Data Preprocessing Steps and Challenges

Data Collection:

The dataset contains 129,879 rows and 23 columns, including features such as customer satisfaction, gender, customer type, age, type of travel, class, flight distance, and various service ratings.

Data Preprocessing:

1. Loading Data:

The data was loaded into a pandas DataFrame from a CSV file.

2. Handling Missing Values:

The “Arrival Delay in Minutes” column had 393 missing values, approximately 0.3% of the total data. These missing values were imputed with the mean of the column. This approach was chosen because the proportion of missing values was tiny, and imputing with the mean is a simple yet effective method to handle missing data without introducing significant bias.

3. Feature Engineering:

New features were created to enhance the predictive power of the model. The rationale behind creating these features was to capture more nuanced information that could influence customer satisfaction:

- Total Delay: Sum of departure and arrival delays. This feature helps understand the overall delay experienced by the customer.

- Delay: Binary feature indicating if there was any delay. This simplifies the delay information into a binary format, making it easier for the model to learn.

- On-Flight Rating: Average rating of in-flight services. This aggregates multiple in-flight service ratings into a single feature.

- Booking Rating: Average rating of booking-related services. This aggregates multiple booking-related service ratings into a single feature.

- Ground Rating: Average rating of ground services. This aggregates multiple ground service ratings into a single feature.

- Rating: Overall average rating. This provides a holistic view of the customer's experience.

- Delay Proportion: Ratio of total delay to flight distance. This normalises the delay by the flight distance, providing a relative measure of delay.

- Age Group: Categorized age into groups. This helps in capturing age-related patterns in customer satisfaction.

- Age_Satisfaction: Product of age and overall rating. This captures the interaction between age and overall satisfaction.

- Comfort to Distance: Ratio of seat comfort to flight distance. This normalises seat comfort by flight distance, providing a relative measure of comfort.

- Arrival-Departure-Delay: Difference between arrival and departure delays. This captures the discrepancy between arrival and departure delays.

4. Encoding Categorical Variables:

OneHotEncoder was used to encode categorical variables, except for the "satisfaction" column, which was encoded by mapping the satisfactions to 0 or 1. One-hot encoding was chosen to convert categorical variables into a format that can be provided to ML algorithms to do a better job in prediction without introducing any ordinality, which could make the model biased.

5. Normalization:

Numerical features were normalised to a range between 0 and 1, as the data did not follow a Gaussian distribution. Normalisation ensures that all features contribute equally to the model and improves the convergence of gradient-based optimisation algorithms.

Challenges:

The main challenge was figuring out the feature engineering to create meaningful new features that could enhance the model's predictive power.

Insights from Exploratory Data Analysis

1. Target Variable Distribution:

The dataset is slightly imbalanced, with a ratio of 0.55 satisfied customers to 0.45 dissatisfied customers. This imbalance was considered during model evaluation to ensure that the model's performance was not biased towards the majority class.

2. Correlation Analysis:

A correlation matrix and heatmap were used to identify features strongly correlated with customer satisfaction. The following features had the highest correlation with customer satisfaction:

- Inflight Entertainment: 0.523496

- Rating: 0.519151

- On-Flight Rating: 0.475953
- Booking Rating: 0.440361
- Ease of Online Booking: 0.431772
- Online Support: 0.390143
- On-Board Service: 0.352047
- Age_Satisfaction: 0.345362
- Online Boarding: 0.338147

These insights guided the feature engineering process and helped identify the most important features to focus on.

Model Evaluation and Results

Models Used:

1. Logistic Regression:

A logistic regression model was trained to predict customer satisfaction. Hyperparameter tuning was performed to find the optimal learning rate. Logistic regression was chosen for its simplicity and effectiveness in binary classification tasks.

2. Linear Discriminant Analysis (LDA):

An LDA model was used for dimensionality reduction and classification. It was influential in separating the classes linearly. LDA was chosen for its ability to reduce dimensionality while preserving class-discriminatory information.

3. Gaussian Discriminant Analysis (GDA):

A GDA model was used to model the distribution of each class using a Gaussian distribution. It provided a probabilistic framework for classification. GDA was chosen for its ability to model the underlying distribution of the data.

4. Custom Ensemble Model:

An ensemble model combining GDA, LDA, and Logistic Regression was created to leverage the strengths of each model. Majority voting was used to aggregate the predictions. The ensemble model was chosen to improve overall performance by combining the complementary strengths of the individual models.

Model Performance:

The ensemble model achieved the highest accuracy, leveraging the complementary strengths of GDA, LDA, and Logistic Regression. The ensemble model's accuracy was higher than two of the individual models, demonstrating the effectiveness of combining multiple models.

Recommendations for Invistico Airlines

1. Enhance In-Flight Entertainment:

Focus on improving in-flight entertainment options, as this feature had the highest correlation with customer satisfaction (0.523496). Consider offering various entertainment options, including movies, TV shows, music, and games.

2. Improve Overall Service Ratings:

Continuously monitor and improve the overall service ratings, including on-flight, booking, and ground services. The overall rating had a strong correlation with customer satisfaction (0.519151). Implement regular training programs for staff to ensure high-quality service.

3. Optimize Online Booking and Support:

Enhance the ease of online booking and support services, as these features showed significant correlations with customer satisfaction (0.431772 and 0.390143, respectively). Simplify the online booking process and provide prompt and effective online support.

4. Focus on On-Board Service:

Improve on-board services such as seat comfort, food and drink, and in-flight entertainment. On-board service had a notable correlation with customer satisfaction (0.352047). Regularly gather feedback from passengers to identify areas for improvement.

5. Personalize Services Based on Age:

Age has a significant impact on satisfaction. Invistico should offer different packages to different demographics to maximise their satisfaction. For example, ages 0-20 and 21-40 have lower correlations with satisfaction than older age groups, suggesting that older people want different kinds/levels of service.

6. Reward loyalty with better service:

A customer's loyalty has a relatively high correlation with satisfaction, suggesting that loyal customers have higher service expectations. As such, loyal customers should receive better service.

7. Customize packages for business class flyers:

Flying business class is highly correlated with service, suggesting that customers who pay more expect better service, which should happen. Special packages should be curated for business-class flyers to enhance their satisfaction.

8. Personalize Services Based on Gender:

Women have a higher correlation with satisfaction than men, so Invistico must investigate which services prefer more and then personalise those to serve the female demographic of their customers.