Airline Passenger Satisfaction Prediction

Sofiane EL FARTASS

Spring 2024

Abstract

This report documents the process and results of a machine learning project aimed at predicting airline passenger satisfaction. Utilizing a dataset from a Kaggle competition, this study employs deep learning techniques to model the complexities of passenger satisfaction based on various flight-related characteristics.

Introduction

Airline passenger satisfaction is influenced by numerous factors, ranging from the booking process to the in-flight experience. Understanding these factors can help airlines improve service quality and enhance customer loyalty. This project applies machine learning techniques to predict passenger satisfaction, providing insights into key drivers of satisfaction.

Dataset Description

The dataset comprises multiple features related to different aspects of air travel. The data was preprocessed to ensure compatibility with machine learning algorithms. Below is a detailed description of each variable:

Variable	**Description**
Gender	Gender of the passengers (Female, Male)
Customer Type	The customer type (Loyal customer, disloyal customer)
Age	The actual age of the passengers
Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance	The flight distance of this journey
Inflight wifi service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient
Ease of Online booking	Satisfaction level of online booking
Gate location	Satisfaction level of Gate location
Food and drink	Satisfaction level of Food and drink
Online boarding	Satisfaction level of online boarding
Seat comfort	Satisfaction level of Seat comfort
Inflight entertainment	Satisfaction level of inflight entertainment
On-board service	Satisfaction level of On-board service
Leg room service	Satisfaction level of Leg room service
Baggage handling	Satisfaction level of baggage handling
Check-in service	Satisfaction level of Check-in service
Inflight service	Satisfaction level of inflight service
Cleanliness	Satisfaction level of Cleanliness
Departure Delay in Minutes	Minutes delayed when departure
Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	Airline satisfaction level (Satisfied, neutral or dissatisfied)

Table 1: Description of Variables in the Airline Passenger Satisfaction Dataset.

Data Preprocessing

Data preprocessing is a critical step in the machine learning pipeline. It ensures that the model trains on data that is clean and well-formatted, thereby improving the model's accuracy and efficiency. In this project, several preprocessing steps were implemented on the dataset as follows:

Removal of Unnecessary Columns

The dataset contained some columns that were not relevant to the analysis and could potentially skew the results. Specifically, columns such as `Unnamed: 0` and `id` were removed. These columns typically represent indexing information that does not contribute to the actual machine learning model.

Encoding Categorical Variables

Many machine learning models, especially those based on mathematical calculations, require input to be numerical. Therefore, categorical variables were encoded as follows:

- **Gender**: Converted into binary values with 'Female' as 1 and 'Male' as 0.
- **Customer Type**: Encoded as 'Loyal Customer' to 1 and 'disloyal Customer' to 0, distinguishing between regular and occasional customers.
- **Type of Travel**: Mapped 'Business travel' to 1 and 'Personal Travel' to 0, indicating the purpose of the travel.
- **Class**: Transformed into ordinal values where 'Business' is 2, 'Eco Plus' is 1, and 'Eco' is 0, reflecting the service class on the flight.
- **Satisfaction**: Changed from categorical ('satisfied', 'neutral or dissatisfied') to binary (1 for 'satisfied', 0 for 'neutral or dissatisfied'), which simplifies the output for binary classification.

Handling Missing Values

Missing data can significantly impact the performance of a machine learning model. In the dataset, the `Arrival Delay in Minutes` column contained missing values which were replaced by the median of the column.

Model Architecture

The predictive model is a deep neural network, details of which are as follows:

- **Input Layer**: Accepts input features with a dimension equal to the number of predictors in the dataset, specified dynamically by `X_train.shape[1]`.
- **Dense Layers**: Three fully connected dense layers with varying numbers of neurons. These layers use the ReLU activation function to introduce non-linearity, helping the model learn complex patterns in the data.
  - First dense layer with 128 neurons.
  - Second dense layer with 64 neurons.
  - Third dense layer with 32 neurons.
- **Dropout Layers**: Positioned after each dense layer with a dropout rate of 0.5, these layers are crucial for preventing overfitting by randomly setting a fraction of input units to 0 at each update during training time.
- **Output Layer**: A single neuron with a sigmoid activation function that outputs the probability of a passenger being satisfied. This setup is typical for binary classification tasks.

Hyperparameter

Hyperparameter	**Value**
Optimizer	Adam
Learning Rate	0.001
Loss Function	Binary Crossentropy
Batch Size	64
Number of Epochs	25
Dropout Rate	0.5

Table 2: Hyperparameters used in the model training process.

Model Evaluation

The model was evaluated using accuracy, precision, recall, and F1-score, providing a holistic view of its performance.

Evaluation Metrics Formulas

The performance of the predictive model is quantitatively assessed using several statistical metrics, each providing unique insights into the model's accuracy and efficacy. The definitions and formulas for these metrics are as follows:

Accuracy Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

Precision Precision, or positive predictive value, measures the accuracy of positive predictions. Formulated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall Recall, or sensitivity, measures the ability of the model to identify all relevant instances (all actual positives). It is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is especially useful when the class distribution is uneven. The formula for F1-score is:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics together offer a comprehensive view of the model's performance across different dimensions of accuracy and reliability.

Model results

Metric	Class 0	Class 1	Macro Avg	Weighted Avg
Precision	0.95	0.97	0.96	0.96
Recall	0.98	0.94	0.96	0.96
F1-Score	0.96	0.95	0.96	0.96

Table 3: Performance Metrics of the Predictive Model

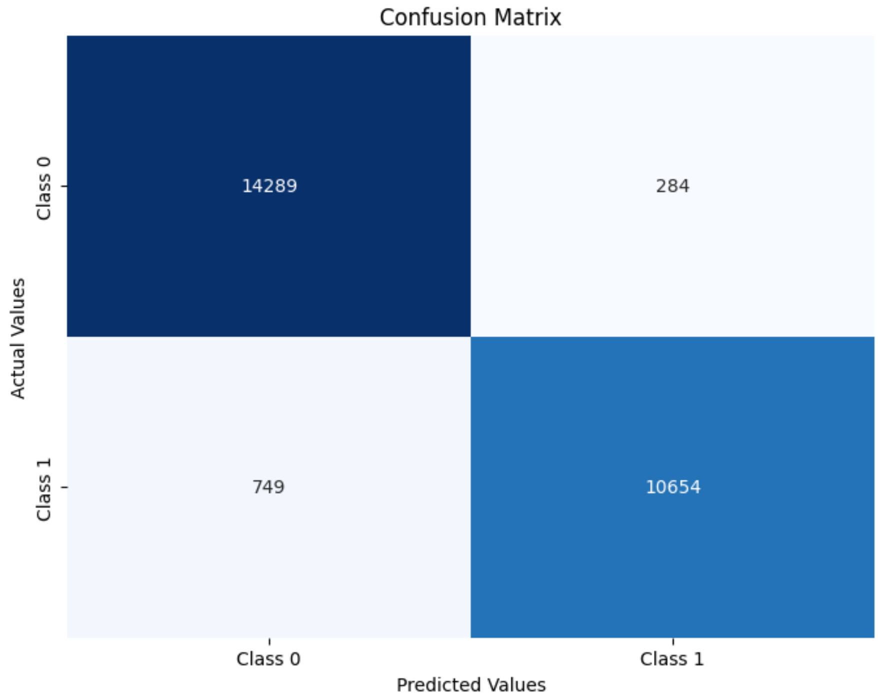

Figure 1: Confusion matrix of the model predictions.

Conclusion

The model demonstrated robust performance with an overall accuracy of 96%, effectively distinguishing between satisfied and dissatisfied passengers. This predictive capability can assist airlines in identifying areas of improvement and enhancing passenger experiences.