A Project Report on

Optimising Flight Booking Decision Through Machine Learning

by

Gunda Rudrani Reddy(Team Leader)-20AT1A05B9
Poldas Lalana Priya -20AT1A0569
Battagari Gnana Amrutha-20AT1A0540

Under the Guidance of

MRS.M.JAYA SUNITHA,M.Tech

Associate Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

G. PULLAIAH COLLEGE OF ENGINEERING AND TECHNOLOGY (Autonomous)

(Approved by AICTE | NAAC Accreditation with 'A' Grade | Accredited by NBA (ECE, CSE, EEE, CE) | Permanently Affiliated to JNTUA)

ABSTRACT

The Flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, destination, duration of flights. In the proposed system a predictive model will be created by applying machine learning algorithms to the collected historical data of flights. Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket. The project implements the validations or contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time frame where the prices are minimum. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum. The scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.

Contents

CHAPTER 1

INTRODUCTION

1.1 Specify The Business Problem

Choosing the wrong booking system for your company's needs you will need (at least) one expert within the business heavily reliant on a good internet connection a target for cyber criminals your business needs to be ready to grow.

1. Choosing the wrong booking system for your company's needs:

   With so many booking system options on the market, it may seem overwhelming to pick the best system for you. This window shopping period is essential though, as you don't want to sign up for a 12 month contract and pay for a system that's not serving you as well as another option could.

2. You will need (at least) one expert within the business :

   No automation system is completely self-sufficient. You will always need at least one person in the business who can spend at least some of their working week managing the system.

3. Heavily reliant on a good internet connection:

   Some things are simply out of our control. Online booking system problems are sometimes unrelated to your system and how you've set it up. If your customers have a poor internet connection you may still need to take telephone or email bookings.

4. A target for cyber criminals:

   The importance of a secure booking system has never been higher. As personal data and bank details are required to complete a booking, your online booking system can become a target for cyber criminals.

5. Your business needs to be ready to grow:

1. Choosing the wrong booking system for your company's needs

   With so many booking system options on the market, it may seem overwhelming to pick the best system for you. This window shopping period is essential though, as you don't want to sign up for a 12 month contract and pay for a system that's not serving you as well as another option could.

2. You will need (at least) one expert within the business

   No automation system is completely self-sufficient. You will always need at least one person in the business who can spend at least some of their working week managing the system.

3. Heavily reliant on a good internet connection

Some things are simply out of our control. Online booking system problems are sometimes unrelated to your system and how you've set it up. If your customers have a poor internet connection you may still need to take telephone or email bookings.

4. A target for cyber criminals:

The importance of a secure booking system has never been higher. As personal data and bank details are required to complete a booking, your online booking system can become a target for cyber criminals.

5. Your business needs to be ready to grow

Our research found that 69% of customers are more likely to book with a company that has an online booking system. If you don't have one, 59% of people will choose a competitor that does instead.Therefore, when your business begins to use an online booking system, you can expect to attract more customers and increase sales. Of course, this is great news! But it can be a problem if your business isn't prepared to keep up with this higher demand for your services.

1.2 Business Requirements

→To create a distributed system that will be used by customers.

→To ensure that ticket reservation is easy and user friendly for customers.

→To satisfy customers need by reaching them at their place.

→Reduce the work load of the ticket officers.

→Provide large number of ticket reservation and cancelation service in a few time

1.3 Literature Survey

Utilizing AI models, [2] connected PLSR(Partial Least Square Regression) model to acquire the greatest presentation to get the least cost of aircraft ticket buying, having 75.3% precision. Janssen [3] presented a direct quantile blended relapse model to anticipate air ticket costs for cheap tickets numerous prior days takeoff. Ren, Yuan, and Yang [4], contemplated the exhibition of Linear Regression (77.06% precision), Naive Bayes (73.06% exactness, Softmax Regression (76.84% precision) and SVM (80.6% exactness) models in anticipating air ticket costs. Papadakis [5] anticipated that the cost of the ticket drop later on, by accepting the issue as a grouping issue with the assistance of Ripple Down Rule Learner (74.5 % exactness.), Logistic Regression with 69.9% precision and Linear SVM with the (69.4% exactness).

1.4 Social Or Business Impact

Corporate flight booking includes the booking of flights tickets for business related activities. The employees travel for various events, client meetings, and other branding purposes regularly. Hence, airlines provide numerous options to book affordable flights for such activities known as corporate flight bookings.

1. Maximize savings: Employees and the company can earn reward points and frequent flier miles to improve the cost efficiency of the business travel program.

2. Simplify corporate air travel bookings: Airlines operating corporate flights offer a dedicated portal to corporate organizations for making and managing corporate bookings, adding corporate travelers, reviewing travel expenses, and redeeming reward points.

3. Larger baggage facilities: Corporate flyers can enjoy a free extra baggage allowance, more than the standard limit. They can board with a larger cabin bag to carry a laptop, and other office accessories during air travel.

4. Seat selection: Corporate flight booking benefits include choice of seats or rows. The feature is free of cost. You can book a seat near the exit, to get extra legroom paying no charges.

5. Faster boarding and exit: Business travelers can access corporate flight booking benefits like priority check-in and check-out. Some airlines may provide priority baggage checking at the bag carousel to begin their business assignments without long delays.

6. Lounge access: Corporate flyers may have lounge access equipped with modern facilities. They can use the lounge to catch up on work during the layover time.

CHAPTER 2

LITERATURE REVIEW

2.1 Collect The Dataset

1. **Airline:** So this column will have all the types of airlines like Indigo, Jet Airways, Air India, and many more.
2. **Date_of_Journey:** This column will let us know about the date on which the passenger's journey will start.
3. **Source:** This column holds the name of the place from where the passenger's journey will start.
4. **Destination:** This column holds the name of the place to where passengers wanted to travel.
5. **Route:** Here we can know about that what is the route through which passengers have opted to travel from his/her source to their destination.
6. **Arrival_Time:** Arrival time is when the passenger will reach his/her destination.
7. **Duration:** Duration is the whole period that a flight will take to complete its journey from source to destination.
8. **Total_Stops:** This will let us know in how many places flights will stop there for the flight in the whole journey.
9. **Additional_Info:** In this column, we will get information about food, kind of food, and other amenities.
10. Price of the flight for a complete journey including all the expenses before onboarding.

Importing the libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error as mse
```

```
from sklearn.metrics import r2_score
from math import sqrt
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import KFold
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from prettytable import PrettyTable
```

Read The Dataset

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.

In pandas we have a function called read_csv() to read the dataset. As a parameter we have to give the directory of the csv file.

```
data=pd.read_csv("Data_Train.csv")
```

```
data.head()
```

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

2.2 Data Preparation

The download data set is not suitable for training the machine learning model as it might have so much of randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

- Handling missing values
- Handling categorical data
- Handling outliers
- Scaling Techniques
- Splitting dataset into training and test set

- **Duplicates**

  There were not many, but a few repetitions in the data collected.

- **Days to departure**

  Our objective is to optimize this parameter. This the difference is the departure date and the day of booking the ticket. We consider this parameter to be within 45 days.

- **Day of departure**

  Intuitively we can say that flights scheduled during weekends will have a higher price compared to the flights on Wednesday or Thursday. Since including this in any of the models we use can be beneficial.

  Duration

  Converting the duration of the flight into numeric values, so that the model can interpret it properly. Also, it will be fair enough to omit flights with a very long duration.

- **Time of departure**

  Similar to day of departure, the time also seem to play an important factor. Hence we divided all the flights into three categories: Morning (6am to noon), Evening (noon to 9pm) and Night (9pm to 6am)

- **Hoppings**

  The data we collected did not give very authentic information about the number of hops a journey takes. Hence, we calculated the hops using the flight ids.

- **Outliers**

  We are focusing on minimizing the flight prices, hence we considered only the economy class with the following conditions:
  a) The minimum value of total fare for all days for a particular flight id is less .

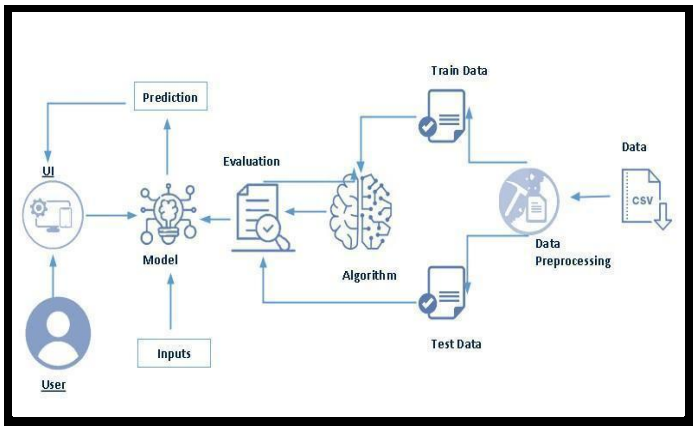
CHAPTER 3

PROPOSED METHOD

3.1 Descriptive Statistical Analysis

People who work frequently travel through flight will have better knowledge on best discount and right time to buy the ticket. For the business purpose many airline companies change prices according to the seasons or time duration. They will increase the price when people travel more. Estimating the highest prices of the airlines data for the route is collected with features such as Duration, Source, Destination, Arrival and Departure. Features are taken from chosen dataset and in the price wherein the airline price ticket costs vary overtime. we have implemented flight price prediction for users by using KNN, decision tree and random forest algorithms. Random Forest shows the best accuracy of 80% for predicting the flight price. also, we have done correlation tests and metrics for the statistical analysis.

Technical Architecture:

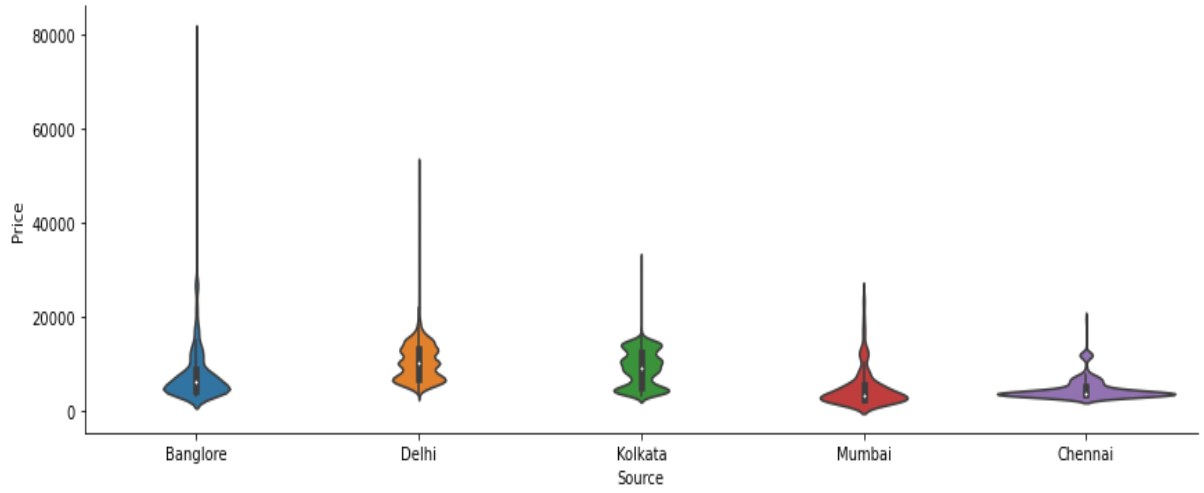


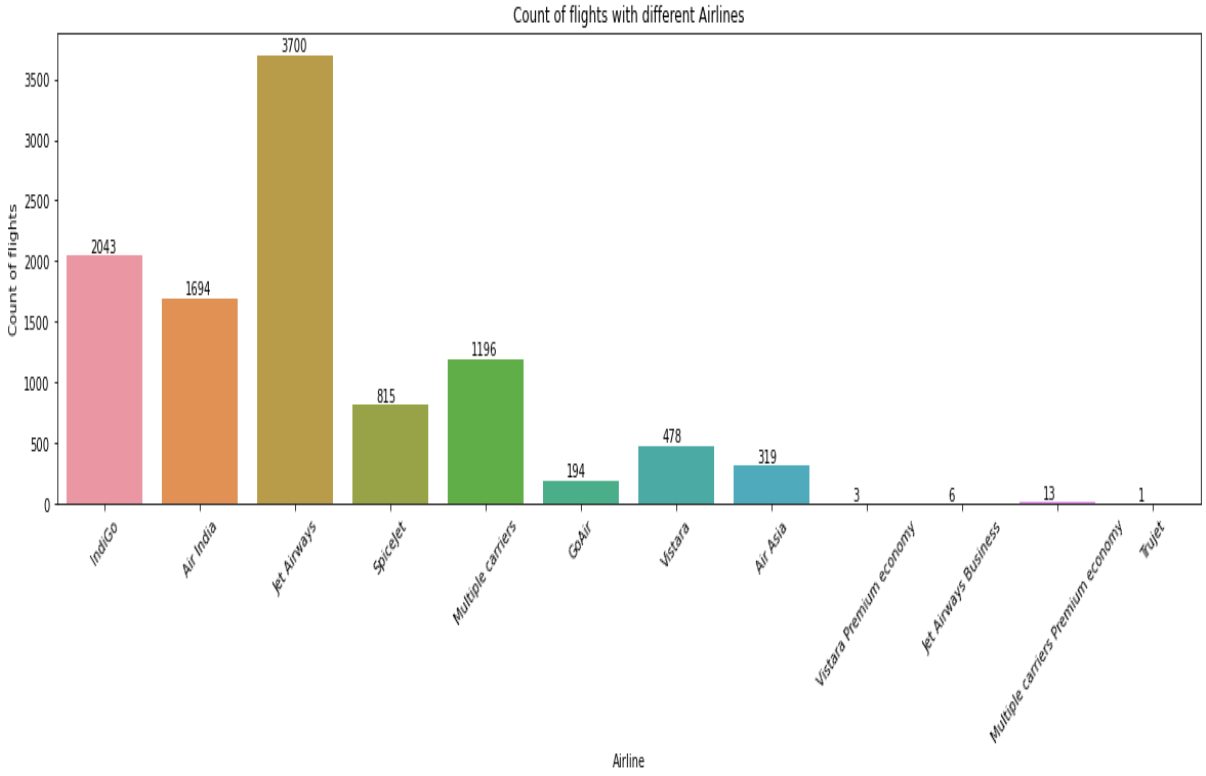
3.2 Visual Analysis

Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

We now plot distribution plots to check the distribution in numerical data (Distribution of 'Price' Column)

- The seaborn.displot() function is used to plot the displot. The displot represents the univariate distribution of data variable as an argument and returns the plot with the density distribution. Here, I used distribution(displot) on 'Price' column.
- It estimates the probability of distribution of continous variable across various data.





3.3 Training The Model In Multiple Algorithms

Checking Cross Validation for RandomForestRegressor

We perform the cross validation of our model to check if the model has any overfitting issue, by checking the ability of the model to make predictions on new data, using k-folds. We test the cross validation for Random forest and Gradient Boosting Regressor.

```
from sklearn.model_selection import cross_val_score
for i in range(2,5):
    cv=cross_val_score(rfr,x,y,cv=i)
    print(rfr,cv.mean())
```

```
RandomForestRegressor() 0.7916634416866438
RandomForestRegressor() 0.7929369032321089
RandomForestRegressor() 0.799914397784633
```

Regression Model

K Neighbors Regressor, SVR, Decision Tree Regressor

A function named KNN, SVR, DecisionTree is created and train and test data are passed as the parameters. Inside the function, KNN, SVR, DecisionTree algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with .predict() function and saved in new variable. For evaluating the model, r2_score, mean_absolute_error, and mean_squared_error is done.

Using Ensemble Techniques

Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor

A function named Random Forest, Gradient Boosting, AdaBoost is created and train and test data are passed as the parameters. Inside the function, Random Forest, Gradient Boosting, AdaBoost algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with.predict() function and saved in new variable. For evaluating the model, r2_score, mean_absolute_error, and mean_squared_error report is done.

3.4 Testing The Model

[tool.poetry]

name = "python-template"

version = "0.1.0"

description = ""

authors = ["Your Name <you@example.com>"]

[tool.poetry.dependencies]

python = ">=3.8.0,<3.9"

numpy = "^1.22.2"

replit = "^3.2.4"

Flask = "^2.2.0"

pandas = "^1.4.3"

openpyxl = "^3.0.10"

[tool.poetry.dev-dependencies]

debugpy = "^1.6.2"

python-lsp-server = {extras = ["yapf", "rope", "pyflakes"], version = "^1.5.0"}

toml = "^0.10.2"

[build-system]

requires = ["poetry-core>=1.0.0"]

build-backend = "poetry.core.masonry.api"

CHAPTER 4

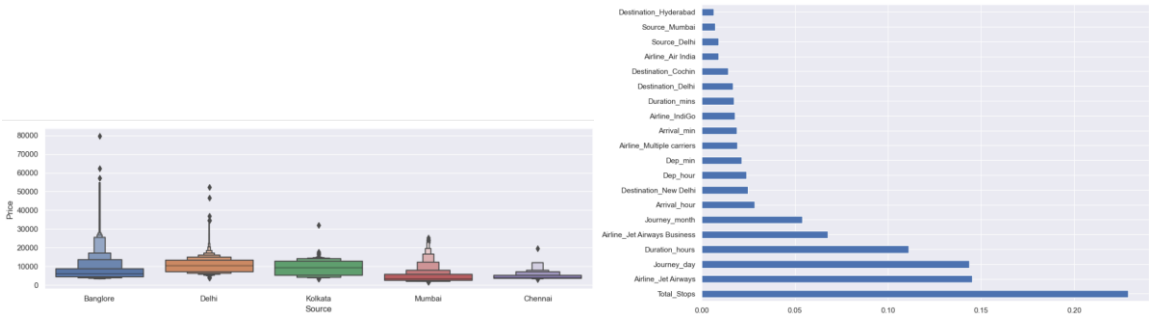
EXPERIMENTAL RESULTS

4.1 Testing Model With Multiple Evaluation Metrics

Hypertuning the model

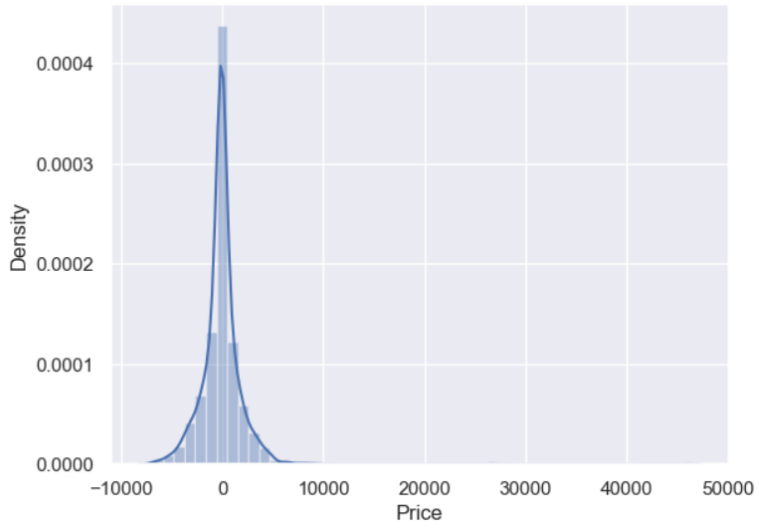
RandomSearch CV is a technique used to validate the model with different parameter combinations, by creating a random of parameters and trying all the combinations to compare which combination gave the best results. We apply random search on our model.

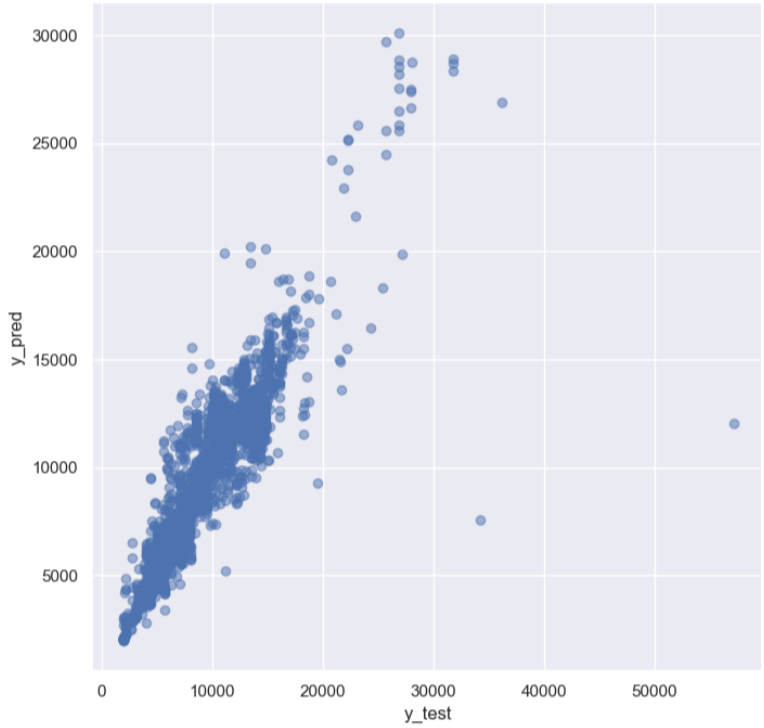
From sklearn, cross_val_score is used to evaluate the score of the model. On the parameters, we have given rf (model name), x, y, cv (as 3 folds). Our model is performing well.



4.2 Evaluate The Results

From sklearn, cross_val_score is used to evaluate the score of the model. On the parameters, we have given rfr (model name), x, y, cv (as 3 folds). Our model is performing well. So, we are saving the model by pickle.dump().





4.3 Saving The Model

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its weights and configuration. This can be useful in avoiding the need to retrain the model every time it is needed and also to be able to use it in the future.

```
import pickle
pickle.dump(rfr,open('model1.pk1','wb')
```

CHAPTER 5

APPLICATIONS/ADVANTAGES

1.Natural language processing :

Natural language processing (NLP) allows machine learning algorithms to process language-based inputs from humans, such as text-based messaging through an organisation's website. With NLP, these algorithms can detect the tone of a message and its topic to better understand what consumers want. An example is the chatbots that many organisations use to respond to consumer queries through their websites.

2. Recognising images:

Machine learning algorithms can learn to recognise images and then classify them into different categories. This means that they can recognise certain objects in an image and even recognise a face. In some cases, the algorithm might even be able to differentiate one person's face from another to identify people. This facial recognition ability is potentially useful for recognising people in photographs and videos, security measures and even product research.

3. Data mining :

Data mining refers to assessing data and finding patterns in it. This usually involves very large datasets containing raw data, which means data that hasn't undergone processing. This requires considerable processing power to allow the algorithm to identify trends in huge quantities of data, but it can help identify useful patterns. Data mining can identify public sentiments, identify spam emails, assess credit risk and detect fraud attempts.

4. Autonomous vehicles :

Machine learning can allow an autonomous vehicle to learn how to safely navigate in the real world. It allows them to identify real-world objects accurately and react to them accordingly, meaning they can avoid collisions or disruptions for other vehicles or pedestrians. The various sensors and cameras in an autonomous vehicle can provide information to the computer, using machine learning algorithms to process this information and make navigational decisions. Some key examples of this technology are self-driving cars and autonomous drones.

5. Better advertising and marketing :

Machine learning algorithms can predict which consumers are the most likely to actually buy a product. This is the process of customer segmentation, and having reliable information on buyer behaviour can make marketing and advertising campaigns much more effective. For instance, an algorithm might process large amounts of consumer data to determine which individuals are the most likely to make a purchase if they receive a prompt in the form of advertising. This allows the company to send advertising specifically to those who are the most likely to respond favourably to it and make a purchase.

6. Better products:

Companies depend on feedback from consumers and reviewers to assess their products. Sales figures can indicate how popular the product is, but other variables like competitor products and marketing can also impact sales. Knowing how to improve a product is key for many businesses, and more information can lead to better decisions. Machine learning algorithms can handle large amounts of data using the same customer segmentation processes for improving marketing.

7. Speech recognition

Speech recognition is similar to natural language processing but focuses solely on verbal communication from humans. Machine learning can help speech recognition applications to better interpret voice-based inputs from consumers and others. One iteration of this is in virtual assistants in smartphones that can understand requests and other voice-based inputs from users and complete tasks based on these inputs. This can also be useful for dictation software, allowing people to take notes without typing or writing. Voice chat applications can also benefit from this.

8.Detecting Fraud:

Fraud detection is an important task for many organisations, especially businesses like banks that issue credit cards. Machine learning algorithms can analyse behaviour and spending patterns to identify potential instances of fraud, such as credit card theft and insurance fraud. The same analytical processes and pattern detection can also be useful for identifying scam messages and other security concerns.

9. More accurate predictions:

Making accurate predictions and forecasts is a key concern for many businesses and policymakers. These can be predictions about the stock market, the economy and

consumer preferences. Using historical data, machine learning algorithms can learn to identify trends and patterns to evaluate possible outcomes. Using this as its reference framework, the algorithm can repeat the process with current data to make predictions about the future. Its ability to learn and process new data as it arrives means it can learn from mistakes and improve its accuracy over time.

10.Medical diagnoses

In the health care industry, machine learning can be useful for identifying patients who are at risk of certain conditions. Based on anonymous patient data from health care system records, machine learning algorithms can analyse patterns and combinations of lifestyle factors, histories and symptoms to assess how likely someone is to be at risk of a particular condition. A key benefit of this is that it can potentially save time and alert medical professionals to at-risk patients sooner, possibly reducing the severity of the intervention necessary to treat the individual. Related: Industry uses of artificial intelligence (plus benefits) Frequently asked questions about machine learning

CHAPTER 6

CONCLUSIONS & FUTURE SCOPE

6.1 CONCLUSION

In conclusion, the project concentrates on the development of a system for student performance analysis. After conducting a comprehensive student performance analysis,the student can know that how to predict the flight ticket cost using machine learning. The data and observations collected highlight several key points that Machine learning is very helpful for predicting values by using past experience e got from task t. A data-driven modeling framework was proposed in this work to estimate airline flight passes prices, based on the flight pass options selected by the costumer. The modeling framework comprises a random forest regression algorithm which outperforms the multiple linear regression analysis currently used by airlines to compute the prices of these passes. Therefore, it can be concluded that airlines would benefit from using data-driven techniques, obtaining more accurate and route dependent estimations for the flight passes. Additionally, the algorithm also gives insights into the importance of the features in predicting the flight prices. This is a relevant information for the airline when defining (standard) single flights and flight passes pricing strategy. For the case study considered, the ticketing lead time was found to be the most relevant feature.

6.2 FUTURE SCOPE

This was the first study done on the topic of airline flight passes. Future steps are necessary to consolidate this research. The most important one would be the validation of the results obtained with data-driven approaches. Since there is no flight pass booking data was available yet, no valid validation could be done. Furthermore, it would be interesting to consider training data from multiple years in order to better capture seasonality effect in the computation of the flight passes prices.