FareSight: Leveraging Machine Learning to Forecast Flight Prices

Aarzoo (2022008) Anushka Srivastava (2022086) Nandini Jain (2022316) Suhani Kalyani (2022511)

Interim Project Report

Abstract

Flight prices can vary significantly, making it difficult for travelers to plan their budgets effectively. This paper presents the development of a machine learning-based predictive model to forecast flight prices. By utilizing a dataset containing various flight details, several regression models were implemented and evaluated based on metrics such as MSE, RMSE, R2 scores and MAE. Through comprehensive analysis, Random Forest and Bagging Regressor models were found to give best prediction accuracy.

Further improvements will include sampling, outlier detection, hyperparameter tuning for enhanced model performance and identifying the day with the minimum price of the ticket.

1. Introduction

Flights have become the preferred choice for long distance travels. However, the cost of flight tickets varies a lot, increasing one day and dropping the very next. This leaves many travelers confused about the factors affecting these variations which is an important financial consideration.

To solve this problem, we are building a predictive model that can help predict the prices of flight tickets and help them identify the day with the most optimal price. This will help travelers make more informed decisions in their travel plans, and allow us to contribute to making air travel more accessible and affordable.

2. Literature Survey

2.1. Dynamic Flight Price Prediction Using Machine Learning Algorithms

The paper describes a machine learning model capable of predicting flight prices. The dataset contains flight prices of various airlines between March and June 2019 across different cities. Three machine learning models are implemented, including Linear Regression, Random Forest, and Decision Trees, achieving an accuracy of 80%. [1]

2.2. Flight Price Prediction for Enhanced Recommendations via Machine Learning Web Application

The paper describes a dependable machine learning model that can accurately estimate flight prices. Six different machine learning algorithms are thoroughly evaluated to ensure the model's accuracy and reliability. Performance metrics such as MSE, RMSE etc. are used for comparison. The Random Forest Regressor outperformed the others, achieving the highest R^2 score. [2]

2.3. A Prediction of Flight Fare Using K-Nearest Neighbors

The study identifies the factors driving airplane price fluctuations, how they influence price changes. The study employed the K-Nearest Neighbors algorithm to model the factors. The model was evaluated using various metrics. The results show that the K-Nearest Neighbors algorithm achieved a solid accuracy of 81.77%. [3]

3. Dataset

We have used the **Flight Prediction Dataset** from Kaggle [4] for this project. The dataset has 300,153 records and 11 columns. From the 11 columns, we have 10 features and 1 label. There are no NULL values in the dataset.

The features `duration` and `days_left` are numeric, while the other features are categorical features. The feature `price` is a continuous target value.

#	Column	Count	Null Values	Dtype
0	airline	300153	0	object
1	flight	300153	0	object
2	source_city	300153	0	object
3	departure_time	300153	0	object
4	stops	300153	0	object
5	arrival_time	300153	0	object
6	destination_city	300153	0	object
7	class	300153	0	object
8	duration	300153	0	float64
9	days_left	300153	0	int64
10	price	300153	0	float64

Table 1. Dataset Column Information

Feature Distribution

1. **Airlines:** There are 6 different airlines in our dataset - Vistara, Air India, Indigo, GO_FIRST, AirAsia, and SpiceJet. SpiceJet has maximum count of flights (with 42.6% of the entire distribution) while Air_India has minimum count of flights (with 3% of the entire distribution).

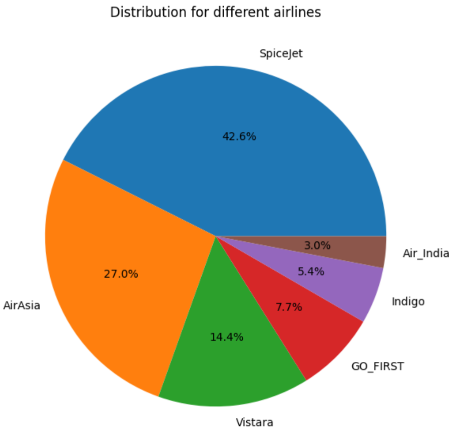


Figure 1. Pie Chart for Airline Distribution

2. **Class:** There are two types of classes - Economy Class and Business Class. The count of Economy class is 206,666 whereas that of Business Class is 93,487.

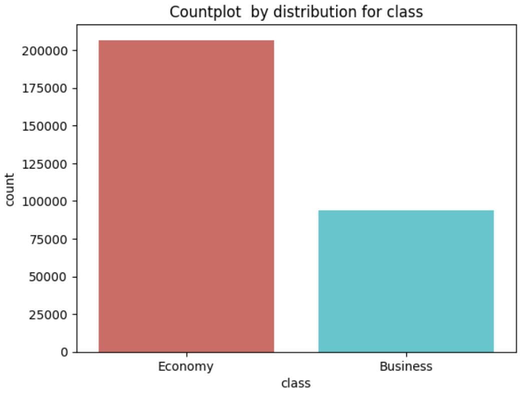


Figure 2. Countplot for Class distribution

3. **Source City and Destination City:** The flights travel to and fro from the following cities: Delhi, Mumbai, Bangalore, Kolkata, Hyderabad, and Chennai. There are maximum departing flights from Delhi (61,343), while there are maximum arriving flights for Mumbai (59,097).

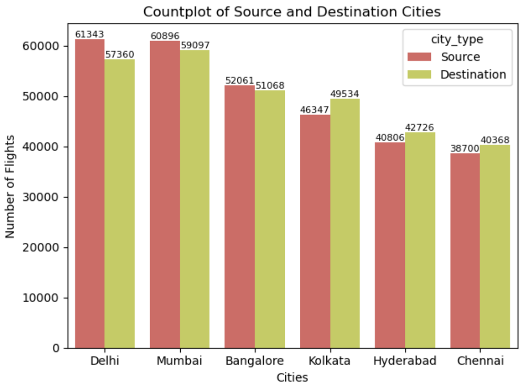


Figure 3. Distribution of Source Cities and Destination Cities

4. **Number of stops:** This feature is a discrete value. Most flights have one stop (250,863) followed by zero stops (36,004), and then two or more (1,286).

EDA Analysis

1.

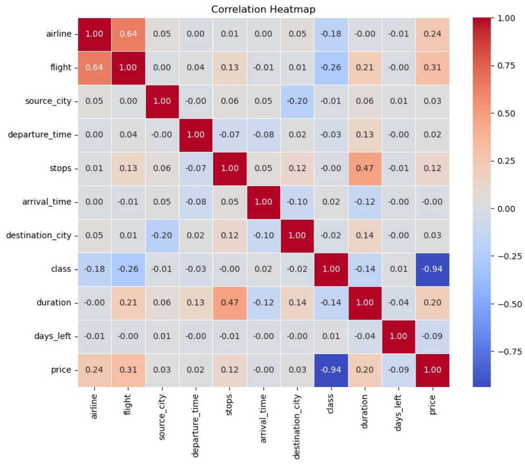


Figure 4. Observation:

The matrix shows that feature `class` and `price` are highly negatively correlated whereas features like `arrival_time`, `departure_time`, `source_city` and `destination_city` do not have significant impact on the flight prices.

2.

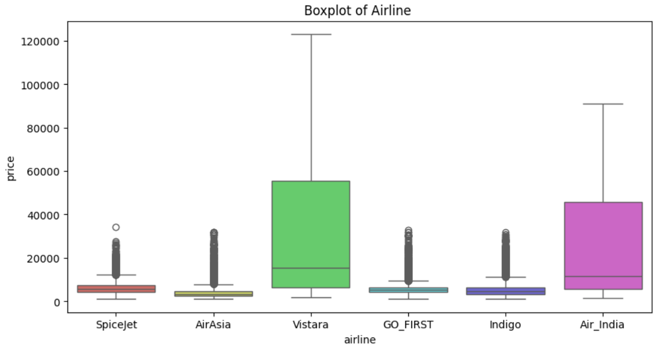


Figure 5. Observation:

From the box plot we observe that Vistara has the highest overall price range of the flights whereas lower-cost airlines like SpiceJet, AirAsia, GO_FIRST and Indigo exhibit lower price ranges. There are many outliers present in these lower-cost airlines.

3.

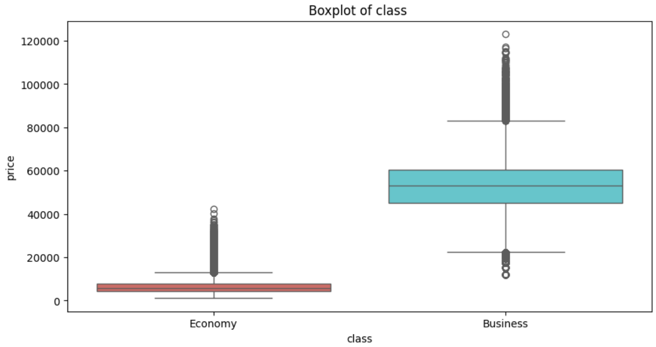


Figure 6. Box plot of Business vs Economy class prices.

The range of prices for Business class is significantly broader, with prices ranging from around 12,000 to above 120,000. Economy class has a much more compressed range, with prices mostly under 45,000.

4.

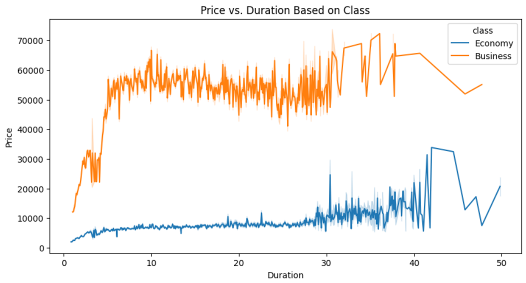


Figure 7. Duration for Economy and Business classes.

Prices for Economy class are relatively low and remain stable across a duration of around 0 to 30 hours. There is a sudden increase in the price around a duration of around 42 hours which may be due to the presence of some outliers in the dataset. Prices for Business class are significantly higher and show considerable fluctuation from around 5 to 30 hours.

5.

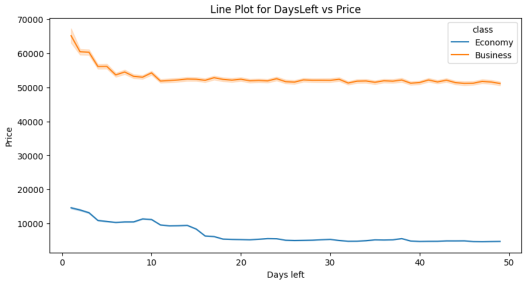


Figure 8. Flight prices vs. days left for departure

The price of both Economy and Business tickets rises significantly as the days left for departure decreases. Business class prices start at a much higher level and spike when the days left are less than 8 whereas Economy class prices start at a lower level and gradually increase as the days left falls below 20.

6.

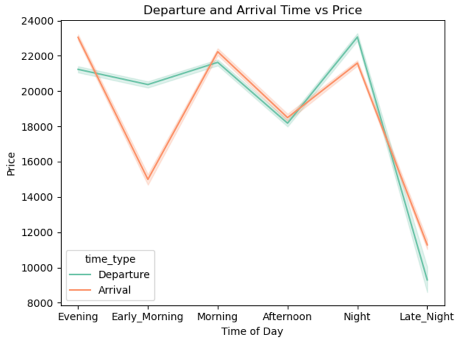


Figure 9. Flight prices based on times.

Flights which arrive in the evening, morning or night have higher prices than those arriving early morning or late night indicating they are more preferred. Similarily, flights which depart in late night are least preferred over others.

Preprocessing

8 out of 10 features are categorical, so we encode them to integers.
From these features, stops, arrival_time, departure_time are ordinal, so we use Label Encoding with the help of manually defined maps to map them to integers. Additionally, we also used maps to map source_city and destination_city to map the same cities in both columns to the same integers. For remaining categorical features like airline, class, and flight, we use label encoding as they are nominal attributes.

After final processing of the dataset, we have converted all features to numerical values, preparing the dataset for for training and analysis by the models.

4. Methodology

We split the data into a 70:30 train-to-test ratio with 210,107 training samples and 90,046 testing samples.

We use regression models because the task involves predicting a continuous target variable - price. Regression models can capture both linear and non-linear relationships, between the target variable and features.

Evaluation Metrics

1. **Mean Squared Error (MSE):** It penalizes the large errors more due to squaring .

2. **Root Mean Squared Error (RMSE):** It expresses the error in the same unit as the target variable, providing easier interpretation.

3. **R-Squared Score:** It helps us understand the proportion of variance in the dependent variable that can be explained by the independent variable.

4. **Adjusted R2 Score:** It adjusts for the inclusion of non-significant predictors, indicating whether adding more predictors enhances the model's performance or not.

5. **Mean Absolute Error (MAE):** It does not penalize the outliers heavily and gives a balanced overview.

Models Used

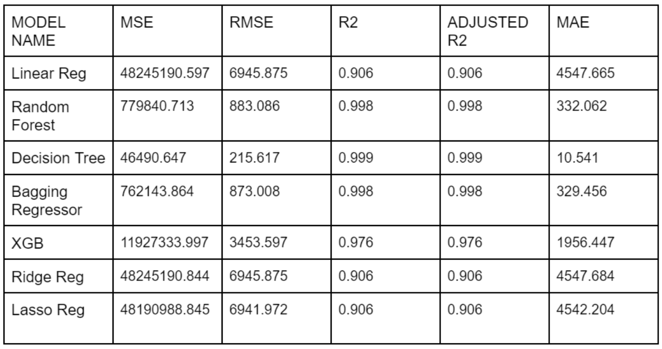
We used a variety of models to capture linear and non-linear relationships in our dataset. The random seed is set to 42 if required in the model to ensure reproducibility of results .
We used **Linear Regression** to explore the basic linear relationship in our dataset. Here, we have used this model to observe the changes in the price of the flights (dependent variable) based on the various other features (independent variables) available in the dataset.
To prevent possible overfitting of models, we used **Lasso Regression** to introduce L1 regularization with a learning rate of 1 and balance the bias-variance tradeoff.
Ridge Regression was also implemented to prevent possible overfitting and rectify multicollinearity problems in regression analysis, we introduced L2 regularization with a learning rate

of 1.
To capture more complex and non-linear relationship between the data, we used **Decision Tree**. It splits data based on the features of the dataset. However, it is prone to overfitting.
An ensemble model that uses Decision Trees as its base model and is resistant to overfitting is **Random Forest**. Hence, to enhance performance, we used Random Forest, which averages the result for 100 decision trees (estimators).
We also tried to improve performance by learning iteratively by correcting previous errors using **XGBoost Regression**. It uses 100 trees (estimators) with learning rate 0.1. The objective used is squared-error.
Another ensemble method used was **Bagging Regression** to improve accuracy as it reduces variance and prevents overfitting of the data. With decision trees as its base estimator, it uses 300 base estimators .

5. Results and Analysis

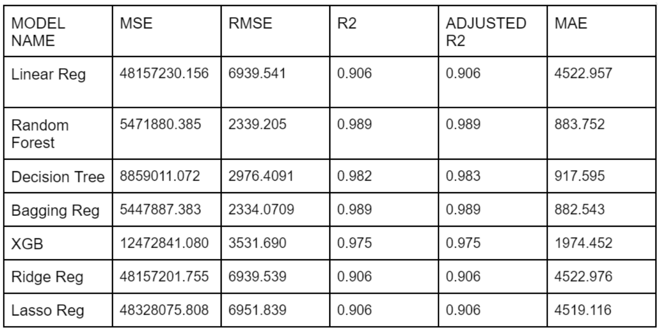
1. **Observation:**

Training errors



MODEL NAME	MSE	RMSE	R2	ADJUSTED R2	MAE
Linear Reg	48245190.597	6945.875	0.906	0.906	4547.665
Random Forest	779840.713	883.086	0.998	0.998	332.062
Decision Tree	46490.647	215.617	0.999	0.999	10.541
Bagging Regressor	762143.864	873.008	0.998	0.998	329.456
XGB	11927333.997	3453.597	0.976	0.976	1956.447
Ridge Reg	48245190.844	6945.875	0.906	0.906	4547.684
Lasso Reg	48190988.845	6941.972	0.906	0.906	4542.204

Testing errors



MODEL NAME	MSE	RMSE	R2	ADJUSTED R2	MAE
Linear Reg	48157230.156	6939.541	0.906	0.906	4522.957
Random Forest	5471880.385	2339.205	0.989	0.989	883.752
Decision Tree	8859011.072	2976.4091	0.982	0.983	917.595
Bagging Reg	5447887.383	2334.0709	0.989	0.989	882.543
XGB	12472841.080	3531.690	0.975	0.975	1974.452
Ridge Reg	48157201.755	6939.539	0.906	0.906	4522.976
Lasso Reg	48328075.808	6951.839	0.906	0.906	4519.116

Figure 10. Model performance metrics (MSE, RMSE, R^2) on testing sets.

- Random Forest and Bagging Regressor consistently perform good on data with lower MSE and RMSE on testing and training data. They are best performers overall with the given evaluation metrics. However, there is a significant gap between training and testing errors.
- Decision tree performs best for training errors with highest R2 score and lower MSE but the same is not true for testing dataset. It shows that model has overfit.
- XGBoost shows higher training and testing error compared to other ensemble methods. However, its performance is very similar in both training and testing error.
- Linear , Ridge and Lasso Regression performs well on the testing set. However, it still shows a relatively high MSE and RMSE compared to more complex models.
- R2 and Adjusted R2 scores are similar in all the models indicating the current number of predictors are optimal.

6. Conclusion

In summary, we observe that the dataset is generally imabalanced, where 4 out of 7 features are imbalanced and only 3 features (`flight`, `source_city` and `destination_city`) are balanced. We also observe that `class` has a significant correlation with the `price` while the correlation between other features and `price` is lower, which shows that `price` is highly dependent on `class`.

Additionally, we observe a significant gap in training and testing errors in models like Decision Tree, Bagging Regressor, and Random Forest. For the remaining models, they show comparable performance in the training and testing dataset. Out of all the models, Random Forest and Bagging Regressor are our best performing models in spite of comparatively higher variance as it performs relatively better on testing dataset compared to other models with lower variance.

Further, we plan to use different sampling techniques (if feasible) to handle the imbalances in the dataset. Next, we will handle outlier detection, aiming to remove any outliers identified through boxplots and the interquartile range. We will also perform hyperparameter tuning to identify the most optimal hyperparameters using grid search or randomized search. We then aim to identify the day for which the price is minimum for the flight with the respect to the booking and departure date.

Member Contribution

All members contributed equally. The tasks listed below are only a representation of the assigned tasks.

- **Aarzoo** - Preprocessing, Model Training
- **Anushka** - Preprocessing, Model Training
- **Nandini** - EDA, Model Training
- **Suhani** - EDA, Model Training

References

[1] “Dynamic flight price prediction using machine learning algorithms,” IEEE Conference Publication — IEEE Xplore, Dec. 16, 2022. Paper 1

[2] “Flight price prediction for enhanced recommendations via machine learning web application,” IEEE Conference Publication — IEEE Xplore, Jul. 10, 2024. Paper 2

[3] “A prediction of flight fare using K-Nearest neighbors,” IEEE Conference Publication — IEEE Xplore, Apr. 28, 2022. Paper 3

[4] ”Flight Price Prediction,” Kaggle. Dataset