Using of Machine Learning Techniques to predict the Flying prices.

Giovanni Magrone Computer Science-University of Bari
Department Computer Science, University of Bari, Bari, Italy

E-mail: g.magrone7@studenti.uniba.it

Abstract

The flying travel industry is a dynamic sector that attracts global attention due to the continuous growth in air travel. With the huge quantity of data available on flights and their characteristics, Machine Learning(ML) techniques can be applied to predict flight prices. In this project, we'll seek for a variety of machine learning models to forecast flight prices based on different features such as departure time, airline, route, and the type of route(economy, standard). The objective is to identify the model that provides the most accurate price predictions. We evaluate the performance of several models, including DecisionTree Regressor, RandomForest Regressor, Extratree Regressor, Linear Regression(Ridge), and a Fully Connected Neural Networks. Our analysis reveals that the RandomForest Regressor model achieves the highest prediction accuracy, demonstrating its effectiveness in forecasting flight prices.

Keywords: Machine learning, Predictive modelling, Flying prices analytics.

1. Introduction

The airline industry is a global sector that facilitates air travel for passengers. It's made up by multiple airlines flights to various destinations. In this project, our objective is to predict flight prices using machine learning techniques. We want to develop models that can estimate the ticket prices for different flights based on factors such as departure location, arrival location, travel dates, airline preferences, and other relevant features. Using machine learning algorithms, we are going to analyze historical flight data and identify patterns to make accurate predictions about future flight prices.We have collected a dataset of Shubham Bathwal (MA Student) that published it on Kaggle and that consists of various flying data of airline and their respective flies, as well as the time before the departure and times of the day in which it depart. This dataset has been extracted from Ease My Trip Repository[1]. We will use a set of regression models, including Decision Tree Regressor[4],Random Forest Regression[5], ExtraTreesRegressor[6],Linear Regression[7] in particular we used the ridge regression[9] and Neural Networks[8], to predict the probability of a country winning a medal. To evaluate the performance of these models, we will use different metrics such as the MSE, MAE, and R2 and the RandomSearchCrossValidation to find the best hyperparameters for the models. This project can have various practical applications, including helping helping the people to get right prediction about the prices of flying ticket prices and allowing even the companies to analyze the trends of the fluctuating prices. And I believe that could be very exciting the predicition of the flying ticket prices due to it's

very useful today to make an estimation of the possible prices for the tickets basing on the company and other such characteristics of it.

2. Dataset

For the scientific project "Flight Price Prediction" we got the open dataset available on Kaggle[2]. This dataset gives access to information about the flying or different airlines, the different departure cities and arrival cities. Additionally, the dataset provides information on the time of the day in which it arrives and depart and moreover it also contain the duration of the flying

2.1 Data pre-processing

Data preprocessing refers to manipulating and changing data raws to make them more correct for a ML purpose. Raw data often is inconsisten and presents errors with missing information; This must be addressed before using it to make predictions.

This dataset has been extracted from the website EasyMyTrip and it was already merged and partially cleaned; Anyway more operation are needed to ensure the data quality, so i applied more preprocessing steps to these data.

To preprocess the data, we did an operation of Feature Encoding[3], where we transformed the Categorical data into Numerical data. then we remved features that was not necessary to our purposes in order to increase the efficency of the models.At the end we randomized the data by a suffle operations in order to remove any systematic patterns that could affect the data, and before to use it with the model always in order to allow to the models to increase the performance we included a MinMax Scaler taken from the Scikit Learn library.

Data encoding: In our dataset the column that regards some information such as the airline, the departure city, arrival city and others, were in a categorical values; In order to perform our regression tasks we had to convert them into numerical value so we applied a factorization on this features, then we saved the mapping of this feature, in order always to perform some analytics then.

Data transformation: To improve the performance of our models, we wanted to apply a MinMaxScaler features transformations.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

This will transform our values in a similar scale and will allow our model to predict correctly.

At the end we computed some statistics to our dataset and we can notice that we got a final dataset proper balanced that could be used for model; Our dataset has 300153 entries.

2.2 Data features.

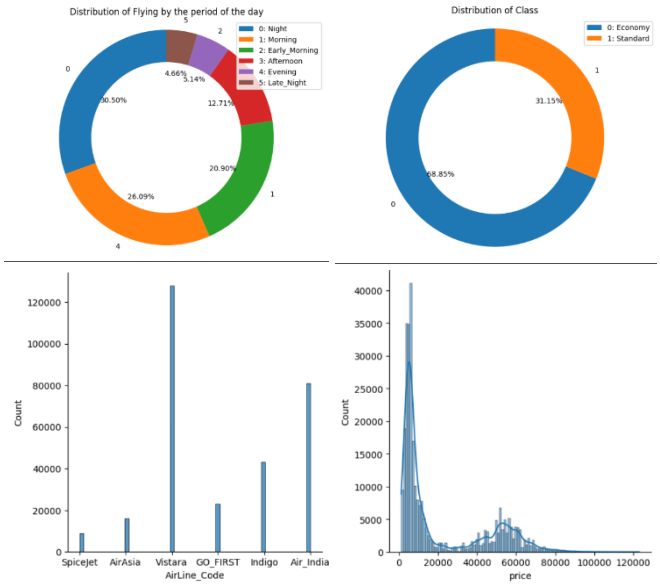
For this dataset 11 features to build out model has been provided.; The features that i used are reported in this table

FEATURES ON FLYING DATASET	
Airline	This represents the name of the airline company
Flight	This feature represents the Flight code that identify the code of flying
Source City	City from which the flight start
Departure Time	Derived categorical feature representing time periods.
Stops	This represents the number of stops between the source and destination cities
Arrival Time	Derived categorical feature representing time intervals.
Destination City	City where the flight will go
Class	Seat class with values "Business" and "Economy".
Duration	Continuous feature indicating the travel duration in hours.
Days Left	Derived feature that represents the remaining days before the flyight.
Price	The target variable representing the ticket price.

In this regression tasks in which we want to predict the price, the vector X of features that describe the flying as been used to predict the target values(y) that is the price of the flying.

2.3 Data analytics

Before to proceed we got some insights of our dataset so using the features and the corresponding mapping(categorical values); So we plotted some analytics of the features.We also did it to analyze the distribution of the data with respect to the relevant aspects that we are taking into account to make the predicition; Here there are some few of them:



From this analytics we learnt different things such as that the most taken company is Vistara, or that the most of flying where Economy or that the most of Flying just started during the night.

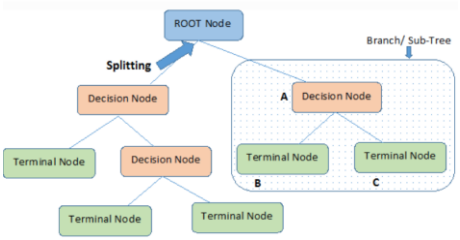
3. Models and Evaluation methodology.

For this problem we decided to use different models and to analyse their behavior, then we will also compare the results.To solve this problem, we decide to treat it as a regression task where we want to predict the price of the airline tickets.

3.1 Models

a) **Decision Tree Regressor:**

A decision tree is a simple tree-shaped structure where each internal node represents a test on one attribute, arcs show the results of a test and leaf nodes reflect classes. They are easy to understand and can be easily converted to a set of rules. Moreover, they can classify both categorical and numerical data and require no priori assumptions about the data. Because of the advantages listed above, the decision tree approach is extensively utilized for both classification and prediction purposes.

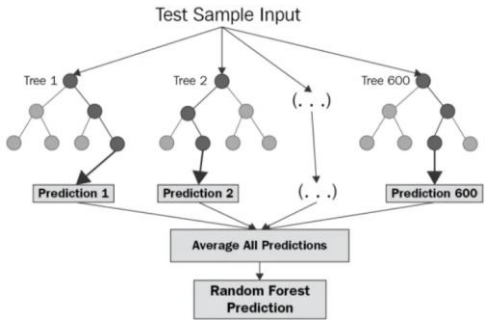


Attribute selection is a crucial step in building decision trees for datasets with multiple attributes. Randomly selecting the root or internal nodes can lead to poor accuracy. To address this problem, we use such criteria suggested such as the entropy, the information gain, etc…

In the scikit-learn library, the DecisionTreeRegressor class uses the mean squared error (MSE) as the criterion for splitting and selecting attributes.

b) **Random Forest Regressor:**

random forest model is a type of ensemble learning method that combines multiple decision trees to make predictions. It can be considered a generalization of the Random Tree Model.



A random forest consists of multiple decision trees that are trained independently. Each decision tree is constructed by recursively splitting the data based on randomly selected subsets of features.

The random selection of features at each node ensures diversity among the trees. The random forest model is trained using a process called "bagging" (bootstrap aggregating).

During the training process, each decision tree is trained on a bootstrap sample, which is a random sampling with replacement from the original dataset. Given a data point to be predicted, the random forest model combines the predictions of all the individual trees.

Each tree independently predicts the outcome for the data point based on its selected features. Output of the random forest model is the average of the predictions from all the trees.

c) **Extra Tree Regressor:**

Extra Trees is an ensemble machine learning algorithm that combines the predictions from several decision trees.

It is a commonly used random forest algorithm. Although it uses an easiest method where the members are the decision trees, it can often give better results than the random forest algorithm.

Both the Random Forest Regressor and the Extra Trees Regressor are tree algorithms. The difference is given by the fact that the Random Forest Regressor uses resampling, and the Extra Trees Regressor uses original data to create the random forest of decision trees.

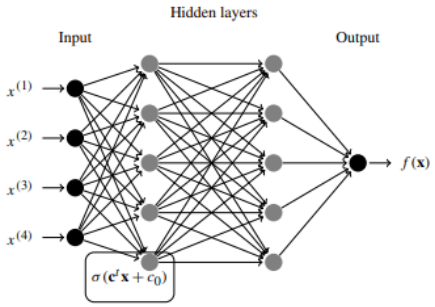
Moreover, Extra Tree Regressor is computationally less expensive as it randomly selects splits without optimizing the split criterion and Extra Tree Regressor, the splitting strategy is slightly different. Instead of evaluating all possible splits for the selected features, Extra Tree Regressor selects random splits without considering the optimal split criterion.

d) **Fully Connected Neural Network:**

A neural network is a collection of interconnected neurons arranged in multiple layers. Each layer takes input either from the previous layer or directly from the data. Within each layer, the data undergoes a linear transformation by multiplying the input values with corresponding weights and summing them together. This weighted sum, known as "z," is then passed through a non-linear activation function, denoted as "f," to produce the output, "y."

$$z = \sum_{i=0}^{n} w_i x_i$$

To make accurate predictions, the neural network needs to learn the optimal values for the biases, represented by "b," and the weights, represented by "W." This learning process is achieved through a technique called backpropagation, which involves updating the parameters using gradient descent. Gradient descent iteratively adjusts the biases and weights based on the calculated gradients until convergence, where the network's predictions align closely with the desired outcomes.



The batch size is a hyper-parameter of the gradient descent algorithm, is the number of training examples used in each iteration to update the parameters. It is a tunable parameter that affects the training process and can impact the convergence speed and computational efficiency of the neural network.

e) **Linear Regressor (Ridge):**

Ridge regression is a regression technique that is used to handle the problem of multicollinearity (high correlation) among predictor variables in a linear regression model. It is a regularized version of linear regression that introduces a penalty term to the cost function.

In ridge regression, the cost function is modified by adding a regularization term that is proportional to the squared magnitude of the coefficient values. The regularization term, also known as the ridge penalty or L2 penalty, helps to shrink the coefficients towards zero, reducing their impact on the regression model.

3.2 Evaluation

After trained the models we decide to evaluate them using standard metrics for regression problem such as the R-2 square, the MSE and the MAE.

MSE = (1/n) * Σ(y - ŷ)^2

where:

- n: Number of samples in the dataset
- y: True values
- ŷ: Predicted values

MAE = (1/n) * Σ|y - ŷ|

where:

- n is the number of samples in the dataset
- y represents the true values
- ŷ represents the predicted values
- Σ denotes the sum of the absolute differences between each true-predicted pair

R-squared (R2) is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model.

R2 = 1 - (Σ(y - ŷ)² / Σ(y - ȳ)²)

where:

Σ denotes the sum

y represents the true values

ŷ represents the predicted values

ȳ represents the mean of the true values

Randomized Search: Is a technique used for tuning the hyperparameters of machine learning models. Hyperparameters are settings that need to be adjusted to optimize the performance of the model on a specific dataset.

Randomized Search defines a search space, which represents the range of possible values for each hyperparameter. Each point within this search space corresponds to a specific configuration of the model.

Unlike Grid Search, which exhaustively evaluates every point on a predefined grid of hyperparameter values, Randomized Search allows for a more efficient search, as it can quickly identify promising configurations.

This exploit the best model performance.

3.3 Results

So I run the several algorithms and we reported the results as showed here and I collected the results of the experiment.

Results on the error

	Decision Tree	Random Forest	Extra Tree	Ridge	Neural Network
MSE	22052780	5893352	6557958	50166074	17879060
R-2	0.95746	0.98863	0.98735	0.90332	0.96551
MAE	2573	985	955	4599	2470

Results on the hyperparameters

As we got, we can see that on the Decision Tree works well with a max depth of 9, while the Random Forest works well with 100 estimators. For Extra Tree Regressor instead 200 estimators are a good choice and at the end the best Alpha for the Ridge model is 0.1.

4. Conclusions

As shown in the table of the error the best results are given by the Random Forest and the Extra tree Regressor, that works even well then, my Fully Connected Neural Network Models. Instead, the Ridge model, even if is faster in execution, it achieved lowest results than all the other models we run.

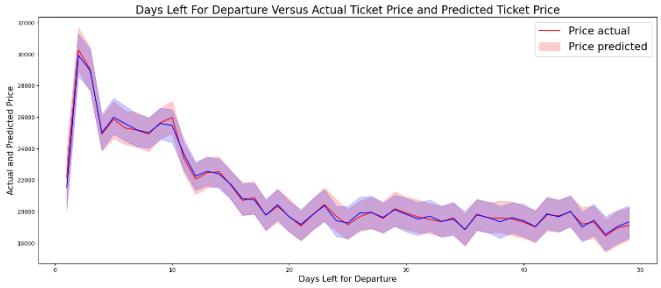


Figure 1 This is the result obtained with the Extra Tree Regressor, compared to the remaining days before the departure.

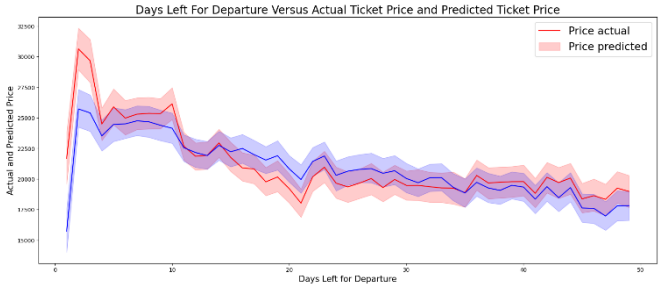


Figure 2 This is the result obtained with the Ridge model, compared to the remaining days before the departure.

References

[1] EaseMyTrip.com. (n.d.). *EaseMyTrip - Flights, Hotels and Bus*. Copyright Â© 2012 easemytrip.com. https://www.easemytrip.com/

[2] *Flight Price Prediction*. (2022, February 25). Kaggle. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction?datasetId=1957837&sortBy=voteCount&searchQuery=dense

[3] Cohen, J. (2021, December 25). Categorical Feature Encoding - Towards Data Science. Medium. https://towardsdatascience.com/categorical-feature-encoding-547707acf4e5

[4] World Academy of Science, Engineering and Technology International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering Vol:2, No:12, 2008 https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4a8fc8bb575f9604685a53447a15c5ea96e29a9a

[5] Segal, M. R. (2004, April 14). *Machine Learning Benchmarks and Random Forest Regression*. https://escholarship.org/uc/item/35x3v9t4

[6] Reza, M., & Haque, M. R. (2020). Photometric redshift estimation using ExtraTreesRegressor: Galaxies and quasars from low to very high redshifts. *Astrophysics and Space Science*, *365*(3). https://doi.org/10.1007/s10509-020-03758-w

[7] Tuning as Linear Regression Marzieh Bazrafshan, Tagyoung Chung and Daniel Gildea Department of Computer Science University of Rochester Rochester, NY 14627 : https://aclanthology.org/N12-1062.pdf

[8] Wang, H., Shi, H., Lin, K., Qin, C., Zhao, L., Huang, Y., & Liu, C. (2020). A high-precision arrhythmia classification method based on dual fully connected neural network. *Biomedical Signal Processing and Control*, *58*, 101874. https://doi.org/10.1016/j.bspc.2020.101874

[9] Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, *6*(S2). https://doi.org/10.1186/1753-6561-6-s2-s10