HIGH SCHOOL EDITION
Journal of Student Research

Volume 12 Issue 4 (2023)

Using Different Machine Learning Algorithms to Predict the Prices of Flight Tickets

Jeremy Rohan Bollack[1] and Joseph Anthony Vincent[2#]

[1]Berlin International School, Germany
[2]Stanford University
[#]Advisor

ABSTRACT

The rising prices of flight tickets and the lack of transparency in the dynamic pricing strategies of airlines have caused many consumers to wonder, what factors actually determine these prices. In order to investigate this question, a large dataset of flight ticket bookings that includes the most price-defining variables was acquired. This data was preprocessed using discretization, normalization, and principal component analysis. This preprocessed data was then used to train 5 different Machine Learning algorithms: Linear Regression, DecisionTree, Ridge Regression, RandomForest, and SVR. The training of the RandomForest and SVR models was not possible due to runtime errors, however, the other models trained as expected. All models performed well, with the Linear Regression and Ridge Regression performing identically. Overall, the DecisionTree model performed the best at predicting the prices of flights, and by adjusting hyperparameters the performance could be further increased. The investigation could be continued by using a larger dataset to investigate how the model performs with more variables and under broader conditions. Additionally, the model could be reappropriated to make a user-friendly flight price prediction tool that helps consumers with their purchasing decisions.

Introduction

This research will investigate **to what extent different Machine Learning algorithms**, hereinafter referred to as “ML algorithms”, **are successful in identifying significant price-defining factors and predicting the prices of flight tickets when provided with a large dataset of flight bookings.**

During the last couple of years, rising rates of inflation and disruptions in the global supply chain due to the COVID-19 pandemic and the war in Ukraine have caused prices to soar, including those of airline flights. These rising prices have made traveling by airplane less affordable for the average consumer. Furthermore, the dynamic pricing strategies of airlines lack transparency, which results in a lack of understanding by the customer as to why a price is set at a certain level. This research aims to use different Artificial Intelligence in the form of ML algorithms and a large dataset to create a model that can determine the significance of different price-defining factors and predict prices, making the decision-making behind ticket prices more transparent and understandable. This type of research can be classified as supervised learning[1], which is the use of labeled datasets to train algorithms that classify data or predict outcomes accurately. Furthermore, the type of Machine Learning can be classified as regression[2], which is a technique for investigating the relationship between independent variables (further explained in the Dataset section) and a

[1] IBM. n.d. “What Is Supervised Learning? | IBM.” Www.ibm.com. Accessed June 8, 2023. https://www.ibm.com/topics/supervised-learning.

[2] Castillo, Dianne. 2021. “Machine Learning Regression Explained.” Seldon. October 29, 2021. https://www.seldon.io/machine-learning-regression-explained#:~:text=Machine%20Learning%20Regression%20is%20a.

dependent variable, in this case, the ticket price. The dataset used to train this model is a combination of both numerical and categorical data.

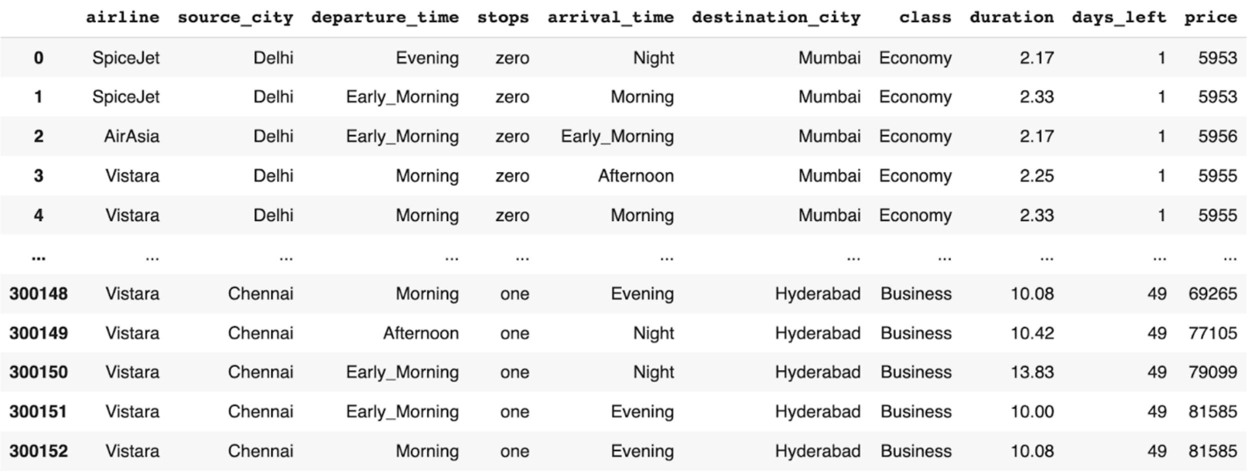
Background

When conducting preliminary research to understand how to investigate this question, some interesting things came to mind. To investigate this topic, it is crucial to understand which factors mainly contribute towards the pricing of flight tickets, in order to find a dataset that best matches the requirements. Identifying these factors proposed a significant challenge, as airlines tend to keep their strategies for dynamic pricing a secret. The next best alternative for identifying these factors was to refer to research conducted by others. There were two main sources to refer to, an online blog post[3] and an online newspaper-article[4]. To summarize the blog post, the main price-defining factors include travel class, customer profiling, time left till flight, current sales volume, length of trip, level of competition, number of layovers, peak travel dates, level of overbooking, and fuel prices. To summarize the article, ticket prices also depend on the platform the ticket was booked on, e.g. a travel agency, ticket machine, or internet browser. However, both articles agree that the prices of tickets mainly depend on the days left until the flight, the length of the flight, the number of layovers, the travel class, the number of seats left, and the oil prices. These factors should be taken into consideration when identifying the most suitable dataset.

Dataset

The most appropriate dataset[5] (seen in **Figure 1**) which is available to the public, was found in an online databank named Kaggle. The dataset is made up of information from over 300,000 past flight bookings from flights between 6 different cities in India. It consists of 9 independent variables, which are both numerical and categorical, and the dependent variable, being the ticket price in Indian Rupees. The independent variables include some of the previously named significant factors, such as the days left until the flight, the length of the flight, the number of layovers, and the travel class. These all provide a good base to train an ML model, as they should correlate well with the final ticket price and allow for accurate comparison between different models. However, a limitation of the dataset is that it does not include the number of seats left, giving no insights into the levels of supply and demand of the flights. Oil prices are also not included in the dataset, however, they may not be as significant in this investigation, as all flights are operated in India and range over a short time, making it safe to assume that prices maintained a steady level for all flights. This means that although not all factors are considered, the dataset includes a majority of price-defining factors, providing an adequate base to train ML models on.

[3] Hayward, Justin, Daniel Martínez Garbuno, and Pranjal Pande. 2020. "How Airline Ticket Pricing Works." Simple Flying. October 22, 2020. https://simpleflying.com/how-airline-ticket-pricing-works/#future-of-airline-pricing.
[4] Mark, Lois Alter. 2021. "This Is the Best Time to Buy Flights." Reader's Digest. December 6, 2021. https://www.rd.com/article/when-to-buy-plane-tickets/.
[5] Bathwal, Shubham. n.d. "Flight Price Prediction." Www.kaggle.com. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction.



	airline	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	SpiceJet	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	SpiceJet	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	AirAsia	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	Vistara	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	Vistara	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955
...	...	...	...	...	...	...	...	...	...	...
300148	Vistara	Chennai	Morning	one	Evening	Hyderabad	Business	10.08	49	69265
300149	Vistara	Chennai	Afternoon	one	Night	Hyderabad	Business	10.42	49	77105
300150	Vistara	Chennai	Early_Morning	one	Night	Hyderabad	Business	13.83	49	79099
300151	Vistara	Chennai	Early_Morning	one	Evening	Hyderabad	Business	10.00	49	81585
300152	Vistara	Chennai	Morning	one	Evening	Hyderabad	Business	10.08	49	81585

Figure 1 (Screenshot of Pandas Dataframe in Google Collaboratory)

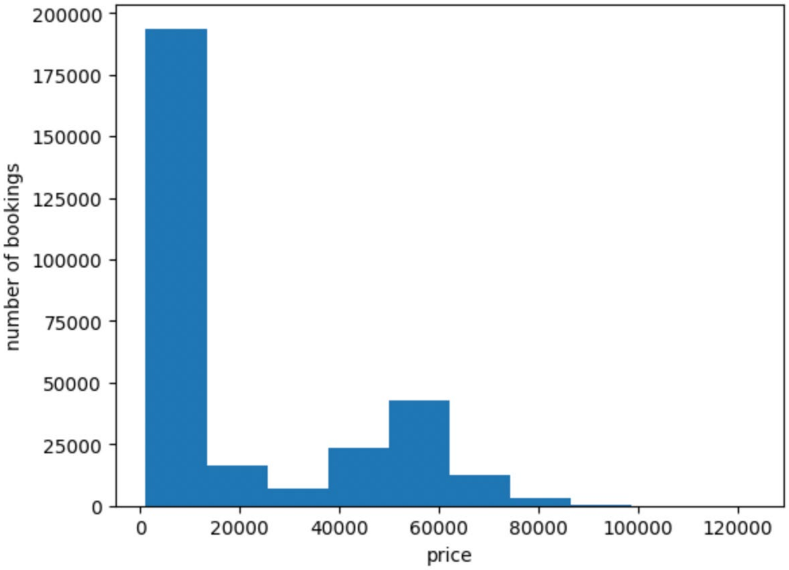


Figure 2 (Histogram made with Matplotlib)

Figure 2 shows the price distribution of the dataset. It is clearly visible that most bookings are at the lower end of the price spectrum, indicating that results in that range may be more accurate because there was more training data.

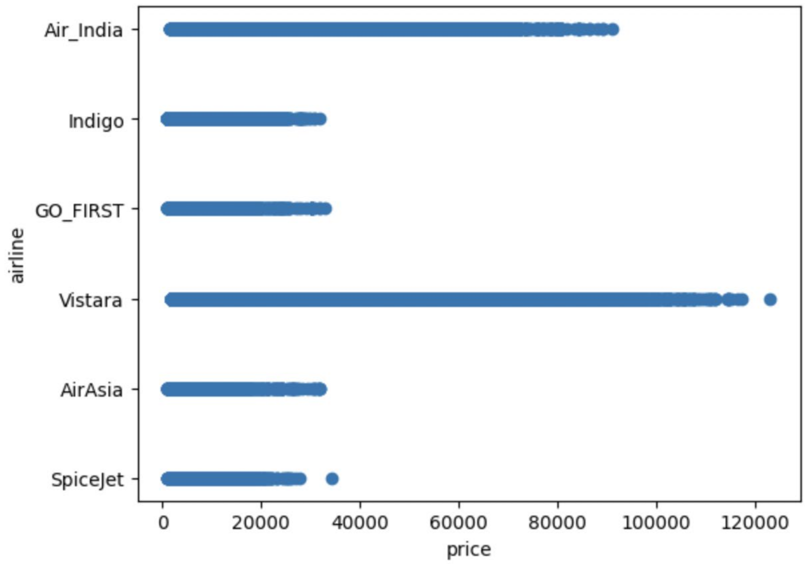


Figure 3 (Histogram made with Matplotlib)

Figure 3 shows that all airlines operate at the low end of the price spectrum, however, only a few operate at higher ends, indicating that the airline itself will influence price predictions.

Methodology/Models

Before training any efficient ML model, the raw dataset needs to be preprocessed[6]. This mainly consists of assessing the quality of the data, transforming it, and cleaning it up, in order to enhance performance and ensure accurate results.

Assessing the quality of data

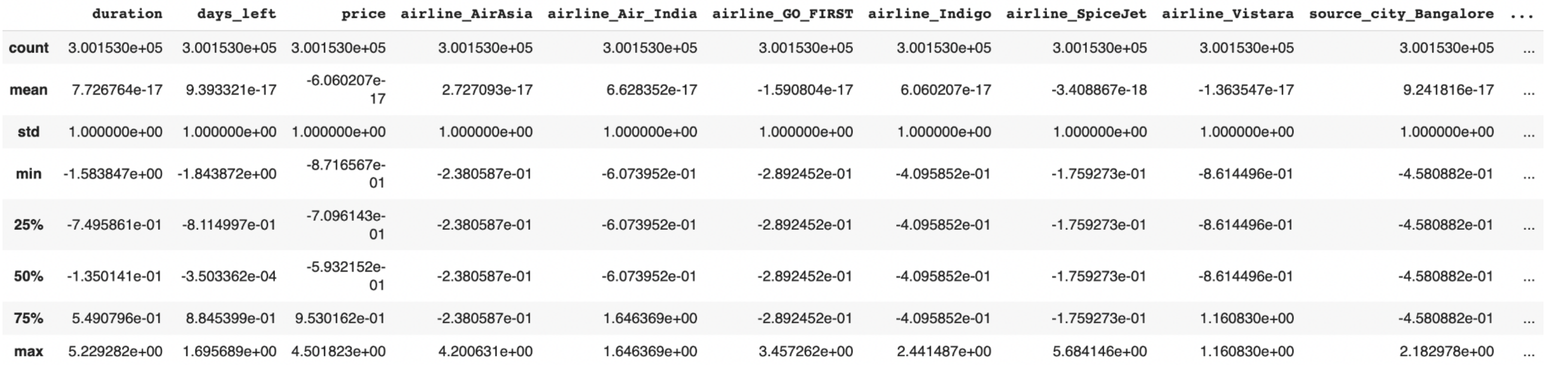
Since the data comes from a single credible source, and there are no Null values in the data, it can be assumed that the quality of the data is adequate.

Transforming the data

Transforming the data turns the dataset into the correct format to train and test an ML model. The dataset is imported as a Pandas[7] dataframe. To allow for computations, all categorical variables have to be discretized by using the pd.getdummies() function. As an example, the "Airline" column would be reformatted into multiple columns, named "Airline_[AirlineName]", with one new column for each airline. By doing this we assign binary values to each column for each datapoint, that indicate whether a specific condition is satisfied or not. Then, we convert these to a Numpy[8] array, which makes the calculations needed for the normalization of data simpler. By normalizing the data, all columns have a mean of 0 and a standard deviation of 1, so that they can be compared at a similar scale to other variables.

[6] Geisler Mesevage, Tobias. 2021. "What Is Data Preprocessing & What Are the Steps Involved?" MonkeyLearn Blog. May 24, 2021. https://monkeylearn.com/blog/data-preprocessing/.
[7] Pandas. 2018. "Python Data Analysis Library" Pydata.org. 2018. https://pandas.pydata.org/.
[8] Numpy. 2009. "NumPy." Numpy.org. 2009. https://numpy.org/.

The normalization[9] of a data point x is explained by this formula:
$x := (x - \mu) / \sigma$,
where μ is the sample mean and σ is the sample standard deviation.



	duration	days_left	price	airline_AirAsia	airline_Air_India	airline_GO_FIRST	airline_Indigo	airline_SpiceJet	airline_Vistara	source_city_Bangalore	...
count	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	3.001530e+05	...
mean	7.726764e-17	9.393321e-17	-6.060207e-17	2.727093e-17	6.628352e-17	-1.590804e-17	6.060207e-17	-3.408867e-18	-1.363547e-17	9.241816e-17	...
std	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	...
min	-1.583847e+00	-1.843872e+00	-8.716567e-01	-2.380587e-01	-6.073952e-01	-2.892452e-01	-4.095852e-01	-1.759273e-01	-8.614496e-01	-4.580882e-01	...
25%	-7.495861e-01	-8.114997e-01	-7.096143e-01	-2.380587e-01	-6.073952e-01	-2.892452e-01	-4.095852e-01	-1.759273e-01	-8.614496e-01	-4.580882e-01	...
50%	-1.350141e-01	-3.503362e-04	-5.932152e-01	-2.380587e-01	-6.073952e-01	-2.892452e-01	-4.095852e-01	-1.759273e-01	-8.614496e-01	-4.580882e-01	...
75%	5.490796e-01	8.845399e-01	9.530162e-01	-2.380587e-01	1.646369e+00	-2.892452e-01	-4.095852e-01	-1.759273e-01	1.160830e+00	-4.580882e-01	...
max	5.229282e+00	1.695689e+00	4.501823e+00	4.200631e+00	1.646369e+00	3.457262e+00	2.441487e+00	5.684146e+00	1.160830e+00	2.182978e+00	...

Figure 4 (Part of the Dataset after Normalization)

As seen in **Figure 4**, the normalization was successful, as the mean is 0, and the standard deviation is 1 for all columns.

Cleaning up the data

Cleaning the data up mainly consists of removing insignificant data. This insignificant data is identified by a process called Principal Component Analysis (PCA), which can be automated using a function from a Python library called scikit-learn[10].

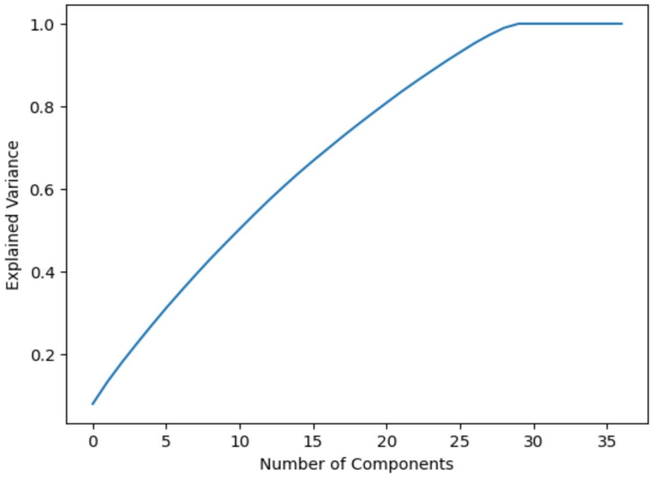


Figure 5 (PCA Explained Variance)

[9] Willaert, Jorrit. 2021. "How to Calculate the Mean and Standard Deviation — Normalizing Datasets in Pytorch." Medium. Towards Data Science. September 24, 2021. https://towardsdatascience.com/how-to-calculate-the-mean-and-standard-deviation-normalizing-datasets-in-pytorch-704bd7d05f4c#:~:text=The%20data%20can%20be%20normalized,channel%20is%20normalized%20this%20way..

[10] scikit-learn. 2019. "Scikit-Learn: Machine Learning in Python." Scikit-Learn.org. 2019. https://scikit-learn.org/stable/.

Figure 5 shows the Explained Variance Ratio for the different variables that could affect the price of tickets. The PCA function rearranges the variables in order of importance, showing that the last 7 variables do not influence the price of flight tickets at all. These variables can then be removed in order to save computational power and increase efficiency.

Then the normalized & cleaned-up data is split into its X and Y components. Lastly, both the X and Y components are further split into the conventional proportions of 80% training data and 20% testing data. The training data is used to train different ML models, and the testing data is used to evaluate how the trained model has performed. After completing the data preprocessing, the most difficult part of creating the model is finished.

Training the Models

For the purpose of this investigation, I wanted to train 5 different prediction models. These were narrowed down to Linear Regression, DecisonTree, Ridge Regression, RandomForest, and SVR, however due to runtime issues RandomForest and SVR had to be excluded from the trials. SVR is a hard to scale model, since the training time is more than quadratic to the number of samples[11].

The Linear Regression model and Ridge Regression Model work in a very similar way. As illustrated below in **Figure 6**, the models create a line of regression, which tries to minimize the sum of the squares of the residuals, being the distance between the line and the actual data points[12]. Ridge Regression works slightly differently, as it also tries to create a linear model with a small slope, whereas there is no slope restriction in the regular Linear Regression model.

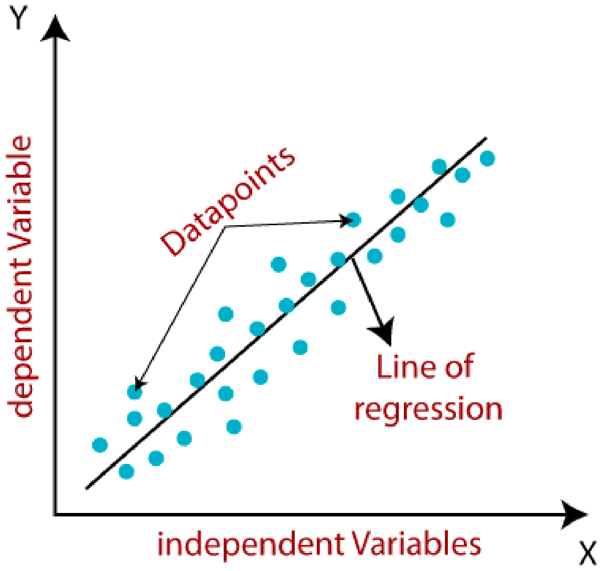


Figure 6[13] (Basic Concept of Regression in ML)

[11] "Sklearn.svm.SVR — Scikit-Learn 0.23.1 Documentation." n.d. Scikit-Learn.org. Accessed July 9, 2023. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.

[12] "How Linear Regression Algorithm Works—ArcGIS pro | Documentation." n.d. Pro.arcgis.com. Accessed July 8, 2023. https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-linear-regression-works.htm

[13] "Linear Regression in Machine Learning - Javatpoint." n.d. Www.javatpoint.com. Accessed July 8, 2023. https://www.javatpoint.com/linear-regression-in-machine-learning.

The DecisionTree model works more like the human brain, asking a series of predefined questions, before making an informed prediction. This is based on how a previous set of questions was answered (training data). A simplified example of how a DecisionTree model works is seen below in **Figure 7**.

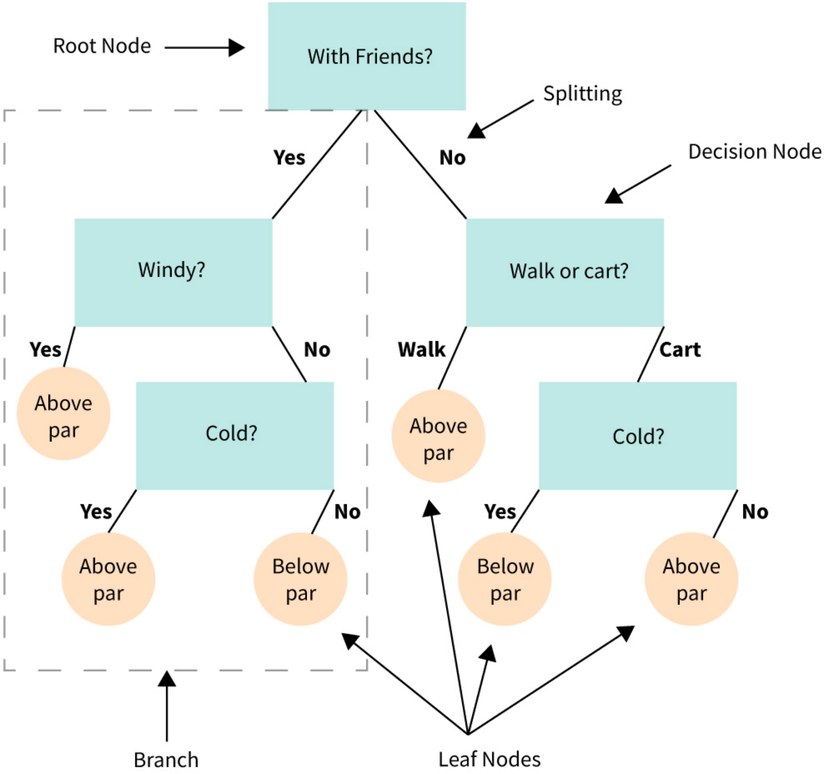
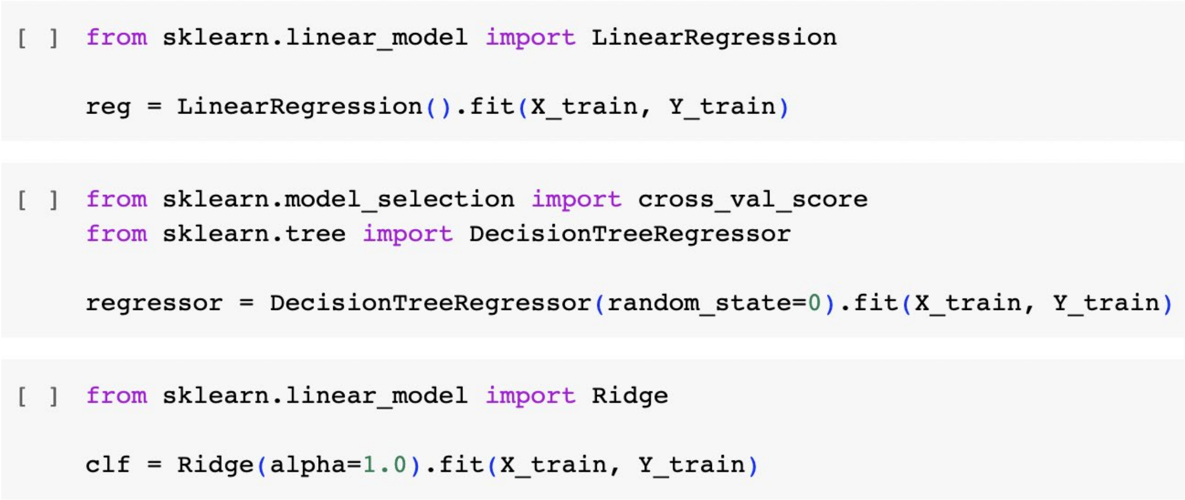


Figure 7[14] (Basic Concept of a DecisionTree in ML)



```
[ ]  from sklearn.linear_model import LinearRegression

     reg = LinearRegression().fit(X_train, Y_train)
```

```
[ ]  from sklearn.model_selection import cross_val_score
     from sklearn.tree import DecisionTreeRegressor

     regressor = DecisionTreeRegressor(random_state=0).fit(X_train, Y_train)
```

```
[ ]  from sklearn.linear_model import Ridge

     clf = Ridge(alpha=1.0).fit(X_train, Y_train)
```

Figure 8 (Training Linear Regression, DecisionTree, and Ridge Regression)

[14] "Decision Tree." n.d. CORP-MIDS1 (MDS). Accessed July 8, 2023. https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/#:~:text=A%20decision%20tree%20is%20a.

As one can see in **Figure 8**, training the models is fairly simple, as all the work is done by scikit-learn's predefined functions. After training the models, the only thing left to do is to decide which one functions best in predicting the price of flight tickets.

Results and Discussion

There are 3 conventional ways to evaluate the effectiveness of regression models in Machine Learning.

1. Mean Squared Error[15]- The average squared error between the actual and predicted values. A lower error represents a better score.
2. Mean Absolute Error[16]- The average absolute error between the actual and predicted values. A lower error represents a better score.
3. R-Squared Value[17]- Indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model. It is a value between 0 and 1, where 1 is a perfect score.

Scikit-learn has analytical tools, which include calculations of these metrics. The results for these metrics are summarized below in **Tables 1 & 2**.

Table 1 (Evaluation of Different ML Models for Predicting Ticket Prices)

	Linear Regression	**DecisionTree**	**Ridge Regression**
Mean Squared Error	0.0887	**0.0313**	0.0887
Mean Absolute Error	0.2006	**0.064**	0.2006
R-Squared	0.9113	**0.9687**	0.9113

Table 2 (Ranking of Different ML Models for Predicting Ticket Prices)

	Mean Squared Error	**Mean Absolute Error**	**R-Squared**
1st	DecisionTree	DecisionTree	DecisionTree
2nd/3rd	Linear/Ridge Regression	Linear/Ridge Regression	Linear/Ridge Regression

[15] Allwright, Stephen. 2022. "MSE vs MAE, Which Is the Better Regression Metric?" Stephen Allwright. July 7, 2022. https://stephenallwright.com/mse-vs-mae/.

[16] Ibid.

[17] Fernando, Jason. 2021. "R-Squared Definition." Investopedia. September 12, 2021. https://www.investopedia.com/terms/r/r-squared.asp.

In all 3 metrics, the DecisionTree model has produced the best results. The Linear Regression and Ridge Regression models have produced identical results, indicating that the regularization term introduced in Ridge has had no impact on the model's performance[18]. This implies that the dataset does not exhibit collinearity between independent variables, which highlights the statistical significance of the independent variables with respect to the dependent variable.

In order to further increase the model's accuracy, a last attempt of increasing the model's performance was conducted with the DecisionTree model by adjusting its hyperparameters. Hyperparameters can be defined as parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning[19].

The main hyperparameter to adjust in a DecisionTree model is the maximum depth, which limits the amounts of nodes and branches in a DecisionTree. **Table 3** shows the DecisionTree's performance with adjusted hyperparameters.

Table 3 (Adjusted hyperparameters vs. DecisionTree performance)

max_depth	Mean Squared Error	Mean Absolute Error	R-Squared
No Adjustment	0.0313	0.0640	0.9687
18	0.0289	0.0804	0.9711
19	**0.0288**	0.0770	**0.9713**
20	0.0290	**0.0743**	0.9710

By adjusting the hyperparameter of the maximum depth, there have been slight, but noticeable changes in the model's metrics. The optimum performance has mainly been achieved at a predefined maximum depth of 19. At this hyperparameter setting, the MSE has decreased from 0.0313 to 0.0288, however, the MAE has increased from 0.0640 to 0.0770. Most importantly, however, the R-Squared Value has increased from 0.9687 to 0.9713, indicating that finetuning the model has increased the performance moderately.

Conclusion

An extension of this research may be to find a larger dataset and see how the model performs with more variables and under broader conditions. Currently, the scope of this dataset is very narrow as it only includes 6 airports in India, making the model hard to implement in a global context. Furthermore, the research may be continued by making this model useful for general consumers. As of this moment, the model can only predict values from an existing data point, which contains a label. An interesting continuation of this might be to allow a user to input their own variables. These could then be preprocessed in the exact same manner as the other data points from the dataset, and then the model could output a justified price for the ticket. This could help the user decide whether it is sensible to buy a flight ticket at a certain price, or whether a ticket offer is overpriced. In conclusion, both the Regression models and the DecisionTree model are appropriate in predicting the prices of flight tickets to a certain extent. However, the DecisionTree

[18] ChatGPT. 2023. "Response to 'What Does It Mean, When My Linear Regression and Ridge Regression Model Perform the Exact Same Way?'" July 8, 2023. https://chat.openai.com.

[19] Nyuytiymbiy, Kizito. 2022. "Parameters and Hyperparameters in Machine Learning and Deep Learning." Medium. January 15, 2022. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac#:~:text=Hyperparameters%20are%20parameters%20whose%20values.

model has performed better than the Regression models in all tested model metrics. An R-Squared value of 0.9713 indicates that there is a low error caused due to the model, which means that the DecisionTree is a suitable model to investigate this highly complex question.

References

Allwright, Stephen. 2022. "MSE vs MAE, Which Is the Better Regression Metric?" Stephen Allwright. July 7, 2022. https://stephenallwright.com/mse-vs-mae/.

Bathwal, Shubham. n.d. "Flight Price Prediction." Www.kaggle.com. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction.

Castillo, Dianne. 2021. "Machine Learning Regression Explained." Seldon. October 29, 2021. https://www.seldon.io/machine-learning-regression-explained#:~:text=Machine%20Learning%20Regression%20is%20a.

ChatGPT. 2023. "Response to 'What Does It Mean, When My Linear Regression and Ridge Regression Model Perform the Exact Same Way?'" July 8, 2023. https://chat.openai.com.

"Decision Tree." n.d. CORP-MIDS1 (MDS). Accessed July 8, 2023. https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/#:~:text=A%20decision%20tree%20is%20a.

Fernando, Jason. 2021. "R-Squared Definition." Investopedia. September 12, 2021. https://www.investopedia.com/terms/r/r-squared.asp.

Geisler Mesevage, Tobias. 2021. "What Is Data Preprocessing & What Are the Steps Involved?" MonkeyLearn Blog. May 24, 2021. https://monkeylearn.com/blog/data-preprocessing/.

Hayward, Justin, Daniel Martínez Garbuno, and Pranjal Pande. 2020. "How Airline Ticket Pricing Works." Simple Flying. October 22, 2020. https://simpleflying.com/how-airline-ticket-pricing-works/#future-of-airline-pricing.

"How Linear Regression Algorithm Works—ArcGIS pro | Documentation." n.d. Pro.arcgis.com. Accessed July 8, 2023. https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-linear-regression-works.htm.

IBM. n.d. "What Is Supervised Learning? | IBM." Www.ibm.com. Accessed June 8, 2023. https://www.ibm.com/topics/supervised-learning.

"Linear Regression in Machine Learning - Javatpoint." n.d. Www.javatpoint.com. Accessed July 8, 2023. https://www.javatpoint.com/linear-regression-in-machine-learning.

Mark, Lois Alter. 2021. "This Is the Best Time to Buy Flights." Reader's Digest. December 6, 2021. https://www.rd.com/article/when-to-buy-plane-tickets/.

Numpy. 2009. "NumPy." Numpy.org. 2009. https://numpy.org/.

Nyuytiymbiy, Kizito. 2022. "Parameters and Hyperparameters in Machine Learning and Deep Learning." Medium. January 15, 2022. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac#:~:text=Hyperparameters%20are%20parameters%20whose%20values.

Pandas. 2018. "Python Data Analysis Library — Pandas: Python Data Analysis Library." Pydata.org. 2018. https://pandas.pydata.org/.

scikit-learn. 2019. "Scikit-Learn: Machine Learning in Python." Scikit-Learn.org. 2019. https://scikit-learn.org/stable/.

"Sklearn.svm.SVR — Scikit-Learn 0.23.1 Documentation." n.d. Scikit-Learn.org. Accessed July 9, 2023. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.

Willaert, Jorrit. 2021. "How to Calculate the Mean and Standard Deviation — Normalizing Datasets in Pytorch." Medium. Towards Data Science. September 24, 2021. https://towardsdatascience.com/how-to-calculate-the-mean-and-standard-deviation-normalizing-datasets-in-pytorch-704bd7d05f4c#:~:text=The%20data%20can%20be%20normalized,channel%20is%20normalized%20this%20way..