Airlines Regression

Team ID: 14_SC

Team members:

- Abdelrhman Mohamed Abdelsalam
- Omar Khaled Ahmed Abdullah
- Ibrahim Youssef Mustafa
- Rehab Hosam Ahmed Mokhtar
- Maivel Maher Isaac

Project points:

- **Preprocessing:**

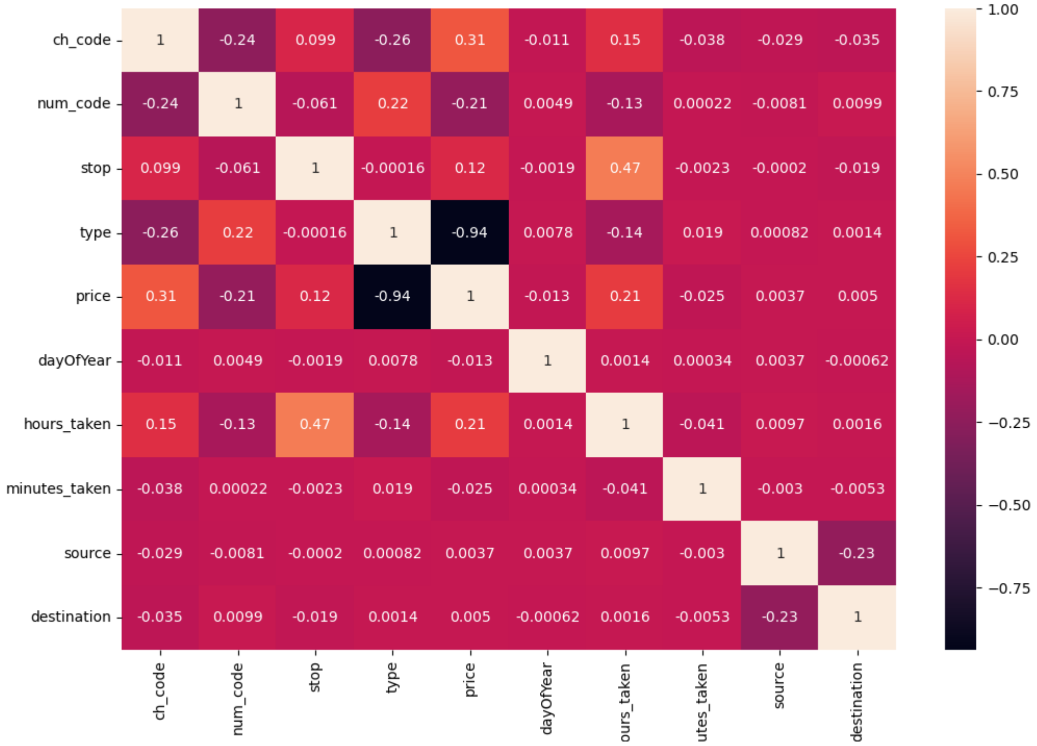
    1. Price: Remove special characters “,” and convert it to Number
    2. Date: convert all the records to the standard format (MM/DD/YYYY), then use dayofyear to represent date.
    3. Ch-code: Encode the column into numeric values.
    4. Time-taken: represent time in total minutes format
    5. Stop: apply string split to get the number of stops and convert non and null valued rows to zero and 2+ to 2.
    6. Type: encoding business and economy into zero and one.
    7. Route: separate it into source and destination columns and apply label encode.

- **Dropped Columns:**

    1. Airline: was dropped due to the being represented in the Ch-code.
    2. Date: was replaced by the dayofyear column.
    3. Dep-time & Time-Taken & arr-time: all were replaced by the Time-Taken-hour and Time-Taken-minute columns.
    4. Route: was replaced by the source and destination columns.

- **Visualization:**

Apply correlation on the Data with the prediction output Y{price} and visualize the correlation output using heatmap.



- **Model training:**

  - **Decision Tree (Test 20%, Train 80%):**
    1. Mean Square Error for testing set 54071278.83793249
    2. Accuracy 0.895812030929177
    3. Training time: 0.14661550521850586s

  - **Polynomial Regression (Test 20%, Train 80%) (degree = 3):**
    1. Mean Square Error for testing set 30782218.52909901
    2. Accuracy 0.940686869240622
    3. Training time: 2.5894250869750977s

  - **XGB Regression (Test 20%, Train 80%) (degree = 3):**

1. Mean Square Error for testing set 929021621.1917028
2. Accuracy -0.7900977749197247
3. Training time: 2.5695159435272217s

- **Polynomial Elastic Net Model (Test 20%, Train 80%) (degree = 3):**

1. Mean Square Error for testing set 33705915.07649131
2. Accuracy 0.9350533053227942
3. Training time: 33.12820339202881s

- **Polynomial Ridge Model (Test 20%, Train 80%):**

1. Mean Square Error for testing set 30783229.181234427
2. Accuracy 0.9406849218519955
3. Training time: 33.12820339202881s

- Conclusion :

1. Label encoder is better than Dummy, One Hot encoder in the used dataset.
2. Polynomial gives a better prediction to time ratio than Linear ,D-tree, Ridge, Elastic.
3. Polynomial degree 3 with more features acts nearly Best.

What did we do new?

- Data Scaling has been applied on Training dataset to improve the models accuracy

- To cover Unknown values we replaced (labelEncoder) to (OrdinalEncoder) by detecting it using (handle_unknown='use_encoded_value') parameter and replace it by a specific value

- ElasticNet Model (80 % Training, 20 % Testing) :-

Mean Square Error for testing set	101070760.64384402
Accuracy	0.8036390894572659

- Lasso Model (80 % Training, 20 % Testing) :-

Mean Square Error for testing set	51406367.09341645
Accuracy	0.9001273861416011