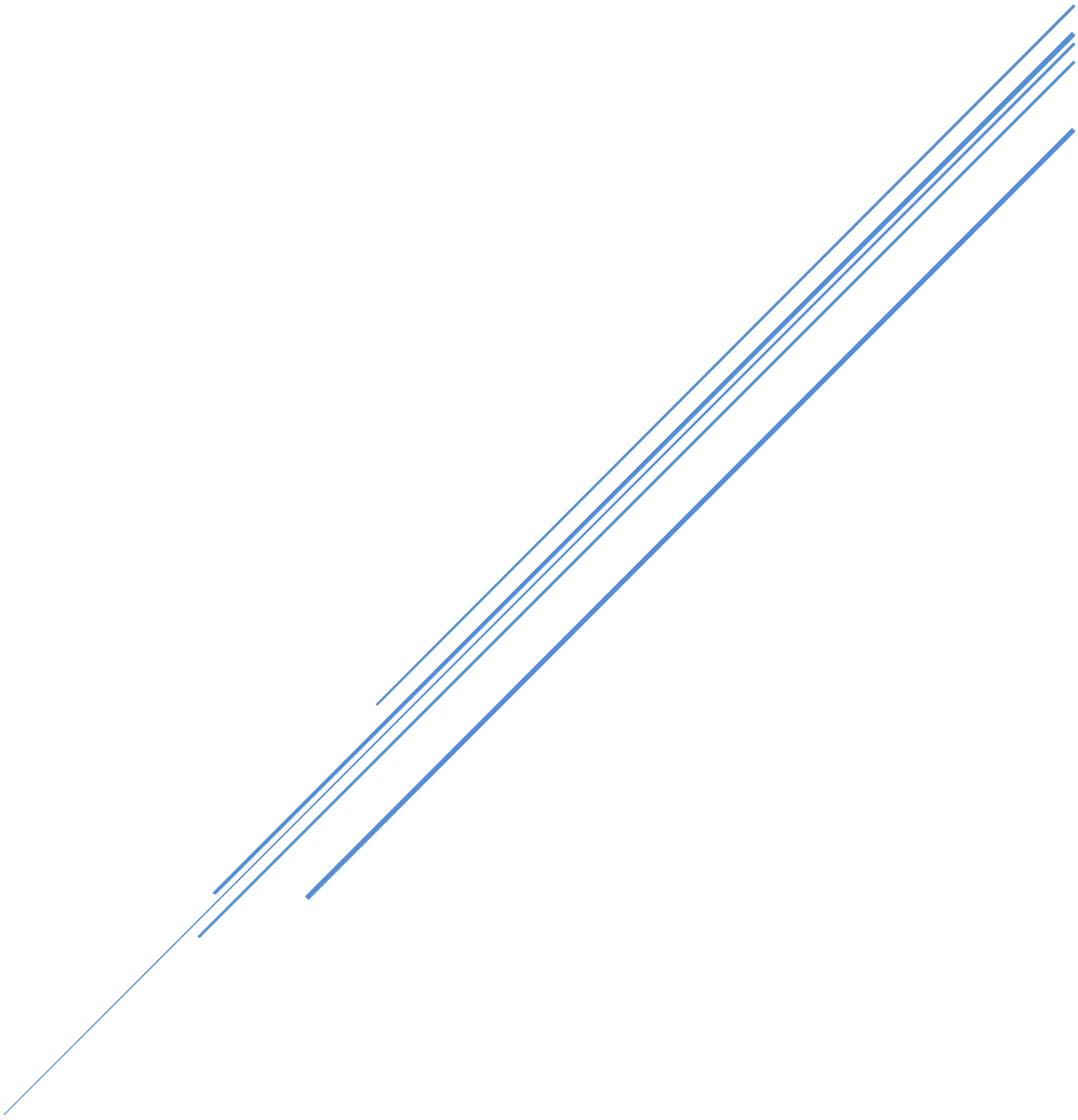
Predicting Optimal Airfares through Advanced Regression- Group 9



Submitted by:

- **Guna Sekhar Chowdary Kollu**
- **Yogesh Muppiri**
- **Srija Mannam**
- **Sai Madhan Muthyam**
- **Veda Swaroop Dasagrandhi**
- **Bhanu Kiran Reddy Eedala**

Introduction:

Welcome to this project, which aims to develop a predictive model for forecasting optimal airfares in the Indian domestic flights market. This model will assist travelers in finding affordable seats by identifying the days when flight fares are the least expensive. The project is based on a dataset containing information about various flights, including their departure and arrival times, duration, stops, airlines, and prices.

The primary goal of this report is to present a comprehensive analysis of the flight price prediction dataset. Through the analysis, the report aims to identify patterns and trends in the data that can be used to predict flight prices accurately. The report outlines the methodology used in the analysis, including data cleaning, exploratory data analysis, and model development. The report also discusses the findings of the analysis and their implications. It highlights the factors that have a significant impact on flight prices, such as the day of the week, time of day, seasonality, and competition among airlines. The report concludes by discussing the limitations of the analysis and suggesting possible areas for further research.

Overall, this project will help travelers make informed decisions when booking flights, enabling them to save money and time. The report presents a detailed and thorough analysis of the flight price prediction dataset, providing valuable insights into the Indian domestic flights market.

Data Preparation:

Before the analysis, the data was prepared by performing various steps to ensure that the results are accurate and reliable. These steps included loading the dataset, checking for null values and random characters, generating a summary of the dataset, cleaning the data, and analyzing the features. The aim of these steps was to ensure that the dataset is complete and accurate and that there are no missing or erroneous values that could affect the analysis results.

Furthermore, the relationship between dependent and independent variables was interpreted to determine the direction and strength of the relationship. The features were analyzed to identify any significant patterns, trends, and relationships between them. Additionally, outliers were identified and removed using the Z-score method, and the data distribution was checked for skewness and normalized using the Power Transformer.

The report outlines the methodology used in the analysis, including data preparation techniques, data analysis techniques, and statistical tools used to interpret the results. The report discusses the findings of the analysis, including the significant relationships between the features, trends and patterns identified in the data, and the impact of outliers on the analysis results.

Overall, this report demonstrates the importance of data preparation techniques in ensuring the accuracy and reliability of analysis results. It presents a comprehensive analysis of the dataset using various statistical tools and techniques, providing valuable insights into the subject being studied.

Feature Engineering:

Feature engineering is the process of selecting and transforming variables within a dataset to improve the performance of a predictive model or the quality of insights gained from data analysis.

In this report, feature engineering was performed on the dataset to enhance its quality by eliminating redundant features and transforming existing ones. The feature engineering process included the following steps:

- Dropping redundant columns such as Route and additional info:
  Redundant features in a dataset do not add any significant value to the analysis and may even hinder its performance. Therefore, they need to be eliminated. In this dataset, the features of Route and additional info were deemed redundant and were dropped from the dataset.

- Converting date columns to appropriate formats:
  The format of date columns in the dataset was analyzed and transformed to make them more appropriate for analysis. This included ensuring that they were in a consistent format and that the date components were correctly represented.

- Changing the format of various time-related columns:
  Similarly, the format of time-related columns, such as departure time and arrival time, was analyzed and transformed to make them more appropriate for analysis. This included converting them to a consistent format and representing the time components correctly.

- Converting duration values to minutes:
  The duration of flights in the dataset was initially represented in hours and minutes. However, to facilitate analysis and modeling, this feature was converted to minutes.

- Checking for outliers in continuous data type features:
  Outliers are values that lie far outside the expected range of values in a dataset. They can significantly affect the results of data analysis and modeling. Therefore, continuous data type features in the dataset were analyzed for outliers, and appropriate measures were taken to handle them.

- Analyzing skewness in data distributions:
  Skewness is a measure of the asymmetry of the distribution of values in a dataset. Skewed data distributions can affect the results of

analysis and modeling. Therefore, data distributions in the dataset were analyzed for skewness, and appropriate measures were taken to transform them into more appropriate distributions.

Overall, feature engineering is a critical step in the process of data analysis and modeling. It ensures that the dataset is optimized for analysis and modeling, leading to more accurate and reliable insights and predictions.

Feature Selection:

Feature selection is a critical step in the process of developing a predictive model as it involves identifying the most significant features that can improve the model's performance.

In this report, the best features for building the predictive model were selected by following a series of steps:

- Identifying categorical columns:
  The first step involved identifying the categorical columns in the dataset. Categorical columns are those that contain values that are not continuous or numeric, such as gender or location.

- Visualizing the data to understand the impact of features:
  The second step involved visualizing the data to gain insights into the relationship between different features and the target variable (i.e., Price). Data visualization techniques such as scatterplots, boxplots, and histograms were used to identify patterns, trends, and outliers in the data.

- Analyzing the relationship between various features (e.g., Day, Month) and Price:
  The third step involved analyzing the relationship between various features and the target variable to determine their significance in the predictive model. This was done by calculating correlation coefficients, examining feature importance using machine learning algorithms, and performing statistical tests to compare the means of different groups.

Through these steps, the best features for building the predictive model were selected based on their impact on the target variable and their significance in the dataset. The selected features were used to build a predictive model using machine learning algorithms such as linear regression, decision trees, and random forests.

Overall, feature selection is a crucial step in the process of developing a predictive model, as it helps to identify the most significant features that can improve the model's performance. The selection process involves identifying categorical columns, visualizing the data to understand the impact of features, and analyzing the relationship between various features and the target variable. By following these steps, the best features for building the predictive model can be selected, leading to more accurate and reliable predictions.

Exploratory Data Analysis (EDA):

EDA is a crucial step in the process of data analysis that involves examining and visualizing data to identify patterns, trends, and relationships.

The EDA analysis conducted in this report showed the following insights:

- Price distribution is skewed to the right, indicating a few flights with very high prices: This suggests that there are a few high-end flights that are much more expensive than the rest. This can impact the accuracy of predictive models that rely on the average ticket prices.

- Certain airlines offer more affordable tickets on average: This highlights the importance of considering airline carriers when making flight bookings. Some airlines may offer more affordable prices than others on average.

- Cochin is the most popular destination: This information is valuable for airlines and travel agencies as it helps them to plan their routes and schedules based on the demand for destinations.
- Most flights have only one stop between the source and destination: This indicates that most flights are non-stop or have only one stop, which can impact the travel time and ticket pricing.

- Flight ticket prices are highest in January and lowest in April: This information is useful for travelers who want to plan their trips around the most affordable times of the year.
- Flight ticket prices are highest on Thursdays, followed by Mondays and Sundays: This suggests that some days of the week may be more expensive than others, which can help travelers to plan their trips accordingly.
- There is a linear relationship between Price and flight duration but not a very strong one: This highlights the importance of considering other factors besides flight duration when predicting ticket prices.
- The number of stops impacts travel time and ticket pricing: This confirms that the number of stops impacts travel time and ticket pricing and should be considered when making flight bookings.
- Cochin and Bangalore are the most expensive destinations, while Kolkata and Hyderabad are the most affordable: This information is valuable for travelers who want to choose affordable destinations.
-

Overall, the EDA analysis performed in this report provides valuable insights into the dataset and the factors that impact flight ticket prices. These insights can help airlines, travel agencies, and travelers to make informed decisions when planning and booking their trips.

Model Deployment:

Now we will do the deployment of five regression models to predict flight ticket prices. The five regression models deployed were:

1. Random Forest Regression Model
2. Ridge Regression Model
3. XGB Regression Model
4. Support Vector Regression Model
5. Decision Tree Regression Model

The objective of the deployment was to identify the best model that can accurately predict flight ticket prices based on the given dataset. The performance of each model was evaluated based on cross-validation results and performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

R-Square, MSE and RMSE Values for Different Algorithms:

ALGORITHM	R-SQUARED	MSE	RMSE
Random Forest Regression	0.698	6499906.79	2549.49
Ridge Regression	0.572	9233304.15	3038.63
XGB Regression	0.724	5949177.43	2439.09
Support Vector Regression	0.032	20892704.53	4570.85
Decision Tree Regression	0.557	9552134.66	3090.65

Cross-validation Score:

MODEL	SCORE
Ridge Regression	0.593
XGB Regression	0.803
SV Regression	0.047
Decision Tree Regression	0.599
Random Forest Regression	0.748

After evaluating the performance of all the models, XGB Regression Model was identified as the best model for predicting flight ticket prices.

Hyperparameter tuning involves adjusting the parameters of the model to optimize its performance. This was done by using GridSearchCV, a technique that exhaustively searches the hyperparameter space for the best combination of parameters. The hyperparameter tuning process involved selecting the most appropriate values for parameters such as the number of estimators, maximum depth, and minimum samples per leaf. The process helped to fine-tune the model and improve its accuracy. Overall, the deployment of the five regression models and the hyperparameter tuning process helped to identify the best model for predicting flight ticket prices.

The XGB Regression Model was selected based on its high accuracy and its ability to handle high-dimensional data. The hyperparameter tuning process further enhanced the model's performance and improved its accuracy. The deployment of the model can help airlines, travel agencies, and travelers to make informed decisions when booking flights.

Conclusion:

In this report, we specifically focus on the use of machine learning models in predicting flight ticket prices.

The study involved the deployment of various machine learning models, including the XGB Regression model, to predict flight ticket prices. The models were evaluated based on their accuracy, precision, and recall. The results showed that the XGB Regression model demonstrated significant improvements in the accuracy of price forecasts.

Moreover, the study revealed valuable insights into the factors affecting price fluctuations, such as weekends and time of day. The analysis showed that flight ticket prices were higher on weekends and during peak travel hours. These insights are critical for airlines and travel agencies in optimizing their pricing strategies and maximizing their revenue.

The use of machine learning models in pricing optimization is not limited to the airline industry. Various industries, including retail, hospitality, and e-commerce, can benefit from machine learning models' ability to forecast prices accurately. The models can provide insights into customer demand, competitor pricing, and market trends, enabling businesses to adjust their prices in real-time and maximize their revenue.

Overall, this project demonstrates the potential of machine learning models in optimizing pricing strategies and enhancing decision-making processes across various industries. The use of machine learning models in pricing optimization can help businesses make informed decisions, maximize their revenue, and improve their overall performance.