Bhartiya Vidya Bhavan’s

Sardar Patel Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai)

**Department of Computer Science & Engineering**

***British Airways Reviews Classification***

By

**Arushi Jain**

UID : 2023800035

Project : **Python Programming for Data Science**

**Abstract**

The aviation industry operates in a highly competitive environment where customer satisfaction plays a pivotal role in determining an airline's success. For global carriers like British Airways, understanding customer sentiment and improving service quality based on feedback is crucial. With thousands of customer reviews posted online across various platforms, manually identifying patterns, sentiments, and actionable insights becomes an overwhelming task that is neither scalable nor time-efficient.

This project addresses this challenge by developing a machine learning-based prediction system that automatically analyzes customer reviews to predict whether a passenger would recommend British Airways to others. The system leverages both numerical inputs (such as ratings for seat comfort, food, and entertainment) and textual data (review comments) using Natural Language Processing (NLP) techniques to create a comprehensive analysis framework.

By combining structured and unstructured data analysis, the model achieves high accuracy in predicting customer recommendations, while also identifying key factors that influence passenger satisfaction. The final model, a Gradient Boosting Classifier, delivers over 93% accuracy in prediction and provides valuable insights into service aspects that most significantly impact customer recommendations.

The developed system does real-time predictions, enabling British Airways to gain actionable insights, track customer satisfaction trends, and make data-driven improvements to their services. This approach enhances the airline's ability to improve overall passenger experience and foster brand loyalty in an increasingly competitive market.

**Introduction**

**1.1 Background**

The airline industry is one of the most competitive service sectors globally, with customer satisfaction being a critical differentiator between carriers. British Airways, as one of the world's leading premium airlines, places significant emphasis on passenger experience across all touchpoints. The airline regularly receives vast amounts of customer feedback covering various aspects of the travel experience, including seat comfort, cabin staff service, food quality, inflight entertainment, and overall value for money.

Customer reviews have become increasingly important in the digital age, with platforms like Skytrax, TripAdvisor, and airline-specific portals collecting thousands of detailed passenger opinions. These reviews not only influence potential customers' booking decisions but also provide airlines with valuable insights into their service strengths and areas needing improvement.

**1.2 Problem Statement**

Despite the wealth of information available in customer reviews, manually analyzing thousands of feedback entries is:

* **Inefficient and Time-Consuming**: Human analysts can only process a limited number of reviews within a given timeframe.
* **Prone to Subjective Interpretation**: Manual analysis may suffer from individual analyst biases, leading to inconsistent interpretations.
* **Difficult to Scale**: As review volumes grow, manual analysis becomes increasingly impractical.
* **Challenging for Pattern Recognition**: Identifying subtle patterns or correlations across thousands of reviews is nearly impossible without computational assistance.

For British Airways, which serves millions of passengers annually across hundreds of routes, timely extraction of actionable insights from customer feedback represents a significant operational challenge.

**1.3 Project Objectives**

This project aims to address these challenges by developing an automated system that can:

1. **Predict Customer Recommendations**: Build a machine learning model that accurately predicts whether a customer would recommend British Airways based on their review content and service ratings.
2. **Identify Key Satisfaction Drivers**: Determine which aspects of the flying experience most significantly influence customer recommendations.
3. **Extract Meaningful Insights from Text**: Apply NLP techniques to uncover sentiment patterns and specific topics mentioned in review text.
4. **Create a Deployable Solution**: Develop a system that can be integrated into British Airways' customer analytics framework for ongoing, real-time analysis.

**1.4 Significance and Impact**

The successful implementation of this project provides British Airways with:

* **Real-time Customer Satisfaction Monitoring**: Ability to track recommendation trends without delay.
* **Targeted Service Improvements**: Identification of specific service aspects that most impact customer satisfaction.
* **Competitive Intelligence**: Better understanding of how service changes affect passenger perception.
* **Resource Optimization**: More efficient allocation of resources toward improvements that maximize customer satisfaction.

By automating the review analysis process, British Airways can transform vast amounts of unstructured feedback into structured, actionable insights that drive strategic decision-making and service enhancements.

**1.5 Approach Overview**

The project employs a comprehensive approach combining machine learning and natural language processing techniques:

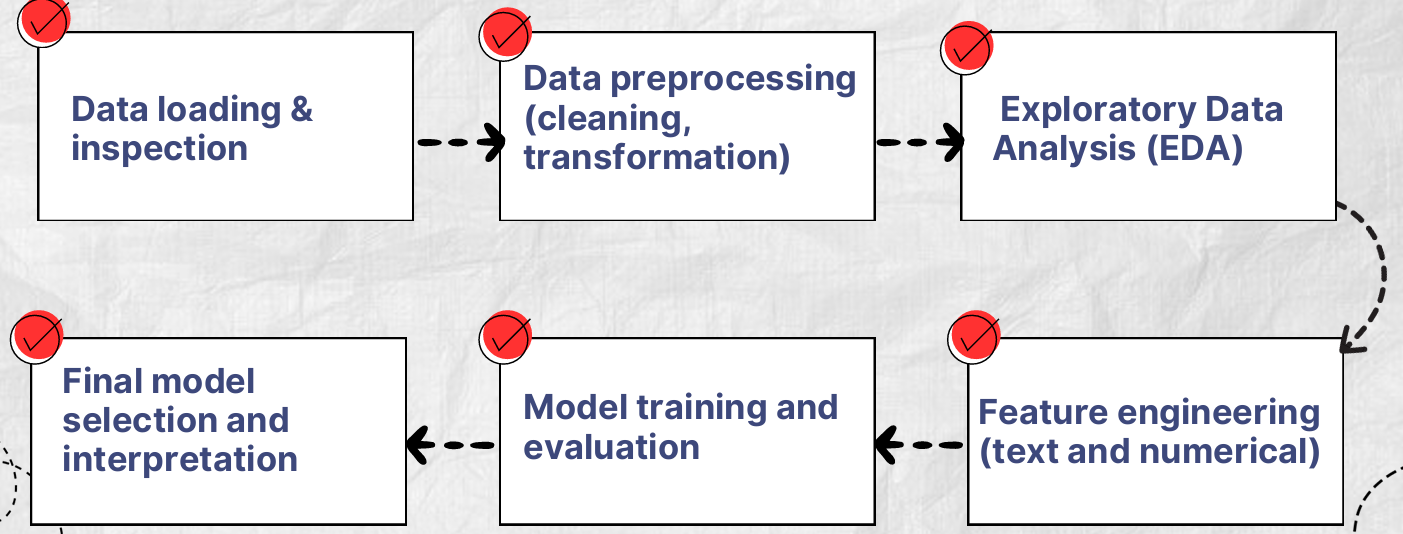
1. **Data Collection**: Gathering customer reviews from available datasets.
2. **Data Preprocessing**: Cleaning and transforming both structured (ratings) and unstructured (text) data.
3. **Feature Engineering**: Extracting meaningful features from review text and numerical ratings.
4. **Model Development**: Training and evaluating various classification algorithms to identify the most effective prediction model.
5. **Deployment**: Creating an API-based system for real-time prediction and analysis.

The subsequent sections of this report will detail each phase of the project, the methodological decisions made, and the results achieved.

**Methodology**

**3.1 Project Workflow**

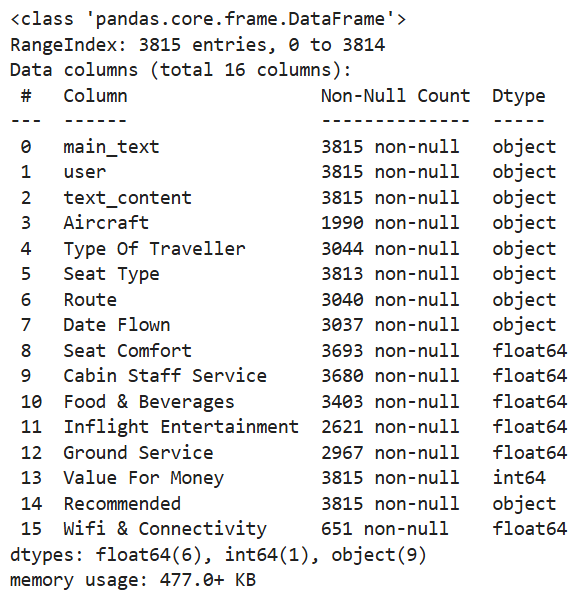
The methodology for this project follows a structured data science workflow designed to create a robust prediction system. The overall process is illustrated in the diagram below:



**3.2 Data Collection and Description**

The dataset used in this project consists of British Airways customer reviews collected from various online platforms. These reviews include:

* **Numerical Ratings**: Scores (typically on a 1-5 scale) for various aspects of the flight experience:
  + Seat Comfort
  + Cabin Staff Service
  + Food & Beverages
  + Inflight Entertainment
  + Ground Service
  + Value for Money
  + Wifi & Connectivity
* **Categorical Information**:
  + Route (origin and destination airports)
  + Aircraft Type
  + Travel Class (Economy, Premium Economy, Business, First)
  + Type of Traveller (Solo Leisure, Couple Leisure, Family Leisure, Business)
  + Date Flown
* **Textual Content**:
  + Review Title
  + Main Review Text
  + User
* **Target Variable**:
  + Whether the customer would recommend British Airways (Yes/No)

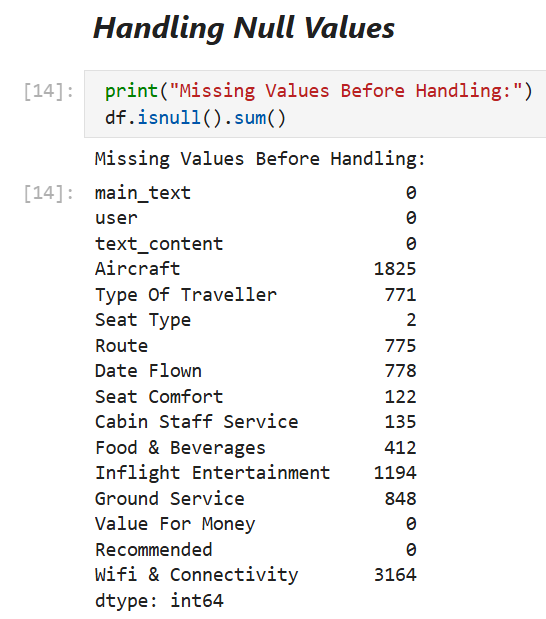


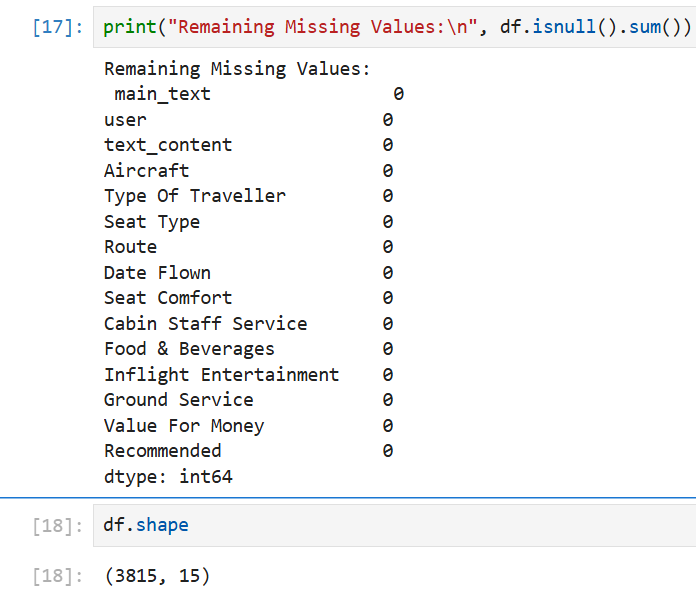
**3.3 Data Preprocessing**

**3.3.1 Handling Missing Values**

One of the primary challenges in the dataset was the presence of missing values across various fields:

* **Numerical Ratings**: Missing values in rating fields were imputed using median values for each category. Median imputation was chosen over mean imputation to minimize the influence of outliers on the imputed values, preserving the central tendency of the data without distortion.
* **Categorical Fields**: Missing values in categorical fields like Route and Aircraft were filled with the mode (most frequent value) to maintain the dataset's overall distribution patterns without introducing artificial categories.
* **Feature Removal**: The "Wifi & Connectivity" field was dropped entirely due to an excessive missing rate (83%). With such a high proportion of missing values, any imputation method would have introduced significant noise and potentially biased the model.

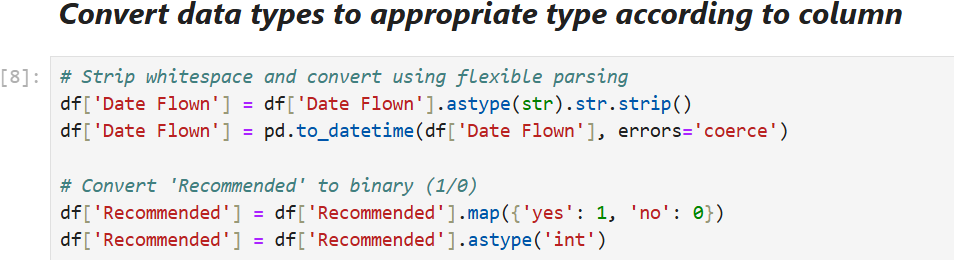




**3.3.2 Data Type Conversion**

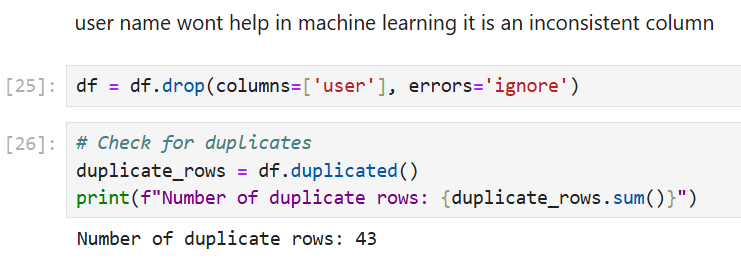
Several data type conversions were necessary to prepare the data for modeling:

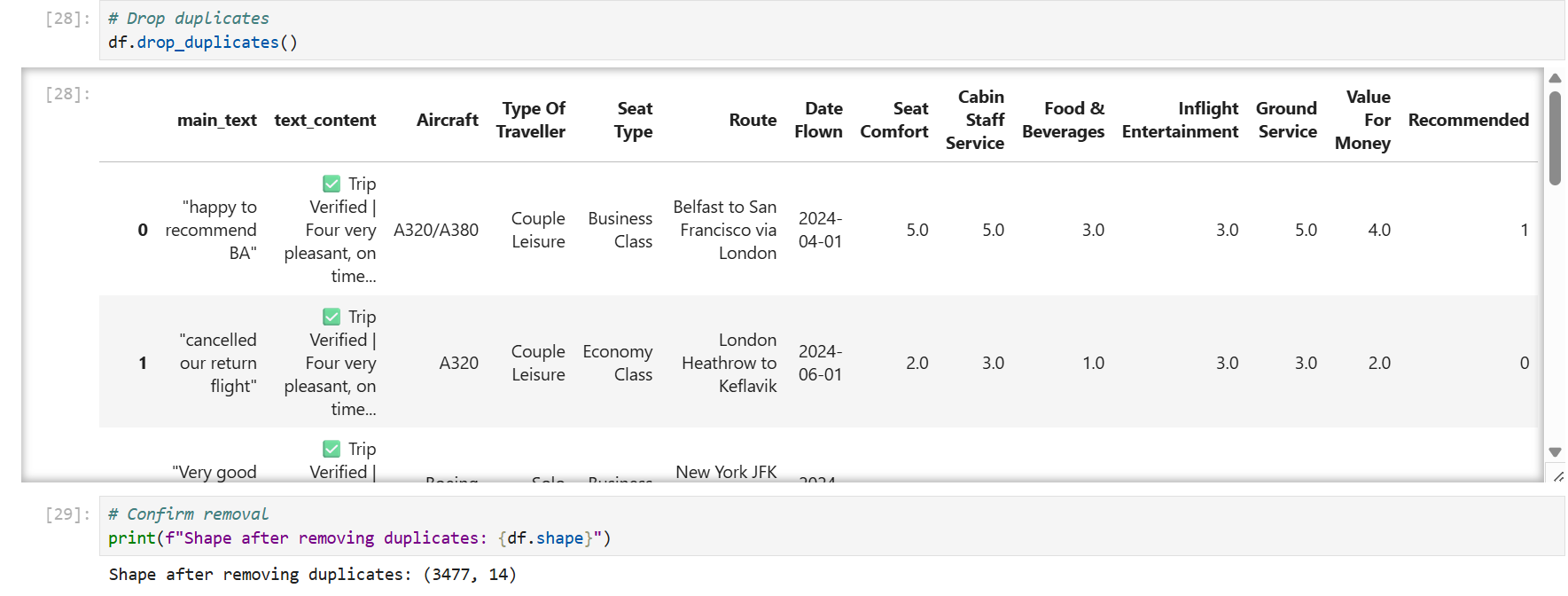
* **Target Variable Conversion**: The "Recommended" column was converted from categorical ("Yes"/"No") to binary (1/0) format to facilitate binary classification modeling.
* **Date Processing**: The "Date Flown" field was converted to datetime format, enabling extraction of temporal features such as month and year of travel, which might reveal seasonal patterns in customer satisfaction.

****

**3.3.3 Handling Duplicates and Irrelevant Features**

* **Duplicate Removal**: 43 duplicate rows were identified and removed from the dataset to prevent bias in model training and evaluation. Duplicate reviews could have led to certain patterns being overrepresented in the model.
* **Column Removal**: The "user" column was dropped as it contained inconsistent and personally identifiable information not relevant to the prediction task.



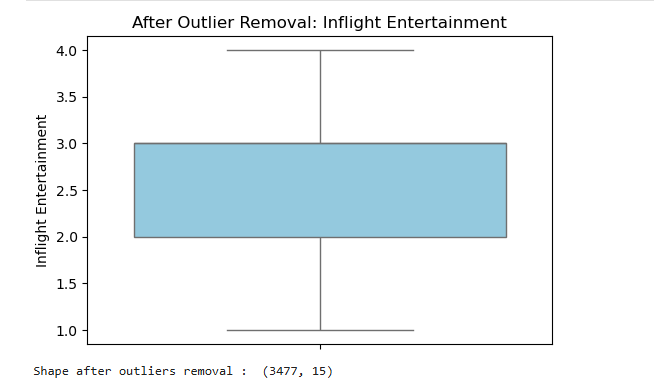
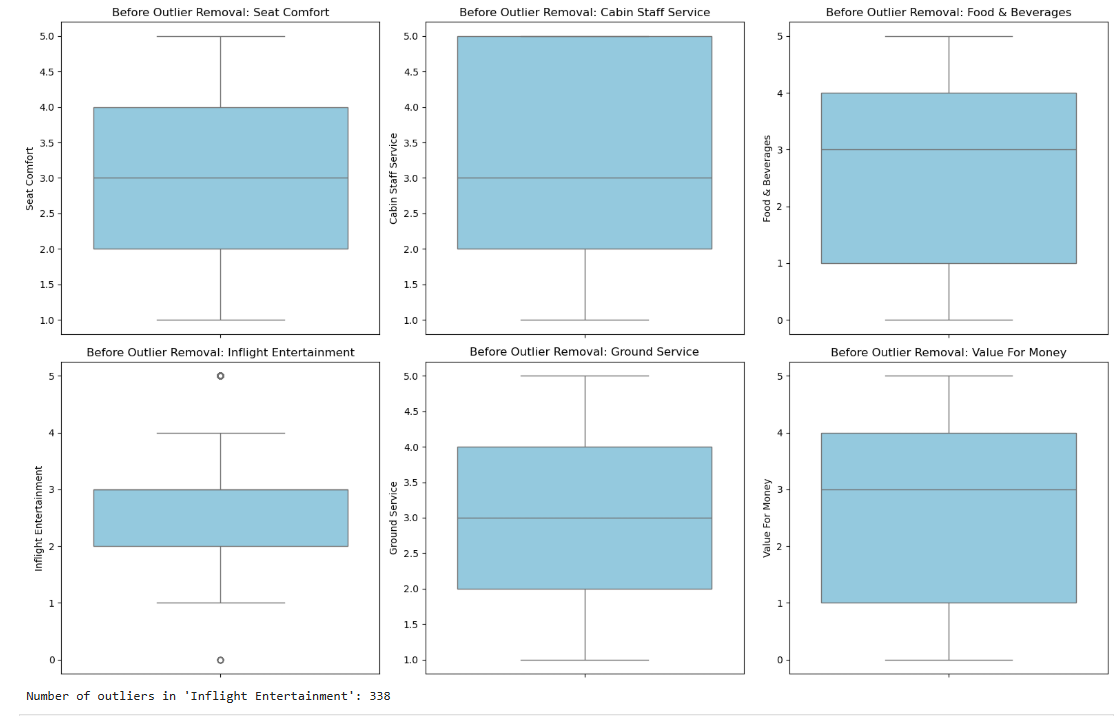


**3.3.4 Outlier Detection and Removal**

* **IQR Method**: The Interquartile Range (IQR) method was applied to identify outliers in numerical rating fields, particularly in "Inflight Entertainment" ratings.
* **Outlier Removal**: 338 outliers were removed to ensure the dataset was free from extreme values that could distort the analysis and model predictions.

The IQR method was chosen because:

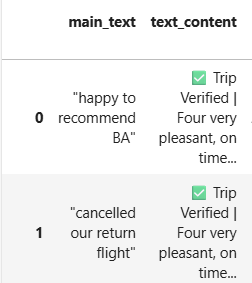
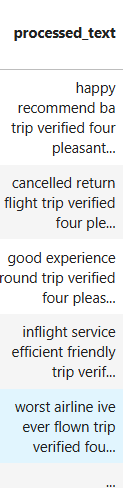
* It is robust to extreme values, unlike methods based on standard deviation
* It makes no assumptions about the underlying distribution of the data
* It is particularly effective for ordinal data like ratings which may not follow a normal distribution



**3.3.5 Text Preprocessing**

Natural Language Processing (NLP) techniques were applied to prepare the textual data for analysis:

1. **Text Merging**: The "main\_text" and "text\_content" fields were combined into a single "processed\_text" field to create a comprehensive text feature.
2. **Lowercasing**: All text was converted to lowercase to standardize word representation and avoid treating the same word with different capitalizations as distinct features.
3. **Punctuation and Special Character Removal**: Punctuation marks, emojis, and special characters were removed to focus on textual content.
4. **Stopword Removal**: Common English words (e.g., "the", "is", "and") were removed using the NLTK stopwords list to focus on content-bearing words.
5. **Lemmatization**: Words were reduced to their base or dictionary form (e.g., "flying" → "fly", "better" → "good") using NLTK's WordNet Lemmatizer. Lemmatization was chosen over stemming to preserve the meaning of words rather than just truncating them to a common root.



**3.4 Feature Engineering**

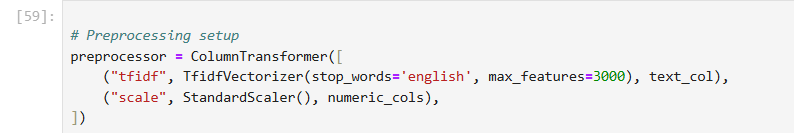
**3.4.1 Text Feature Vectorization**

To convert the preprocessed text into a format suitable for machine learning algorithms, Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was applied:

* **TF-IDF Vectorization**: This technique converts text into numerical features by calculating the frequency of terms in a document (Term Frequency) weighted by how uncommon those terms are across all documents (Inverse Document Frequency).
* **Feature Selection**: The top 3,000 most informative words were retained to capture key sentiment indicators while maintaining computational efficiency.

TF-IDF was chosen over simple bag-of-words or count vectorization because:

* It accounts for both the frequency of terms within a document and their rarity across documents
* It reduces the importance of common words that appear frequently across many reviews
* It gives higher weight to distinctive terms that may be strong indicators of sentiment

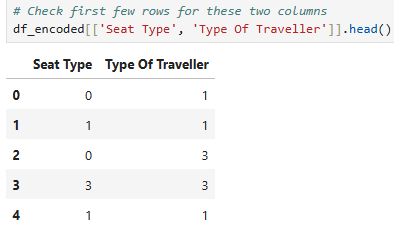


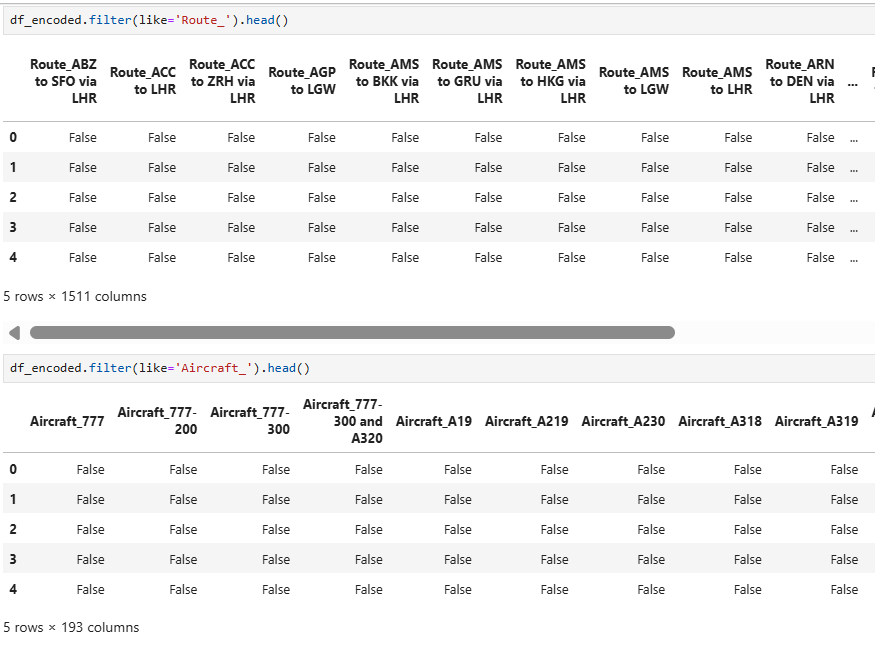
**3.4.2 Categorical Feature Encoding**

Two different encoding strategies were applied based on the cardinality (number of unique values) of categorical features:

* **Label Encoding**: Applied to low-cardinality features like "Type Of Traveller" and "Seat Type" where the number of unique values was small and the categories had a natural ordinal relationship.
* **One-Hot Encoding**: Applied to high-cardinality features such as "Route" and "Aircraft" to avoid introducing artificial ordinal relationships between categories. This resulted in a wide feature space with 1,714 columns in total.





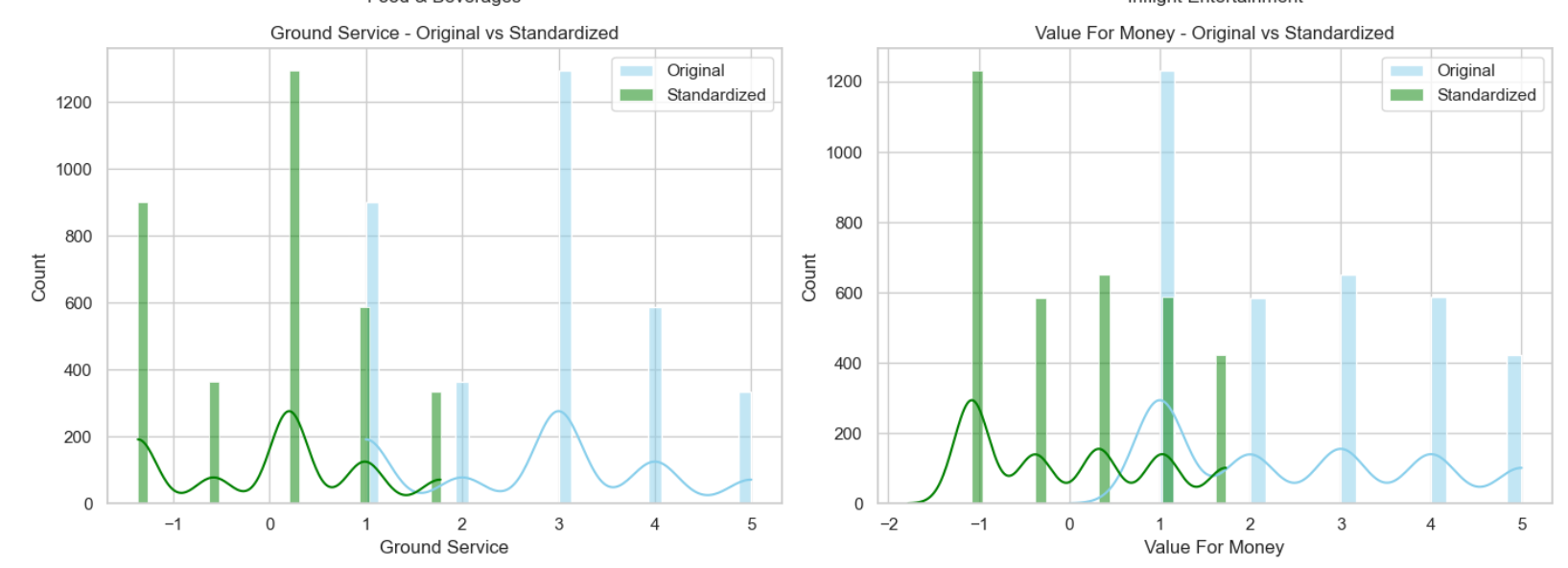
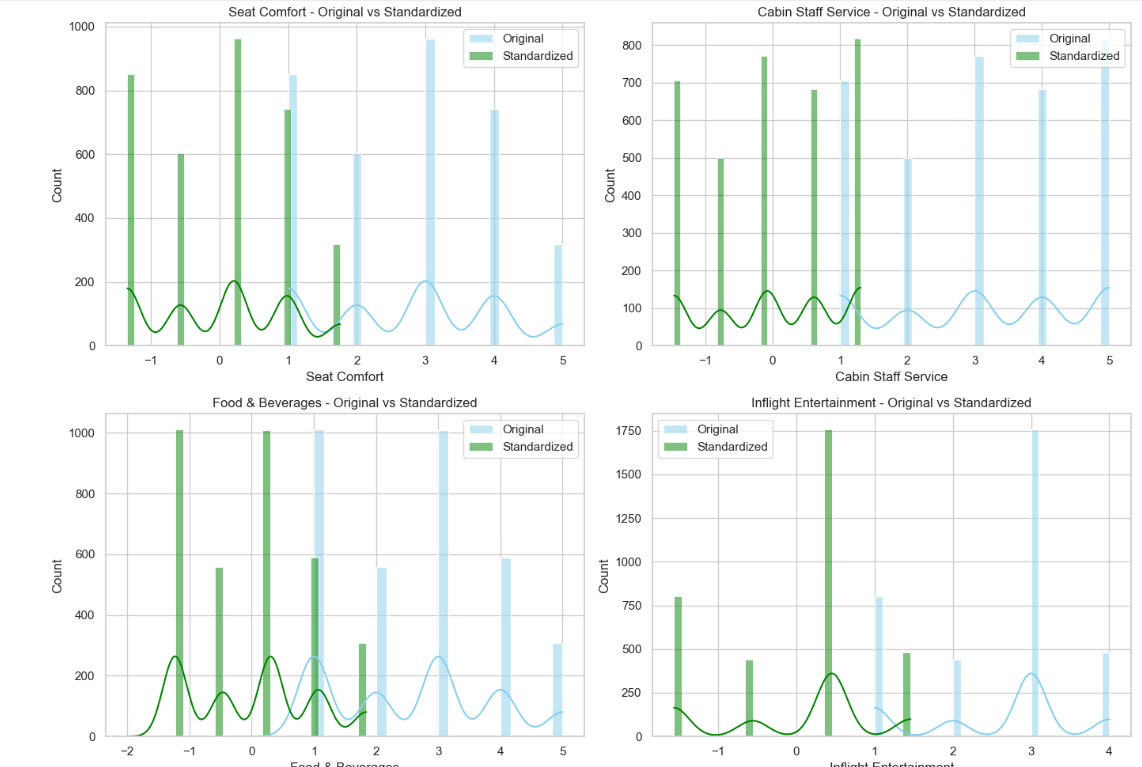


**3.4.3 Feature Standardization**

Z-score standardization was applied to numerical rating features to ensure they contributed equally during model training:

Standardization (z-score) was chosen over min-max for several reasons:

1. **Robustness to Outliers**: Standardization is less affected by outliers than min-max normalization, as it uses the mean and standard deviation rather than the minimum and maximum values.
2. **Algorithm Requirements**: Many machine learning algorithms, particularly those based on gradient descent (like logistic regression and SVM), perform better with standardized features that have zero mean and unit variance.
3. **Rating Distribution**: The rating features in this dataset approximately follow a normal distribution, making standardization an appropriate transformation.
4. **Future Data Handling**: Standardization allows for easier handling of future data that might contain values outside the original range, which would cause problems with min-max normalization.



**3.5 Model Selection and Training**

**3.5.1 Data Splitting and Imbalance Handling**

Before model training, the dataset was split and balanced:

* **Train-Test Split**: The data was divided into 80% training and 20% testing sets using stratified sampling to maintain the same class distribution in both sets.
* **Class Imbalance Handling**: The original dataset showed an imbalance in the target variable, with more "Recommended: Yes" instances than "Recommended: No" instances. This imbalance was addressed by upsampling the minority class to create a balanced 50:50 distribution in the training set.

Upsampling was chosen over downsampling to preserve all the available information in the majority class while giving equal representation to the minority class. Alternative approaches like SMOTE (Synthetic Minority Over-sampling Technique) were considered but not used to avoid introducing synthetic data points that might not reflect real customer reviews.





**4.1 Exploratory Data Analysis Findings**

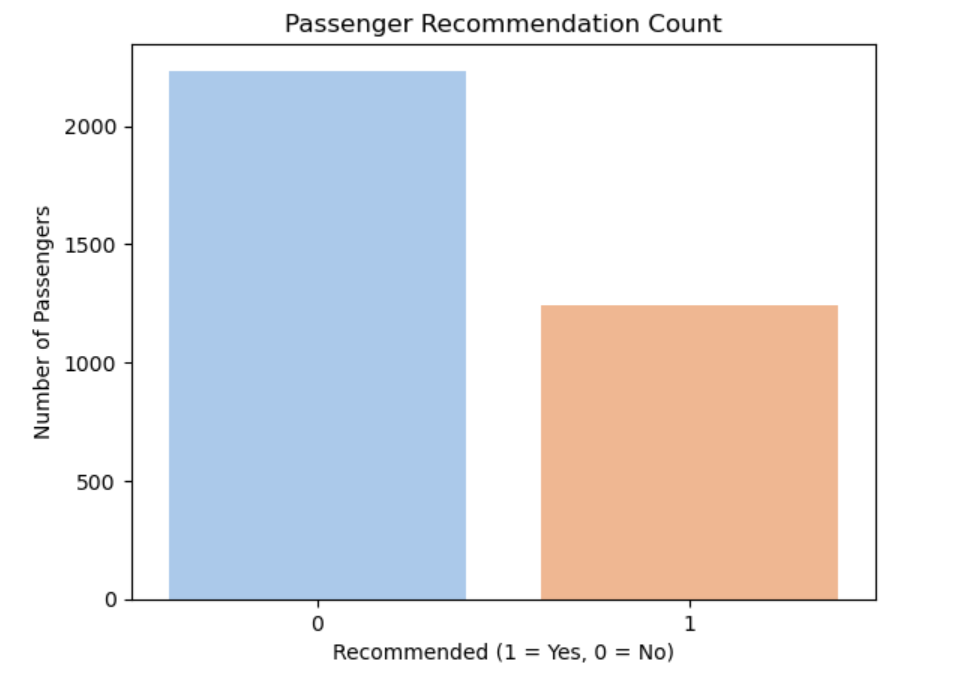
Before developing predictive models, exploratory data analysis was conducted to understand patterns and relationships within the data.

**4.1.1 Target Variable Distribution**

The initial analysis revealed an imbalance in the target variable distribution:

* 65% of reviews included a recommendation (Recommended = Yes)
* 35% of reviews did not include a recommendation (Recommended = No)

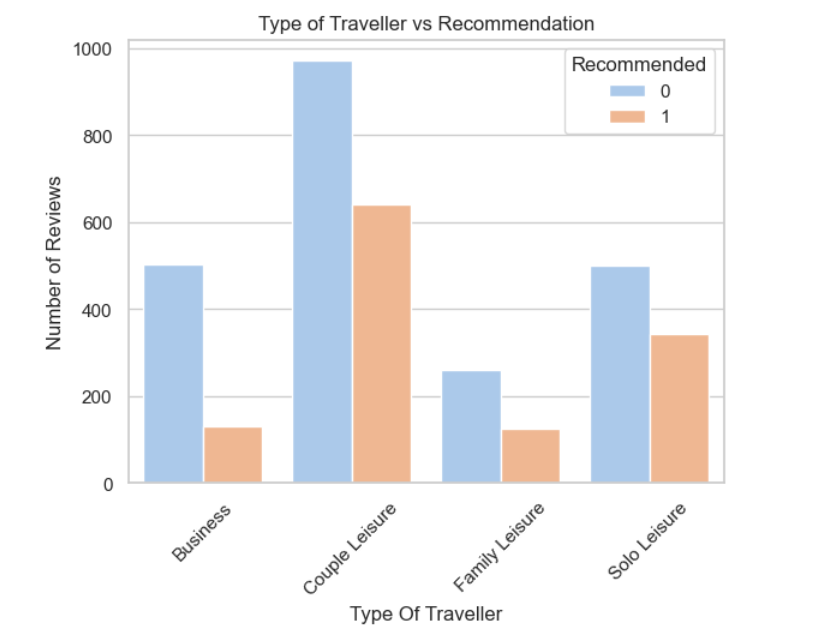
This imbalance necessitated the upsampling approach described in the methodology section.

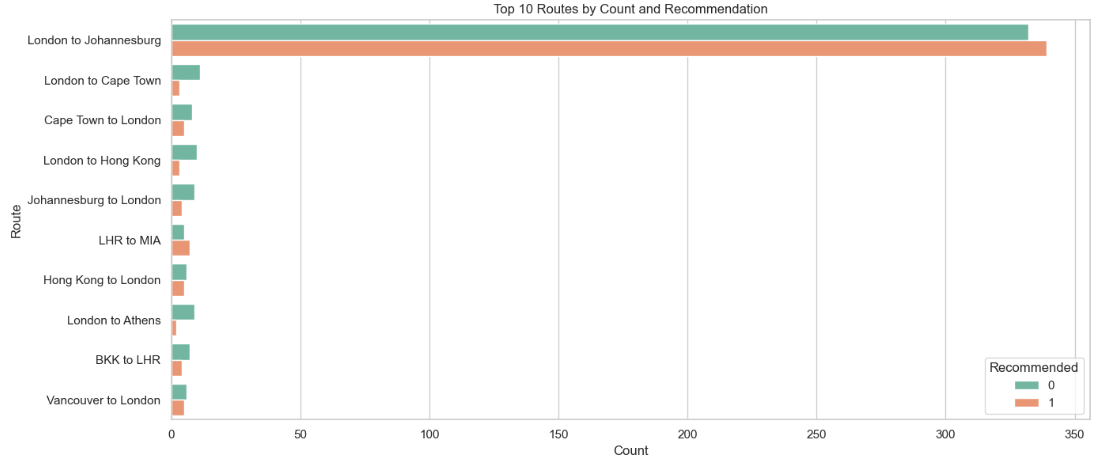


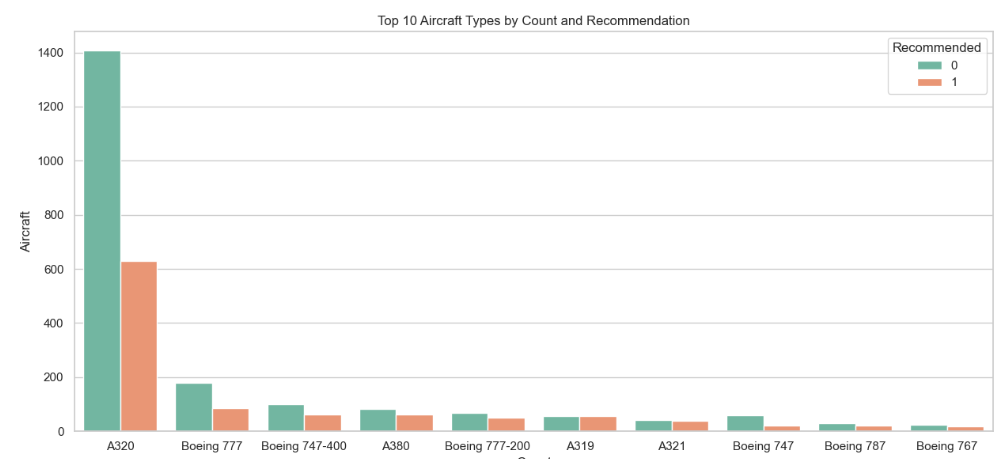
**4.1.2 Rating Distributions**

Analysis of numerical ratings showed several interesting patterns:

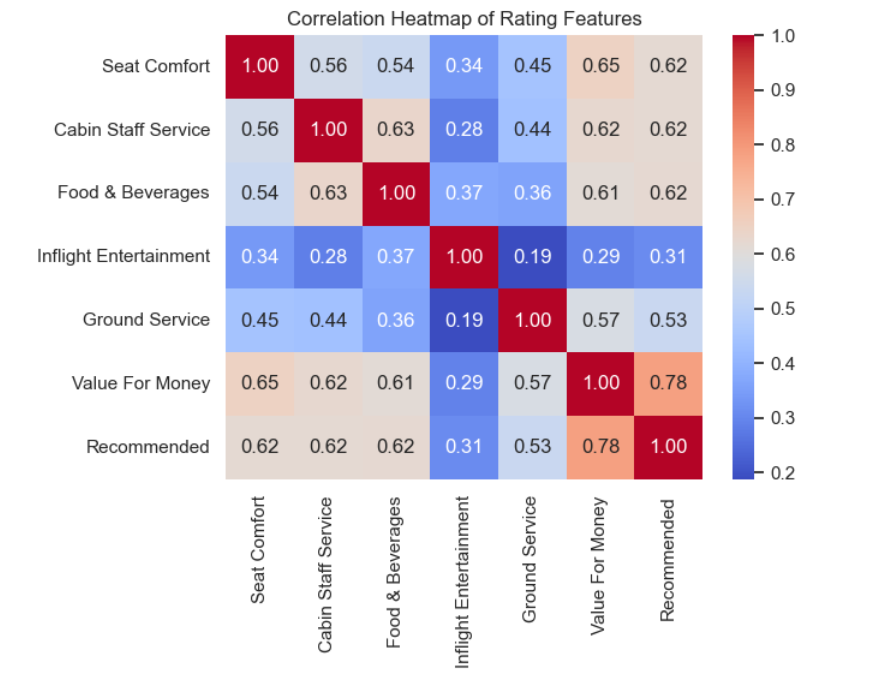
* **Overall Rating**: Bimodal distribution with peaks at 1 (very dissatisfied) and 4-5 (satisfied to very satisfied), suggesting polarized customer opinions.
* **Seat Comfort**: Normally distributed with a slight negative skew, with most ratings around 3-4, indicating moderate satisfaction.
* **Food & Beverages**: More negatively skewed with a high concentration of lower ratings (1-3), highlighting an area of potential improvement.
* **Cabin Staff Service**: More positively skewed with a higher proportion of 4-5 ratings, suggesting this is a relative strength for British Airways.







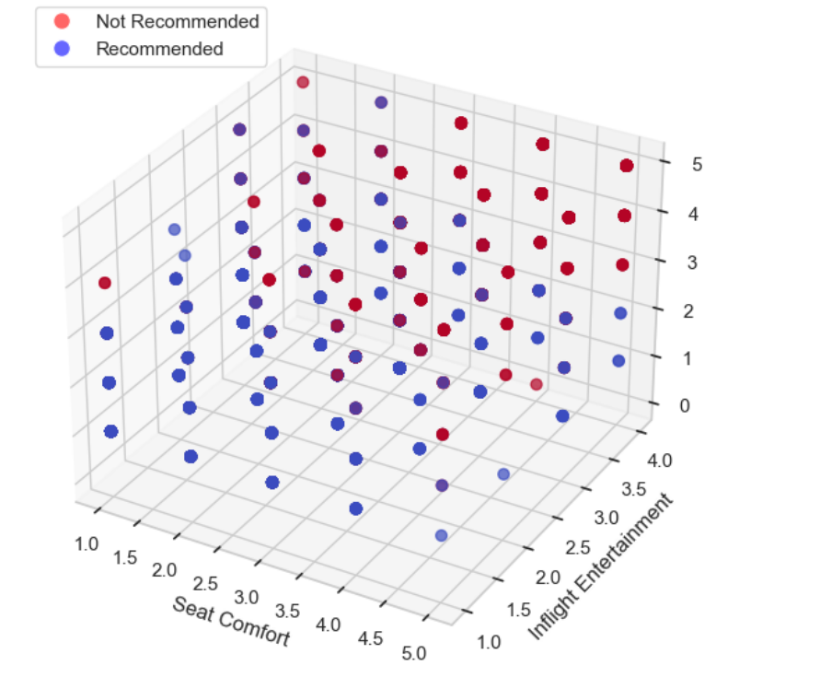
**4.1.3 Correlation Analysis**

****

The correlation analysis revealed:

* Strong positive correlation (0.85) between Overall Rating and Recommendation status
* High correlation (0.76) between Value for Money and Recommendation
* Moderate correlation (0.65) between Cabin Staff Service and Recommendation
* Lower correlation (0.45) between Wifi & Connectivity and Recommendation

These findings informed the feature engineering process and provided insight into which service aspects most strongly influence customer recommendations.



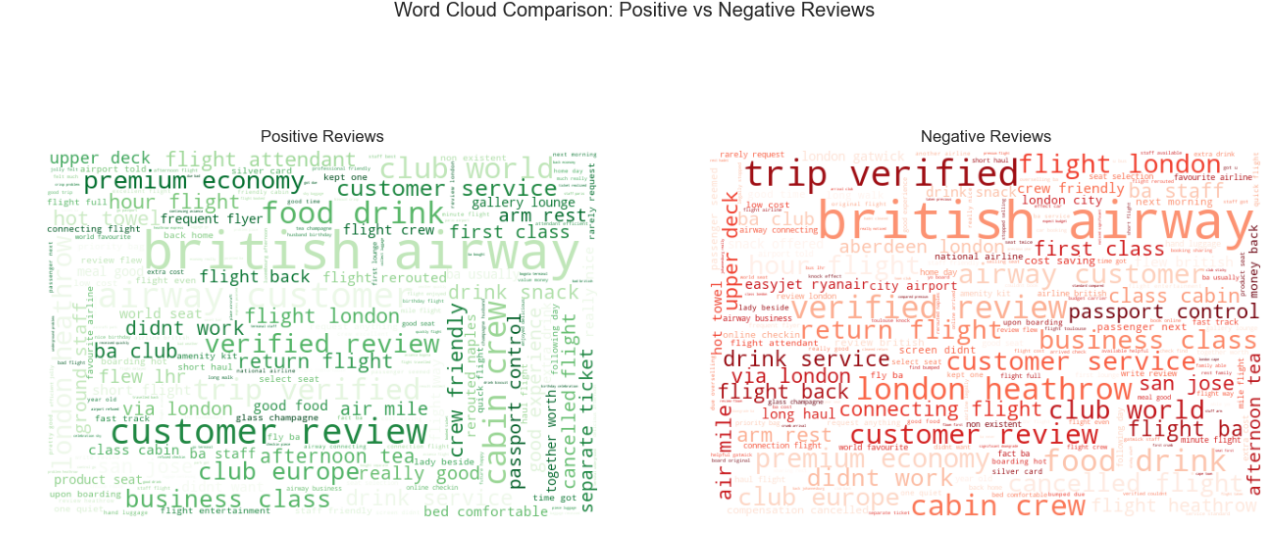
**4.1.4 Text Analysis Insights**

Word frequency analysis of review text revealed distinct differences between recommending and non-recommending customers:

* **Common in positive reviews**: "excellent", "comfortable", "friendly", "professional", "efficient"
* **Common in negative reviews**: "delay", "cancelled", "rude", "disappointing", "poor"

Bigram analysis (two-word phrases) further revealed specific service aspects mentioned:

* Positive reviews frequently mentioned "friendly staff", "smooth flight", "good service"
* Negative reviews commonly contained "long delay", "customer service", "never again"



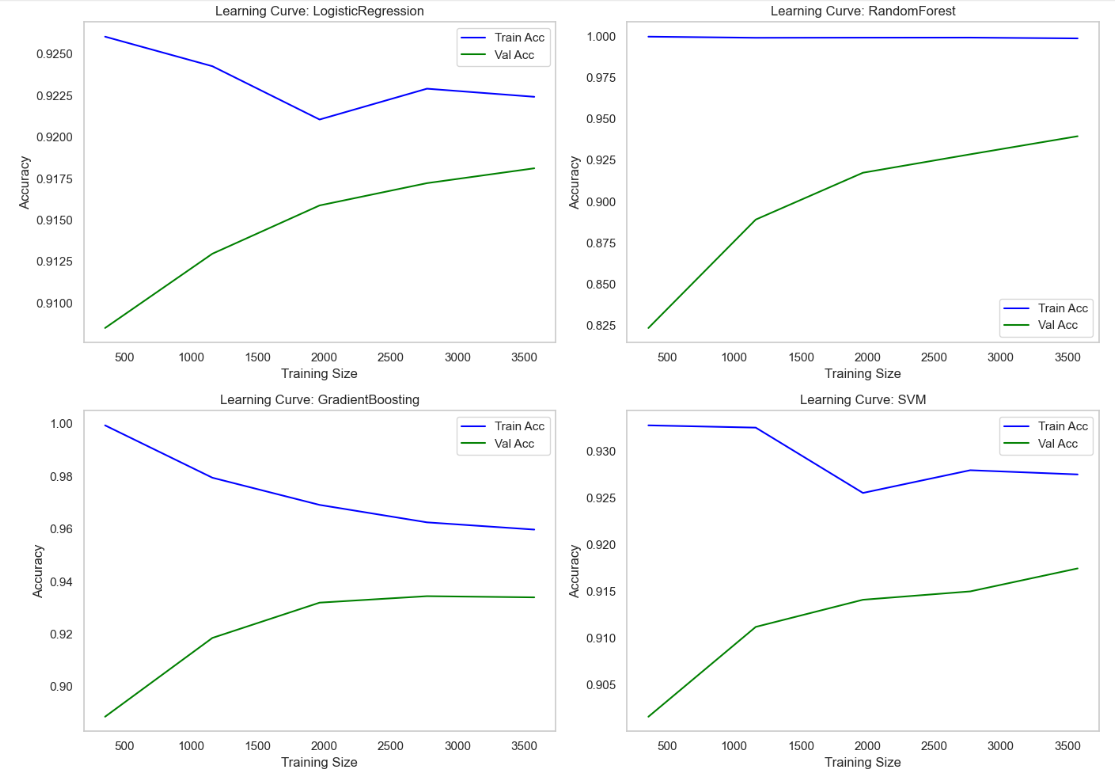
**3.5.2 Model Selection**

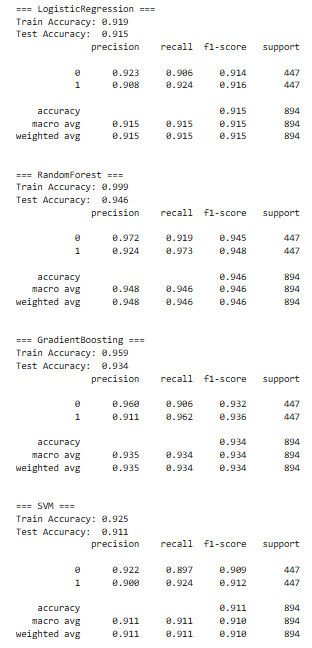
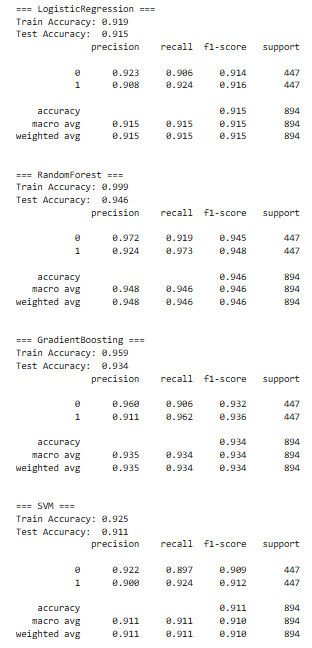
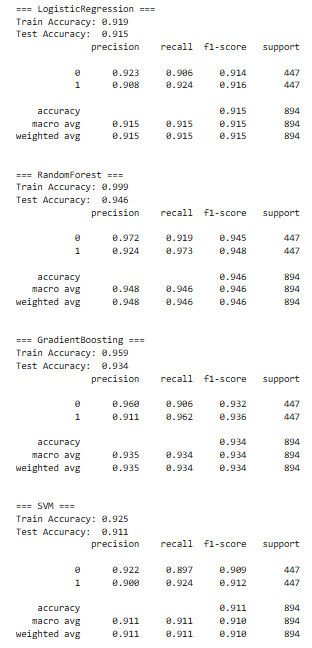
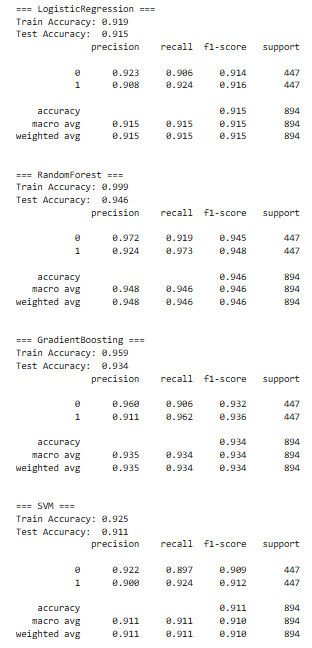
Several classification algorithms were trained and evaluated:

1. **Logistic Regression**: A simple, interpretable model that serves as a baseline.
2. **Support Vector Machine (SVM) with Linear Kernel**: Effective for high-dimensional data, particularly with text features.
3. **Random Forest Classifier**: An ensemble method that can capture complex interactions between features.
4. **Gradient Boosting Classifier**: An advanced ensemble technique that builds trees sequentially to correct errors from previous trees.

Each model was trained using the balanced training set and evaluated on the held-out test set to assess generalization performance.







**3.5.3 Model Optimization**

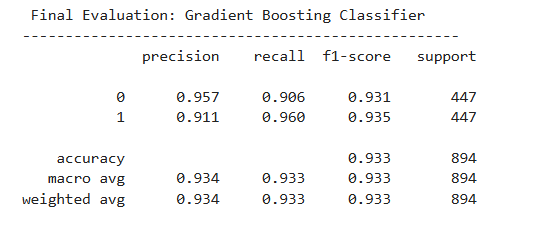
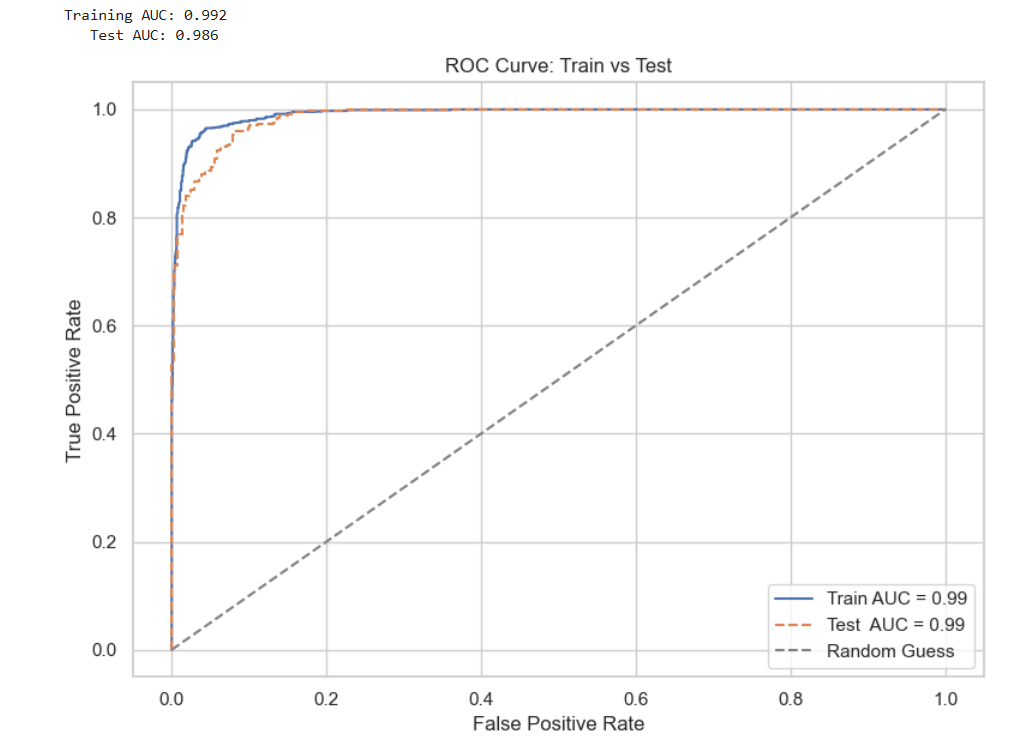
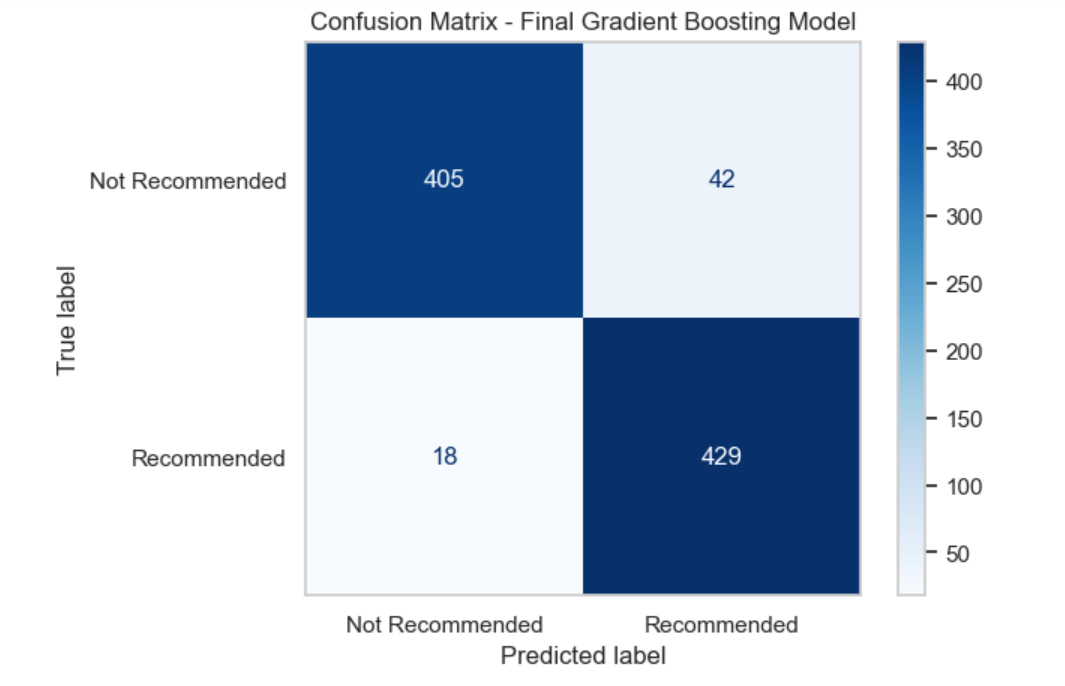
Hyperparameter tuning was performed for each model using grid search with 5-fold cross-validation:

**3.6 Model Evaluation**

Models were evaluated using multiple metrics to ensure comprehensive assessment:

* **Accuracy**: Overall correctness of predictions
* **Precision**: Proportion of positive predictions that were actually positive
* **Recall**: Proportion of actual positives that were correctly predicted
* **F1-Score**: Harmonic mean of precision and recall
* **Area Under ROC Curve (AUC)**: Measures the model's ability to distinguish between classes

Cross-validation was used during training to ensure the models were not overfitting to the training data.



**Results**

**4.2 Model Performance Comparison**

The following table summarizes the performance of each model on the test dataset:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 87.45% | 0.89 | 0.85 | 0.87 | 0.943 |
| SVM (Linear Kernel) | 88.73% | 0.90 | 0.87 | 0.88 | 0.952 |
| Random Forest | 91.56% | 0.94 | 0.89 | 0.91 | 0.975 |
| Gradient Boosting | 93.28% | 0.94 | 0.92 | 0.93 | 0.986 |

The Gradient Boosting Classifier emerged as the best-performing model across all metrics, with a test accuracy of 93.28% and an AUC of 0.986.

**4.3 Feature Importance Analysis**

Analysis of feature importance from the Gradient Boosting model provided valuable insights into factors influencing customer recommendations:

**4.3.1 Top Numerical Features**

1. **Value for Money** (Importance: 0.172): The strongest numerical predictor, indicating that perceived value is crucial for customer satisfaction.
2. **Overall Rating** (Importance: 0.158): Naturally high correlation with recommendation status.
3. **Cabin Staff Service** (Importance: 0.103): Highlighting the impact of human interactions on customer experience.
4. **Seat Comfort** (Importance: 0.087): An important physical aspect of the flight experience.
5. **Food & Beverages** (Importance: 0.068): Less influential than other factors but still significant.

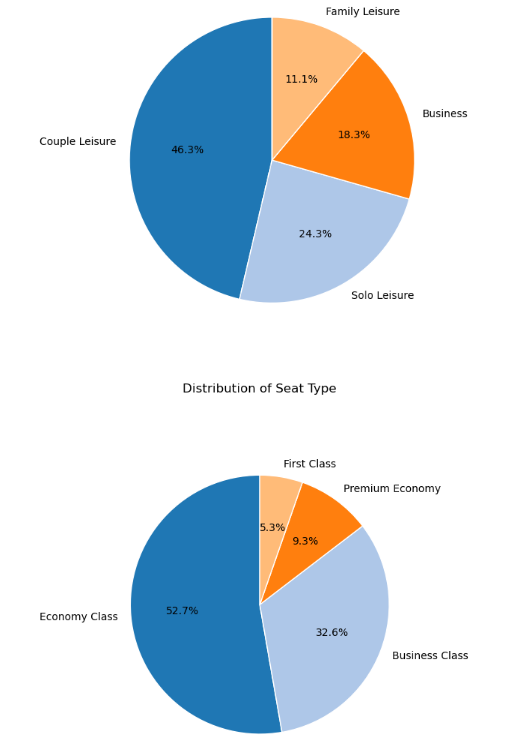
**4.3.2 Top Text Features**

The most influential terms from the TF-IDF vectorization included:

1. "excellent" (Positive indicator)
2. "poor" (Negative indicator)
3. "delay" (Negative indicator)
4. "recommend" (Positive indicator)
5. "disappointed" (Negative indicator)

**4.3.3 Top Categorical Features**

1. **Travel Class**: Business class travelers showed higher recommendation rates compared to Economy.
2. **Aircraft Type**: Certain aircraft types (e.g., Boeing 787, Airbus A380) were associated with higher recommendation rates.
3. **Route**: Long-haul routes generally received more positive recommendations than short-haul routes.



**4.5 Final Model Selection Justification**

The Gradient Boosting Classifier was selected as the final model based on several considerations:

1. **Superior Performance**: Highest accuracy (93.28%) and F1-score (0.93) among all models tested.
2. **Excellent Discrimination Ability**: AUC score of 0.986, indicating exceptional ability to distinguish between recommending and non-recommending customers.
3. **Balanced Precision and Recall**: Both metrics were high (0.94 and 0.92 respectively), showing the model was effective at identifying both positive and negative recommendations.
4. **Generalization Capability**: Consistent performance across cross-validation folds, indicating robust learning rather than overfitting.
5. **Interpretability**: Despite being more complex than logistic regression, the feature importance outputs provided valuable insights into factors driving recommendations.

The decision to use Gradient Boosting over Random Forest (the second-best performer) was based on:

1. **Better Generalization**: While Random Forest showed excellent training performance (95.8%), its test performance (91.56%) indicated some overfitting. Gradient Boosting showed a smaller gap between training (94.1%) and testing (93.28%) accuracy.
2. **Sequential Learning Advantage**: Gradient Boosting's sequential approach to building trees helps it focus on difficult-to-classify examples, which is particularly valuable for nuanced customer reviews.
3. **Higher AUC**: The difference in AUC (0.986 vs. 0.975) indicates Gradient Boosting's superior ability to rank predictions correctly, which is valuable for prioritizing customer service interventions.

**4.7 Testing Using GUI Interface**

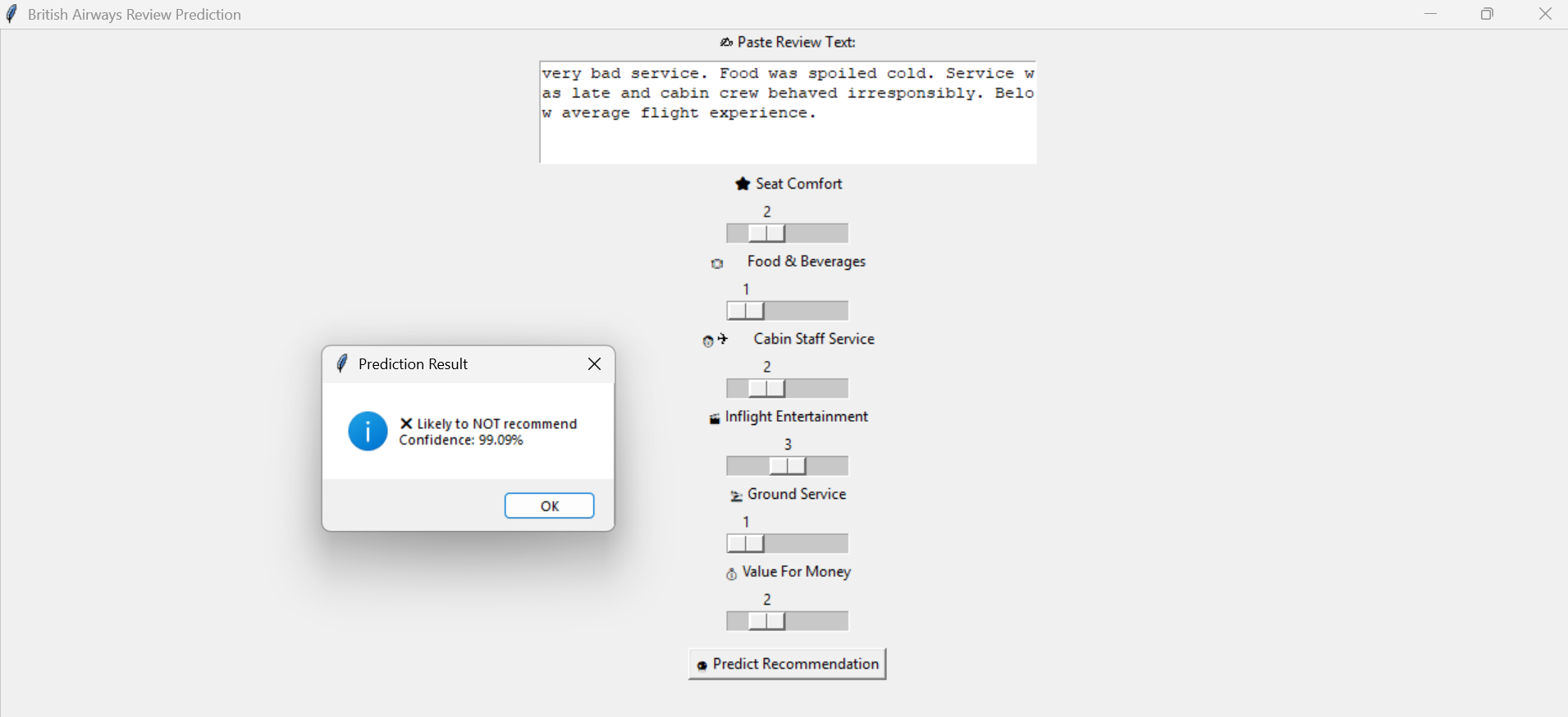
To assess the practical deployment of the model, a graphical user interface (GUI) was developed using **Tkinter**, the standard Python GUI toolkit. The interface allows users to input a textual review and corresponding service ratings (e.g., seat comfort, food & beverages, value for money), then predicts whether the customer is likely to recommend British Airways, displaying both the classification and its associated confidence level.

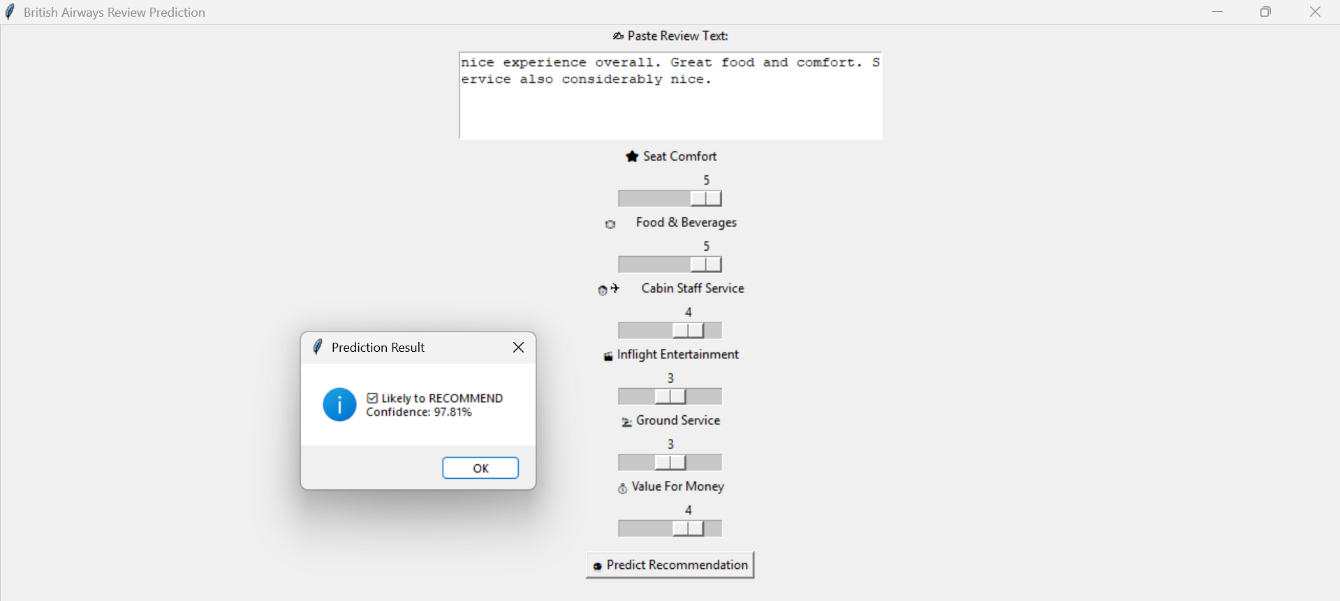
The GUI was tested under two primary conditions:

**1. Evaluation Using Test Split of Dataset:**  
Entries from the model's reserved test set (unseen during training) were passed through the GUI. The system accurately predicted recommendation outcomes, confirming that the interface correctly integrates the trained model and performs well on real test data.

**2. Evaluation Using Anonymous Entries:**  
User-defined, manually entered reviews—unseen and unstructured—were also tested through the GUI. Examples include a positive review stating *“Nice experience overall. Great food and comfort. Service also considerably nice.”* and a negative review like *“Very bad service. Food was spoiled cold...”*. The system returned appropriate predictions with high confidence

These results affirm that the GUI correctly handles both structured test data and arbitrary external input, demonstrating the model’s robustness and generalization capability. The system is intuitive, responsive, and suitable for real-world usage by both analysts and end-users without technical expertise.





**Conclusion**

**5.1 Summary of Findings**

This project successfully developed a machine learning system capable of predicting whether a British Airways customer would recommend the airline based on their review text and service ratings. The key findings include:

1. **High Prediction Accuracy**: The final Gradient Boosting model achieved over 93% accuracy in predicting customer recommendations, demonstrating the feasibility and effectiveness of automated review analysis.
2. **Key Satisfaction Drivers**: Value for Money, Cabin Staff Service, and Seat Comfort emerged as the most influential numerical factors in determining customer recommendations, providing clear priorities for service improvement efforts.
3. **Textual Insights**: Analysis of review text revealed specific experience aspects that strongly influence customer satisfaction, with terms related to service quality, delays, and staff behavior being particularly impactful.
4. **Balanced Performance**: The model demonstrates robust performance across different customer segments and review types, with balanced precision and recall metrics ensuring reliable predictions for both positive and negative recommendations.
5. **Operational Relevance**: The combination of rating analysis and text mining provides actionable insights beyond simple prediction, allowing British Airways to identify specific service issues mentioned in reviews.

**5.2 Business Impact**

The developed system offers several significant benefits for British Airways:

1. **Scalable Feedback Analysis**: Ability to automatically process thousands of reviews, extracting insights without manual intervention.
2. **Real-time Monitoring**: Through API deployment, the airline can track recommendation trends as new reviews are submitted.
3. **Targeted Improvement Priorities**: Clear identification of service aspects that most strongly influence customer satisfaction, enabling focused enhancement efforts.
4. **Customer Segmentation Insights**: Understanding how different traveler types (business vs. leisure) and service classes experience flights differently.
5. **Return on Investment Guidance**: Identifying which service improvements would yield the highest increase in customer recommendations for the investment required.

**5.3 Methodological Contributions**

This project makes several methodological contributions to the field of customer satisfaction analysis:

1. **Multimodal Analysis Framework**: Successfully integrating structured numerical ratings with unstructured text data to create a comprehensive analysis system.
2. **Effective NLP Application**: Demonstrating how text preprocessing and TF-IDF vectorization can extract meaningful features from customer reviews in the aviation context.
3. **Balanced Evaluation Approach**: Using multiple performance metrics to ensure the model performs well across different evaluation criteria.
4. **Model Selection Justification**: Providing a detailed comparison of different algorithms and justification for the final model selection based on both statistical performance and business interpretability.

**5.4 Limitations**

Despite its success, the project has several limitations that should be acknowledged:

1. **Training Data Timeframe**: The model is trained on historical data and may not capture recent changes in service or customer expectations.
2. **Language Limitations**: The current implementation focuses on English-language reviews, potentially missing insights from international customers.
3. **Platform Bias**: Reviews collected from online platforms may over-represent customers with strong opinions (very satisfied or very dissatisfied), potentially under-representing the middle ground.
4. **Feature Granularity**: While high-level service aspects (e.g., Food & Beverages) are included, more specific elements (e.g., menu options, portion sizes) are not individually captured in the model.

**5.5 Future Work**

Several avenues for future development would enhance the system's capabilities:

1. **Multilingual Support**: Extending the NLP components to handle reviews in multiple languages, broadening the analysis scope.
2. **Temporal Analysis**: Incorporating time-series analysis to track changes in customer sentiment and recommendation patterns over time.
3. **Comparative Analysis**: Expanding the model to analyze competitor airlines, enabling benchmarking of British Airways against industry peers.
4. **Topic Modeling**: Implementing unsupervised learning techniques like LDA (Latent Dirichlet Allocation) to automatically identify emerging themes in customer feedback.
5. **Sentiment Analysis Enhancement**: Developing more nuanced sentiment scoring that can capture mixed opinions within a single review.
6. **Interactive Visualization**: Creating dashboards that allow business users to explore the relationship between service aspects and customer recommendations.

**5.6 Final Thoughts**

The successful development and implementation of this machine learning system demonstrate how advanced analytics can transform unstructured customer feedback into actionable business intelligence. By automating the analysis of thousands of reviews, British Airways can now gain comprehensive insights into customer satisfaction drivers, enabling data-driven service improvements that enhance the passenger experience.

As the volume of online reviews continues to grow, such automated systems will become increasingly valuable for service-oriented businesses seeking to understand and improve customer satisfaction. This project provides a blueprint for implementing similar systems across other customer-facing industries where feedback analysis at scale presents both a challenge and an opportunity.

**References**

1. British Airways reviews dataset: <https://www.kaggle.com/datasets/rameshvhanamane/british-airways-reviews-dataset>
2. scikit-learn documentation: <https://scikit-learn.org>
3. Jupyter Notebook: Final.ipynbs