PATTERN RECOGNITION AND MACHINE LEARNING

BONUS COURSE PROJECT : Flight Ticket Price Prediction

ABSTRACT :

This paper reports my experience with predicting Flight Ticket prices. I will be analyzing the flight fare prediction using Machine Learning dataset using essential exploratory data analysis techniques then will draw some predictions about the price of the flight based on some features such as what type of airline it is, what is the arrival time, what is the departure time, what is the duration of the flight, source, destination and more.

INTRODUCTION :

Flight ticket prices can be something hard to guess, today we might see a price, check out the prices of the same flight tomorrow, it will be a different story. This project predicts the price of flight tickets.

DATASET :

The dataset is used as the training dataset. The dataset has been split into train and test with test size of 0.3. Columns :-

1. Airline: Have all the types of airlines like Indigo, Jet Airways, Air India, and many more.
2. Date_of_Journey: Lets us know about the date on which the passenger's journey will start.
3. Source: Holds the name of the place from where the passenger's journey will start.
4. Destination: Holds the name of the place to where passengers wanted to travel.
5. Route: About 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: Whole period that a flight will take to complete its journey from source to destination.
8. Total_Stops: How many places flights will stop there for the flight in the whole journey.
9. Additional_Info: Information about food, kind of food, and other amenities.
10. Price: Price of the flight for a complete journey.

METHODOLOGY :

CONTENTS OF COLAB NOTEBOOK

- Importing libraries
- Importing dataset

- Exploratory Data Analysis
- Target Distribution
- Data Preprocessing
- Categorical Data Distribution
- Implementing different models
- Comparing different models
- Building Final model

OVERVIEW

There are various regression algorithms present out of which I shall implement the following

- Logistic Regression
- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Naive Bayes
- K Neighbors Regressor

Exploring the dataset and pre-processing

- From the basic data exploration I inferred that the two features Route and Total_Stops have 1 missing value.
- I have shown **target distribution**.
- In **categorical data distribution** I have made bar plots describing the relation of a weekday with the price and a month with the price.
- The “Date_of_Journey”, “Route”, “Dep_Time”, “ArrivaL_Time” attributes are dropped from the dataset as they dont hold seem to hold much importance in prediction.
- I made a new dataframe with the available features.
- For **numerical data distribution**, I have shown a **heatmap correlation** of columns.
- For **data preprocessing,**
  - I have dropped the missing instance of “Total_Stops” and modified that column to integer values.
  - Label encoding of “Airline”, “Source”, “Destination” and “Additional_Info” columns.
  - For the train test split , we have 30% testing data and 70% training data.
  - For feature scaling, we have used a normalizer.

Implementation of classification algorithms

- *Logistic Regression* : It is a predictive analysis algorithm and based on the concept of probability.

- *Linear Regression* : The supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

● *Decision Tree Regressor* : Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

● *Random Forest Regressor :* Random forest operates by constructing a multitude of decision trees at training time and outputting the class that's the mode of the classes (classification) or mean prediction (regression) of the individual trees.

● *Naive Bayes* : Naive Bayes classifiers are a collection of algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Three types of Naive Bayes were used :
■ Gaussian NB
■ Bernoulli NB
■ Multinomial NB

● *KNN (k - nearest neighbors)* : KNN are supervised algorithms which classify on the basis of distance from similar points. Here k is the number of nearest neighbors to be considered in the majority voting process.

Ensemble model (max_features=5, n_estimators=200) and cross validation used to build the final model.

EVALUATION OF MODELS :

COMPARING MODELS

MODELS		R2 SCORE	MAE	MAPE
Logistic Regression		0.070208	3450.917174	0.517768
Linear Regression		0.408899	2504.714885	0.309570
Decision Tree Regressor		0.606642	1456.873513	0.164159
Random Forest Regressor		0.736838	1269.358818	0.148968
Naive Bayes	**GaussianNB**	0.452230	2029.251755	0.221446
	BernoulliNB	0.305129	2607.831072	0.306433

	MultinomialNB	-0.60635	3768.561535	0.623403
K Neighbors Regressor		0.704360	1454.286008	0.167963

RESULTS AND ANALYSIS :

For the final model we have used Random Forest regression as it has highest accuracy (73.683805 %) alongwith n_estimators = 200 and max_features = 5.

MODEL	**RMSE**	**MAE**	**MAPE**
Random Forest	6109.022923726992	5060.498549953204	0.92638373664840

CONCLUSION :

So I have done a complete EDA process, getting data insights and data visualization so after all these steps I can go for the prediction using machine learning model-making steps. By comparing all the models (Logistic regression, Linear regression, Decision Tree regressor, Random Forest regressor, Naive Bayes and K Neighbors regressor), I can conclude that Random Forest Regressor performs the best.