**AIRLINE TICKET FARE PREDICTION**

**A PROJECT REPORT**

***Submitted by TEAM 9***

MOHAMMED KHAJA MUJAHIDDIN – TEAM LEAD

SANA TABASSUM SHAIK – DATA SCIENTIST

KATAM SAI DEEPAK REDDY – DATA COLLECTION AND PREPROCESSING

***Under the guidance of***

**Professor Ms. Ardiana Sula**

**For the Course DSCI-6007-01**

**DISTRIBUTED AND SCALABLE DATA ENGINEERING**

Logo

Description automatically generated

**UNIVERSITY OF NEW HAVEN**

**WEST HAVEN, CONNECTICUT**

**FALL 2024**

**ABSTRACT**

The Airline industry is a significant component of the global travel and transportation sector. It encompasses the buying and selling of airline tickets for passenger travel, both domestic and international. Travelers look for best deals and the Airlines try to make profit by selling their tickets. As the industry evolves, building a flight ticket fare prediction model will be a valuable tool for travelers and the airline industry. These models can benefit both consumers looking for the best deals and airlines striving to optimize their pricing strategies.

Airline ticket prices are influenced by a multitude of factors, making them a complex and dynamic pricing challenge. However, with the power of data and machine learning, we can create models that provide accurate fare predictions. As the industry evolves, these models can adapt to changing market dynamics, ultimately benefiting both consumers and airlines.

Data science and machine learning have a wide range of applications across various domains and industries. These technologies have the potential to extract valuable insights, make predictions, and automate tasks, leading to improved decision-making and efficiency. Predicting airline ticket fares is a complex task that involves multiple variables and factors. Data Science and Machine learning techniques can be used to build an effective solution for predicting airline ticket fares.

Dataset has been collected for the above requirement and preprocessing has been performed on it using various python libraries. Feature extraction has been performed to get valuable insights from the dataset. Machine Learning models like Linear Regressor and Random Forest regressor have been build and the models performance has been evaluated and it is observed that the model is able to perform well with a good prediction rate and can be deployed for real time usage.

**CHALLENGE**

How to predict the price of an Airline ticket based on certain features?

Buying an airline ticket depends on various factors like source, destination, number of stops between source and destination, price range, flight duration, airline company etc. Travelers look for the best deal at a low price with certain specifications while buying a flight ticket. Building a model considering all these factors is a challenging task. The model has to include various requirements like:

What is the source and destination?

What is the price range?

Number of tickets needed?

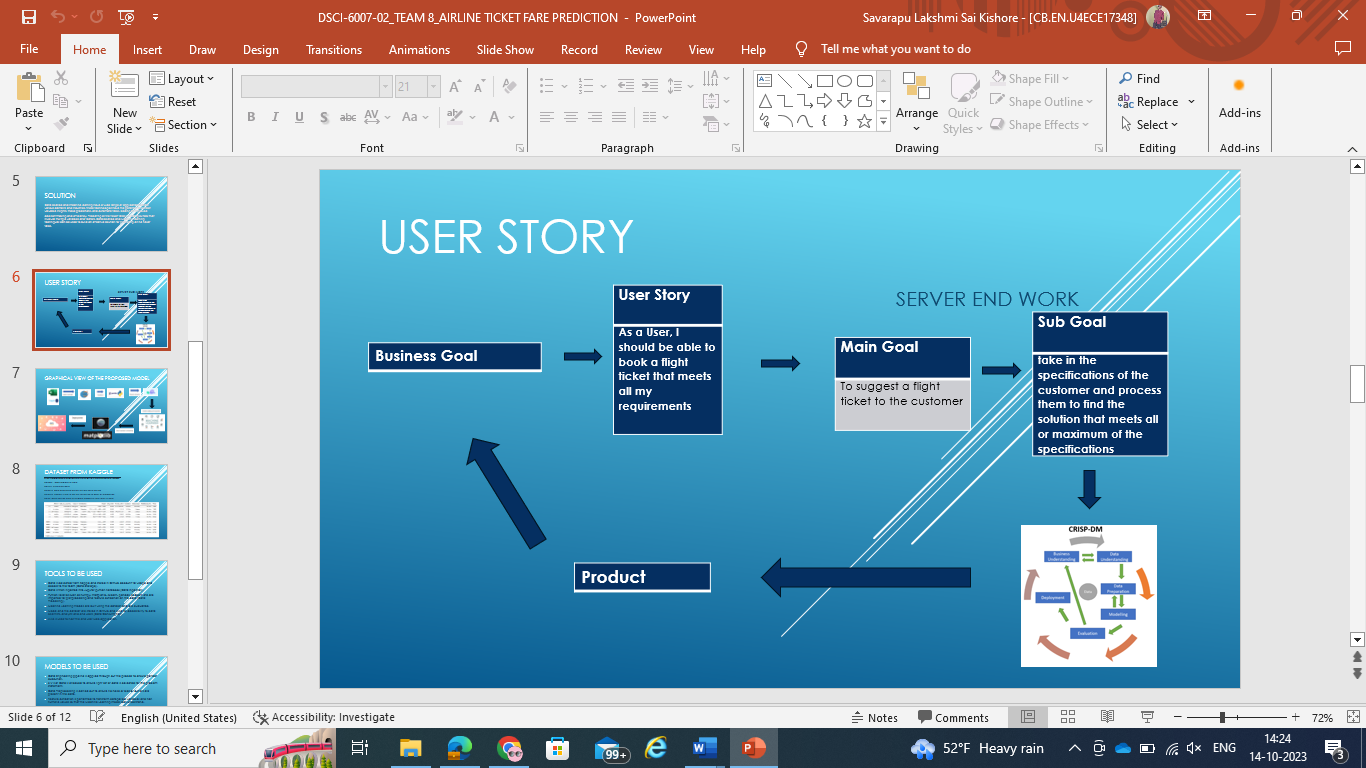
Number of stops in between source and destination?

What is the flight duration?

**SOLUTION**

Considering all sorts of requirements like the above, a Machine Learning model can be built that can predict the price of an airline ticket. Data Science and Machine Learning techniques have been used to solve these kinds of problems in various domains and industries. Many industries have implemented Data Science and Machine Learning techniques in real time cases and have achieved great yield. Fraud detection, Recommendation systems, E-commerce pricing, Healthcare diagnostics are some examples of real time applications of Data Science and Machine Learning techniques. With boundless number of applications of Data Science and Machine Learning Techniques, it is possible to build an efficient model that can be used in real time for predicting Airline ticket price.

**USER STORY**



Title: Flight ticket Fare Prediction

User story: As a user, I should be able to get a predicted price of a flight ticket depending on my requirements so that I can make a decision on my travel plans.

Acceptance Criteria:

Input Parameters:

Source and destination, Number of stops in between, Flight duration, Airline Company, etc.

Historical data is required for the model building so that the input parameters can be considered while predicting the ticket price and suggesting it to the user.

Output: Machine Learning model is built that takes input of all the parameters and predicts the ticket price to the customer.

**DATASET**

<https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh>

5 V’s of Data is checked to ensure right set of data is collected for the problem statement.

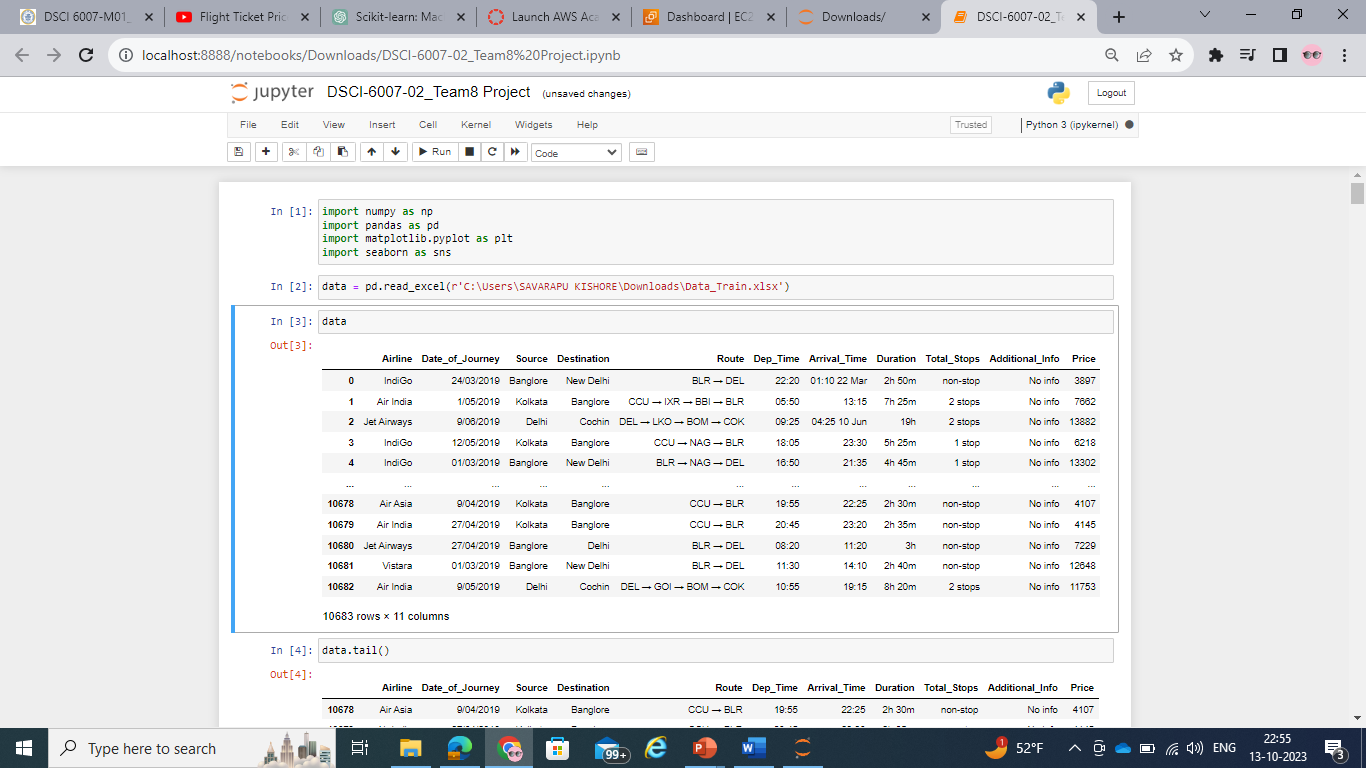
Volume: 10683 samples of data

Variety: Structured Data

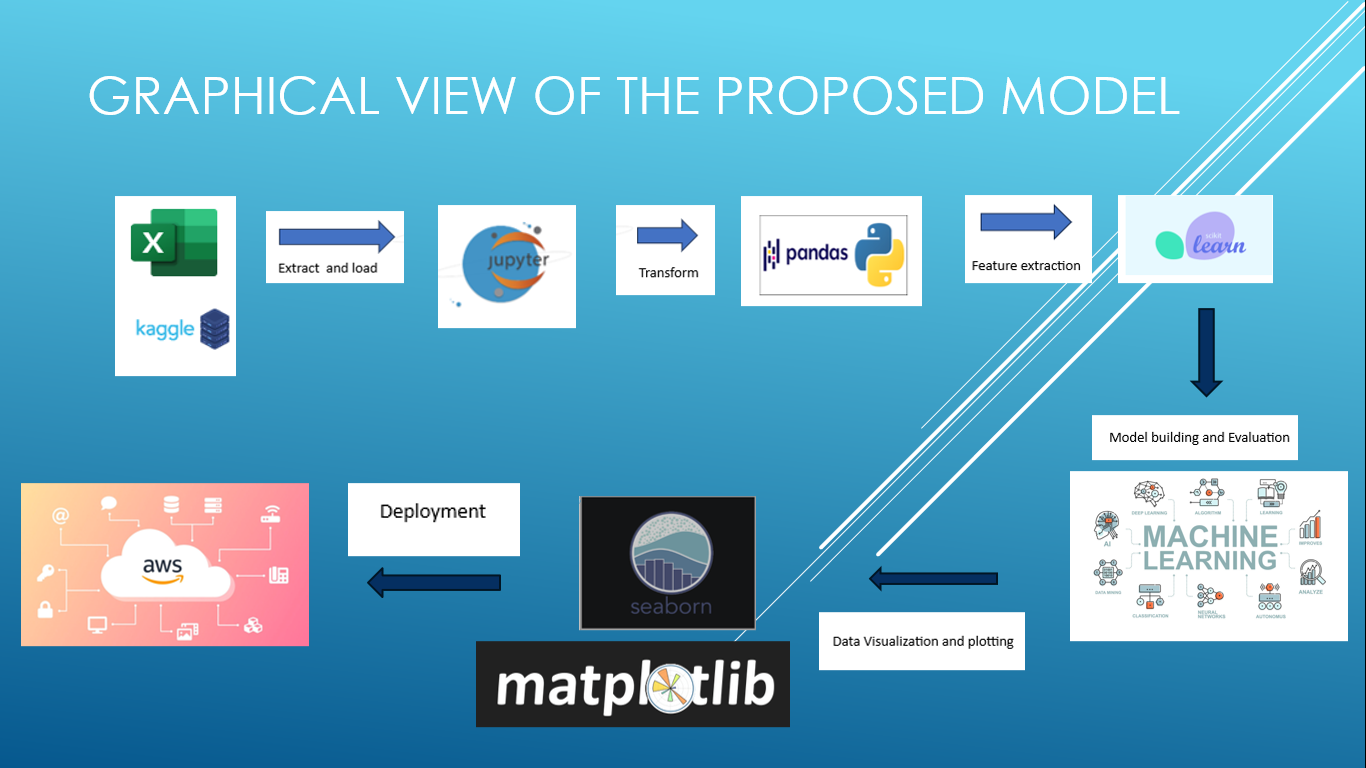
Velocity: Data generated during the pre-covid period

Veracity: Kaggle is said to be the trusted go to point of information

Value: Data can be used to estimate domestic flight fares in India



**GRAPHICAL VIEW OF THE PROPOSED MODEL**



Data Engineering pipeline will be followed while building the model.

Data Engineering pipeline will have the following steps:

* Data Ingestion
* Data Storage
* Data Processing
* Data Consumption

**TOOLS AND TECHNIQUES TO BE USED**

* Data is collected from Kaggle and stored in Github account for usage and access to the team (Data Storage).
* Data is then ingested into Jupyter python notebook (Data Ingestion).
* Python libraries such as numpy, matplotlib, sklearn, pandas, seaborn etc are imported for pre-processing and feature extraction on the data (Data Processing).
* Machine Learning models are built using the dataset and are evaluated.
* Model and the dataset are stored in Github and AWS for accessibility to data scientists, analysts and end users (Data Consumption).
* AWS is used to host the end user web application.
* Data Engineering pipeline is applied through out the process to ensure perfect execution.
* 5 V’s of Data is checked to ensure right set of data is collected for the problem statement.
* Data Preprocessing is carried out to ensure no noise or bias or outliers are present in the data.
* Feature extraction is performed to transform categorical variables and non numeric values so that the Machine Learning model can understand.
* Linear Regression and Random Forest Regressor models are build using the dataset.
* Models are evaluated by calculating errors and accuracy.
* Models are deployed successfully.

**REFERENCES**

* <https://youtu.be/WjLou_0YVMo?si=O_MsRoI5q4ooTXeC>
* <https://youtu.be/46bRwRvW0PI?si=KIchjcJvDOZkfBMq>
* <https://youtu.be/KFPX0FLhscs?si=VSFVpvoKYssh3epv>