Proposal

# **Machine Learning - Group Project Proposal**

Team 12B

(Dennis Wu, Kenjee Koh, Vy Nguyen, Cyan Huang, Becky Wang)

[dennisw8@uci.edu](mailto:dennisw8@uci.edu), [ckkoh@uci.edu](mailto:ckkoh@uci.edu), [vynn7@uci.edu](mailto:vynn7@uci.edu), [hsianghh@uci.edu](mailto:hsianghh@uci.edu), [beckyw5@uci.edu](mailto:beckyw5@uci.edu)

# **Introduction**

In the airline industry, airline fares fluctuate based on multiple factors, from booking time to seasonal demand variations. Our team will try to build a model that will forecast airline ticket prices, taking into account features like flight details and demand trends. This project will provide insights into various influences on fare pricing, enabling a better understanding of pricing dynamics in the airline market.

# **Project Objectives**

The aim of this project is to understand how airlines make pricing decisions depending on the contexts and answer the following questions: 1) Which features have the most impact on the pricing model? 2) Does the time of booking or flight departure significantly impact fare? 3) How do route characteristics (e.g., Origin Destination, Source) influence fare?

# **Data Sources**

To collect data regarding Airline fares, we will be leveraging the Amadeus Flight Offers Search API. This API will provide airline fares meeting the search criteria that we set (e.g. destination, source, # of passengers, etc) from 250+ airlines (however we will only be focusing on multiple fares from a singular airline).

To generate the dataset that we will be using to train and test our models, we will make a wide variety of API calls with randomized features (e.g. different destinations, different sources, different $ of passengers, etc) for one specific airline. This will then output fare offers from that airline with additional details such as number of bookable seats, pricing options, and so on.

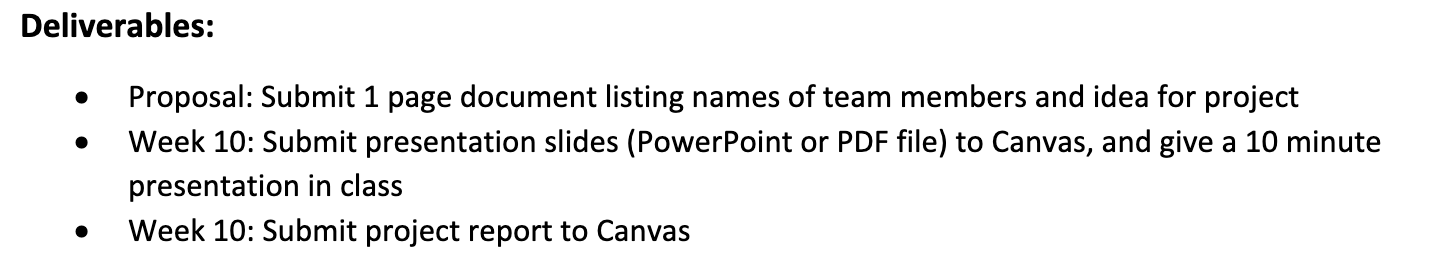
# **Potential Analytical Approach**

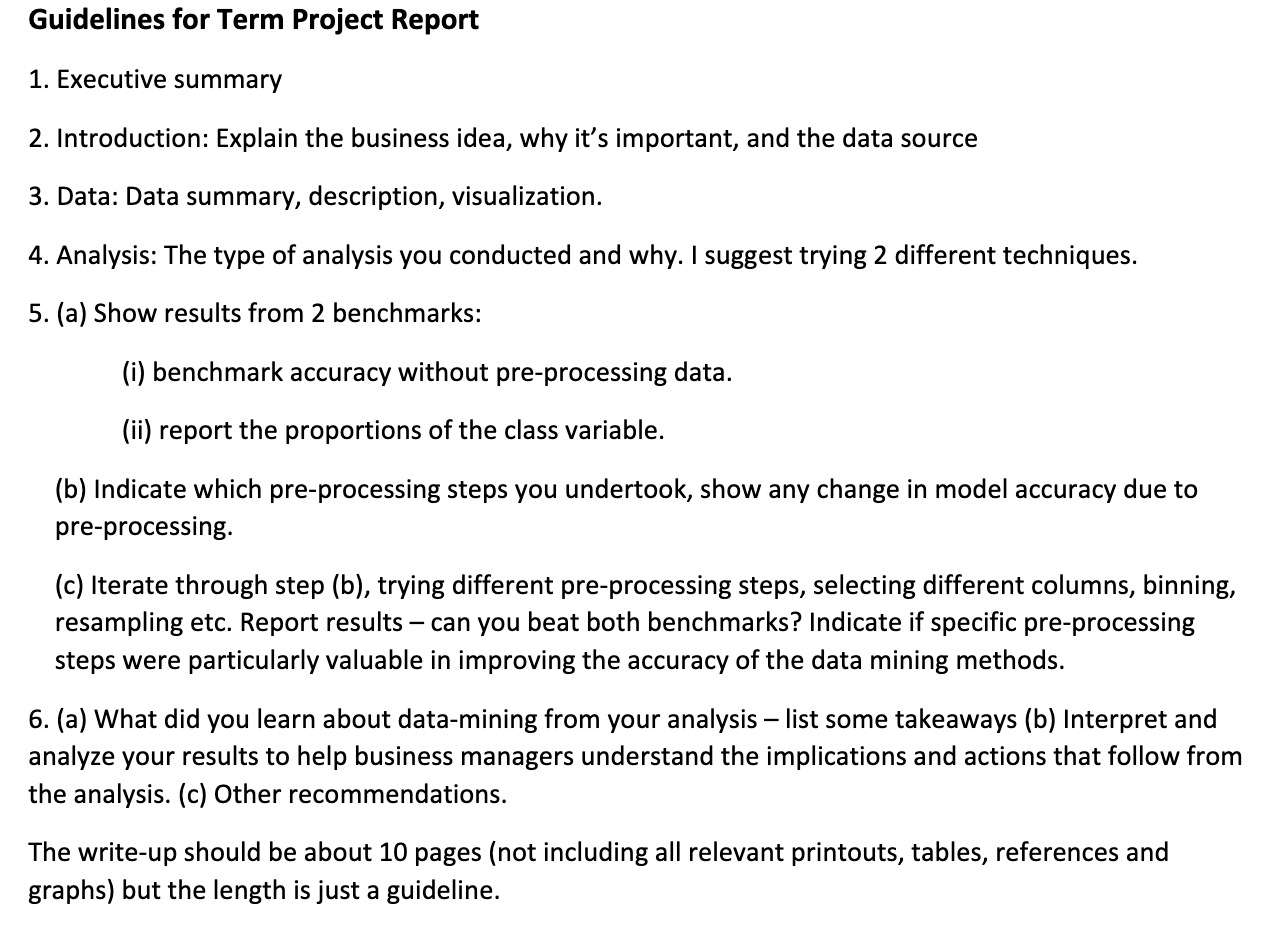
Our analytical approach will start with data cleaning and feature engineering to prepare the data for modeling, analyzing key features such as ”last ticket price“, “days to departure” and “route distance” to improve prediction accuracy. We will conduct exploratory data analysis (EDA) to uncover relationships and trends, such as the effects of seasonality and route popularity on pricing. Following EDA, we’ll train both baseline and advanced machine learning models, using linear regression for initial insights and Random Forest Regression as the primary model due to its ability to capture non-linear relationships and feature interactions. Model tuning and evaluation will involve cross-validation and metrics like MAE, RMSE, and R² to ensure robustness.

# **Machine Learning Models**

We plan to utilize 2-3 machine learning models to predict flight ticket prices and understand the underlying pricing mechanisms. Our approach involves training and comparing different models’ accuracy to determine important features and capture non-linear relationships between features (booking information, flight details, etc.) and the target variable (predicted ticket price). We aim to train Decision Tree and Random Forest Regression models as they can handle complex interactions between features.

Requirement





* **10-page**
* At least 2 techniques, comparing accuracy with 2 benchmarks (1 w/o preprocessing data & 1 from class distribution etc-80% yes)
* What have been done to improve accuracy? (processing data, sampling, parameter tuning,...)
* No need to attach codes, just visualization of data & results

Ideas

1. It looks like you plan to call an API to download data in batches instead of using an existing dataset. I suggest you look for and use an existing dataset. There are many **airline-related datasets on Kaggle**.

Kaggle Airline Dataset:

1. <https://www.kaggle.com/datasets/iamsouravbanerjee/airline-dataset> : don’t have prices,..

2. <https://www.kaggle.com/datasets/yyxian/u-s-airline-traffic-data>: year/revenue/passenger

3.<https://www.kaggle.com/code/sasakitetsuya/airfare-price-prediction-model-by-catboost-r2-0-91/input> Date of journey, price, TA said can’t do

4. <https://www.kaggle.com/datasets/yashdharme36/airfare-ml-predicting-flight-fares> - TA said can’t do

5. <https://www.kaggle.com/datasets/anandshaw2001/airlines-booking-csv/data>

2. Your plan seems to involve some time series problems. Our course does not cover time series problems at all. Keep this in mind.

3. In your research question, you must have a clear dependent variable object with predictive value that can be classified(not a continuous data). For example: satisfied or not, high or low ticket price.

If you plan to preprocess the data to get a dependent variable, you need a more detailed plan: for example, how do you define the range of high price and low price.

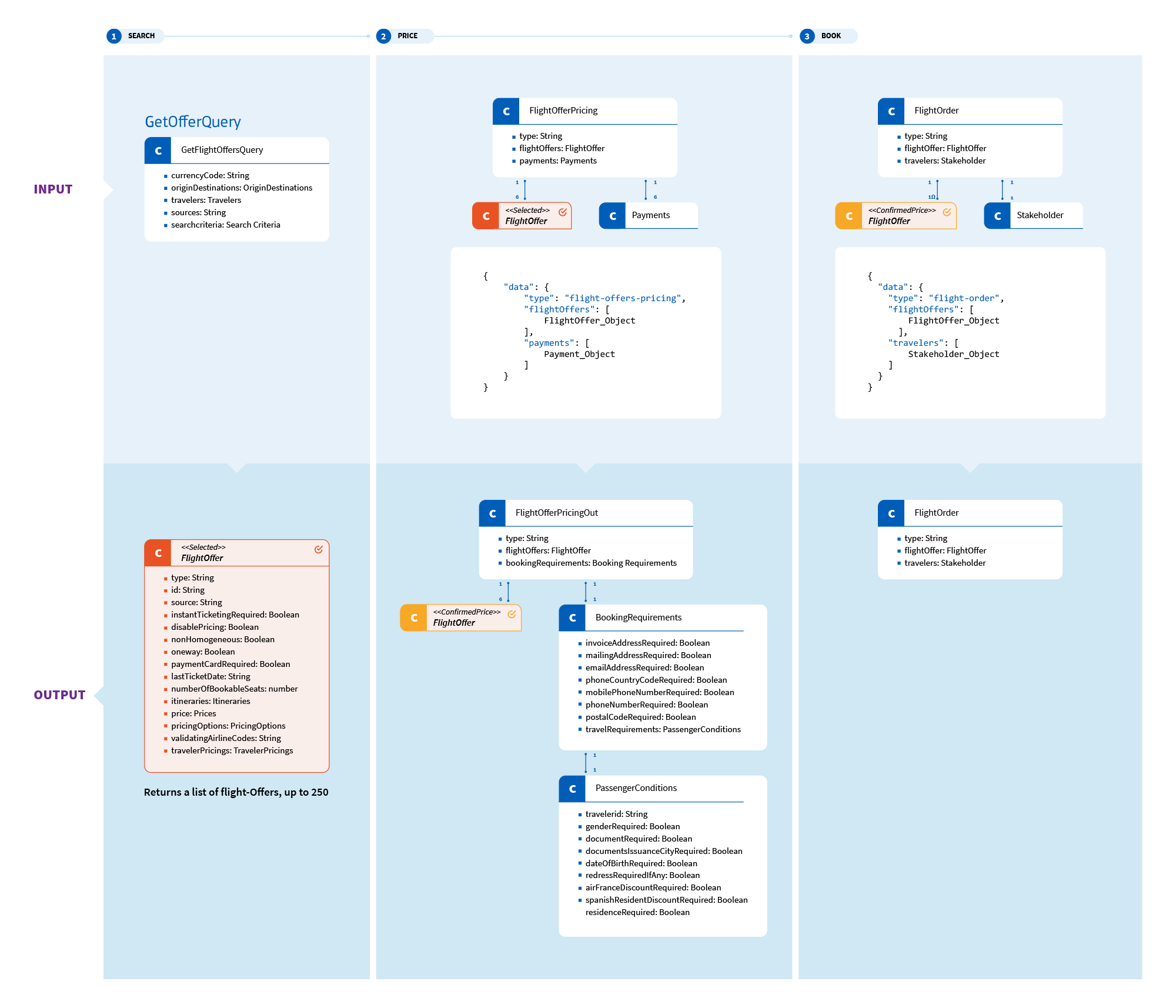
Data-set research:

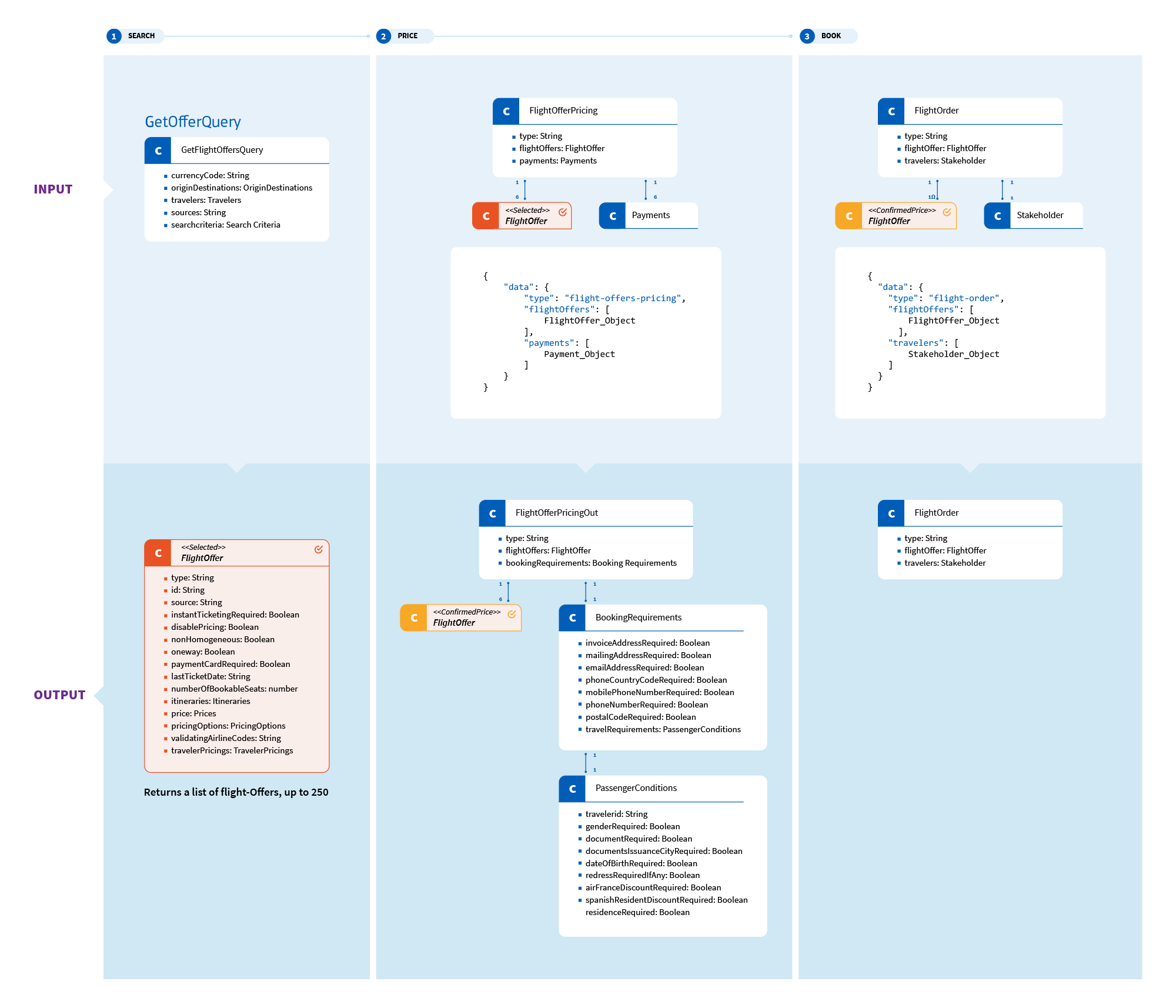
* What dataset?

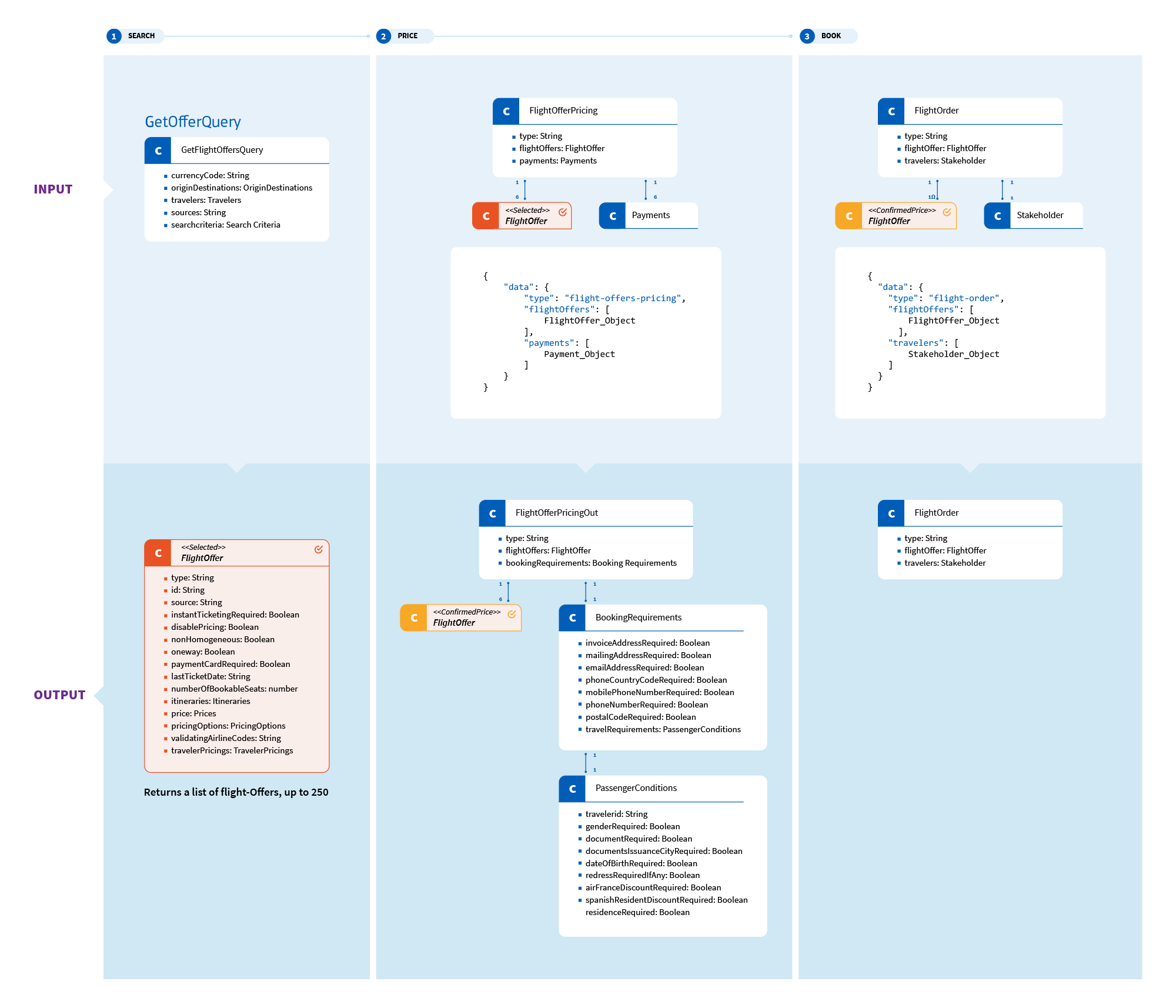
[Bureau Stats Release](https://www.bts.gov/statistical-releases?field_effective_date_value=&field_effective_date_value_1=&combine=&field_editorial_type_target_id=376)

Kenjee………………

Flight







* Target Variable:
  + **Input:**

**Primary**

* + Booking date and time (days to departure, time of day, day of the week, time of the year)
  + Flight details (origin, destination, airline, class, number of stops, duration, distance, departure/arrival time, day of the week, month/season, public holidays/events)

**Secondary**

* Demand indicators: search trends for flights?
  + Economic indicators: fuel prices, inflation?
  + Competitor price: Average prices on the same route from other airlines?
  + **Output:**
  + Price of the Ticket
  + Price Tier
* ML Model

For non-linear & predict continuous target price

* 1. Decision Tree Regressor & Linear Regression: find feature importance
* 2. Random Forest Regression

For predict price classes

* 1. Logistic Regression
* 2. Decision Tree Classifier
* 3. Random Forest Classifier

Questions:

* Time

1. How will the price be according to the booking time(Morning/Afternoon/Night/Midnight)?

* Seasonality

1. How does seasonality, including holiday periods, affect fare prices?

* Airlines: How does the brand name or reputation of different airlines impact ticket pricing?
* Geographical

—------------------------------------------------------------

Dataset Research

What Dataset?

To predict airline ticket prices effectively, we require a comprehensive dataset that includes:

• Historical Booking and Fare Data: Details on booking behavior, flight characteristics, and fare prices over time.

• Sources: Possible sources include airline booking APIs, open-source repositories (e.g., Kaggle), and government or industry data (e.g., Bureau of Transportation Statistics for U.S. domestic flights).

Target Variable

• Price of the Ticket: The actual fare amount, which will be our main target variable for prediction.

• Price Tier: A categorical variable indicating the fare level (e.g., low, medium, high) for easier segmentation and tier-based pricing strategies.

Input Variables

To accurately predict airline fares, we will use the following input variables:

1. Booking Date and Time:

• Days to Departure: The number of days between booking and flight departure.

• Time of Day: Whether the booking was made in the morning, afternoon, or evening.

• Day of the Week: Day of the week the booking was made, as price patterns vary across weekdays and weekends.

• Time of the Year: Season or quarter (e.g., summer, winter) to capture seasonal price variations.

2. Flight Details:

• Origin and Destination: Route-specific price differences due to demand and airline competition.

• Airline: Each airline has unique pricing strategies that impact fare.

• Class: Cabin class (e.g., economy, business) directly impacts pricing.

• Number of Stops: Direct flights vs. flights with stopovers, as non-direct flights tend to be cheaper.

• Duration: Total flight time, often correlated with fare prices.

• Distance: Distance between origin and destination, which can also be a key factor.

• Departure/Arrival Time: Flights departing during peak or off-peak hours show price variance.

• Day of the Week, Month/Season, Public Holidays/Events: Flight date specifics that influence fare due to increased demand during certain periods.

3. Demand Indicators:

• Search Trends for Flights: Search volumes, potentially from Google Trends or other sources, which could indicate heightened demand for certain routes or dates.

4. Economic Indicators:

• Fuel Prices: Fuel costs impact airline expenses and, by extension, fare pricing.

• Inflation: Inflation affects operational costs, influencing ticket pricing over time.

5. Competitor Price:

• Average Prices on the Same Route: Competitor pricing for the same routes and dates, which influences fare adjustments.

Output

• Price of the Ticket: A continuous variable representing the fare price.

• Price Tier: A categorical variable representing the fare level.

ML Model

1. Decision Tree Regressor & Linear Regression: Use these models to assess feature importance, revealing the key drivers behind fare prices.

2. Random Forest Regression: A robust choice for handling non-linear relationships and interactions between variables in fare prediction.

Potential Analytical Approach

1. Data Collection and Cleaning: Gather and clean data, handling missing values, standardizing formats, and preparing for analysis.

2. Feature Engineering: Create new features like “days to departure” and se

asonal indicators to enhance the model’s accuracy.

3. Model Building and Hyperparameter Tuning: Use cross-validation to select the best parameters for each model, focusing on Random Forest as the primary model.

4. Model Evaluation: Evaluate model performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

5. Feature Importance Analysis: Identify which variables are most influential on fare pricing using the Decision Tree model’s feature importance output.

6. Exploratory Data Analysis (EDA): Explore patterns across factors such as season, route, and airline to understand fare distribution better.

Questions to Address:

1. Time: Does the time of booking or flight departure significantly impact fare?

2. Seasonality: aa

3. Airlines: What impact does airline choice have on fare, and are there distinct pricing strategies?

4. Geographical Factors: How do route characteristics (e.g., distance, route popularity) influence fare?

Final Report Draft

Executive Summary

The primary goal of this project was to analyze customer reviews from a Kaggle dataset to:

1. Predict whether a customer would recommend a flight based on various aspects of their flight experience.
2. Understand how each aspect of the flight experience influences overall satisfaction and customer recommendations.

To achieve this, we employed Decision Tree and Random Forest models. Preprocessing techniques included feature engineering, feature selection, SMOTE for handling class imbalance, and hyperparameter tuning. By building both benchmark and fully preprocessed models, we aimed to ensure fairness in evaluation and maximize predictive accuracy.

#### **Key Findings**

1. **Most Influential Features**:
   * "Value for Money" emerged as the most significant factor influencing customer satisfaction and recommendations.
   * "Ground Service" and "Cabin Staff Service" were the next most impactful factors.
2. **Model Performance**:
   * The Random Forest model with full preprocessing demonstrated the most robust performance, offering superior predictive power compared to the Decision Tree model.

#### **Recommendations**

Based on the findings, the following strategies are recommended:

1. **Focus on Price-Performance Optimization**: Airlines should prioritize initiatives such as promotional pricing, bundling, or loyalty programs to enhance the perceived value of flights.
2. **Enhance Ground Service**: Improving the quality of ground services will help deliver a more seamless end-to-end customer experience, complementing other satisfaction drivers.
3. **Maintain Adequate Standards**: While "Cabin Staff Service," "Seat Comfort," and "Inflight Entertainment" contribute modestly to satisfaction, maintaining a baseline standard in these areas is advisable.
4. **Strategic Resource Allocation**: Features like "Seat Type" and "Wifi & Connectivity" have minimal impact on recommendations. Airlines should avoid overinvesting here unless targeting specific customer segments or premium offerings.

#### **Conclusion**

This analysis demonstrates the importance of focusing on factors that significantly influence customer recommendations, such as perceived value and service quality. By leveraging the insights from this project, airlines can strategically allocate resources to optimize customer satisfaction and loyalty.

Introduction

**Business Idea**

The airline industry is one of the most competitive and customer-centric markets in the world. With countless options available to travelers, understanding what truly drives customer satisfaction and loyalty is not just an advantage—it's a necessity for standing out as a competitor in the market.

In this project, we set out to answer a critical question for the industry:

***“How do the various aspects of a flight experience influence a customer’s overall satisfaction and their likelihood to recommend the airline?”***

This question is not just relevant for airlines seeking to optimize their services and maximize customer loyalty, but also for travelers who increasingly rely on reviews and recommendations to guide their decisions. By applying advanced analytical techniques and leveraging real-world customer review data, we aimed to uncover actionable insights that could help airlines enhance the customer journey and gain a competitive edge.

**Why It’s Important**

Customer service and satisfaction is at the core of success in the airline industry, where loyalty and customer recommendations can heavily influence an airline’s reputation, and thus their profitability. In an era where online reviews are increasingly accessible and can significantly influence consumer behavior and outlooks, understanding the factors that drive customer satisfaction has never been more critical.

This project aims to bridge the gap between raw customer feedback and actionable business insights by harnessing machine learning models and techniques. By identifying the key aspects of the flight experience that most impact customer satisfaction and recommendation likelihood, airlines can:

* Strategically allocate resources to the areas that matter most to customers.
* Enhance the overall customer experience, fostering loyalty in a highly competitive market.
* Make data-driven decisions to design services and packages that align with customer priorities, improving operational efficiency and profitability.

Additionally, this analysis contributes to a broader understanding of consumer behavior in the airline industry, offering insights that can benefit not only airlines but also passengers and travel agencies.

Data

The dataset we are using for this project was retrieved from Kaggle, posted by user Juhi Bhojani. The data was originally scraped from the website airlineequality.com, which serves as an incredible online resource for exploring data and gathering insights related to airlines, airports, and customer flight satisfaction.

This dataset provides in-depth information about the various details associated with a flight, with the focus being on whether or not a particular reviewer recommended their flight or not.

At A Glance:

23,171 reviews (rows)

19 flight details (columns/features)

Attribute Descriptions:

**Categorical Attributes**

| Airline Name  Name of the airline | Overall Rating  General rating given to the airline. |
| --- | --- |
| Type of Traveller  Solo Leisure, Couple Leisure, Business, Family Leisure | Seat Type  Economy Class, Business Class, Premium Economy, First Class |
| Verified  Whether the review is verified or not. | **Recommended (Target Variable)**  Whether or not the reviewer ultimately recommended their airlines (Yes or No) |

**Numerical Attributes** (each rating on a scale of 1-5)

| Seat Comfort  Rating of seat comfort during the flight | Cabin Staff Service  Rating of the cabin crew’s service quality |
| --- | --- |
| Food & Beverages  Rating of the quality and variety of food and drinks | Ground Service  Rating of services at the airport |
| Inflight Entertainment  Rating of available entertainment options | Wifi & Connectivity  Rating of in-flight Wi-Fi quality and speed |
| Value for Money  Rating of how passengers perceive the value of the flight for the price paid |  |

**Textual/Date Attributes**

Note: We dropped these attributes in our final models since we aren’t using NLP/text analysis or Time-Series Models.

| Review\_Title  Title of the review | Review Date  Date the review was posted |
| --- | --- |
| Review  Text of the review | Route  Flight route (origin-destination pair) |
| Date Flown  When the flight occurred |  |

Exploratory Data Analysis Visualizations

During the EDA phase, we generated various visualizations to help us develop a better understanding of our raw data and how we might need to preprocess and clean our data for our final models.

Below are some of the most insightful visualizations produced during this phase of EDA:

| **Distribution of Numerical Ratings**    These histograms provided an overview of the frequency distribution of each numerical column, showing how ratings are spread.  Additionally, it also helped us see which rating attributes had “error” entries that came as 0’s. With this, we knew to remove rows that contained 0 values in any of their rating features. |
| --- |

| **Relationship Between Ratings and Recommendation**    These boxplots visualized how ratings differed for reviews where customers recommended the airline vs where they did not. |
| --- |

Preprocessing

This section explores the preprocessing steps taken for all of our models:

1. Decision Tree
   1. Benchmark Model
   2. Full Model
2. Random Forest
   1. Benchmark Model
   2. Full Model

Minimal Preprocessing (for Benchmark Models)

Although a major project requirement was that benchmark models were supposed to run off the raw dataset, doing so caused many issues with our models, preventing them from running at all.

To go around this, we applied very minimal preprocessing measures for our benchmark models to ensure that our dataset was compatible with the tree models. These were the bare minimum preprocessing steps that ensured that our benchmark models weren’t as processed as our full models, but still allowed the benchmark models to run.

**Step 1: Dropping attributes that caused errors**

Since tree models in scikit-learn use only numerical features, many string-based attributes caused errors that prevented our benchmark models from running at all.

*It is also important to note that although we could have applied feature engineering (encoding) to some of these variables to make use of them, we also dropped them by intuition as we knew that they would have little to no impact on customer satisfaction based on the situation context.*

*There were columns that were:*

1. *invalid in the problem analysis context*
   1. *e.g. “Overall\_Rating” was just another target variable similar to “Recommendation”*
2. *contain insights that wouldn't have an influence on the customers' Recommendation*
   1. *E.g. “Review Title” wasn’t an actual aspect of the flight experience*
3. *contain string/text-based insights (since we aren't using NLP or text analysis)*
   1. *E.g. “Review” (which is just the actual text review that each reviewer included as a part of their overall flight experience review) could have included great insights for us to use, but deriving meaningful insights from the Reviews of such a huge dataset would have required using text analysis methods/NLP*
4. *time-sensitive (since we aren't using a time-series model)*
   1. *E.g. “Review Date” and “Date Flown”*

To solve this, we dropped the following attributes:

Airline Name

Review\_Title

Review Date

Review

Aircraft

Route

Date Flown

**Step 2: Use label encoding for string values in “Type of Traveller” and “Seat Type”**

For the string-based attributes “Type of Traveller” and “Seat Type”, we decided to use feature engineering (label encoding) to make them compatible with our tree models. In doing this, we wouldn’t have to drop these attributes, which have significant influence on the Recommendation variable.

Both features exhibit a natural hierarchy or order that label encoding effectively preserves.

* For instance, in 'Type of Traveller', categories like 'Solo Leisure' (0), 'Couple Leisure' (1), 'Business' (2), and 'Family Leisure' (3) reflect an increasing degree of group size or trip complexity.
* Similarly, in 'Seat Type', categories such as 'Economy Class' (0), 'Business Class' (1), 'Premium Economy' (2), and 'First Class' (3) represent progressively higher levels of luxury or service.
* Label encoding helps maintain these inherent relationships between the categories.

Label encoding is computationally simple and provides a numeric representation that facilitates faster model training while ensuring compatibility with our chosen algorithms. It strikes a balance between preserving the structure of the data and keeping the feature space manageable.

**Step 3: Transform values in Recommended column as (Yes = 1) and (No = 0).**

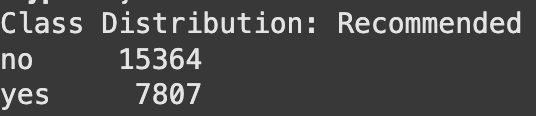
We transformed the 'Recommended' column values from 'Yes' and 'No' to **1** and **0** because tree-based models, like Decision Trees and Random Forests, require numerical inputs. These models cannot process categorical values like 'Yes' and 'No' directly for splitting and decision-making. By encoding them as 1 (for "Yes") and 0 (for "No"), we enable the model to perform efficient calculations and make binary classifications, improving model performance.

Full Preprocessing (for Full Models)

In addition to the minimal preprocessing steps used for benchmark models, our full preprocessing plan for our full models included feature engineering, feature selection, SMOTE resampling, and hyperparameter tuning.

**Step 0: Exploratory Data Analysis (EDA)**

Before we began full preprocessing, we utilized various pandas functions to further explore our raw dataset, develop a deeper understanding of the data available to us, and get a better idea of what specific areas needed attention during the full preprocessing stages.

* Some columns, like **Aircraft** and **Wifi & Connectivity**, had a high percentage of missing values.
* Rating columns (e.g. seat comfort, cabin staff service, etc) are numeric, with some missing data.
* The dataset has a few redundant columns, such as **Unnamed: 0**.
* There is a vast imbalance in class distribution:
  + 

**Step 3: Transform null values if applicable**

Many reviews had null values in rating columns where the reviewer may have just felt neutral

* (e.g. Wifi & Connectivity, Inflight Entertainment, etc)
* If there was a null rating, just set it to "3.0"
  + Rationale: Assuming they just felt neutral, 3.0 was the neutral value on the 1-5 scale

**Step 4: Drop all rows with null values in non-transformed attributes**

Many rows still had null values in various columns outside of the transformed values mentioned in Step 3.

**Step 5: Drop all rows that are not Verified**

To get more reliable insights, we dropped reviewers where their “Verified” column value was “False”.

According to airlinequality.com, the “editorial staff have inspected a copy of an e-ticket, booking details or a boarding pass, with the customer name confirming the trip written about in the review”.

Note: With the inclusion of this preprocessing step, it is important to note that our full models are trained and tested to predict the Recommendation values for “Verified” reviews.

**Step 6: Resampling with SMOTE**

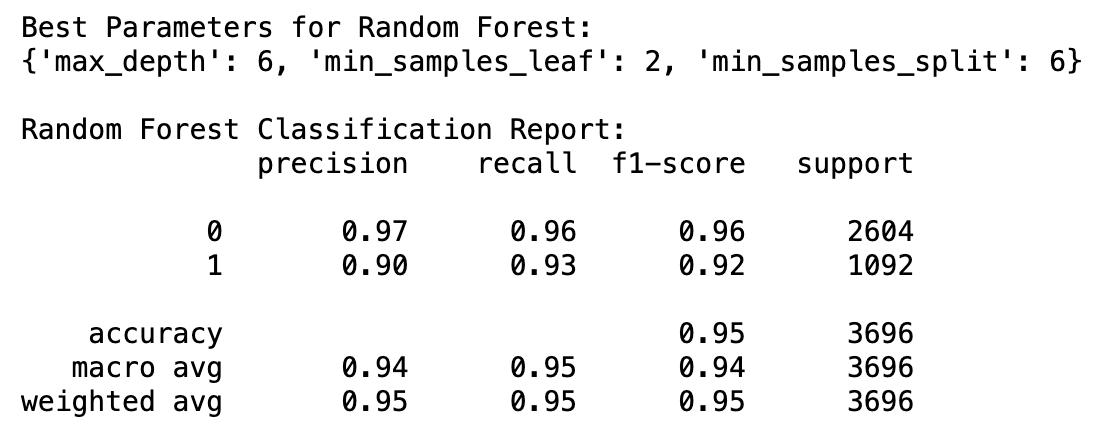
During our EDA, we noticed a vast imbalance in class distribution for the target variable.

To improve the performance and accuracy of our model, we decided to target this issue by utilizing SMOTE resampling.

* To Balance the Target Variable Distribution
* To Improve Model Performance on the Minority Class

**Step 7: Employing Hyperparameter Tuning**

This ensured that our full Decision Tree and Random Forest models utilized the best possible parameters to achieve best performance.



Model I : Decision Trees

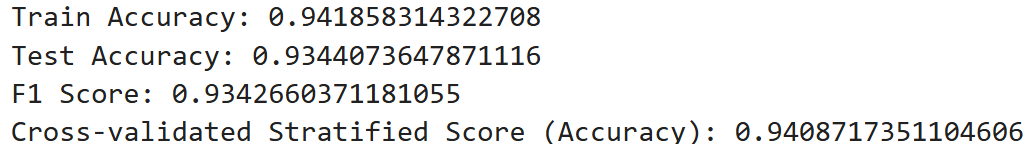
Why this model?

Decision Tree models are simple to understand, interpret, and visualize, making them a great choice for datasets where explainability is important. They can handle both categorical and numerical data effectively and are robust to outliers. Additionally, Decision Trees naturally manage non-linear relationships in the data, making them versatile for various types of problems.The model’s ability to handle missing values ensures that we can achieve reliable results without extensive preprocessing.

Benchmark model

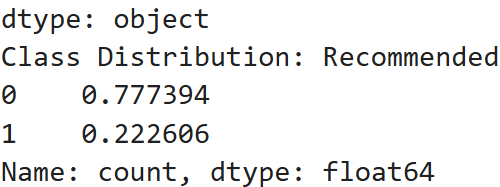
**Accuracy**

The benchmark accuracy of decision tree model is shown below:



**Proportion**

The benchmark proportion of decision tree is shown below:

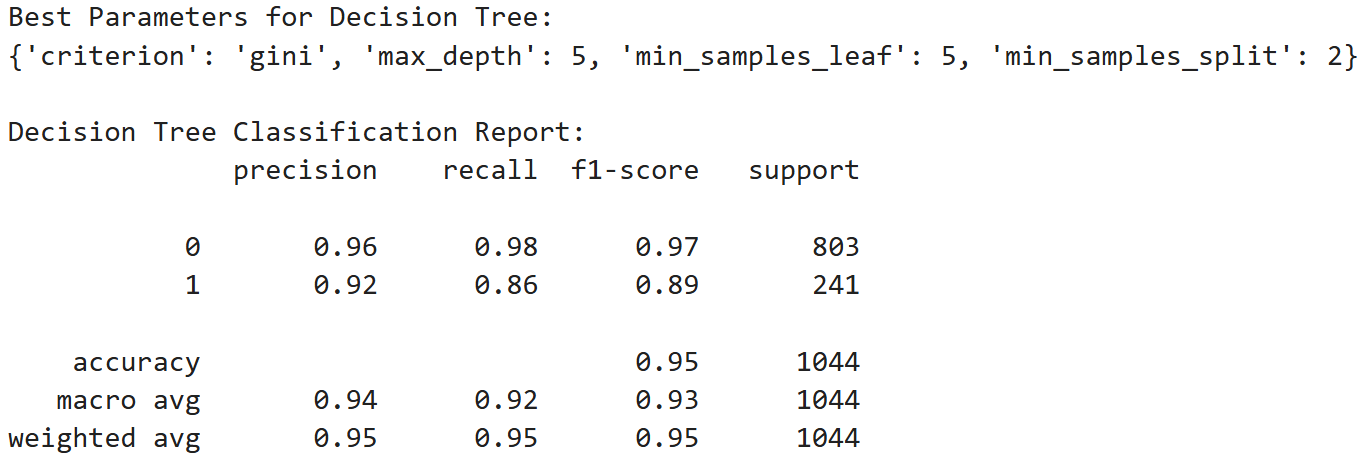


The result shows that the dataset is imbalanced significantly, we evaluate the confusion matrix to understand how well the model handles both classes.

Preprocessing

**Steps, and the changes of model accuracy**

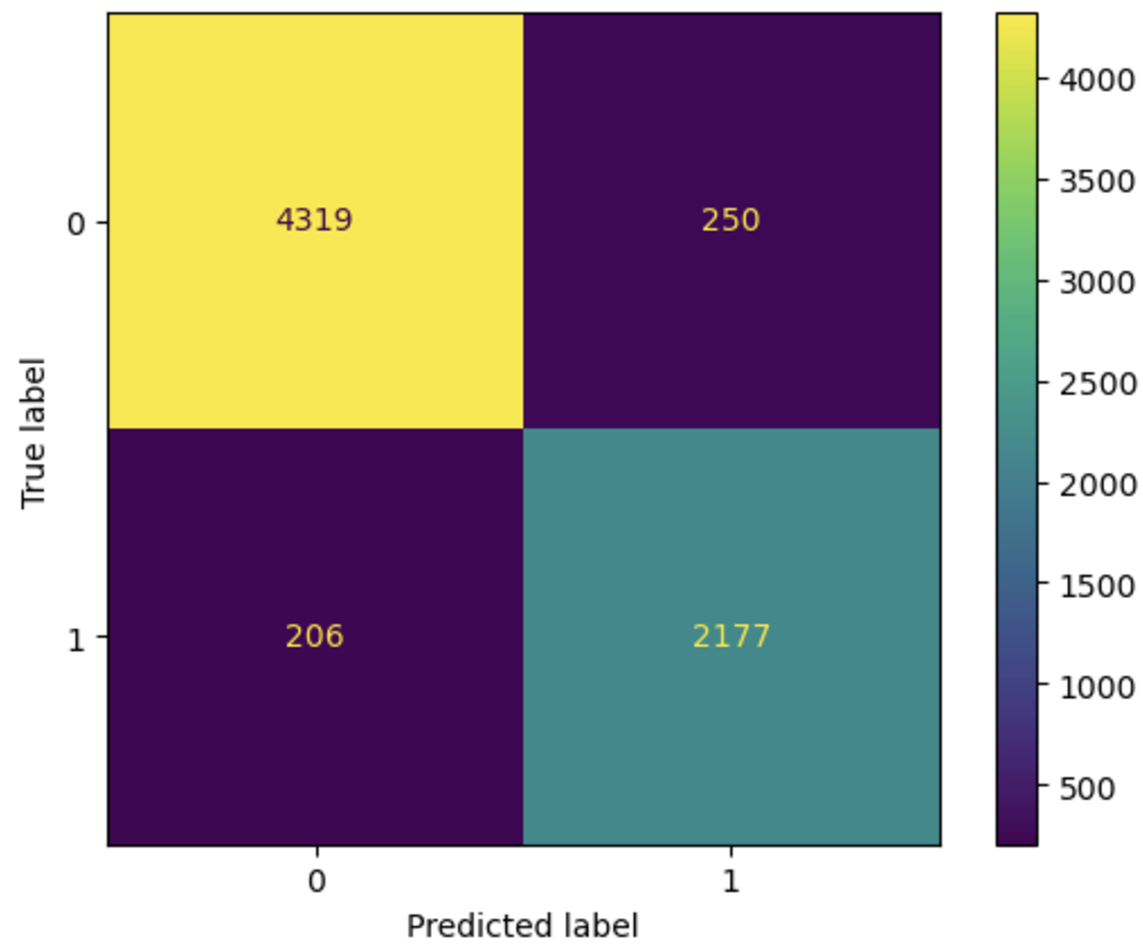
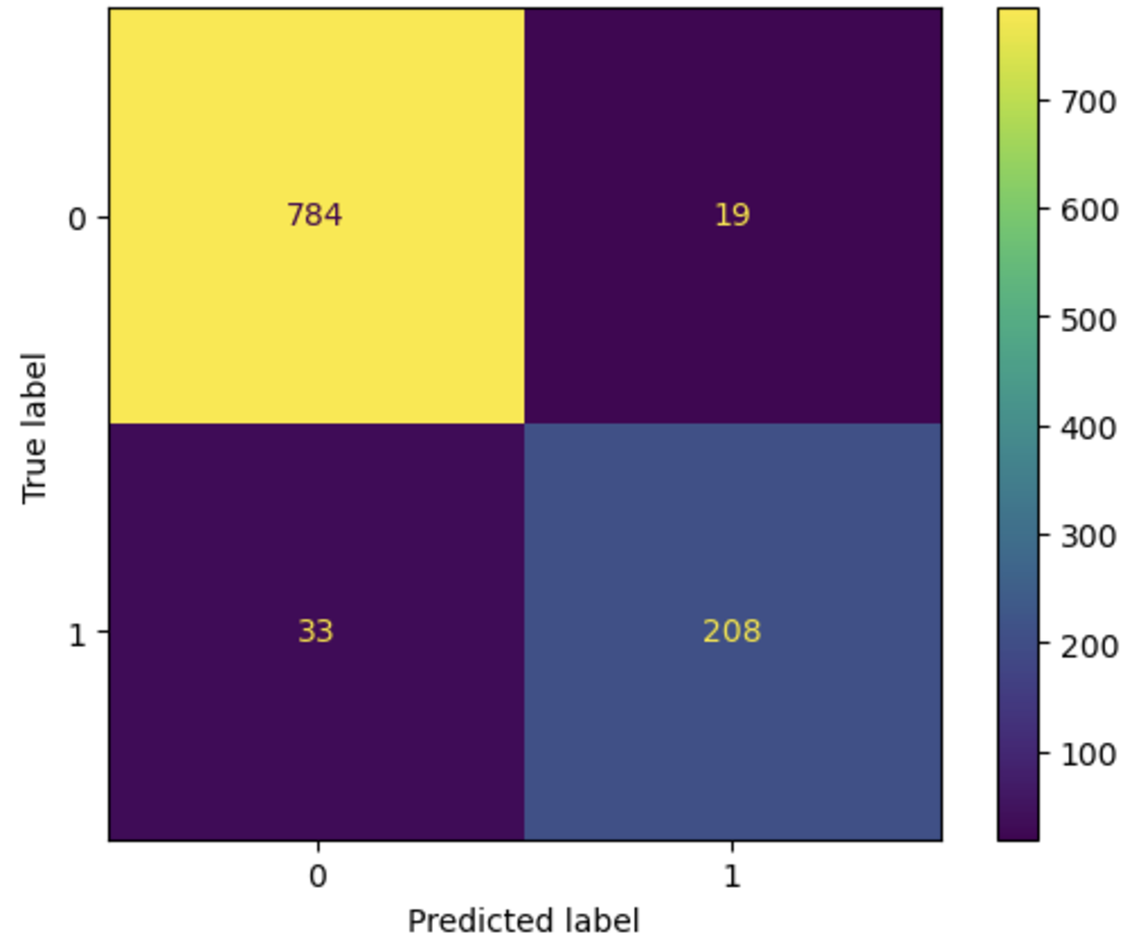
After the preprocessing and hyperparameter tuning that mention in the previous part, we get the best parameters for random forests show below:



Predicted Result

**Confusion Matrix**

Based on the confusion matrix results, the benchmark model achieves an accuracy of 93%, while the final model reaches 95%. The model with full preprocessing performs significantly better compared to minimal preprocessing. Additionally, the error rate decreases notably, indicating that the hyperparameter adjustments were effective.

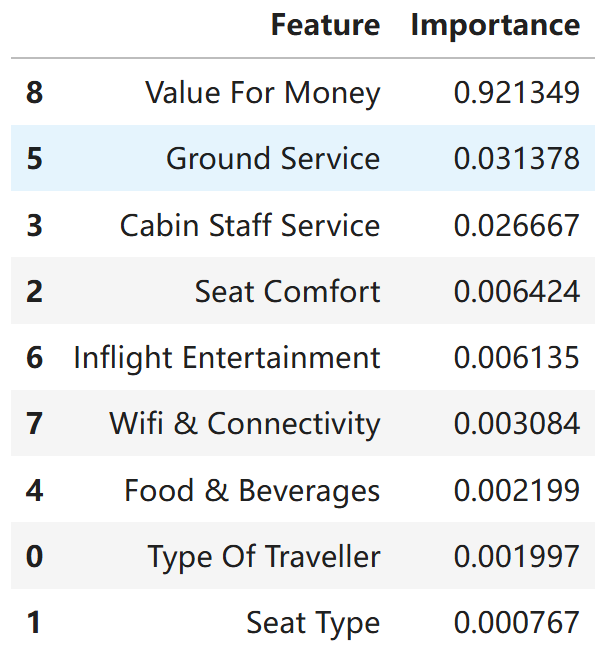
 

Benchmark Final decision tree model

**Accuracy**

| Accuracy Comparison | Accuracy | F1 Score | Cross-Validated Stratified Score |
| --- | --- | --- | --- |
| Benchmark | 0.9344 | 0.9343 | 0.9409 |
| With Preprocessing | 0.9502 | 0.9507 | 0.9613 |

Feature Importance



The top three contributors to customer satisfaction were identified as “Value for Money”, “Cabin Staff Service”, and “Ground Service”. These findings are consistent with general expectations in the airline industry, highlighting the critical roles of cost-effectiveness and service quality in shaping customer experiences.

Overall, the decision tree proved to be an effective baseline model. It delivered reasonable accuracy while providing valuable insights into the key factors driving customer recommendations.

Model II: Random Forest

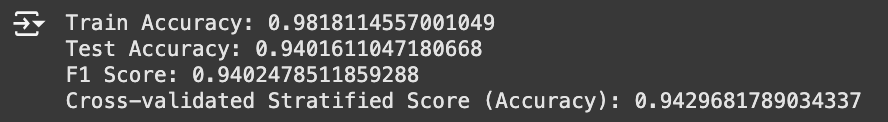
Why this model?

Random Forest has the advantage of avoiding overfitting, flexibility to handle both regression and classification with high accuracy, and easy to determine feature importance. Our dataset contains both categorical and numerical data, and many missing values. Random Forest also can effectively manage missing values.

Benchmark model

* Accuracy

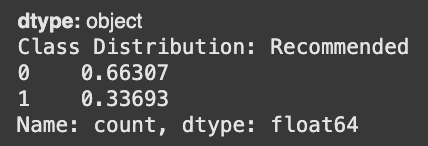
The benchmark accuracy of random forest model is shown below:



Both our train accuracy and test accuracy is high, which means our model captures the patterns in the sets well. The difference between train accuracy and test accuracy is only 4%, suggesting that our model is not overfitting and generalizes well to the unseen data. While the minor tuning might further optimize the test performance, there’s no major concerns of overfitting or underfitting based on the accuracy.

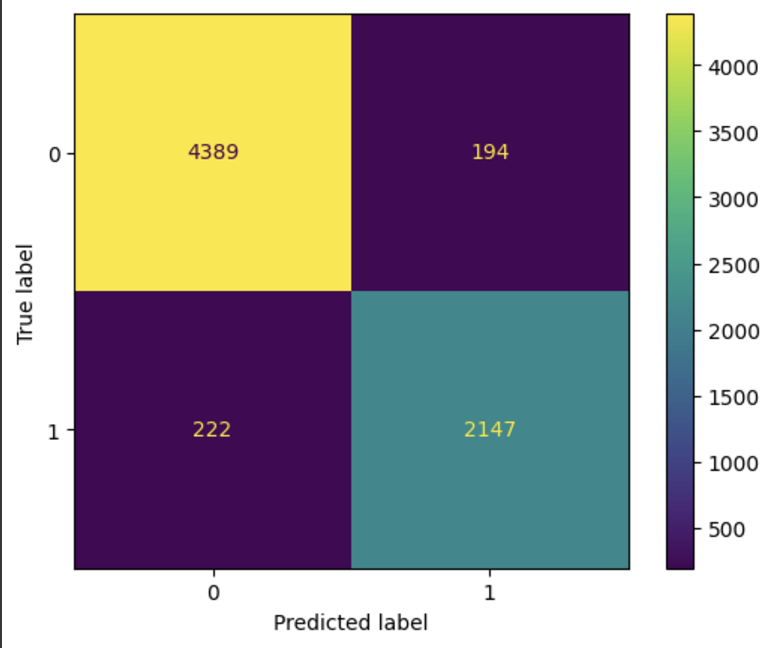
* Proportion

The benchmark proportion of random forest is shown below:



The result shows that the dataset is imbalanced significantly, biased toward the majority class (Class 0) and it may lead to predicting the majority class (Class 0) more often and poor performance for the minority class (Class 1).

Therefore, we evaluate the confusion matrix to understand how well the model handles both classes. Below is the confusion matrix of the benchmark:

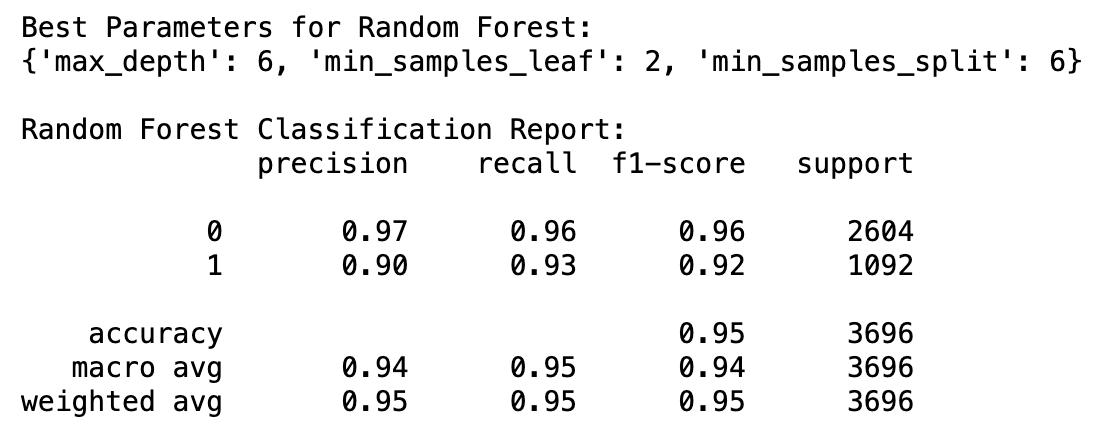


According to the confusion matrix of benchmark, the precision of Class 1 is 91% and the overall accuracy is 94%, which performs well in identifying both classes and there are false positives (194) and false negatives (222) that may impact decision making. In order to improve performance, we should do some adjustments.

Preprocessing

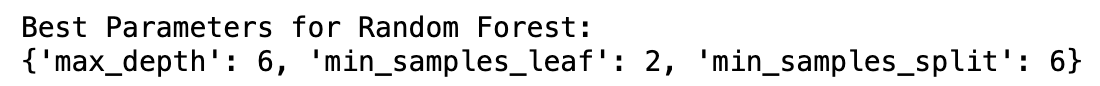
Steps, and the changes of model accuracy

After the preprocessing and hyperparameter tuning that mention in the previous part, we get the best parameters for random forests show below:

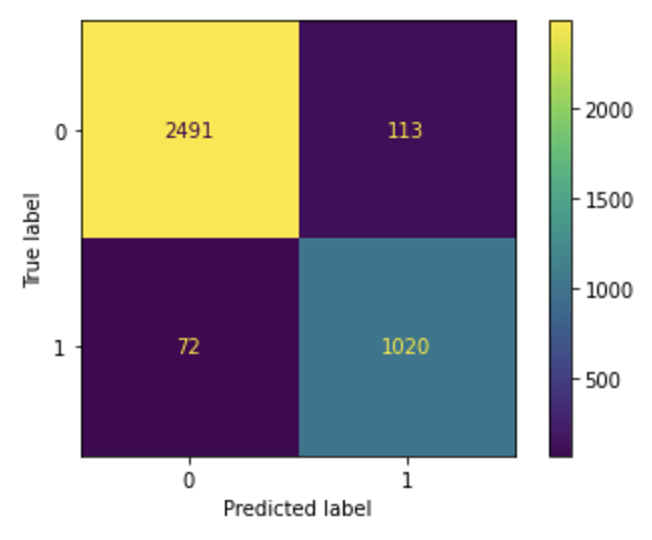


Predicted Result

Moreover, we get a higher Cross-validated Stratified Score of 95%, which is higher than the one of the decision tree (\_\_).



Below is the confusion matrix of the final model:

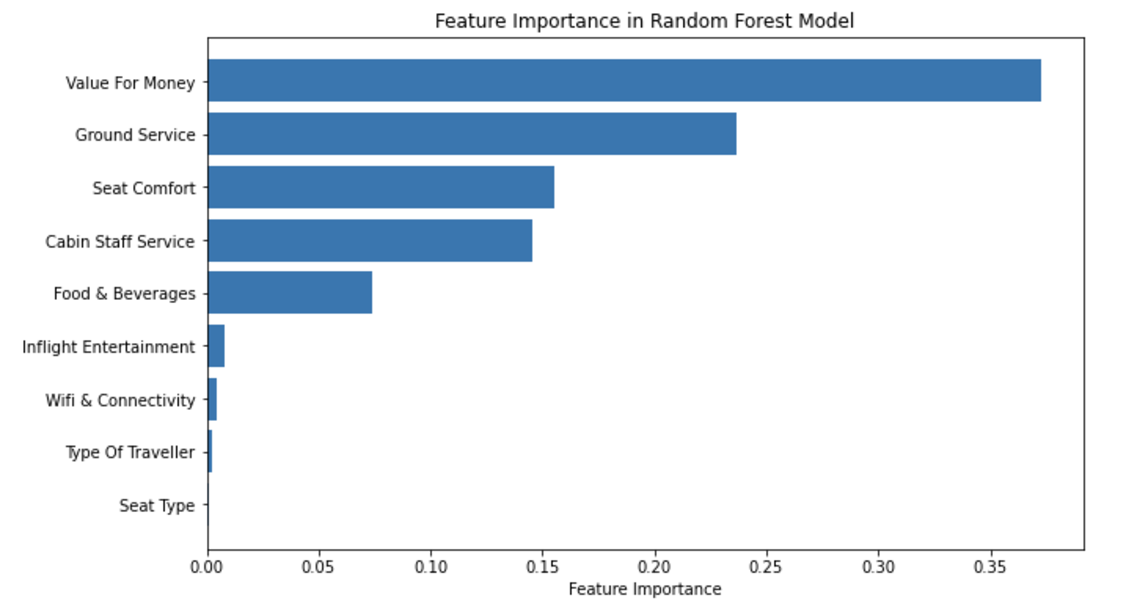


According to the confusion matrix of the final model, the overall accuracy is 95%, which is much better than the one of the decision tree. Also, the error rate decreases significantly, meaning that the adjustments of hyperparameters are useful.

Feature Importance

In the end, we can conclude how much each feature improves the accuracy.

The feature importance of the final model is shown below:



According to the graph above, “Value For Money” plays an important role in predicting our target variable, contributing 37% to the customer decisions. It makes sense considering that

The feature importance of random forest’s top 2 ranking and the last one is the same as the one of the decision tree, which is “Value For Money”, “Ground Service” and the last one is “Seat Type”. However, the ranking from 3-8 is different. In the model of random forest, “Cabin Staff Service”, “Food & Beverages”, “Inflight Entertainment”, “Wifi & Connectivity” and “Type of Traveller” are ranked from3 to 8. But the model of decision tree starts from “Cabin Staff Service”, “Seat Comfort”, “Food & Beverages”, “Type of Traveller”, “Wifi & Connectivity”, and “Inflight Entertainment”.

The table below shows much clearer of the difference between random forest and decision tree:

|  |  |
| --- | --- |
| Random Forest | Decision Trees |

Takeaways

According to our result, “Value For Money” are the most important attributes for airline customers’ decisions. To address this, airlines should focus on strategies that optimize the price-performance ratio. Example of implementing promotional pricing, offering bundle service packages and introducing loyalty rewards programs. Second one is “Ground Service”, improving its quality is essential for creating a better end-to-end customer experience. Examples include streamlining the check-in and boarding processes, enhancing baggage handling efficiency, and providing more responsive and courteous customer support at the airport. The third is “Cabin Staff Service” in the decision tree and “Seat Comfort” in the random forest. For “Cabin Staff Service”, since it doesn’t affect too much for the customer satisfaction, we would not suggest the airlines reinvest too much on it, maintaining consistent service standards but provide comprehensive training, offer personalized assistance, and use passenger feedback to drive improvements. For “Seat Comfort”, we suggest that airlines can offer more legroom options, improving seat design and cushioning, and providing adjustable features.