IPSA
ÉCOLE D'INGÉNIEURS
DE L'AIR, DE L'ESPACE ET DE LA MOBILITÉ DURABLE

AERO 4 - MATHEMATICAL TOOLS FOR DATA SCIENCE

ANASTASIA ATTOS AERO4 SET

Introduction

The global pandemic has significantly impacted the airline industry, leading to a sharp decline in passenger satisfaction and overall business performance. As the industry strives to recover, it becomes imperative to understand the key factors influencing passenger satisfaction. This project aims to predict passenger satisfaction levels using a machine learning approach, leveraging a dataset split into training and testing sets. The goal is to classify passengers into categories of satisfaction, neutral, or dissatisfied, and identify the factors that are most indicative of these classifications.

Data Description

The dataset provided includes survey responses from air passengers. It consists of 103,904 entries with 25 columns, which include both categorical and numerical features. Key categorical features include Gender, Customer Type, Type of Travel, Class, and the target variable satisfaction. Numerical features include Age, Flight Distance, and various service ratings. It was observed that the Arrival Delay in Minutes column contained missing values, which needed to be addressed.

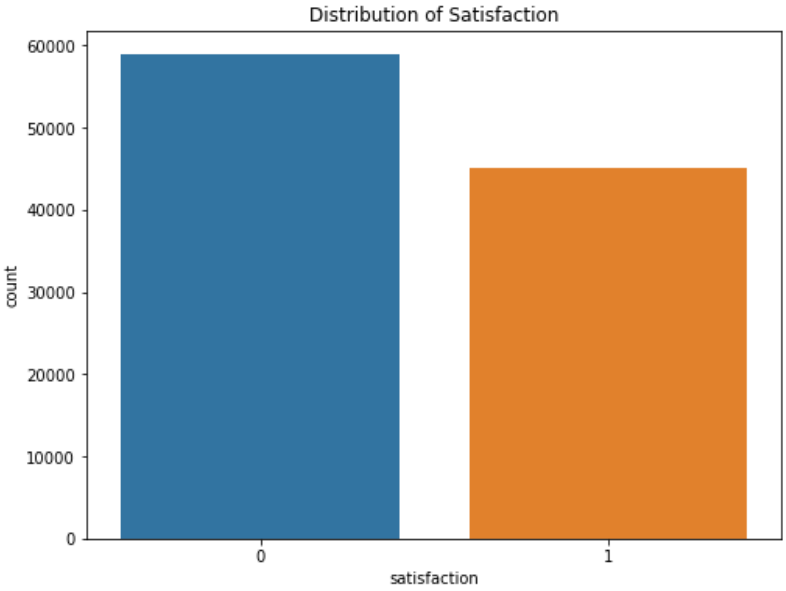
Data Preprocessing

To prepare the data for modeling, several preprocessing steps were performed:

- Handling Missing Values: The column `Arrival Delay in Minutes` had missing values. These were imputed using the median value of the column to ensure no data was left out during analysis.
- Encoding Categorical Variables: Categorical variables were encoded using `LabelEncoder` to convert them into numerical values suitable for machine learning algorithms. This included encoding columns like Gender, Customer Type, Type of Travel, Class, and the target variable satisfaction.
- Normalizing Numerical Features: Numerical features were normalized using `StandardScaler` to ensure they have a mean of zero and a standard deviation of one. This normalization is crucial for algorithms that are sensitive to the scale of data.

Exploratory Data Analysis

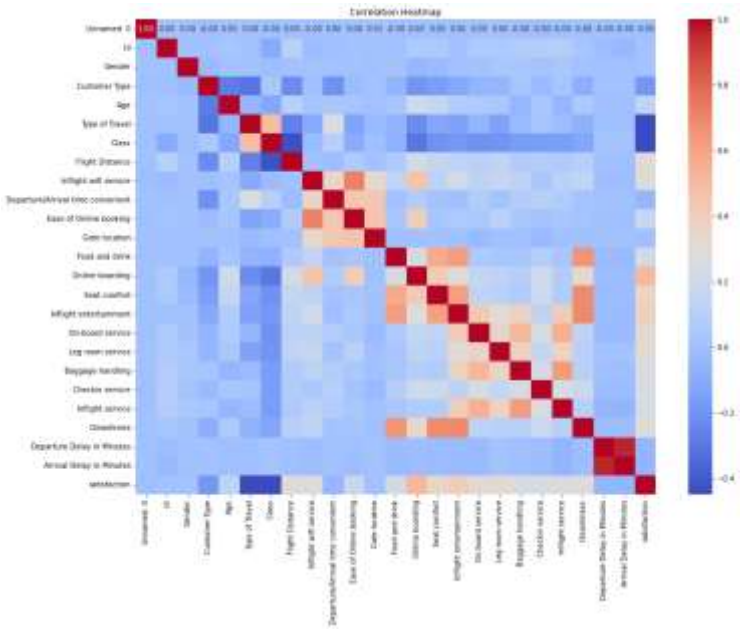
Distribution of Satisfaction:



The graph illustrates the distribution of the target variable "satisfaction" in the form of a bar graph. It shows a significant class imbalance: Class 0 (not satisfied) has about 60,000 samples, while Class 1 (satisfied) has about 45,000. This imbalance can affect the performance of some classification algorithms, as models may favour the majority class. To mitigate this bias, it is crucial to use measures such as class weighting (as implemented in the code with ‘class_weight='balanced’’) or other techniques such as oversampling or subsampling.

Correlation Heatmap:

A correlation matrix was computed and visualized using a heatmap to understand the relationships between various features. This helped identify which features were most strongly correlated with passenger satisfaction.



The correlation matrix uses red and blue squares to indicate positive and negative correlations, respectively, with color intensity proportional to the strength of the correlation. For example, a strong positive correlation between in-flight service ("Inflight service") and customer satisfaction ("satisfaction") suggests that good in-flight services are associated with greater customer satisfaction. The variables most correlated to satisfaction, located at the bottom right of the matrix, are probably the most relevant to predict satisfaction. However, it is important to note the high correlations between independent variables, as they may indicate a multicollinearity problem, impacting the performance and interpretability of the model. The heat map is annotated for easy understanding of correlation values, and the use of the coolwarm palette helps to clearly distinguish positive and negative correlations.

Model Development

For the classification task, a Random Forest classifier was selected due to its robustness and its ability to handle both categorical and numerical features effectively. The initial step involved splitting the training data into training and validation sets to evaluate the model's performance. This was achieved using the `train_test_split` function, with stratified sampling ensuring that the class distribution was maintained in both sets.

Next, a Random Forest classifier was instantiated with `class_weight='balanced'` to manage the issue of class imbalance. The model was then trained on the training data, fitting it to the features (X_train) and the target variable (y_train). This step ensured that the model learned from the provided data, balancing the influence of each class to improve its ability to predict both satisfied and unsatisfied customers accurately.

Finally, the trained model was applied to the validation set to predict satisfaction levels. The model's performance was evaluated using several metrics, including a confusion matrix, classification report,

and accuracy score. These metrics offered comprehensive insights into the model's precision, recall, F1-score, and overall accuracy, providing a clear picture of its effectiveness in predicting customer satisfaction.

Results

The model's performance on the validation set can be interpreted based on several key metrics:

```
 Confusion Matrix:
 [[11491   285]
 [  505  8500]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97     11776
           1       0.97      0.94      0.96      9005

    accuracy                           0.96     20781
   macro avg       0.96      0.96      0.96     20781
weighted avg       0.96      0.96      0.96     20781


Accuracy Score: 0.9619845050767528
```

The Confusion Matrix provides a detailed breakdown of the model's predictions. In this case, the matrix indicates that the model correctly predicted 11,491 instances of the "0" class (non-satisfied) and 8,500 instances of the "1" class (satisfied). It made 285 incorrect predictions where the true class was "0" but predicted as "1" and 505 incorrect predictions where the true class was "1" but predicted as "0". This matrix highlights where the model is making errors, giving insights into its strengths and weaknesses in distinguishing between satisfied and non-satisfied customers.

The Classification Report offers a comprehensive overview of the model's performance across different metrics for each class. The model achieved a precision of 0.96 and a recall of 0.98 for the "0" class, resulting in an F1-score of 0.97. For the "1" class, the model achieved a precision of 0.97 and a recall of 0.94, leading to an F1-score of 0.96. These metrics illustrate that the model performs exceptionally well in identifying both satisfied and non-satisfied customers, with high precision and recall values indicating that the model makes very few false positive and false negative errors.

The Accuracy Score of 0.96 suggests that the model correctly classified approximately 96.2% of the instances in the validation set. This high accuracy indicates the model's overall effectiveness in predicting customer satisfaction. The macro average and weighted average precision, recall, and F1-scores are all consistent at 0.96, further demonstrating the model's balanced performance across both classes.

Conclusion

The results of this project highlight the significant impact of various factors on passenger satisfaction in the airline industry, particularly in the context of recovery from the global pandemic. The Random Forest classifier proved to be a robust tool for predicting passenger satisfaction levels, effectively handling both categorical and numerical features.

By preprocessing the data meticulously, including handling missing values, encoding categorical variables, and normalizing numerical features, we ensured that the dataset was well-prepared for modeling. Exploratory data analysis, including the distribution of satisfaction and a correlation heatmap, provided valuable insights into the relationships between features and their impact on passenger satisfaction.

The Random Forest model, trained with balanced class weights to address the class imbalance, demonstrated high accuracy and balanced performance across both satisfaction categories. The confusion matrix and classification report revealed the model's strengths in correctly predicting satisfaction levels, with high precision and recall scores indicating a low rate of false positives and false negatives.

Overall, the model's accuracy score of 96.2% underscores its effectiveness in classifying passenger satisfaction. These results suggest that key factors, such as in-flight service quality, are crucial for improving passenger satisfaction. As the airline industry continues to recover, leveraging such predictive models can help airlines enhance their services and better meet passenger expectations, ultimately leading to improved business performance.