PATTERN RECOGNITION AND MACHINE LEARNING

BONUS PROJECT : Predicting flight ticket price

Name: Niharika Manhar
Roll no.: B20CS038

ABSTRACT

The paper reports experience while predicting flight ticket price. We had a dataset with some features that might affect flight ticket prices. I implemented different regression models like decision tree, random forest, knn, etc. and compared the results in this report.

INTRODUCTION

Each flight has a different price pattern. When we buy a plane ticket, the higher probability is to get a cheap ticket, but this is not always true. It changes everyday and becomes difficult for customers to buy the tickets at the right time. Our flight price predictor tracks region, route and timing fluctuations of a flight and recommend the best moment to book

METHODOLOGY

The dataset has 10 input parameter columns which shows: Name of airline, source, destination, number of stops, arrival and departure timings, etc.

Content of the collab notebook:
- Importing libraries
- Importing dataset
- Data preprocessing
- Implementing models
- Conclusion/ result

I implemented 4 regression models:
- Linear regression
- Decision tree regression
- Random forest regression
- K nearest neighbor regression

In data preprocessing:
- I converted the "date of journey" to day and month to use it for predictions.
- Then I converted departure, arrival time and duration to correct format and recorded it in the dataset.
- Then performed categorical encoding of the required columns.
- I have also shown heatmap correlation which shows relation between various columns.

Implementation of regression models:
- **Linear regression**: the model is used for linear data.
- **Decision tree regression**: it is one of the most used regression models, and performs well in almost all cases.
  Two types of decision tree regression are performed:
  - ❖ With default parameters
  - ❖ With max depth that gives the best score.
- **Random forest regression**: it is a boosting ensemble method to train various decision trees and give the best result..
  Two types of random forest regression are performed
  - ❖ With default parameters.
  - ❖ With max depth that gives the best score.
- **K nearest neighbor regression**: KNN are supervised algorithms which classify on the basis of distance from similar points.

EVALUATION OF MODELS

Models		MSE	Testing score	R2 score
Linear regression		10631348.853	0.46243	0.46243
Decision tree regression	default	4511372.351	0.77188	0.77188
	max_depth = 17	4078902.814	0.79375	0.79375
Random forest regression	default	2589326.088	0.86907	0.86907
	max_depth = 16	2457370.794	0.87574	0.87574
KNN		6970696.811	0.64753	0.64753

RESULT

The best prediction is given by random forest regression with the score of **0.87574 .**

REFERENCES

[1] For implementing models | sklearn.org

[2] Pattern Classification -Book by David G. Stork, Peter E. Hart, and Richard O. Duda

[3] Link for dataset: Provided