Machine learning Project – Milestone 1 report

**1st: Preprocessing techniques:**

- Reading data:

* And specifying the target labels.

Text

Description automatically generated

A screen shot of a computer

Description automatically generated with low confidence- Previewing data’s head:

- Previewing data info:

* Noticed from below that there are non-null elements.
* Price is object, which indicates there is something wrong with target.
* Some columns like date, dep\_time, time\_taken, and arr\_time they are all relevant to the same thing which is time, and need to be adjusted.

Text

Description automatically generated

- For date:

* The date is split into 3 different sections: day, month, and year. All took the format int8 except year took int16 to save memory. Then dropped date.

Text

Description automatically generated

- For stop:

* It had a problem with its format is previewed above in the data.head()



- For route:

* It was in dictionary format, so the keys are columns, and the values are the elements of columns.

A screenshot of a computer

Description automatically generated with medium confidence

- For dep\_time:

* It’s split into hours and minutes then dropped.



- For arr\_time:

* It’s split into hours and minutes then dropped.



- For time\_taken:

* It needed to be adjusted since it had the “h” and “m”, then we converted the hours to minutes to be added in the same column.



- For rows:

* It was found in the data that there is wrong format in hours like 1.03 and 1.01, so it was dropped.

Text

Description automatically generated

- Previewing the data after adjusting:

A picture containing text, monitor, screen, black

Description automatically generated

Text

Description automatically generated- Previewing info:

- Find the data statistics and sensibility:

* As it’s found below, all the columns are in sensible and correct range. No furthermore adjustments.

Graphical user interface, text

Description automatically generated

- Knowing number of unique elements in each column:

Table

Description automatically generated with medium confidence

- Finally dropping the target “price”.

- For preprocessing:

* train\_test\_split with ratio=0.2 for test data
* For each categorical column, each of them had <10 unique elements, so LabelEncoder was used on them.
* Used MinMaxScaler to make the values vary in the range 0 to 1.

A screenshot of a computer

Description automatically generated

- Dropping columns:

* Since year had one value, it won’t add anything to the model.
* For both dep\_time and arr\_time, the reason is because:

From this equation we transformed the time\_taken as illustrated above. So they won’t be so useful features.



- Previewing X\_train:

A black screen with white text

Description automatically generated with low confidence

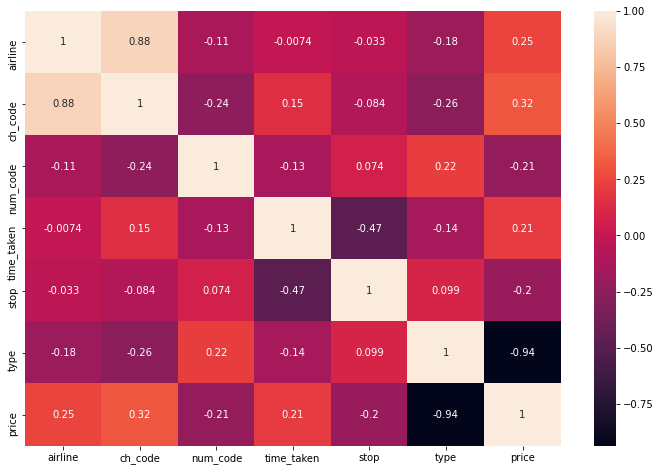
- Previewing X\_test:

Graphical user interface, text

Description automatically generated

**2nd: Analysis of dataset:**

* To analyze the dataset we used Data Correlation to understand the relationship between multiple variables and attributes in the provided dataset.
* As shown in figure below, there is a **high correlation** between **airline** and **ch\_code** features.
* While studying the relation between features that affect on price, we noticed that there is **high correlation** between **type** feature and **price.**
* There are several features that has **lower** **affect** on predicting the price such as **airline, ch\_code, num\_code, time\_taken** and **stop**.
* There are several features that has **very small affect** on predicting the price such as **date** and **route (source, destination)**.



**3rd : Regression techniques used:**

* We used **two** techniques which are **Polynomial Regression** and **Multivariable Regression**.

**4th : Differences between the used models:**

* In the following table there as a **comparison** between **Polynomial Regression** and **Multivariable Regression** in terms of:

1. Mean square error.
2. Accuracy.
3. Training time.
4. Prediction time.
5. Interception with y-axis.
6. Degree used.

|  |  |  |
| --- | --- | --- |
| **Point of comparison** | **Polynomial Regression** | **Multivariable Regression** |
| Mean square error | 34013736.500354946 | 52668893.69151366 |
| Accuracy | 93.4% | 89.8% |
| Training time | 1.4031610488891602 | 0.020010948181152344 |
| Prediction time | 0.46058130264282227 | 0.005001068115234375 |
| Interception with y-axis | -942992211009089.1 | 49032.947688995744 |
| Degree used | 4th degree | - |

**5th : The features used and discarded:**

* **Used:** (airline, ch\_code, num\_code, time\_taken, stop, type).
* **Discarded:** (date, route, arr\_time, dep\_time).

**6th : training, testing and validation sets:**

* **Training set:** 70%
* **Testing set:** 30%
* **Validation set:** not used.

**7th: conclusion:**

* After we see a dataset we notice that there's a wrong data between our data so we do preprocessing.
* we expected the feature of (route) will affect on our model but after we finish it we find it isn't have any effect, and we expect the feature of (type) will have big effect on the price of flight and after we finish the model we find that we are right and it have 0.94 correlation it nearby 1 (high correlation), we Doubt the feature of ( Time\_taken) will have a medium effect but after we finish we find that it have weak effect because correlation was low, we expect the feature of (Date) will have weak effect on the price of flight and after we finish the model we find that we are right.
* we expect the polynomial regression will be better than multiple regression and we find we are right in that, and we find if we increase the degree over 4 the MSE will be increase and it will be overfitting.
* We split our data to 20% testing set and 80% training set and we increase the testing set to 30% and we find the error less than before then we try to increase it again and we find the error increase again so we make sure testing set (20%) more suitable.