Cairo University - Faculty of Engineering

Computer Engineering Department

Big Data

Flight Price Prediction

Team 6

Name	Section	B.N.
Abdallah Hussien	2	3
David Raafat	1	24
François Adham	2	10
Kareem Mohamed	2	11

Submitted to:

Eng./ **Omar Samir**

Academic Year

2021 / 2022

1- Problem Description

Analyse the flight booking dataset obtained from the “Ease My Trip” website to predict the flight prices based on some features.

2- Dataset

Flight Price Prediction:

https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

3- Project Pipeline

1. Analyse and visualise the features
2. See the correlation between features and flight prices
3. Try to find out and eliminate the features that do not affect the flight prices
4. We will implement Multivariate Linear Regression using MapReduce
    a. Split the dataset to m splits
    b. Map phase:
        i. Each mapper will decompose the feature vector using QR decomposition i.e. $X_j = Q_j^{(1)} R_j^{(1)}$ where j is the block number.
        ii. Map phase output will be a key-value pair where key will be NONE and the value will be (Q, R)

c. Reduce phase:

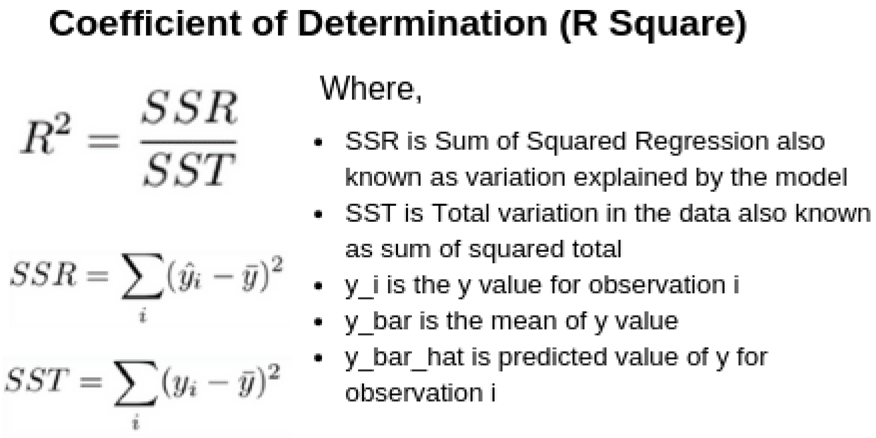
   i. Decompose each R matrix using QR decomposition i.e. $R_j^{(1)} = Q_j^{(2)} R_j^{(2)}$

   ii. Compute $Q^{(3)}$ where $Q_j^{(3)} = Q_j^{(2)} Q_j^{(1)}$

   iii. Compute V_j where $V_j = Q_j^{(3)} y_j$ where y_j is a vector of the predicted values

   iv. Compute the weights for each feature $\beta = [\, R^{(2)} \,]^{-1} \sum_{j=1}^{m} V_j$

5. Test our model performance using Coefficient of Determination.



Coefficient of Determination (R Square)

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

Where,

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- y_bar is the mean of y value
- y_bar_hat is predicted value of y for observation i

4- Analysis and solution of the problem:

- **Preprocessing**

1. First we show data shape and the features in it.
2. Check if data has any NAN values, but fortunately we hadn't.
3. Then we see how many features are categorical in our dataset.
4. Check how many unique values in these categorical features.
5. After seeing that we removed the "flight" feature from the data.

- **Visualisation**
    1. First we show a histogram of our target ‘price’.
    2. Show a histogram for each categorical feature and see how it affects the price.
    3. For numerical features we show the correlation between it and the price to see how it affects the price.

- **Extracting insights**
    1. After seeing ‘class’ histogram and its effect on price we can see that it’s unbalanced and business class has more price than economy.
    2. From the price histogram we can see the more the price is the less the count of flight tickets be.
    3. From the correlation between ‘days_left’ and price we can see that when there are a few days for the flights the prices increase.
    4. From the histogram of ‘stops’ we can see that when the number of stops increases the price increases.
    5. After visualising all categorical features we can see that each categorical feature has at least 2 features which can affect the price significantly.
    6. For (source_city, destination_city) features we found that they don’t affect the price significantly, so we extract new features from them by combining each unique value pair together which showed that they can have effect on the price.
    7. Finally we convert all categorical variables into numerical ones.

- **Model/Classifier training**
    1. First we split the independent variables together and the dependent variable alone.
    2. Then we split our data into 80% training data and 20% testing data.
    3. Our problem is to predict a continuous variable (Price), so we used a linear regression model.
    4. After training the model, we test it by the 20% testing data.

5- Results and Evaluation

After building the model we have the following results:

1. 90.97 % on train data.
2. 91.09% on test data.

6- Unsuccessful trials

1. After Seeing that the source_city and destination_city don't affect the price significantly we tried to remove them from the data, but we found that the accuracy decreased to 84%. That's Why we construct the new unique pair features which show us that they can affect the price.
2. After seeing the correlation of 'duration' feature we thought that it's not very important but we found that it affects the price.

7- Future Work

1. We can work with the fully-distributed mode of hadoop.
2. We can find more data entries in the dataset.
3. We can add more important features to the dataset which can affect the price.