Report

Madiba Hudson-Quansah

Data Processing Steps and Challenges

The steps I took in preprocessing the dataset were mainly focused on these areas: encoding categorical values, standardizing numerical values, and handling missing values.

From my initial overview of the dataset, I noticed that the data was primarily categorical with some usual numerical data like age, flight distance, departure delay, etc. Looking closer at the nature of the categorical data, I noticed that many features had few unique values. I decided to encode the data using label encoding as the numbers produced by the label encoding would be manageable. I anyway opted to scale the label-encoded categories using z-score standardization to ensure that the data was centred around zero and had a standard deviation of 1 so that the more significant label numbers would not introduce any bias in the model.

Focusing on the numerical data, I noticed some values that needed to be added, specifically in the arrival delay feature. To rectify this, I used the median of the arrival delay feature to fill in the missing values. Then, I scaled the numerical features again using z-score standardization to ensure that the data was centred around zero and had a standard deviation of 1.

I then separated the target variable, satisfaction, from the rest of the data and split the data randomly into training and testing sets using a 70/30 split.

Insights from Exploratory Data Analysis

My exploratory data analysis was focused on understanding the distribution of the data and the relationship between the features and the target variable. For the numerical data, I first viewed the distribution using histograms, which showed, as expected, that the age of passengers was normally distributed. At the same time, departure and arrival delays were heavily right-skewed, indicating the presence of high outliers. Viewing their relationship with the target variable, I found that the satisfaction of passengers was not significantly affected by most of the numerical features, except for departure delay, which showed a moderate negative correlation with satisfaction.

My expiration of the categorical features using stacked bar charts, however, showed significant correlations with the target variable, with the apparent correlation with customer loyalty and satisfaction, a significant correlation with, flight class, seat comfort, in-flight entertainment, online support and online boarding.

Model Evaluation and Results

I selected a Logistic Regression model as my final model as it is easy to interpret and could quickly provide actionable insights to Invistico Airlines. Also, as it is a simple model, it is less likely to overfit the data. Looking at the dataset size, I opted to use Stochastic Gradient Descent as the optimization algorithm for the Logistic Regression model as it is faster and

more efficient for large datasets. I ran the model on the training dataset at a learning rate of 0.01 and $20,000$ iterations to ensure that the model converged.

This resulted in a model with an accuracy of 0.83 on the training dataset and an accuracy of 0.82, indicating minimal overfitting. When computing the precision and recall of the model, I found that the model had a precision of 0.85, a recall of 0.84, and an F1 score of 0.84. This means that the model can correctly predict 85% of the time when a passenger is satisfied and 84% of the time when a passenger is dissatisfied. The F1 score of 0.84 indicates that the model can correctly predict 84% of the time when a passenger is satisfied and 84% of the time when a passenger is dissatisfied.

Looking at the model's feature importance, I found that the model placed a significant positive importance on in-flight entertainment and seat comfort while placing significant negative importance on Customer Type, i.e. Loyal and Disloyal customers and Gender, with the model placing men as more likely to be dissatisfied than women.

Recommendations

From the insights gained from the model, I would recommend that Invistico Airlines focus on improving the in-flight entertainment and seat comfort for their passengers, as these are the features that have the most significant impact on their satisfaction. I recommend that Invistico Airlines focus on improving the satisfaction of their male passengers as the model indicates that they are more likely to be dissatisfied than women.