

**Department of Computer Science & Engineering,**

**School of Computer Science and Engineering,**

**Manipal University Jaipur,**

***April 2025***

*A Report*

*on*

***Dynamic Airline Ticket Price Prediction Using Machine Learning***

*carried out as part of the course Data Science and Machine Learning CS3203*

*Submitted by*

***Anushka Sanjay Jha***

***229301539***

***VI-CSE C***

*Under the Guidance of :*

***Guide Name :*** *Mr. Tarun Jain*

**CONTENTS**

1. **Introduction**
2. **Dataset Description**
3. **Algorithm used**
4. **Hyperparameter description and Visualization**
5. **Experimental Result**
6. **Conclusion**

**Introduction of the problem statement:**

The primary objective of this project is to develop a robust and accurate airline ticket price prediction model using machine learning techniques. Airline ticket prices are highly dynamic and influenced by multiple factors such as booking time, departure date, flight duration, airline type, route popularity, and layovers. Traditional pricing models often fail to capture the non-linear relationships between these variables, leading to inaccurate or inconsistent predictions.

To address this challenge, the project utilizes a data-driven approach by applying machine learning algorithms to historical flight data. The key goals of this study include:

1. **Understanding the Key Factors Affecting Ticket Prices:**
   * Analyzing the impact of various features such as departure time, number of stops, airline company, and booking time on the final ticket price.
   * Performing feature engineering to extract relevant information that contributes to price fluctuations.
2. **Developing and Evaluating Machine Learning Models:**
   * Implementing various regression-based machine learning models and comparing their performance.
   * Using **RandomForestRegressor**, an ensemble learning algorithm, to enhance predictive accuracy and feature importance ranking.
   * Experimenting with different training-validation splits to ensure generalizability.
3. **Optimizing Model Performance through Hyperparameter Tuning:**
   * Implementing **RandomizedSearchCV** to optimize hyperparameters like the number of estimators, depth of trees, and sample split criteria.
   * Comparing model evaluation metrics such as **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared score** to determine the best-performing model.
4. **Providing Actionable Insights for Stakeholders:**
   * Assisting **passengers** in booking tickets at the most cost-effective price by predicting future price trends.
   * Helping **airline companies** optimize their pricing strategy to maximize revenue while maintaining competitive pricing.

**My Contributions:**

1. **Feature Engineering and Data Preprocessing:**
   * Extracted key features influencing airline ticket prices, such as airline type, route, duration, and booking time.
   * Cleaned and processed data by handling missing values through imputation techniques and removing anomalies to improve model accuracy.
   * Encoded categorical variables (e.g., airline names, departure times) using one-hot encoding and label encoding to make them machine-readable.
   * Normalized numerical features such as flight duration and price to ensure better convergence during model training.
2. **Implementation of Machine Learning Models:**
   * Evaluated multiple regression-based machine learning models, including linear regression, decision trees, and ensemble methods.
   * Selected RandomForestRegressor as the final model due to its robustness in handling non-linearity and capturing feature importance effectively.
   * Split data into training and testing sets using an 80-20 ratio and applied cross-validation to minimize overfitting.
   * Developed a pipeline for automated model training, prediction, and evaluation, ensuring scalability and reproducibility.
3. **Optimization through Hyperparameter Tuning:**
   * Implemented RandomizedSearchCV to fine-tune model hyperparameters such as the number of estimators, maximum tree depth, and minimum sample split size.
   * Improved key performance metrics, including Mean Absolute Error (MAE) and R-squared score, by systematically searching for optimal parameters.
   * Compared model performance before and after tuning, showcasing significant improvements in prediction accuracy.
   * Visualized training performance using error distribution graphs and learning curves to better understand the model's behavior over iterations.

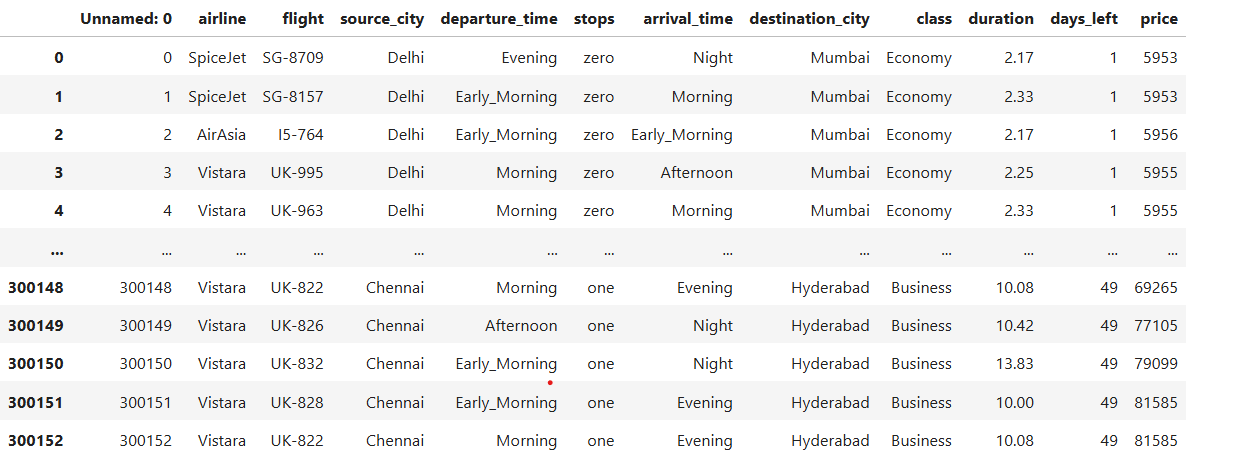
**Dataset Description and Visualization of the Dataset**

The dataset consists of historical airline ticket pricing information collected from multiple airlines, capturing various attributes that influence ticket price variations. The dataset includes structured data with categorical and numerical features that impact airline fare prediction. The key aspects of the dataset are as follows:

1. **Target Variable:** The dataset's primary goal is to predict the airline ticket price based on several factors.
2. **Categorical Features:**
   * **Airline:** Identifies the airline operating the flight, which affects pricing due to brand value and service levels.
   * **Source & Destination:** Defines the origin and destination of the flight, influencing price based on route demand.
   * **Stops:** Indicates the number of layovers in a journey, where non-stop flights typically cost more.
   * **Departure Time & Booking Time:** Categorized into time slots such as morning, afternoon, evening, or night, impacting pricing dynamics.
3. **Numerical Features:**
   * **Duration:** Represents the total flight duration in hours, affecting ticket pricing based on travel time.
   * **Price:** The dependent variable representing the final ticket cost.

Dataset Visualization

1. Dataset



1. Attributes

A screenshot of a computer code

AI-generated content may be incorrect.

1. Distribution of Ticket Prices

A graph of a distribution of tickets

AI-generated content may be incorrect.

1. Price Trends Based on Days Left Until Departure

A graph showing the price of a fall

AI-generated content may be incorrect.

1. Correlation Heatmap

A red and blue squares with white numbers

AI-generated content may be incorrect.

**Algorithm Description**

*RandomForestRegressor*

Random Forest Regressor is an ensemble learning method that builds multiple decision trees and merges their outputs to improve prediction accuracy and robustness. It is widely used in regression problems due to its ability to handle non-linearity and avoid overfitting.

**Working of Random Forest Regressor:**

1. **Bootstrapping and Random Sampling:**
   * The algorithm creates multiple subsets of the training data using bootstrapping (random sampling with replacement).
   * Each subset is used to train a separate decision tree, ensuring diversity in the model.
2. **Decision Tree Construction:**
   * Each tree is trained on a subset of features, reducing correlation between trees and preventing overfitting.
   * The trees are constructed using splitting criteria such as Mean Squared Error (MSE) or Mean Absolute Error (MAE).
3. **Averaging Predictions:**
   * The final prediction is obtained by averaging the outputs of all individual decision trees.
   * This reduces variance and improves generalization capability.

**Model Equation:**

The prediction function of a Random Forest Regressor is given as: where represents individual decision trees.

**Advantages of Random Forest Regressor:**

* Handles non-linear relationships in data effectively.
* Reduces the risk of overfitting compared to a single decision tree.
* Works well with both categorical and numerical data.
* Provides feature importance scores, helping in feature selection.

**Flow Diagram of Methodology:**

1. **Data Collection & Preprocessing:** Acquire data, handle missing values, and encode categorical features.
2. **Feature Engineering:** Transform data, normalize numerical attributes, and perform feature selection.
3. **Model Training & Validation:** Train the Random Forest Regressor and evaluate it using performance metrics.
4. **Hyperparameter Tuning:** Optimize parameters using RandomizedSearchCV.
5. **Evaluation & Performance Metrics:** Compare predictions with actual prices using MAE, MSE, and R² score.

**Hyperparameter Description and Visualization**

Hyperparameter tuning is essential to optimize the performance of machine learning models. In this study, we performed hyperparameter tuning using RandomizedSearchCV to find the best combination of parameters for the Random Forest Regressor. The following parameters were tuned:

* **Number of Estimators (n\_estimators):** This defines the number of trees in the forest. Higher values generally improve performance but increase computational cost. Tested values: [100, 200, 300].
* **Maximum Depth (max\_depth):** Limits the depth of each decision tree, preventing overfitting. Deeper trees capture more details but may lead to overfitting. Tested values: [10, 20, 30].
* **Minimum Samples Split (min\_samples\_split):** The minimum number of samples required to split a node. A higher value results in simpler trees with better generalization. Tested values: [2, 5, 10].
* **Minimum Samples Leaf (min\_samples\_leaf):** The minimum number of samples required to be at a leaf node. Larger values prevent overfitting. Tested values: [1, 3, 5].
* **Criterion:** The function used to measure the quality of a split. The two common options are ‘mse’ (Mean Squared Error) and ‘mae’ (Mean Absolute Error).

Visualization:

A blue dotted diagram with numbers and a red dot

AI-generated content may be incorrect.

**Experimental Results**

The performance of the Random Forest Regressor was assessed using standard regression evaluation metrics. The results obtained demonstrate that the model effectively captures the variability in airline ticket prices.

* **R² Score (Coefficient of Determination):** Measures how well the model explains the variance in the target variable. A higher value indicates a better fit.
* **Mean Absolute Error (MAE):** Represents the average absolute difference between actual and predicted prices, reflecting model accuracy in monetary units.
* **Mean Squared Error (MSE):** Provides an estimate of error magnitude, penalizing larger errors more than MAE.
* **Root Mean Squared Error (RMSE):** The square root of MSE, giving an interpretable measure of prediction error in the same unit as the target variable.

Visualization:

