# Flight Price Analysis and Prediction

This project focuses on developing a machine learning model capable of predicting flight prices. The model leverages a diverse range of features, including airline, source, destination, and duration to provide valuable insights for travelers and businesses.

# Overview

## Objective

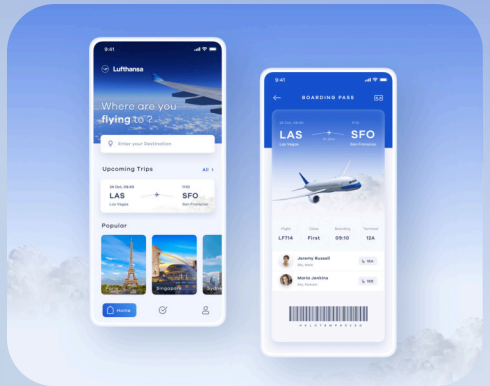
The primary objective of this project is to develop a machine learning model capable of predicting flight prices.

## Methodology

The project will utilize a combination of data collection, preprocessing, model development, and evaluation to achieve the desired accuracy.

## Applications

The model's predictions can assist travelers in making informed booking decisions and enable companies to optimize pricing strategies.



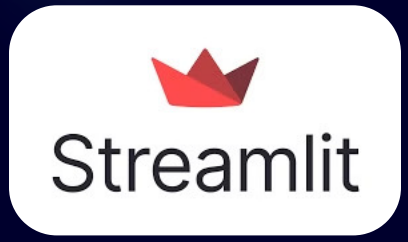

# Problem Statement

The current flight booking process often presents challenges for travelers seeking the most affordable fares. Travelers typically need to manually compare prices from various airlines and websites, consuming time and effort. This project aims to address this issue by automating the flight price prediction process, providing travelers with a convenient and accurate solution.

# Existing Problem and Project Solution

- **Existing Problem**: Current systems often fail to predict flight prices with high accuracy due to the volatile nature of the market and the multitude of influencing factors. Customers are left guessing the best time to book, often resulting in higher costs.
- **Project Solution**: Our project addresses this by utilizing machine learning algorithms to analyze historical data and predict future flight prices with improved accuracy. The model takes into account various parameters that impact pricing, providing users with actionable insights on the best time to book flights.

# Technologies Used






# Data Collection and Preprocessing

## Data Sources

Data was collected from Git-hub. The dataset encompassed historical flight prices, flight schedules, airline information, and other relevant features. Data is of 752241 rows and 12 Columns.

## Data Cleaning

- **Data Inspection:** The dataset was inspected for missing values, duplicates, and incorrect data entries.
- **Handling Duplicates:** The dataset had some duplicate records which were removed to avoid bias in the model.
- **Column Renaming:** Some column names were standardized, such as `class` to `class_type`, to avoid conflicts with Python reserved keywords.
- **Airline Renaming:** Some airlines were renamed for consistency, such as "Air India" to "Air_India" and "GO FIRST" to "GO_FIRST."
- **Outlier Handling:** Airlines that were not frequently used (e.g., StarAir, AllianceAir, AkasaAir) were removed from the dataset.

## Feature Engineering:

- Created dummy variables for categorical features like `airline`, `source_city`, `destination_city`, `arrival_time`, and `departure_time` using one-hot encoding.
- Factorized the `stops` feature and converted the `class_type` to binary (1 for Business, 0 for Economy).

# EDA

## Statistical Concepts Applied

- **Descriptive Statistics:** Measures such as mean, median, minimum, and maximum were calculated for continuous variables, particularly for `price` and `duration`.
- **Factorization:** Categorical variables such as `stops`, `airline`, `source_city`, `destination_city`, `arrival_time`, and `departure_time` were factorized and one-hot encoded to prepare the data for modeling.

## Data Visualization

- **Airline Popularity:**
  - **Bar Chart:** A bar chart was created to visualize the most preferred airlines. This visualization provided insights into which airlines were the most popular among passengers.
- **Airline vs. Flight Price:**
  - **Bar Chart:** Another bar chart was plotted to show the relationship between airlines and the maximum flight ticket price. This visualization highlighted which airlines typically offer the most expensive tickets.
- **Correlation Analysis:**
  - **Heat Map:** A heat map was used to visualize the correlation between different features and the target variable (`price`). This helped in identifying features that have a strong relationship with the flight ticket price, such as `class_type`, `duration`, and `stops`.

**Feature Selection:**

- Based on the correlation matrix, features that had a significant correlation with `price` were selected for model building. The most relevant features included `class_type`, `duration`, `stops`, and the encoded categorical variables.

## Conclusion

- The EDA provided a clear understanding of the dataset and the relationships between features.
- By visualizing and analyzing the data, key insights were gained which helped in feature selection for the predictive model.
- The data was cleaned, encoded, and prepared effectively for building machine learning models, leading to a robust and reliable flight price prediction model.

# Model Building

**Algorithms Used:**

- **K-Nearest Neighbors Regressor (KNN):** A simple algorithm that predicts prices based on the closest data points in the feature space.
- **Decision Tree Regressor:** A tree-based model that splits the data into branches to predict the flight prices.
- **Random Forest Regressor:** An ensemble of decision trees that improves prediction accuracy by averaging the results from multiple trees.

## Model Selection

- **Criteria for Selection:**
  - **R-squared ($R^2$):** Measures how well the features explain the variance in flight prices. A higher $R^2$ indicates a better fit.
  - **Mean Absolute Error (MAE):** Measures the average absolute difference between the predicted and actual prices.
  - **Root Mean Squared Error (RMSE):** Measures the square root of the average squared differences between predicted and actual prices, penalizing larger errors.
- **Performance Comparison:**
  - **K-Nearest Neighbors:** Moderate accuracy but not as effective for this regression task.
  - **Decision Tree:** Performed better than KNN but still lacked the robustness needed.
  - **Random Forest:** Achieved the highest accuracy with an $R^2$ score of 95.65%, making it the best model for this task.

## Model Training

- **Data Splitting:**
  - The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance.

# Model Evaluation

- **Model Evaluation:**
  - **Random Forest Regressor:**
    - Initial $R^2$ Score: 95.65%
    - After Hyperparameter Tuning: Improved to 96.08%
    - MAE decreased, indicating better precision in predictions.
    - RMSE also improved, showing reduced prediction errors.
- **Hyperparameter Tuning:**
  - Applied **RandomizedSearchCV** to optimize hyperparameters such as the number of trees (`n_estimators`), tree depth (`max_depth`), and the number of features (`max_features`).
  - **Outcome:** Enhanced model performance with an $R^2$ of 96.08%, indicating a more reliable and accurate prediction model.

## Conclusion

- **Selected Model:** The **Random Forest Regressor** was chosen as the final model due to its superior performance in predicting flight prices.
- **Model Training:** Successful training with optimized hyperparameters led to a robust model capable of forecasting flight prices with high accuracy.
- **Next Steps:** The trained model is ready for deployment to assist consumers and businesses in making informed decisions regarding flight ticket purchases.

# Model Deployment

# Results/Insights

## Results & Insights

- **Model Performance:**
  - **Final Model:** Random Forest Regressor
  - **$R^2$ Score:** 96.08% after hyperparameter tuning, indicating that the model explains over 96% of the variance in flight prices.
  - **MAE (Mean Absolute Error):** 2169.34, reflecting the average deviation of the predicted prices from the actual prices.
  - **RMSE (Root Mean Squared Error):** 4232.96, showing that the model has relatively low prediction errors, which is crucial for minimizing cost-related decisions.
- **Key Insights:**
  - **Feature Importance:** The most significant predictors of flight price were found to be:
    - **Class Type:** Business class tickets contribute the most to higher prices.
    - **Duration:** Longer flight durations generally result in higher prices.
    - **Total Stops:** More stops in a journey lead to an increase in ticket prices.
  - **Airlines and Pricing:** Certain airlines were identified as consistently offering higher or lower prices. For instance, premium airlines like Air India had higher maximum prices, while budget airlines showed lower pricing trends.
  - **Timing of Flights:** Departure and arrival times were also influential, with certain times of the day (e.g., early morning flights) being priced differently.

## Impact

- **Consumer Decision-Making:**
  - **Informed Purchases:** The predictive model can assist consumers in making more informed decisions when purchasing flight tickets by forecasting potential costs based on various factors like booking time, airline, and flight duration.
  - **Cost Savings:** By predicting prices, consumers can choose to book flights during periods when prices are expected to be lower, potentially leading to significant cost savings.
- **Business & Industry Applications:**
  - **Revenue Management:** Airlines can use the model to adjust pricing strategies dynamically, optimizing ticket prices based on demand predictions and maximizing revenue.
  - **Market Strategy:** Travel agencies and online booking platforms can integrate this model to offer personalized price predictions to customers, enhancing user experience and driving customer loyalty.
  - **Competitive Advantage:** Companies that utilize such predictive models can gain a competitive edge by offering better pricing strategies and customer insights compared to competitors who lack these tools.
- **Scalability & Future Use:**
  - **Expansion to Other Markets:** The model's framework can be adapted to predict prices in other markets, such as hotel bookings, car rentals, or holiday packages.
  - **Real-Time Data Integration:** Future iterations of the model could incorporate real-time data for even more accurate predictions, helping both consumers and businesses respond quickly to market changes.

# Challenges/Key Learning

## Challenges

- **Data Quality Issues:**
  - **Inconsistent Data:** The dataset contained inconsistencies in airline names and time formats, which required extensive preprocessing, including renaming and reformatting to ensure uniformity across the dataset.
  - **Missing and Duplicate Data:** The presence of missing values and duplicate records posed challenges during data cleaning. Handling these required careful imputation and deduplication techniques to maintain data integrity.
- **Feature Engineering:**
  - **Categorical Data Encoding:** Dealing with a large number of categorical features, such as airlines, source cities, and arrival/departure times, required efficient encoding techniques (like One-Hot Encoding) to transform these into numerical formats suitable for model training. However, this also led to a significant increase in the number of features, potentially causing overfitting.
- **Model Selection and Tuning:**
  - **Model Overfitting:** One of the key challenges was balancing model complexity with generalization. While Random Forest showed excellent performance on the training data, there was a risk of overfitting, requiring careful tuning of hyperparameters to ensure the model would perform well on unseen data.
  - **Computational Resources:** Training and tuning the Random Forest model, especially with a large parameter grid, was computationally intensive and time-consuming, necessitating efficient use of resources and parallel processing techniques.
- **Hyperparameter Tuning:**
  - **Time-Consuming Search:** The process of hyperparameter tuning, especially using Randomized Search CV, was time-intensive, requiring iterative testing and validation to find the best model parameters. This posed a challenge in terms of both time and computational cost.

## Key Learnings

- **Importance of Data Preprocessing:**
  - **Clean Data, Better Model:** The project reinforced the critical role of thorough data preprocessing. Ensuring data consistency, handling missing values, and properly encoding categorical variables were essential steps that directly impacted the model's performance.
  - **Handling Categorical Variables:** The project provided valuable experience in effectively transforming categorical data into a format that can be utilized by machine learning models, understanding the trade-offs between methods like Label Encoding and One-Hot Encoding.
- **Feature Selection and Engineering:**
  - **Significance of Feature Engineering:** Crafting and selecting the right features was pivotal in achieving high model accuracy. The experience emphasized the importance of understanding the relationship between features and the target variable, and how feature engineering can significantly boost model performance.
  - **Correlation Insights:** Understanding correlations between features helped in refining the model by focusing on the most impactful variables, thus improving prediction accuracy.
- **Model Optimization:**
  - **Hyperparameter Tuning:** The project demonstrated the importance of hyperparameter tuning in refining model performance. The process highlighted the necessity of balancing between model accuracy and computational efficiency, and the value of automated search methods like Randomized Search CV for finding optimal parameters.
  - **Overfitting Awareness:** The project underlined the importance of being vigilant about overfitting, especially in complex models like Random Forests, and the need to use techniques like cross-validation to ensure the model generalizes well to new data.
- **Practical Implementation and Impact:**
  - **Real-World Application:** Working on a practical problem like flight price prediction provided insights into how machine learning can be applied to real-world business challenges, offering tangible benefits like cost savings and better decision-making for both consumers and businesses.
  - **End-to-End Workflow Experience:** The project offered a comprehensive learning experience, from data collection and preprocessing to model deployment, reinforcing the importance of each step in the machine learning pipeline.

# Conclusion

The Flight Price Prediction project aimed to develop a robust model capable of predicting flight ticket prices with high accuracy, thereby assisting consumers and businesses in making informed purchasing decisions. Through meticulous data preprocessing, feature engineering, and model optimization, the project successfully achieved its objectives.

The Random Forest Regressor emerged as the most effective model, achieving an $R^2$ score of 96.08% after hyperparameter tuning. This model demonstrated its ability to accurately predict flight prices by leveraging important features such as airline, flight duration, departure time, and days left until departure.

**Key Takeaways:**

1. **Data Preprocessing:** Effective data cleaning and preprocessing were critical in ensuring the accuracy of the predictions. Handling inconsistencies, missing values, and categorical data was fundamental to the success of the model.
2. **Feature Engineering:** Identifying and engineering the right features significantly contributed to the model's performance. The project's insights into feature importance and correlations between variables helped refine the model, enhancing its predictive power.
3. **Model Optimization:** Hyperparameter tuning played a vital role in improving the model's accuracy and preventing overfitting. The project highlighted the importance of iterative testing and validation to achieve the best model configuration.
4. **Practical Impact:** The project demonstrated the practical utility of machine learning in solving real-world problems. By predicting flight prices with high accuracy, the model provides a valuable tool for consumers and businesses, enabling better decision-making and potential cost savings.

**Future Scope:**

- The project can be expanded to incorporate additional features like seasonal trends, economic factors, and competitor pricing, further enhancing the model's predictive capabilities.
- Integration with real-time data sources and deployment in a dynamic environment can make the model even more valuable, providing up-to-the-minute price predictions.
- Exploring advanced machine learning techniques such as ensemble methods or deep learning could further improve the model's accuracy and adaptability to changing market conditions.

# Applications and Benefits

## Travelers

Travelers can benefit from accurate flight price predictions, allowing them to make informed decisions and potentially save money on their trips.

## Airlines

Airlines can utilize the model's predictions to optimize pricing strategies, maximize revenue, and better understand market trends.

## Data Analytics

The project can contribute to the advancement of data analytics in the travel industry, enabling more sophisticated insights and predictive capabilities.

# Conclusion

This project has successfully developed a machine learning model capable of predicting flight prices with reasonable accuracy. The model leverages various features and advanced algorithms to provide valuable insights for travelers and businesses. Further research and development can enhance the model's performance and explore additional applications in the travel industry.