Flight Price Prediction

Daniel Escobar, Juan Esteban Garcia

I. THE MACHINE LEARNING PROBLEM

A. Description of the problem

During these times the aviation industry has been growing despite the closures of borders due to the pandemic, this has led people to look for apps to get flights as cheap as possible. In this case, we have focused on a database with flights in the Republic of India, with which we intend to create a predictive regression model to estimate the value of a ticket as accurately as possible. It is important to clarify that the purpose is to estimate the price of the plane ticket, not its rise or fall.

B. Database

The database used in the project can be found at the following link [4]

The total number of samples is 10683.

C. Variables

In this database we can find the following variables:

- Airline (Categorical)
- Date
- Origin (Categorical)
- Destination (Categorical)
- Departure Time (It is not the take off time, this means the time when the airplane leaves the gate)
- Arrival Time
- Duration
- Total Stops (Categorical)
- Additional Information (Categorical)
- Price (Dependent Variable)

The model is going to use the first 9 variables to make a prediction for the last variable (Price).

C. Empty values

The dataset has 2 empty values, one in the route column and one in the total stop's column, since these empty values are minimal, we drop both rows.

D. Cleaning Data

- The variable Date is defined as an object, then we convert it into a Date value to make a proper prediction.
- The following process creates two new variables, Day of Travel and Month of Travel. (There is no need for year because the model is supposed to work for every year and only depend on the day and month of the year in course)
- The variables Departure Time and Arrival Time are defined as an object, then we convert them into a time value. After this conversion, we extract the values of hour and time for each variable, creating then Departure Hour, Departure Minute, Arrival Hour, Arrival Minute.
- The variable Duration is an object, so we extract the values from the object, creating then two new variables, Duration Hours and Duration Minutes, these ones are defined as integers.
- For this "Extraction" we create a function to get duration values (Hours and minutes)

D. Handling Categorical Data

The first thing we have to talk about is data encoding, what is data encoding? and how it could help to solve the problem?

Encoding is the process of converting the data (In this case categorical data) or a given sequence of characters, into a specified format, for the secured transmission of data. There are multiple ways to encode variables, for this project we are going to use one hot encoding and label encoder.
One hot encoding is used for categorical variables where there's no order in the data, for example colors, countries or in this case, airlines. This encoding process let us create new columns for each variable, were we are going to fill with 0 (This mean there is no presence of the variable in the sample) and 1 (This mean there's presence of the variable in the sample)

Label encoding is used for categorical variables where there is order in the data, in this case we have the number of layovers (nonstop , 1 stop, 2 stops, 3 stops), then after use the label encoder the values of this variables are going to change in the following way:

- Nonstop → 0
- 1 Stop → 1
- 2 Stops → 2
- 3 Stops → 3

Additional Information is also a categorical value, so we need to convert the values to numbers. At first we are replacing the values for its name, but this lets us know that No info value is repeated, for that reason, we have to replace these values.

After this process this variable is still categorical, so we can use the process that we used before to convert this categorical value into a numerical one.

E. Validation Methodology

The validation methodology used is cross-validation, with 80% of the data for training and the remaining 20% for testing.

F. Performance Metrics

We chose 2 performance metrics, the coefficient of determination (R2) and the root mean square error (RMSE), where we took the R2 as the main one, so this is a metric that tells us the performance of the models.

To choose the best 3 models, we must evaluate those that have a higher R2 whose range is between [0-1]. The best models will be those whose r2 is closest to 1. We also consider the RMSE, where the situation is the opposite of the above, i.e. the best models evaluated by RMSE will be those that have this metric closest to 0.

It is worth mentioning that RMSE measures the error taking into account the same unit of the independent variable, and given that this is between [0-79000], where 79000 is the largest value for a ticket in the database, for this reason the values obtained for RMSE in the models are "large" values, since we are talking about a fairly wide interval in the independent variable.

G. Related Articles

1. A Framework for Airfare Price Prediction: A Machine Learning Approach [1]
    - The first article talks about a very similar problem, but in this case, the data used has information about different countries and its different flight information. In order to solve the problem the technique applied was supervised learning.
    - ***What were the results of the article mentioned above?***
      Based on the different origins and destinations within the database a high prediction precision could be obtained with an R-squared score of 0.869 in the test data set.
2. Airline ticket price and demand prediction: A survey [2]
    - This article was more a guide of how the problem can be approached. showing the best models to solve the problem, the metrics that can be used and the kind of validation available. In that case, cross validation,leave-one-out and bootstrapping.
    - ***What were the results of the article mentioned above?***
      In most of the cases, the models obtained accurate prediction results of more than 80%, considering that there was a wide variety of regression models.
3. Flight Fare Prediction System [3]
    - The only article or blog that we found with the same database. It contains the decision tree, random forest and linear regression evaluated with the same database.
    - ***Which validation methodology did they use?***
      Cross validation
    - ***What were the results of the article mentioned above?***
      Of three models evaluated, the most efficient in score aspect, was the random forest with a r2_score of 0.85.

II. Model Building

We created a definition that will help us show all information needed in the evaluation of a model. For example, the RMSE, MAE or MSE score, finally we show a graphic using the seaborn library which shows how much percentage of predictive success was obtained.

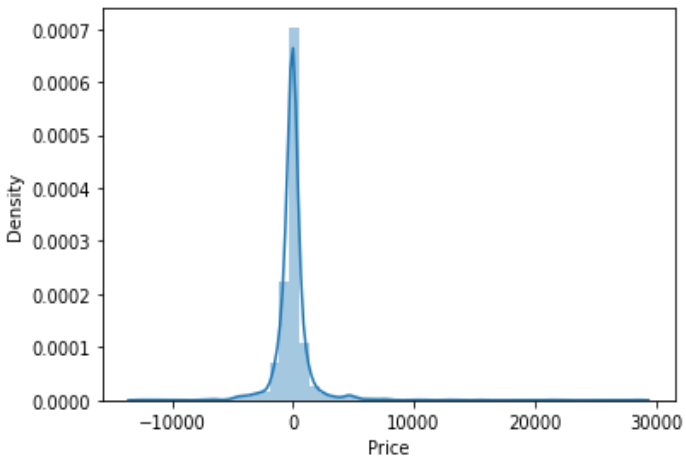
A. Random Forest Regression.

The first model that we are going to build is the RandomForestRegressor, for its building we have to do the next steps:

The first step is to separate the dataset in X and Y columns. Then we split the dataset for training and test dates (20% of total data).

Then, we import the model's library.

Finally, we use the defined function to evaluate the model and see the results.



R2 SCORE	0.974803
RMSE SCORE	768.900313

For this model we got

- r2 score is: 0.974802
- RMSE: 768.900313

	Actual	Predicted
10507	14781	14684.60
7705	5636	5950.25
7700	3597	3625.76
1437	16757	16203.39

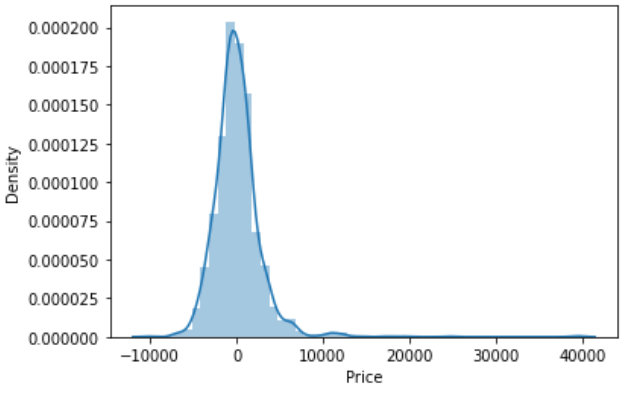
Some comparison between the real data and the predicted by the model.

B. Multiple Regression Model.

R2 SCORE	0.668549
RMSE SCORE	2788.706251

For this model we got

- r2 score is: 0.668549
- RMSE: 2788.706251



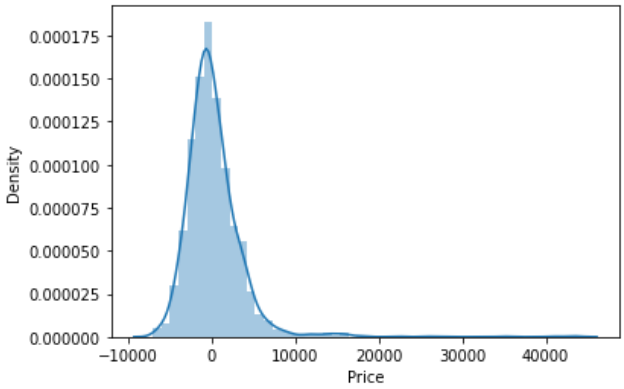
	Actual	Predicted
10507	14781	13152.270134
7705	5636	6934.399262
7700	3597	2603.550453
1437	16757	16028.946303

Hyper tuning is not possible on this model. There are some variants for this model but they were not requested by the report.

C. Neural Network

For this model we got

- r2 score is: 0.472930
- RMSE:3516.63518



Hyper Tuning the model

Hypertunning the model with GridSearchCV to find out the best parameters for the model.

- Fitting 5 folds for each of 10 candidates, totalling 50 fits
- Train R^2 Score : 0.868
- Test R^2 Score : 0.817
- Best R^2 Score Through Grid Search : 0.822
- BestParameters:{'activation':'relu','hidden_layer_sizes':(50,100),'learning_rate':'constant', 'max_iter': 4000, 'solver': 'adam'}

Then use the obtained parameters, and the model score is much better than the previous one.

For this model we got

- r2 score is: 0.8248479345239864
- RMSE:2027.2206486588368

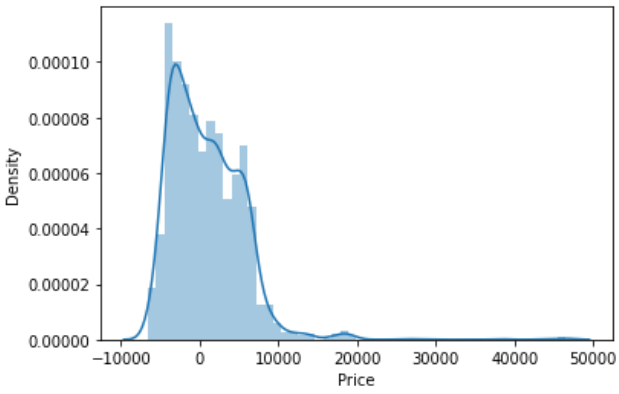
	Actual	Predicted
10507	14781	14706.873383
7705	5636	6520.033715
7700	3597	3345.829078
1437	16757	19355.378121

C. SVR model (Support Vector Regression with kernel rbf)

We reuse the GridSearchCV to use the best parameters in order to obtain the best result.

For this model we got

R2 SCORE	0.770595
RMSE SCORE	2320.035240



	Actual	Predicted
10507	14781	13807.891504
7705	5636	5636.099831
7700	3597	3807.274455
1437	16757	13807.891504

IX. Conclusion

In conclusion, after we analyzed the models with the data, the random forest model is the model with best performance according to the metrics that we chose at the very beginning of this document.

	RandomForest	Neural Network	SV
R2 SCORE	0.974	0.824	0.77
RMSE SCORE	768.900	2027.220	2320.03

Most of the articles that we looked at had a R square value close to 0.85, so if we compared 0.85 against the values we got, it is clear that they are close to each other, and actually our Random Forest model has a better score than the best models of these articles.

If you want to see the code please refer to [5]

References

[1] Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. (n.d.). A Framework for Airfare Price Prediction: A Machine Learning Approach. Retrieved September 7, 2021, from https://ieeexplore.ieee.org/abstract/document/8843464/references#references

[2] Abdella, J. A., Zaki, N., Shuaib, K., & Khan, F. (2019, February 05). Airline ticket price and demand prediction: A survey. Retrieved September 8, 2021, from https://www.sciencedirect.com/science/article/pii/S131915781830884X

[3] Kimbahune, V., Donga, H., Trivedi, A., Mahajan, S., & Mahajan, V. (2021, May 19). Flight Fare Prediction System. Retrieved September 8, 2021, from https://www.google.com/url?q=https://easychair.org/publications/preprint_download/htzQ&sa=D&source=editors&ust=1631089747137000&usg=AOvVaw0i_hCIbbSh8ZQYIU1nltOn

[4] Mittal, N. (2019, March 15). Flight Fare Prediction MH. Retrieved October 9, 2021, from https://www.kaggle.com/anshigupta01/flight-price-prediction/notebook

[5] Escobar, D., & Garcia, J. (n.d.). FlightPricePrediction. Retrieved October 11, 2021, from https://github.com/DanielEscobar01/FlightPricePrediction