Airline Passenger Satisfaction Analysis & Predictive Modeling

1. Introduction
This report details the process of analyzing an airline passenger satisfaction dataset, performing exploratory data analysis (EDA), preprocessing the data, building a predictive model, and evaluating its performance. The primary goal was to develop a reliable model to predict passenger satisfaction, which can inform targeted improvements in airline services.

2. Data Loading and Initial Cleaning

Dataset Source: The dataset was downloaded from KaggleHub (`teejmahal20/airline-passenger-satisfaction`).

Initial Steps:
The `train.csv` file was loaded into a Pandas DataFrame.
Irrelevant identifier columns (`Unnamed: 0`, `id`) were dropped to ensure that the model focuses on actionable features rather than unique identifiers.
Missing values in `Arrival Delay in Minutes` were imputed using the median, as delay data can be skewed.

3. Exploratory Data Analysis (EDA)

Key visualizations and insights:

A. Overall Satisfaction Distribution
The initial pie chart showed that approximately 56.7% of passengers were 'Neutral or dissatisfied', while 43.3% were 'Satisfied'. This indicates a significant portion of the customer base requires attention.

B. Impact of Service Ratings (0-5) on Satisfaction

- Seat Comfort: A strong positive correlation was observed. Passengers with higher seat comfort ratings (4-5) were predominantly satisfied, while those with lower ratings (0-2) were largely dissatisfied or neutral.

- Inflight Entertainment: Similar to seat comfort, higher ratings for inflight entertainment (4-5) correlated with higher satisfaction, suggesting its importance as a key service driver.

- Ease of Online Booking : This also showed a clear trend: higher ratings for online booking ease correlated with higher passenger satisfaction.

C. Influence of Travel Demographics and Type on Satisfaction

- Type of Travel : Business travelers were overwhelmingly more satisfied than personal travelers. This highlights business travel as a high-satisfaction segment that could significantly impact overall satisfaction metrics.

- Travel Class : Business Class passengers showed a substantially higher satisfaction rate compared to Eco Plus and Eco Class passengers, consistent with the higher service levels typically offered.

- Gender : Satisfaction distribution by gender showed relatively similar patterns for both males and females, indicating gender might not be a primary differentiator for overall satisfaction.

- Customer Type : Loyal customers exhibited higher satisfaction levels compared to disloyal customers, which is expected and emphasizes the importance of customer loyalty programs.

D. Correlation with Numerical Factors and Delays

Correlation Matrix : The heatmap revealed strong positive correlations between satisfaction and most service-related features (e.g., Online boarding, Inflight entertainment, Seat comfort, Inflight wifi service). Conversely, `Departure Delay in Minutes` and `Arrival Delay in Minutes` showed negative correlations, indicating that delays negatively impact satisfaction.

Departure Delay Impact : A histogram of departure delays (under 60 minutes) showed that 'neutral or dissatisfied' passengers experienced a higher average delay (16.50 minutes) compared to 'satisfied' passengers (12.61 minutes). This quantitatively confirms that delays are a significant detractor from passenger satisfaction.

4. Data Preprocessing Pipeline

To prepare the data for machine learning, the following steps were executed:

1. Feature and Target Separation: The `satisfaction` column was designated as the target variable (y), and all other columns were considered features (X).

2. Target Encoding: The categorical `satisfaction` target was converted into a binary numerical format (`satisfied`: 1, `neutral or dissatisfied`: 0).

3. **Train-Test Split**: The dataset was split into 80% training and 20% testing sets to evaluate model generalization (`random_state=42` for reproducibility).

4. **One-Hot Encoding**: Nominal categorical features (`Gender`, `Customer Type`, `Type of Travel`, `Class`) were converted into numerical format using one-hot encoding to avoid imposing any false ordinal relationships.

5. **Numerical Feature Scaling**: Continuous features were standardized using `StandardScaler` to ensure they contribute equally to the model, preventing features with larger scales from dominating the learning process.

5. Predictive Modeling and Evaluation

A. Model Selection

The Random Forest Classifier was chosen due to its robustness, ability to handle various data types, and strong predictive performance in classification tasks.

B. Initial Model Performance (with Data Leakage)

Initially, the model achieved an accuracy of 1.00 and perfect precision, recall, and F1-scores. This artificially high performance was a strong indicator of data leakage.

Data Leakage Identification: Upon inspection, it was discovered that the `satisfaction_numerical` column (a direct numerical representation of the target variable) was inadvertently included in the feature set (`X_train`). This provided the model with the answer during training, leading to unrealistic results, as clearly shown by its overwhelming feature importance .

C. Correcting Data Leakage

Fix Implemented: The `satisfaction_numerical` column was explicitly removed from the feature set (`X`) before the train-test split and subsequent preprocessing steps.

Impact: This crucial step ensured that the model genuinely learned patterns from independent features.

D. Re-evaluated Model Performance

After retraining the Random Forest Classifier with the corrected data, the model's performance on the test set is as follows:

Accuracy on Test Set: 0.9589 (approximately 95.89%)

Classification Report:
Class 0 (neutral or dissatisfied): Precision: 0.95, Recall: 0.97, F1-score: 0.96
Class 1 (satisfied): Precision: 0.97, Recall: 0.94, F1-score: 0.95
Macro Avg: Precision: 0.96, Recall: 0.96, F1-score: 0.96
Weighted Avg: Precision: 0.96, Recall: 0.96, F1-score: 0.96

Comparison: The accuracy dropped from an artificial 1.00 to a realistic and robust 0.9589. The F1-scores for both classes are strong (0.95-0.96), indicating the model effectively predicts both satisfied and dissatisfied passengers without significant bias towards one class.

6. Conclusion and Next Steps

The revised Random Forest Classifier is a highly effective and trustworthy model for predicting airline passenger satisfaction, achieving an accuracy of **95.89%** and strong F1-scores across both satisfaction categories. The successful identification and correction of data leakage ensures that the model's predictions are based on genuine underlying patterns within the data, making it suitable for practical application.

7. Next Steps & Recommendations:

- Feature Importance Analysis: Re-running the feature importance analysis on the corrected model will provide reliable insights into the most influential factors driving passenger satisfaction, enabling data-driven strategic decisions for service improvement.

- Hyperparameter Tuning: Explore advanced hyperparameter tuning techniques (e.g., GridSearchCV, RandomizedSearchCV) for the Random Forest model to potentially enhance performance further.

- Model Interpretability: Investigate tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to gain deeper insights into individual feature contributions for specific predictions, aiding in understanding why a passenger might be predicted as satisfied or dissatisfied.

- Deployment Consideration: Given the high performance, consider the steps for deploying this model to a production environment to provide real-time satisfaction predictions or to integrate it with operational dashboards.