Vivek Ponnala
Maggie Qin
Hoai An Nguyen
Team Name: Data Mining Developers

Predicting Flight Prices: An Approach Using Random Forest

Project Description: The goal of this project is to create a data mining model that classifies flight prices based on various features like flight routes, times, and more. We are going to use classification techniques, and possibly dimensionality reduction for pre-processing, depending on the number of dimensionality we selected.

Proposed Methodology:
We are going to preprocess the data and filter according to the data with the necessary columns: search date, end destination, non stop, segmentCabinCode, remainingSeats and totalSeatsTravelled. Then we are going to split the data into 80% train, 20% split. Then we are going to apply the Random forest Classifier on the train data. Then we will evaluate how our model does with the test data with the following metrics: accuracy, precision, recall, and F1-score. Lastly, we will plot the data with either seaborn or matplotlib.

Dataset Structure and Relevant Features:

Main Dataset link: https://www.kaggle.com/datasets/dilwong/flightprices
Dataset Subset link: dataMay.zip

For our project, we'll be working with the Flight Prices dataset from Kaggle, which provides various features that are important for predicting flight prices. This data set provides flight ticket prices found on Expedia from April 2022 to October 2022. There are 5999739 unique values and 27 columns. Due to the high complexity of this dataset, our group decided to extract a subset to conduct our research. We will also not consider the values of the features that don't have effect on flight prices (isBasicEconomy segmentsEquipmentDescription, segmentsCabinCode, etc), duplicate features that are represented in different format (segmentsDepartureTimeEpochSeconds, segmentsArrivalTimeEpochSeconds, segmentsDepartureTimeEpochSeconds, segmentsArrivalTimeEpochSeconds, segmentsAirlineCode and segmentsAirlineName).

Choosing the Subset of Data:

For the scope of this project we will be focusing on flights departing from San Francisco International Airport (SFO), we have narrowed down the dataset to include only non-stop flights.

Time Period:

We plan to use data from May 2022 as our training data, which is 252594 data entry that gives us a clear and manageable dataset. We will split data into training and testing.

Features to Focus On:

We'll likely use features such as search date, end destination, non stop, segmentCabinCode, remainingSeats and totalSeatsTravelled to classify flight prices based on low, medium and high. Additionally, we should consider variables like search date and flight date (weekday vs. weekend), as they can significantly influence flight prices.

Classification Techniques:

To classify flights into price ranges (e.g., low, medium, high), we can create categories based on the price distribution within the dataset. We'll then apply Random Forests to classify flights into these ranges.

Evaluation Metrics:

We'll measure performance using metrics such as accuracy, precision, recall, and F1-score to evaluate how well the models classify flights into price ranges.

Resources:

In our case, Python will be mostly used for developing our flight price prediction project because of the availability of numerous libraries. Some of the most important are pandas for data loading, data handling and data cleaning as well as numerical analysis with NumPy. For the classification model we will be using scikit-learn for random forest. The libraries for the data visualization and exploration are Matplotlib and Seaborn.