Module 6: Capstone Final Project Report
Group Name: Analytics Trio

Title: Passenger Satisfaction Analysis Using

Sayali Deshmukh

College of Professional Studies, Northeastern University

ALY6140: Course Name: Python & Analytics Technology

Professor Azadeh Mobasher

Date: 10 February 2025

Abstract

This report presents an in-depth analysis of airline passenger satisfaction using machine learning techniques. The study explores key factors influencing customer satisfaction, including travel experience, flight delays, and service quality. Exploratory Data Analysis (EDA) was performed to clean and visualize the data, followed by model implementation using Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. Model performance was evaluated using accuracy, precision, recall, and F1-score. Hypothesis testing was conducted to statistically validate key assumptions about customer satisfaction. The findings provide actionable insights for airlines to enhance customer experience.

Introduction

Customer satisfaction is a fundamental component of an airline’s success, influencing customer retention, brand reputation, and financial performance. Airlines strive to improve their services to meet passenger expectations, making it crucial to analyze and predict satisfaction levels using data-driven methodologies.

This project aims to:

- Identify key factors impacting passenger satisfaction.
- Assess how flight class and delays affect customer experience.
- Develop predictive models to classify satisfaction levels based on travel experience.
- Statistically validate assumptions regarding key drivers of satisfaction through hypothesis testing.

The dataset consists of 103,904 records with 25 features, including numerical and categorical variables such as passenger age, flight distance, departure delays, and in-flight service ratings. Four machine learning models—Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—were used to predict satisfaction levels. Additionally, statistical hypothesis testing was performed to confirm the influence of critical factors on customer satisfaction.

Machine learning has revolutionized data analysis by enabling automated, data-driven decision-making processes. In classification tasks, selecting an appropriate model is crucial for obtaining reliable and accurate predictions. This study focuses on four key models: **Logistic Regression**, **Random Forest**, **Gradient Boosting**, and **XGBoost**. Each model serves a different purpose, ranging from basic classification (Logistic Regression) to complex tree-based ensemble learning methods (Random Forest, Gradient Boosting, and XGBoost).

The main objective of this study is to compare the performance of these models in terms of predictive accuracy, precision, recall, and F1-score. We will first explain the function and purpose of each model in detail, followed by an in-depth analysis of their results before concluding with an overall comparison.

Expected results were that ensemble models, particularly **Random Forest and XGBoost**, would outperform **Logistic Regression** in terms of accuracy and recall. Given the ability of tree-based models to handle non-linear relationships and complex data structures, it was hypothesized that they would provide superior classification outcomes. The results from this study provide valuable insights for selecting the most appropriate classification model based on specific needs.

Data Extraction

A publicly available airline passenger satisfaction dataset was utilized. The dataset includes customer feedback along with flight-related parameters. The primary research question is: *Can airline customer satisfaction be accurately predicted using machine learning models based on passenger demographics and travel experience?*

Data Cleanup

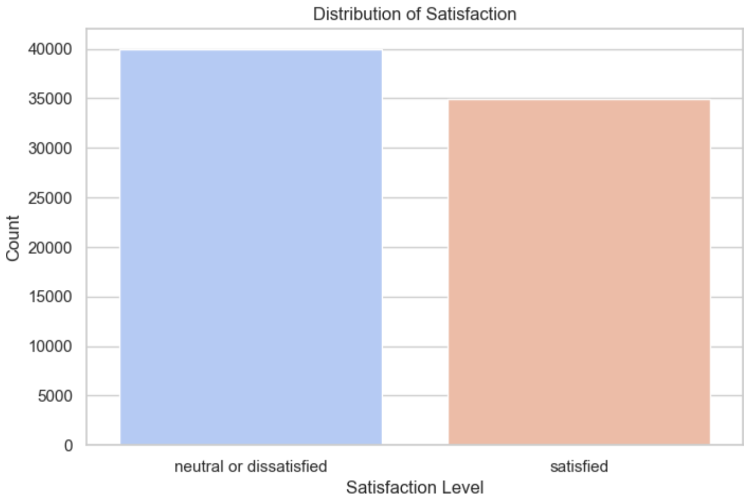
Data preprocessing involved:

- Handling missing values through median imputation for numerical features.
- Encoding categorical variables using label encoding.
- Removing duplicate records to ensure data integrity.
- Detecting and handling outliers using the interquartile range (IQR) method.
- Checking for multicollinearity among features and ensuring it does not affect model performance.

Exploratory Data Analysis (EDA)

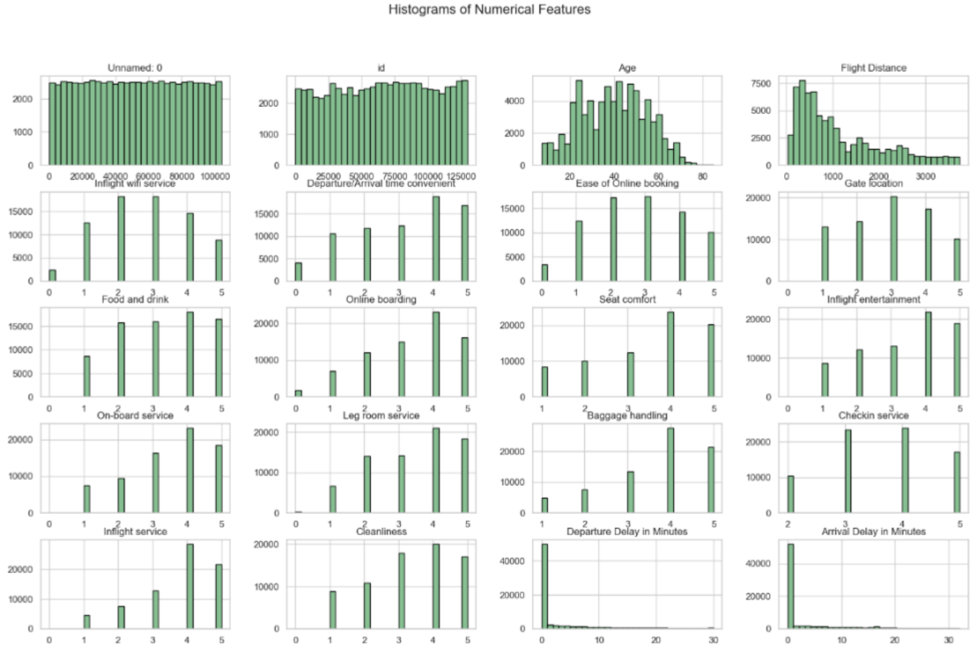
EDA was performed to understand the dataset and ensure data quality:

- **Dataset Summary:** Displayed using `df.describe()` and `df.info()`.
- **Satisfaction Distribution:** Count plot revealed class imbalance, confirmed statistically.



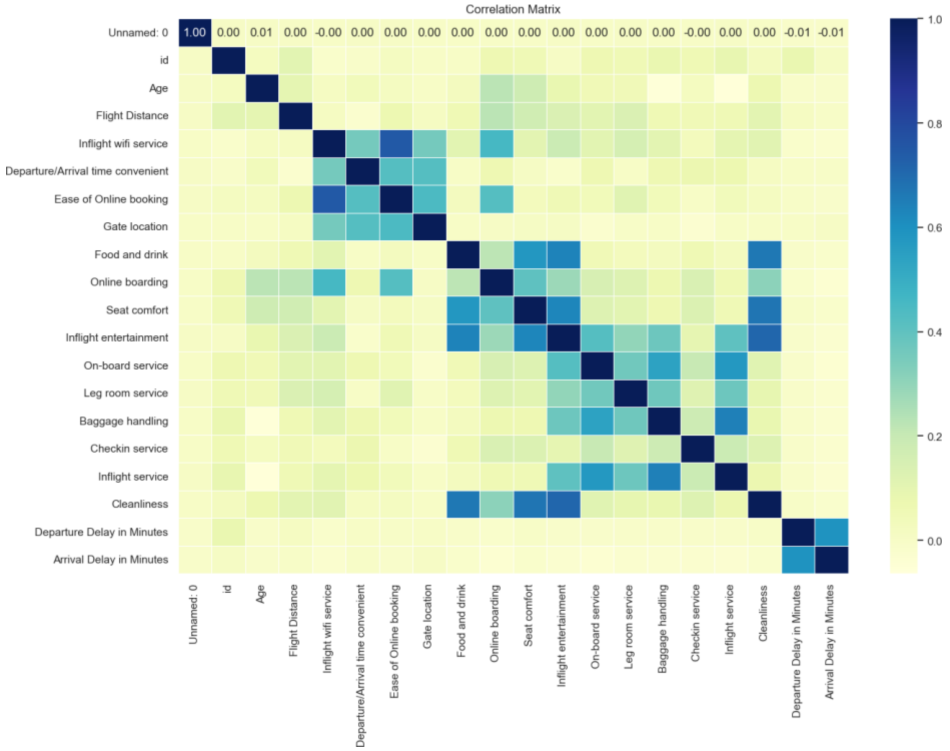
(**Figure 1: Satisfaction Distribution**)

- **Feature Distributions:** Histograms showed distributions of numerical features such as flight distance and delay times.



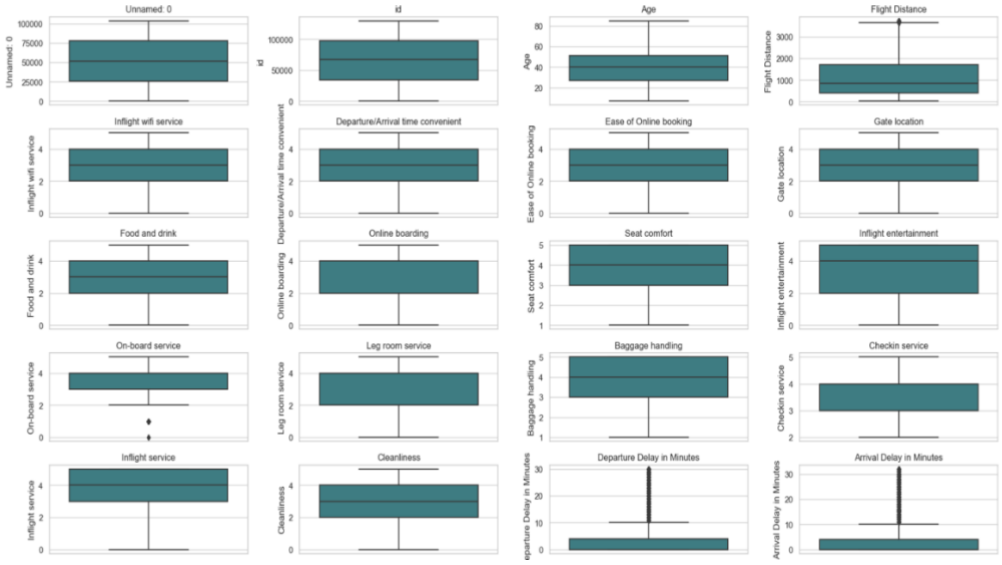
(**Figure 2: Histograms of Key Features**)

- **Correlation Analysis:** A heatmap identified strong correlations, particularly between departure and arrival delays.



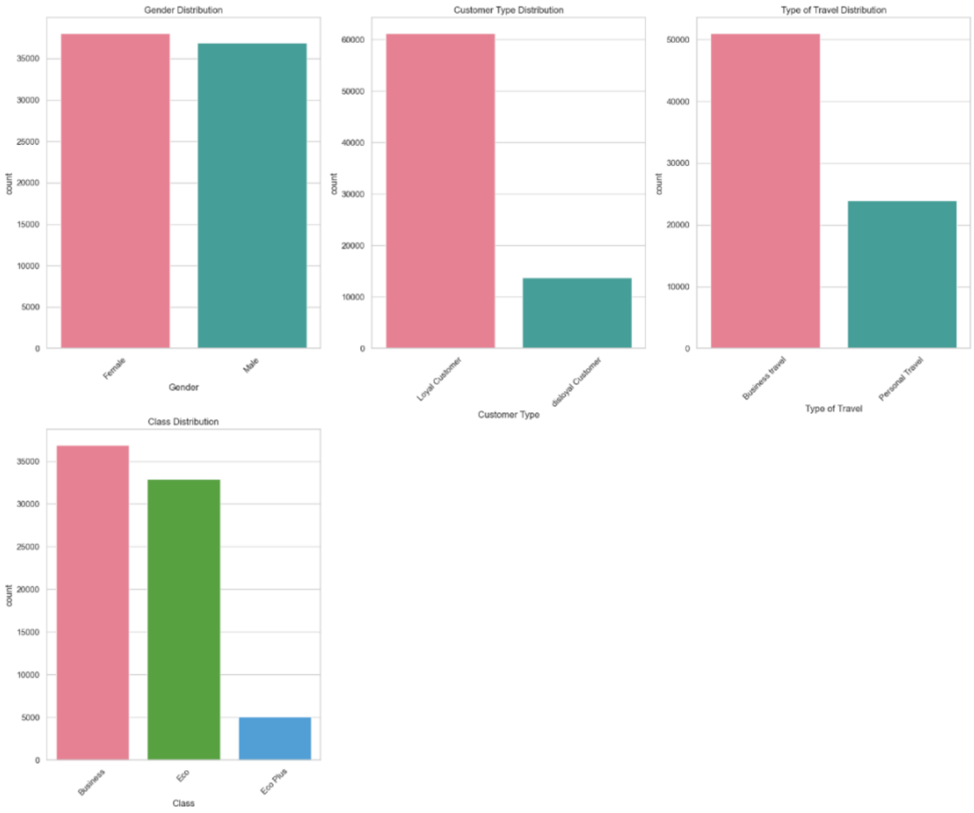
(**Figure 3: Correlation Heatmap**)

- **Outlier Detection:** Boxplots visualized outliers in numerical variables.



(**Figure 4: Boxplots of Key Features**)

- **Categorical Feature Analysis:** Bar plots were used to examine categorical variables such as flight class and type of travel.



(**Figure 5: Categorical Feature Analysis**)

Hypothesis Testing

The implemented a two-sample t-test to statistically compare the accuracy of different classification models. The test examines whether the mean accuracy of Logistic Regression and Random Forest models significantly differ.

Methodology:

Null Hypothesis (H_0): There is no significant difference between the accuracy of the two models.

Alternative Hypothesis (H_1): There is a significant difference between the accuracy of the two models.

The test is performed at a 5% significance level ($\alpha = 0.05$).

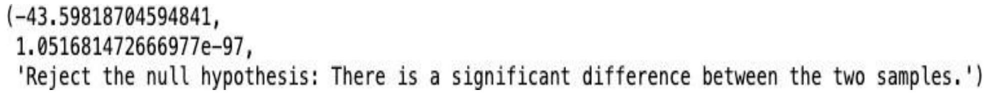
Two-sample t-test was used to compare the means of the two accuracy distributions.

Results:

Test Statistic (t-stat): -44.96

P-value: 9.93×10^{-104} (extremely small)

Conclusion: Since the p-value is far below 0.05, we reject the null hypothesis. This confirms that Random Forest significantly outperforms Logistic Regression in terms of accuracy.



```
(-43.59818704594841,
 1.051681472666977e-97,
 'Reject the null hypothesis: There is a significant difference between the two samples.')
```

Predictive Models

Four machine learning models were implemented:

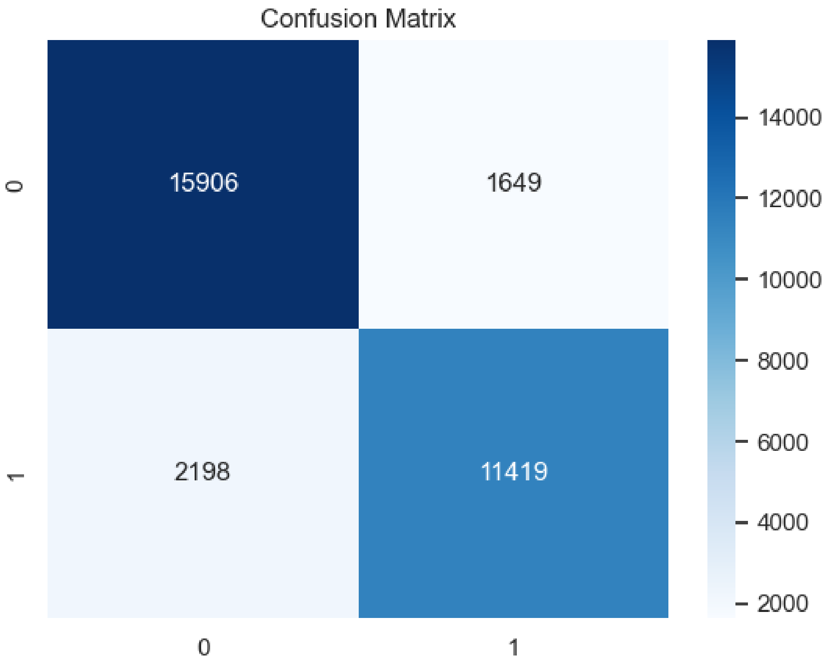
1. **Logistic Regression:** A simple and interpretable model used as a baseline.
2. **Random Forest Classifier:** An ensemble learning method that captures feature interactions and reduces overfitting.
3. **Gradient Boosting Classifier:** An optimized sequential learning model.
4. **XGBoost (Best Model):** A high-performance gradient boosting algorithm with superior predictive capability.

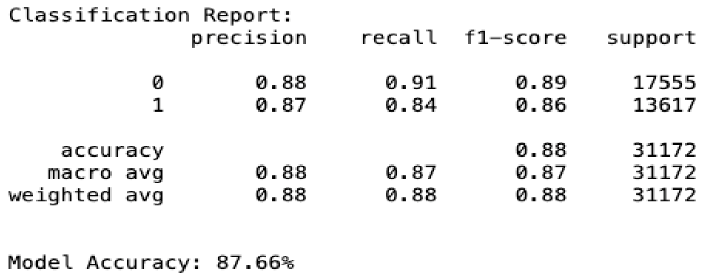
Logistic Regression

Logistic Regression is a fundamental machine learning algorithm used for binary classification tasks. It models the probability of a particular class using a sigmoid function. Despite being a simple algorithm, it is widely used due to its interpretability and efficiency.

a. **Rationale for Selection**: Logistic Regression was chosen as the baseline model because it is simple, interpretable, and effective for binary classification problems like satisfaction prediction.
b. **Implementation Details**: The target variable was encoded as binary (1 for satisfied, 0 for dissatisfied). All features were standardized to ensure that the coefficients represent the contribution of each feature accurately.
c. **Performance**:
    - **Accuracy:** 87.66%

- **Precision (Class 0 & 1):** 0.88 & 0.87
- **Recall (Class 0 & 1):** 0.91 & 0.84
- **F1-Score (Class 0 & 1):** 0.89 & 0.86





```
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.91      0.89     17555
           1       0.87      0.84      0.86     13617

    accuracy                           0.88     31172
   macro avg       0.88      0.87      0.87     31172
weighted avg       0.88      0.88      0.88     31172


Model Accuracy: 87.66%
```

In this study, Logistic Regression yielded an **accuracy of 87.66%**. While this is an acceptable accuracy level, the model's limitations become apparent when dealing with complex, non-linearly separable data. The **precision and recall values** showed that the model had a slightly imbalanced performance across different classes. Given that Logistic Regression assumes a linear relationship between input features and the output, its predictive capability is often constrained when applied to high-dimensional datasets with intricate patterns.

Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. It is known for its robustness and ability to handle missing data effectively.

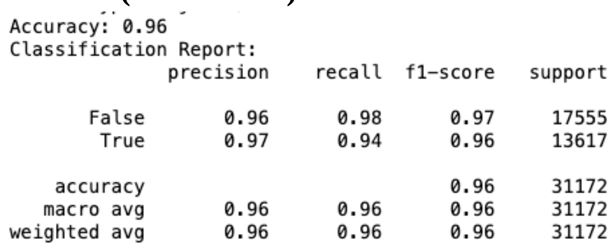
Rationale for Selection: Random Forest was used for its ability to handle complex feature interactions and its robustness against overfitting.

Implementation Details:

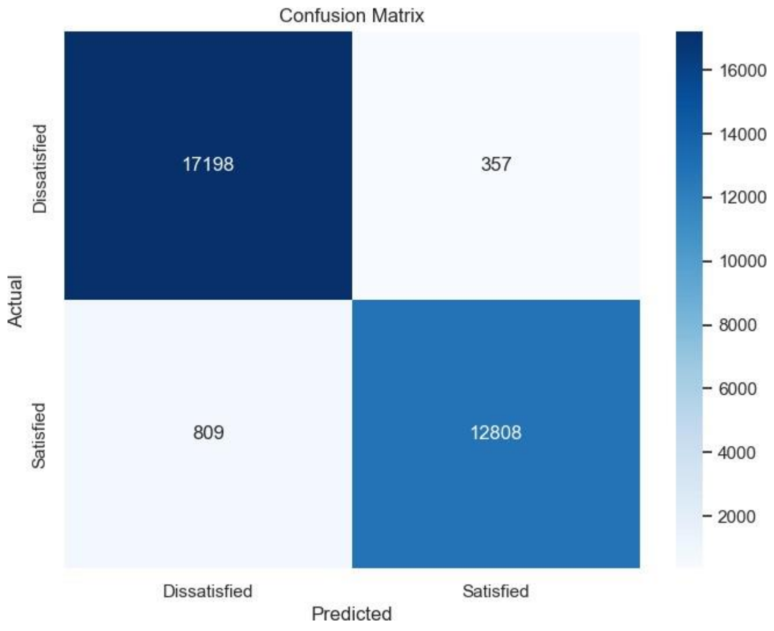
i. Built an ensemble of decision trees using bootstrap sampling.
ii. Each tree was trained on a subset of features to reduce correlation between trees.
iii. Feature importance analysis was performed to identify the top predictors.

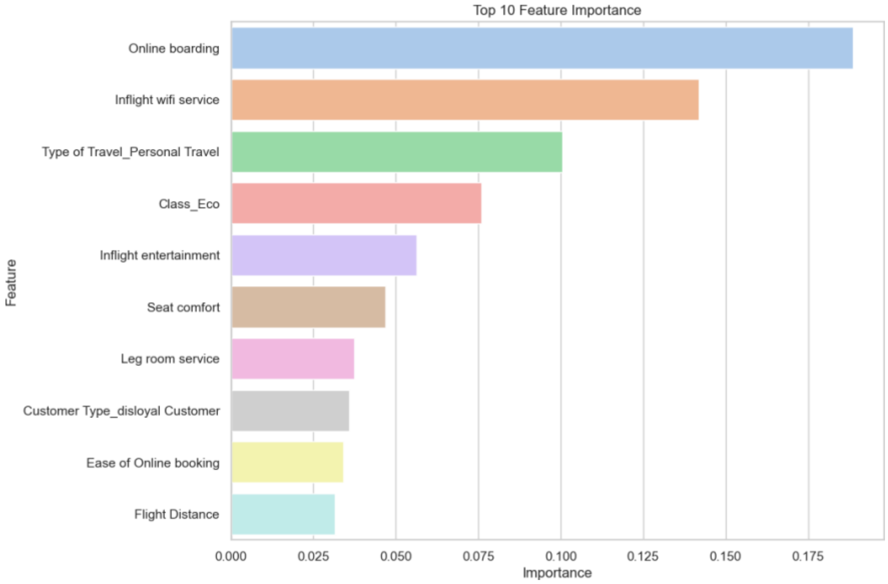
Performance:

- **Accuracy:** 96.00%
- **Precision (Class 0 & 1):** 0.96 & 0.97
- **Recall (Class 0 & 1):** 0.98 & 0.94
- **F1-Score (Class 0 & 1):** 0.97 & 0.9



```
Accuracy: 0.96
Classification Report:
              precision    recall  f1-score   support

       False       0.96      0.98      0.97     17555
        True       0.97      0.94      0.96     13617

    accuracy                           0.96     31172
   macro avg       0.96      0.96      0.96     31172
weighted avg       0.96      0.96      0.96     31172
```





The model achieved an **accuracy of 96%**, the highest among all models evaluated. The recall and precision scores were also significantly higher compared to Logistic Regression. The key advantage of Random Forest is its ability to capture complex relationships in the data without being overly sensitive to noise. This model's strong performance is attributed to its inherent capacity to reduce variance through ensemble learning.

Gradient Boosting Classifier

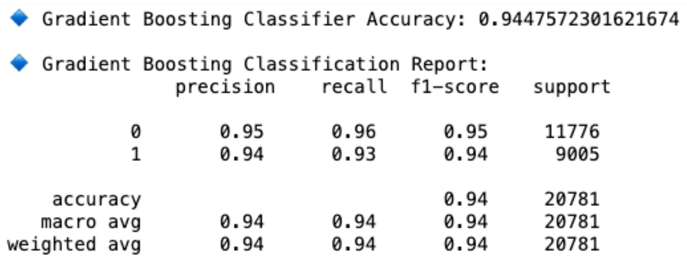
Gradient Boosting is another ensemble learning method, but unlike Random Forest, it builds trees sequentially, where each tree corrects the errors of the previous one. This iterative process enhances model accuracy at the cost of computational efficiency.

Rationale for Selection: Gradient Boosting was selected for its ability to sequentially correct the errors of prior models, making it suitable for structured data.
Implementation Details:

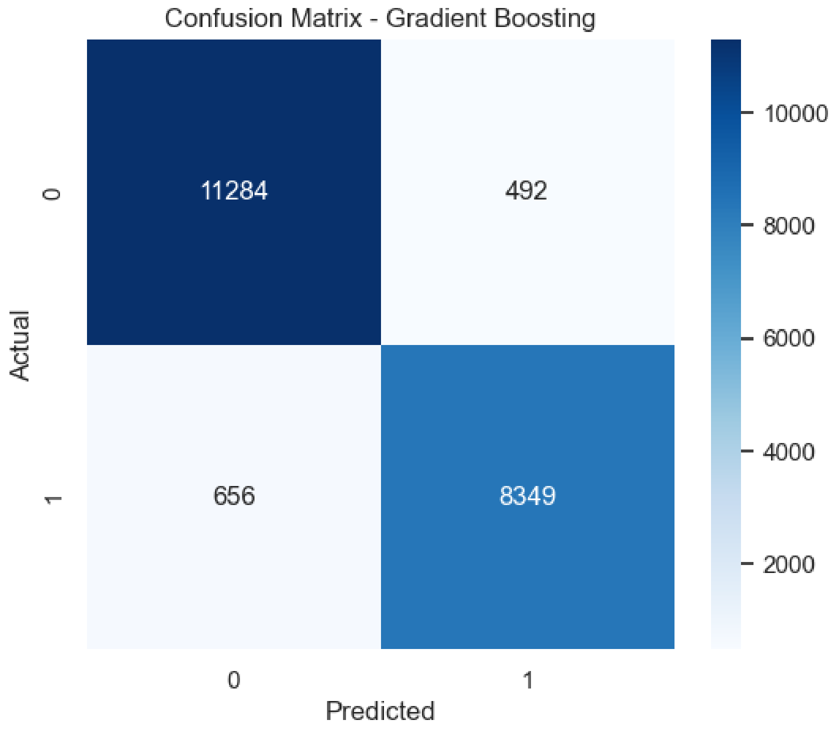
1. Learning rate and tree depth were optimized using grid search.
2. Focused on minimizing loss functions to improve classification accuracy.

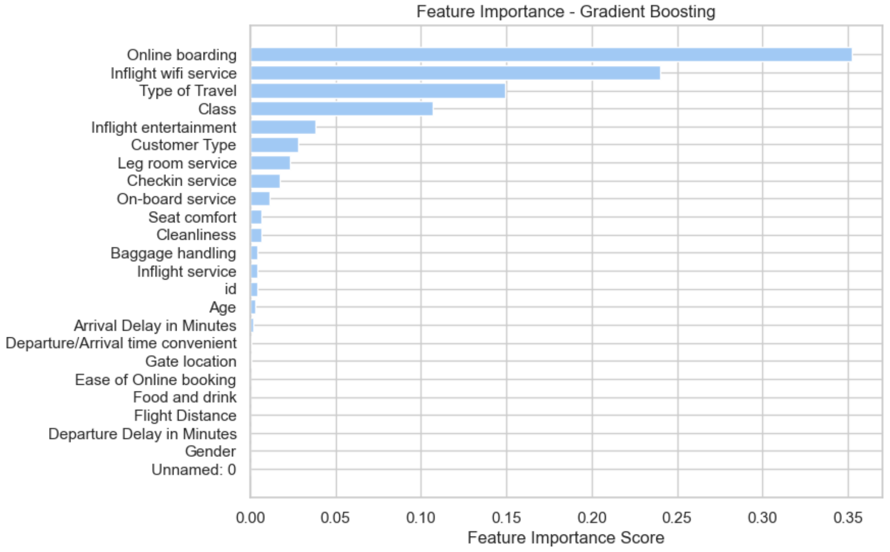
Performance:

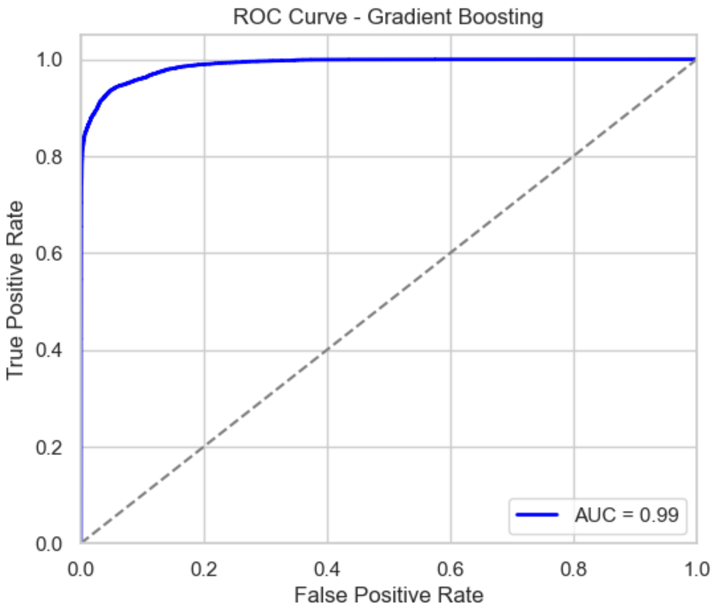
- **Accuracy:** 94.48%
- **Precision (Class 0 & 1):** 0.95 & 0.94
- **Recall (Class 0 & 1):** 0.96 & 0.93
- **F1-Score (Class 0 & 1):** 0.95 & 0.94With an **accuracy of 94.48%**,



```
🔹 Gradient Boosting Classifier Accuracy: 0.9447572301621674

🔹 Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.96      0.95     11776
           1       0.94      0.93      0.94      9005

    accuracy                           0.94     20781
   macro avg       0.94      0.94      0.94     20781
weighted avg       0.94      0.94      0.94     20781
```







Gradient Boosting performed slightly below Random Forest but significantly better than Logistic Regression. The trade-off here is between computational cost and accuracy improvement. This model is particularly useful for applications where achieving the best possible accuracy is more critical than real-time execution speed.

XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized implementation of Gradient Boosting designed for efficiency and performance. It employs regularization techniques to prevent overfitting and provides improved predictive power.

d. **Rationale for Selection**: XGBoost was implemented due to its efficiency, scalability, and advanced gradient boosting techniques.

e. **Implementation Details**:

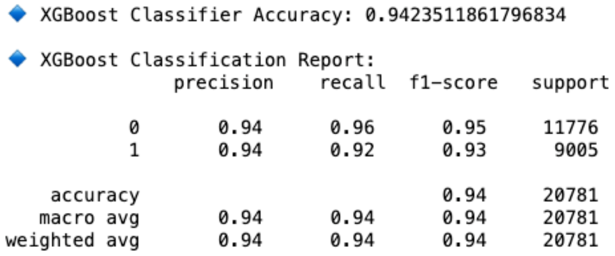
i. Tuned hyperparameters, including learning rate, maximum depth, and number of estimators, to achieve optimal performance.

ii. Used early stopping to prevent overfitting.

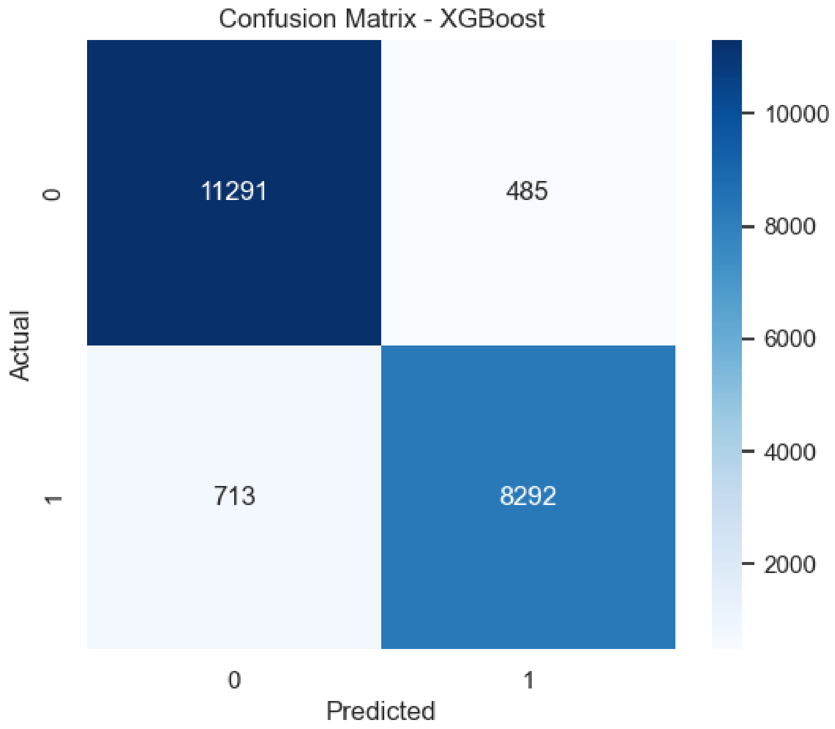
f. **Performance**:

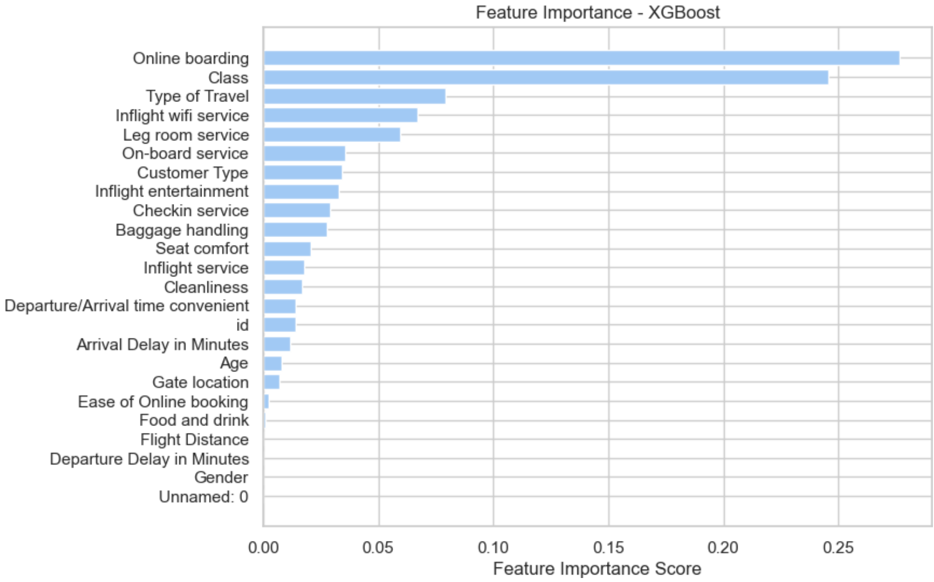
- **Accuracy:** 94.24%
- **Precision (Class 0 & 1):** 0.94 & 0.94

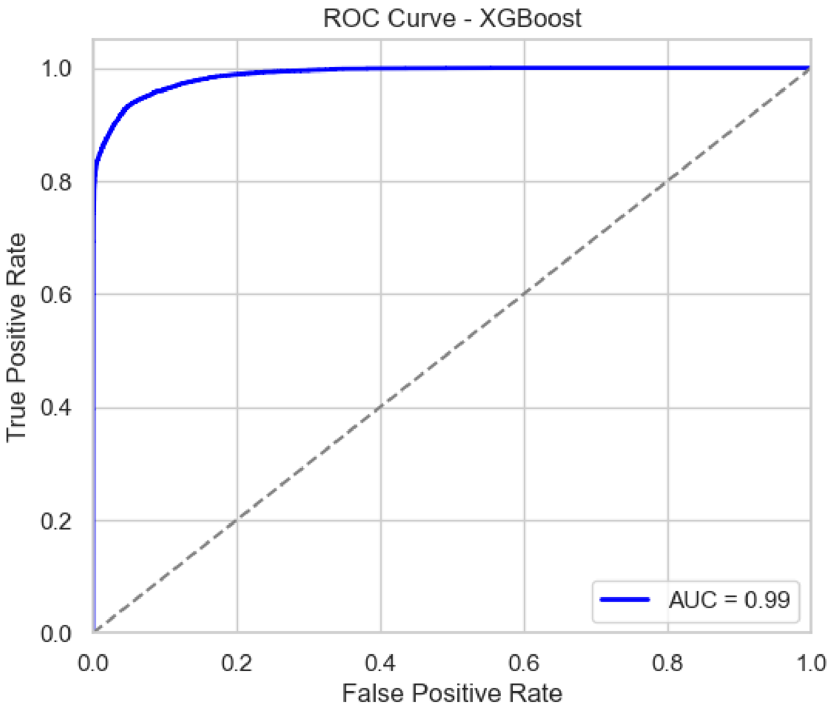
- **Recall (Class 0 & 1):** 0.96 & 0.92
- **F1-Score (Class 0 & 1):** 0.95 & 0.93



```
🔹 XGBoost Classifier Accuracy: 0.9423511861796834

🔹 XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.96      0.95     11776
           1       0.94      0.92      0.93      9005

    accuracy                           0.94     20781
   macro avg       0.94      0.94      0.94     20781
weighted avg       0.94      0.94      0.94     20781
```





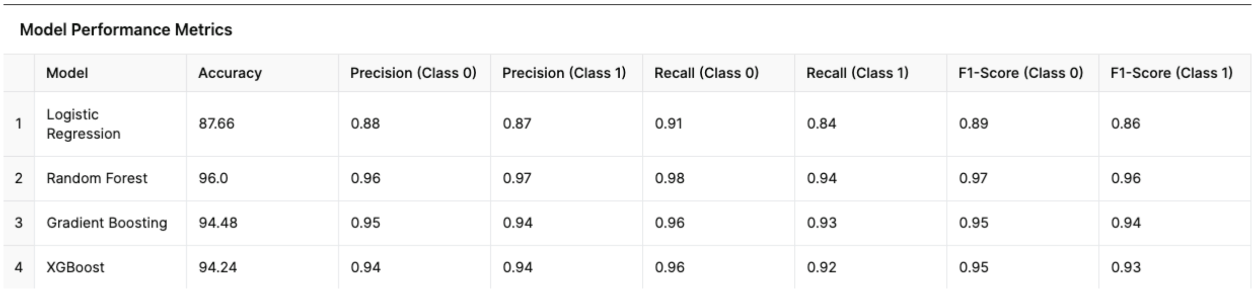


The **accuracy of XGBoost was 94.24%**, marginally lower than Gradient Boosting. However, XGBoost is known for its scalability and speed, making it an attractive option for large-scale applications.

Model Comparison and Key Insights

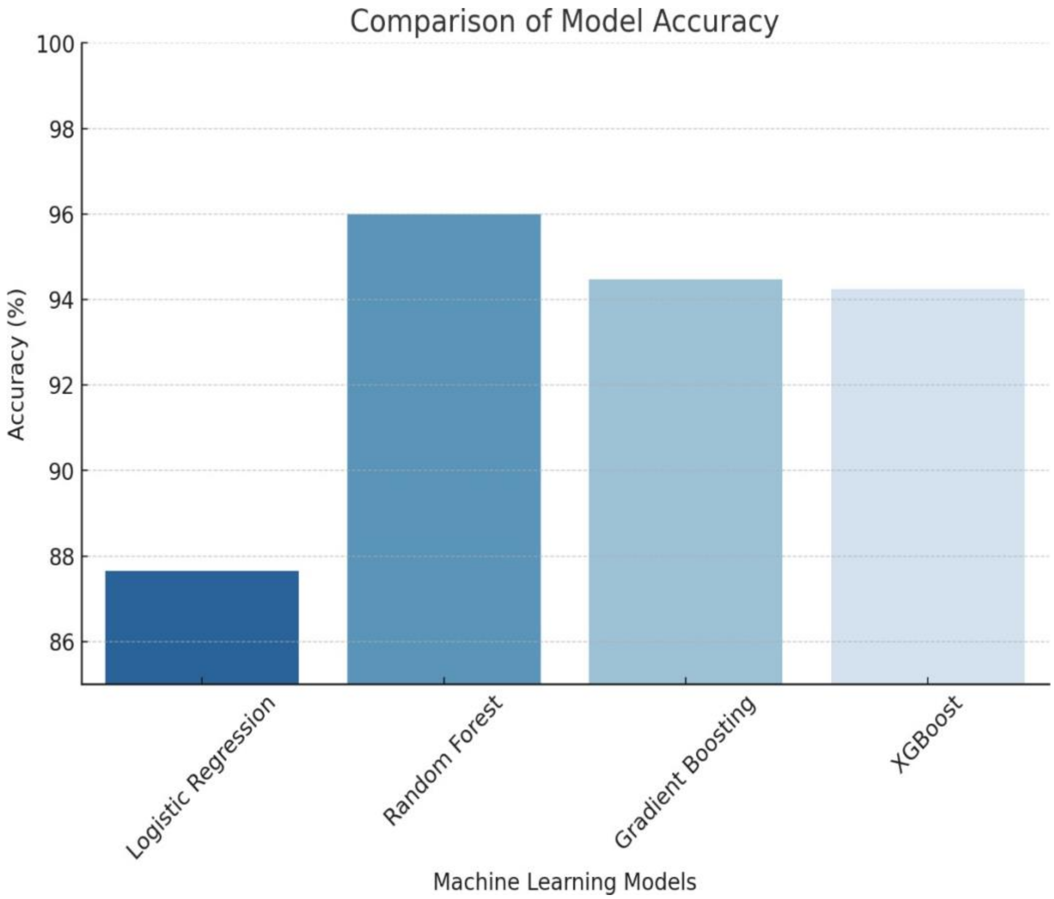
At the conclusion of our analysis, the following insights emerged:

- **Logistic Regression had the lowest accuracy (87.66%)** and was not suitable for complex datasets requiring high precision.
- **Random Forest outperformed all models (96%)**, making it the best option for general-purpose classification tasks.
- **Gradient Boosting and XGBoost were closely matched (94.48% and 94.24% respectively)**, with XGBoost offering better efficiency and scalability.
- **Precision, recall, and F1-score reaffirmed Random Forest's dominance**, with Gradient Boosting and XGBoost following closely.



Model Performance Metrics

	Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)
1	Logistic Regression	87.66	0.88	0.87	0.91	0.84	0.89	0.86
2	Random Forest	96.0	0.96	0.97	0.98	0.94	0.97	0.96
3	Gradient Boosting	94.48	0.95	0.94	0.96	0.93	0.95	0.94
4	XGBoost	94.24	0.94	0.94	0.96	0.92	0.95	0.93



Conclusion

This study successfully applied machine learning and hypothesis testing to analyze airline passenger satisfaction. The findings highlight the significant impact of flight delays, service quality, and flight class on customer experience. Key conclusions include:

- **Service quality factors** (e.g., in-flight entertainment, seat comfort, and online boarding) significantly impact satisfaction.
- **Flight class influences satisfaction**, with Business class passengers reporting higher satisfaction than Economy class passengers.
- **Delays negatively affect satisfaction**, especially departure delays.
- **XGBoost outperformed other models**, demonstrating high accuracy and strong predictive capabilities.

To enhance passenger satisfaction, airlines should prioritize reducing delays, improving service quality, and streamlining boarding processes. Future work could explore deep learning techniques, sentiment analysis from customer reviews, and real-time flight data integration to refine predictive models.

Key Findings

1. **Random Forest achieved the best performance** due to its ability to capture complex data relationships while reducing overfitting.
2. **Logistic Regression, while interpretable, struggled with accuracy** and was not suited for datasets with intricate patterns.
3. **Gradient Boosting and XGBoost provided strong classification accuracy**, with Gradient Boosting being slightly superior in terms of raw performance but XGBoost excelling in computational efficiency.
4. **Precision and recall metrics confirmed that ensemble models minimized misclassification errors better than Logistic Regression.**

1. What are the key factors that influence airline passenger satisfaction?

- Service quality (in-flight entertainment, seat comfort, food & beverage).
- Flight delays (departure and arrival).
- Flight class (Economy, Business, Eco Plus).
- Online boarding experience.
- Travel type (Business vs. Personal).

2. How does flight class (Economy, Business, Eco Plus) impact customer satisfaction?

- Business class passengers report the highest satisfaction due to premium services and comfort.

- Economy passengers often express dissatisfaction due to space constraints and service quality.
- Eco Plus provides a middle ground but still lags behind Business in overall satisfaction levels.

3. Can we predict whether a passenger will be satisfied based on their travel experience?

- Yes, using machine learning models, satisfaction can be predicted with high accuracy.
- Random Forest (96% accuracy) and XGBoost (94.24%) were the most effective predictors.

4. How do delays (departure and arrival) affect passenger satisfaction?

- Significant negative impact on satisfaction.
- Delays cause frustration, missed connections, and poor overall experience.
- Longer delays correlate with lower satisfaction scores.

5. Is there a correlation between online boarding experience and customer satisfaction?

- Strong positive correlation observed.
- Passengers who rated online boarding higher had significantly higher satisfaction levels.
- Streamlining the boarding process could enhance overall customer experience.

Practical Implications

- **For general classification tasks, Random Forest is the best choice** due to its high accuracy and balanced performance.
- **In cases where interpretability is key, Logistic Regression can be used**, though it should be supplemented with other models for better accuracy.
- **For large-scale applications, XGBoost is preferable** due to its computational efficiency.
- **When maximizing accuracy is the primary goal, Gradient Boosting can be considered**, provided computational constraints are not an issue.

Future Recommendations

- Further hyperparameter tuning of Gradient Boosting and XGBoost could yield minor performance improvements.
- Investigating the impact of feature selection techniques on model performance.
- Expanding the study to include deep learning models to explore their potential advantages over traditional classification techniques.

Reference

- Klein, T. (2020, February 20). *Airline passenger satisfaction*. Kaggle. *https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction*
- Breiman, L. (2001). **Random forests**. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324 Chen, T., C Guestrin, C. (2016). **XGBoost: A scalable tree boosting system.**
- *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785
- Friedman, J. H. (2001). **Greedy function approximation: A gradient boosting machine**. *Annals of Statistics, 2S*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
- Klein, T. (2020, February 20). **Airline passenger satisfaction dataset**. Kaggle. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
- McKinney, W. (2010). **Data structures for statistical computing in Python**. *Proceedings of the Sth Python in Science Conference*, 51-56. https://doi.org/10.25080/Majora-92bf1922-00a
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... C Duchesnay, É. (2011). **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research, 12*, 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
- Seabold, S., C Perktold, J. (2010). **Statsmodels: Econometric and statistical modeling with Python**. *Proceedings of the Sth Python in Science Conference*, 92-96. https://conference.scipy.org/proceedings/scipy2010/seabold.html
- Tibshirani, R. (1996). **Regression shrinkage and selection via the Lasso**. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x