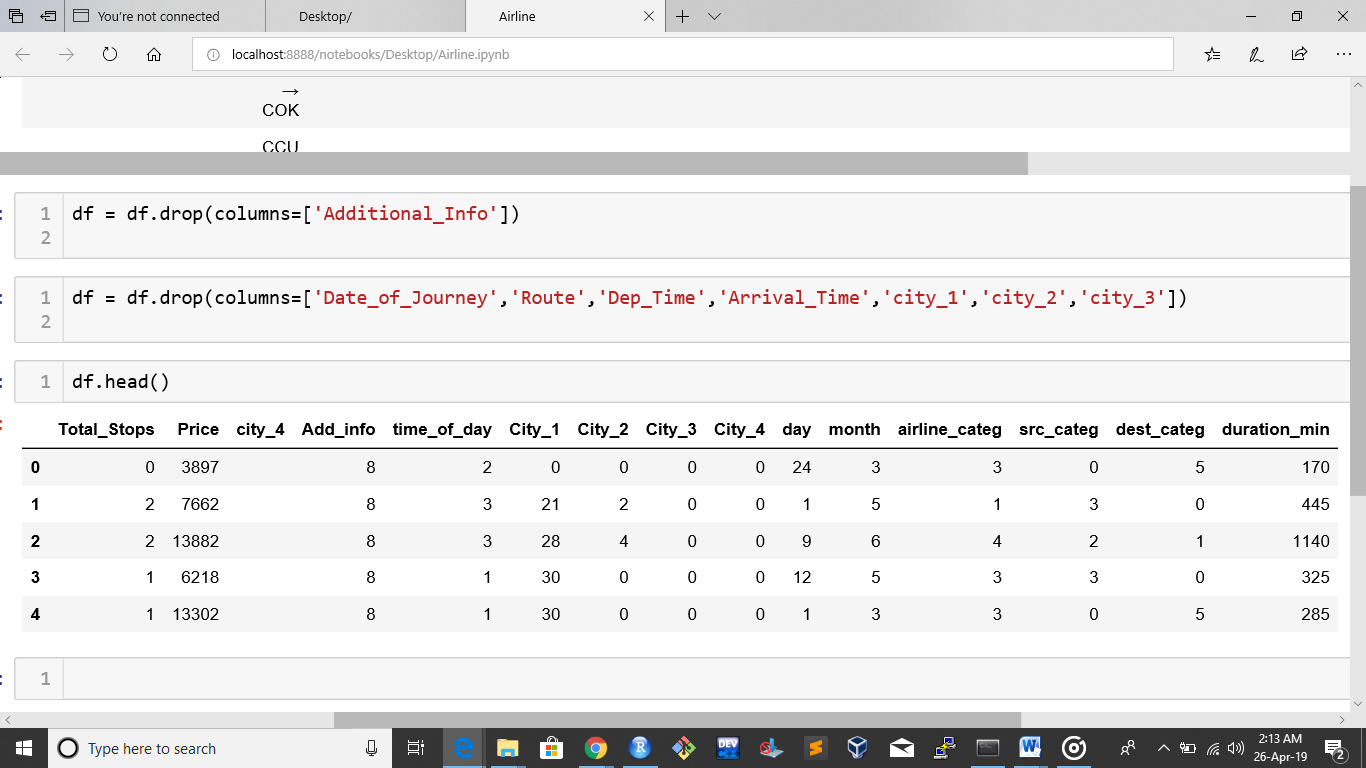
Our dataset had various categorical, time and date related features in addition to numeric data.



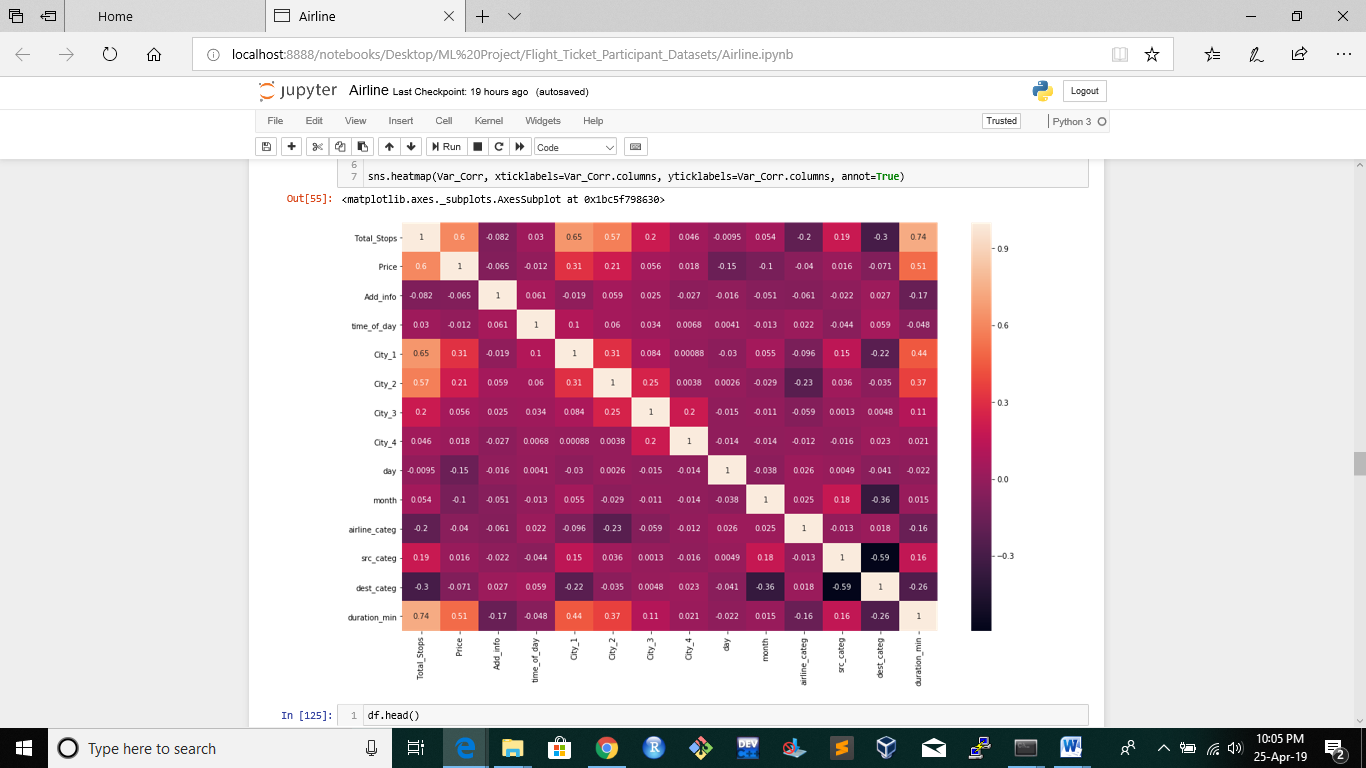
In order to use our features we converted the categorical features such as Airline name, Source, Destination, Total\_Stops and Additional\_Info to categorical data.

We also had to handle Date\_of\_Journey, Dep\_Time and Duration as they were not numeric data, and therefore cannot be used in any ML model. So we divided the Date\_of\_Journey into day and month as different columns.

The Dep\_Time was converted into numeric code depending upon time of the day. The duration was converted into minutes depending upon the format.



We decided to construct a heatmap to see the correlation between the target and input features, and thought of using only the most correlated feature i.e. No\_of\_stops, but that resulted in very less accuracy. So instead we used all the features as input features and used various regression models to predict the price.



# Environment Requirements

## Hardware Requirements

Since our data is not Big Data, there are no extra hardware requirements, like GPU or High RAM. But 4 GB RAM is preferable for speedy execution of models.

Software Requirements

There are no special OS requirements; any OS Ubuntu or Windows will work.

However we do require some python packages preinstalled.

1) sklearn.model\_selection : Split arrays or matrices into random train and test subsets.

Version:scikit-learn v0.20.3

2) pandas : It is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the programming language.

Version : v0.23.4

3) plotnine : plotnine is an implementation of a grammarofgraphics in Python, it is based on ggplot2. The grammar allows users to compose plots by explicitly mapping data to the visual objects that make up the plot.

Version : plotline 0.5.1

4) seaborn : Seaborn is a Python data visualization library based on matplotlib.

Version: seaborn 0.90

5) RandomForestRegressor : A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Version : scikit-learn v0.20.3

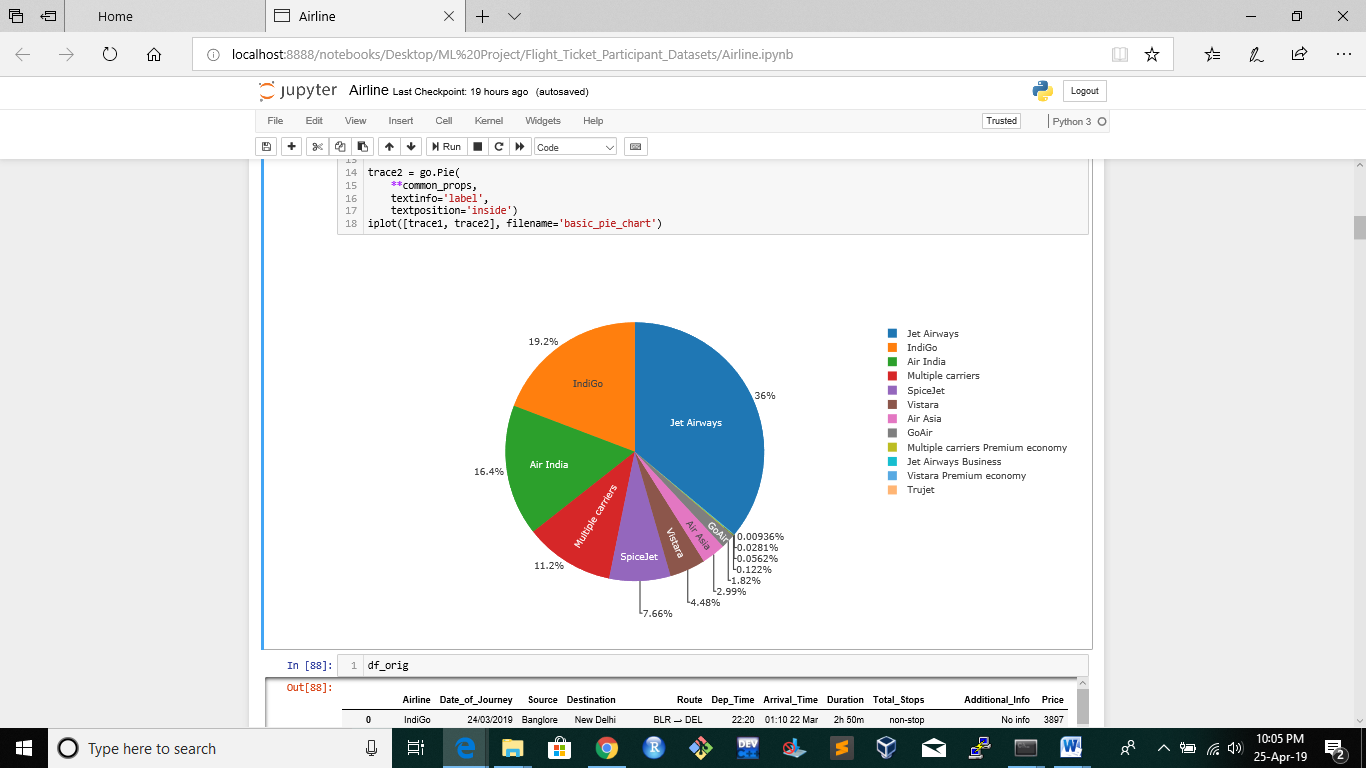
7) AdaBoostRegressor: The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.

Version : scikit-learn v0.20.3

8) GradientBoostingRegressor: GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function

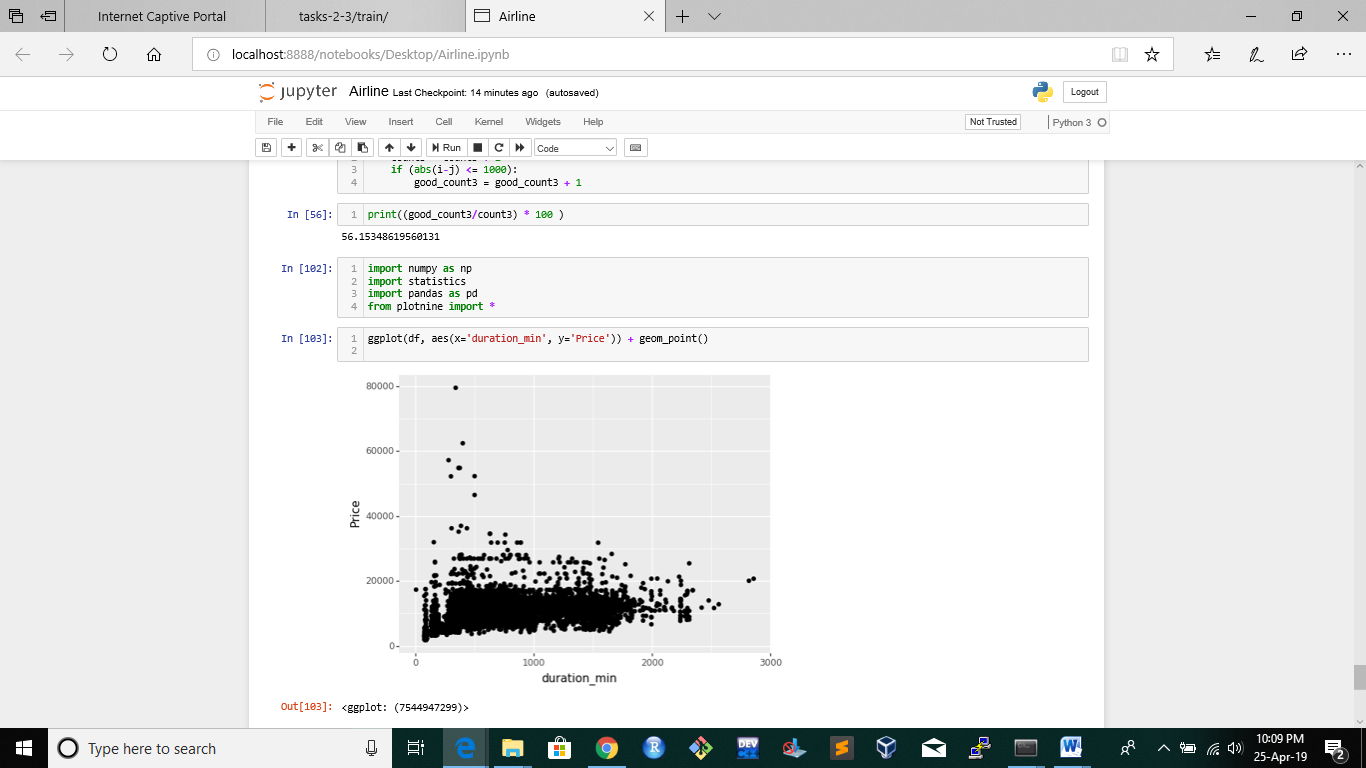
Version : scikit-learn v0.20.3

# Some Visualizations

A pie plot to see the percentage usage of different airlines

Scatter plot showing the dependency of

duration\_min vs Price

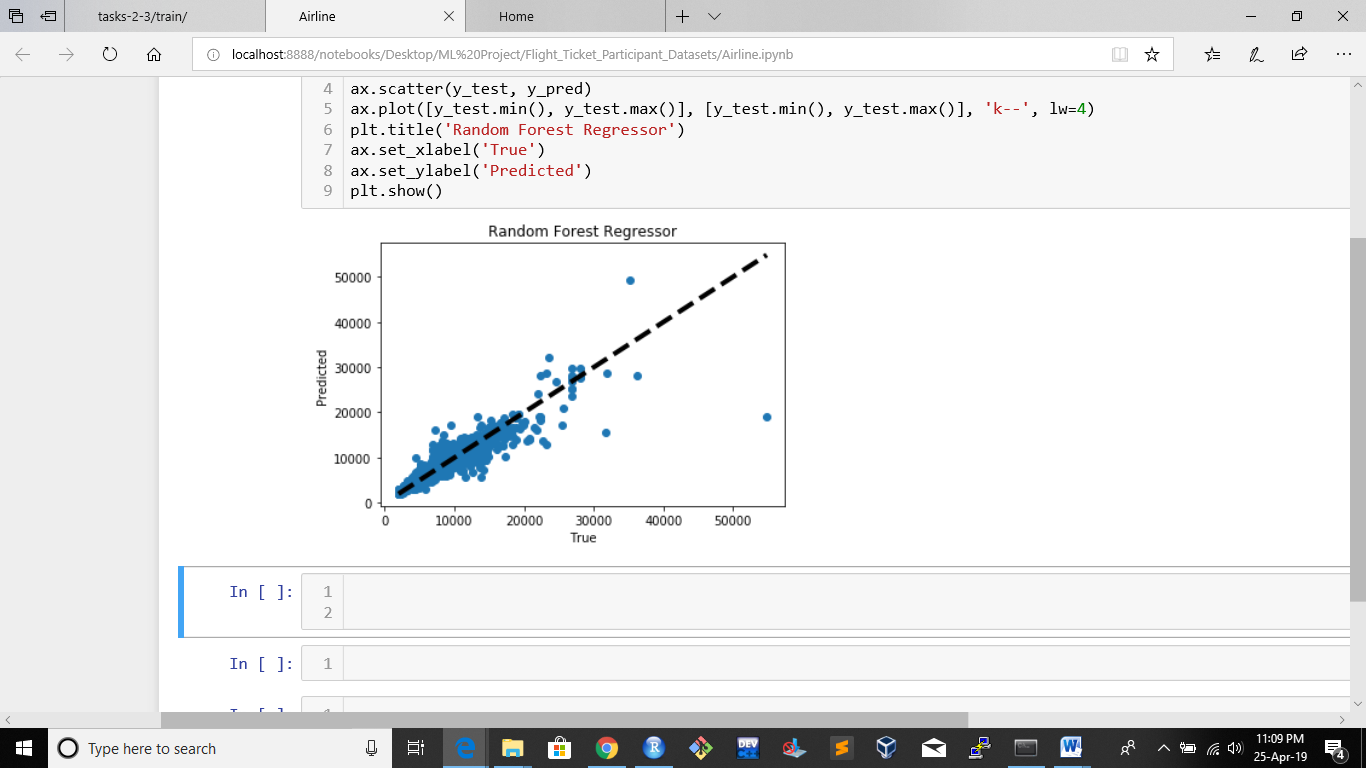


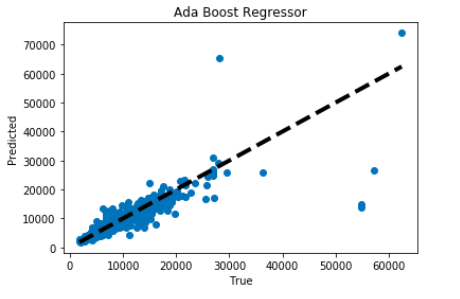
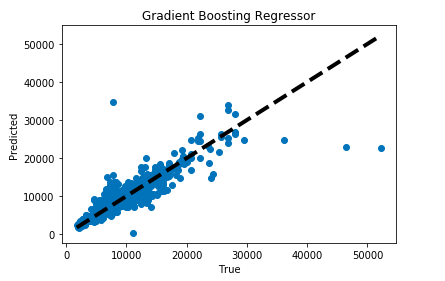
# Results

After applying 3 different regression models, we calculated an accuracy metric.

Accuracy\_score = (good\_count) / (count)

Where count is the total number of samples and good\_count is the number of samples where the predicted and actual ticket price differ by 1000 Rs or less.



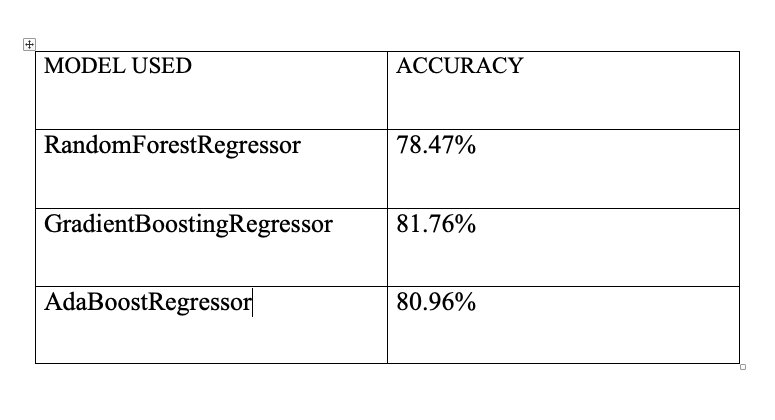


# Conclusions

GradientBoostingRegressor gave us the best accuracy 81.76%

Since we have used all of the input features we got a decent accuracy.

Now if later on our dataset increases, and taking all the features becomes computationally extensive, then we can use PCA or SVD to reduce the feature space, but it might lead a small drop in accuracy.



Future Work

We got approx. 82% accuracy just by considering internal features, in later stage we can use external features like festive season, competition between airline company, weather conditions , economic activity.

Also since nowadays many prediction tasks is using social media data, we can also refer to it to help in our prediction.

We can apply Deep Learning Models to extract features from dataset and predict using ANN.

References

Papers used for Literature survey –

* Airline ticket price and demand prediction: A survey

Juhar AhmedAbdella, Nazar Zaki,Khaled Shuaib, Fahad Khan

* When to Book: Predicting Flight Pricing

Qiqi Ren (Stanford University)

* A Review on Flight Delay Prediction

Alice Sternberg, Jorge Soares,

( Diego Carvalho, Eduardo Ogasawara∗ )

CEFET/RJ

( Rio de Janeiro, Brazil )