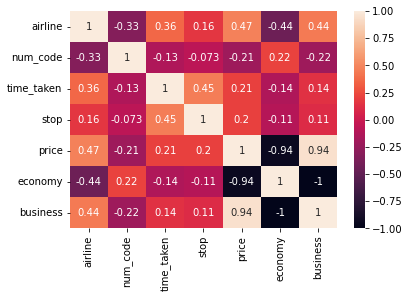
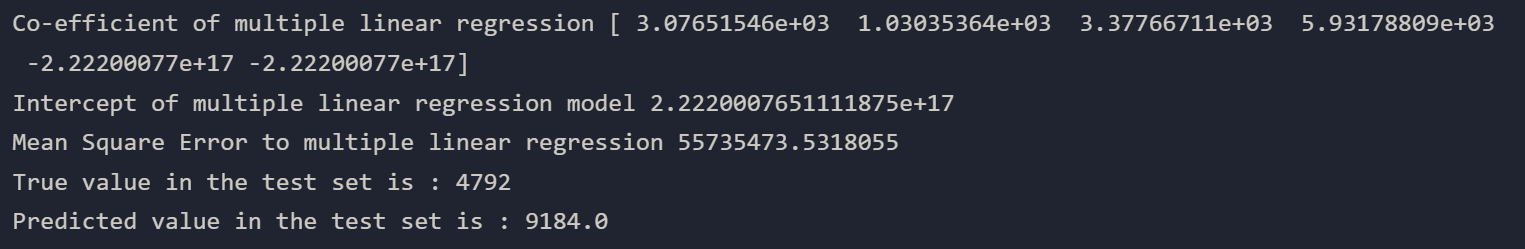
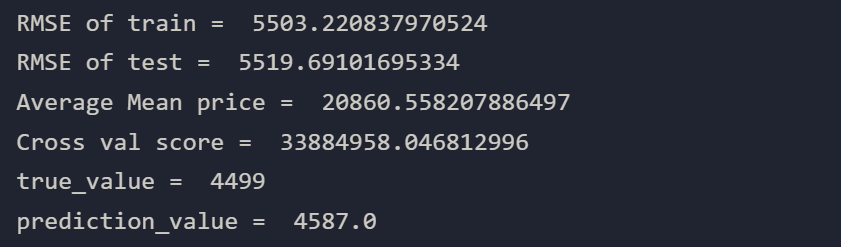
Airline ticket price prediction

|  |  |  |
| --- | --- | --- |
| Section | Id | Name |
| Sec 2 | 20191700169 | إياد عبده محمد اسماعيل |
| Sec 2 | 20191700168 | إيهاب محمد عبدالمنعم محمد |
| Sec 5 | 20191700127 | الفت ناصر على محمد |
| Sec 2 | 20191700146 | انس احمد محمد عبدالعزيز |
| Sec 2 | 20191700197 | جمال احمد عبدالحكم |

* **Preprocessing: -**
  + we applied different types of encoding for textual features such as label encoding, one-hot encoding, and average encoding.
  + For categorical features such as “ type ”, we used one-hot encoding so as to not give different categories different weights during the learning process.
  + For other textual features, we tried multiple methods and choose the one that gave the best correlation results.
  + Date and time data such as “ date ” and “ dep\_time ”, were converted into DateTime then to numerical data so that we can check for correlation and apply normalization.
  + “ Time taken “ was converted into minutes so that we can check for a correlation between the length of the trip with the price and apply normalization.
  + The “ stops ” column was analyzed to have a few values all contributing to one category under different names so they were encoded to be part of that category.
  + We converted the “ price ” column from string to numerical value.
  + We applied a min max scaler on the data to normalize the range of the data.
  + We dropped the “ ch\_code ” column as it represented the same thing as airline names.
* **Data analysis: -**
  + We dropped the “ ch\_code ” column as it represented the same thing as airline names.
  + We calculated the correlation between all the column and dropped the columns with correlation lower than | 0.2 | relative to the price.



* + We dropped the date, arrival time, departure time, and route as they have a very low correlation with the price ( lower than 0.2 ).
  + The figure above shows that type, stop, time taken, airline, and num\_code have the highest correlation with the price.
  + We figured that business trips are more expensive, also the longer the trip the more expensive it gets, the fewer stops a trip has the more expensive the trip gets, and some airlines are more expensive than the others.
* **Regression: -**
  + We used 2 regression models, multivariable linear regression, and polynomial regression using multiple degrees of polynomials.
  + In our testing the polynomial model fit our data a lot better than the linear model, but it took longer to train as we increase the complexity of the model
  + We tried many degrees for the polynomial model, and the best results we got were at degree 7, which took more than 3 minutes in the training, when we increase the degree further the model may over fit the data.
  + We used a regular 80 / 20 train test split, as the data was plenty and a 20% of it would suffice as a test sample.
* Linear model
* Polynomial model
* **Conclusion: -**

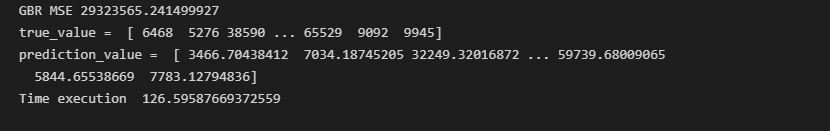
We can conclude that given the length of the trip in minutes, its type, number of stops, the airline name, and the flight code we can safely predict the price of the trip.

Some data that will not aid in the prediction process are date, arrival and departure time, and the source and destination of the flight as they have little influence on the price of the ticket.

Using a polynomial model yielded better results with our dataset.

* **Updates**
  + some updates on prepressing to handle null values
  + trying two new algorithms:

1. **Gradient Boosting Regression**



1. **Extreme Gradient Boosting Regression**

