Hotel Reservations – Classification using Python

Introduction

Classification is a supervised learning technique in data mining used to identify a category or class to which new observations belong based on their characteristics. In classification, a model is trained on a labeled dataset (where the classes are known) and then used to predict the class of unseen observations.

In a Random Forest Classifier, multiple decision trees are created on random subsets of the training data, and each tree votes for the class of the input data. The final prediction is then made based on the majority vote of all the trees.

k-Nearest Neighbor (kNN) model is a non-parametric and lazy learning algorithm that relies on a similarity measure between instances to make predictions. 'In kNN classification, the idea is to find the k-nearest neighbors of a new data point based on a distance metric such as Euclidean distance, and then classify the new point based on the majority class of its k-nearest neighbors. The value of k is a hyperparameter that can be tuned to optimize the performance of the model.

Scenario-

The booking options and consumer behavior have been significantly altered by internet hotel reservation channels. Due to changes in plans or no-shows, many hotel reservations are canceled.

Changes in plans, scheduling issues, and other common causes of cancellations are listed below. This is frequently made simpler for hotel visitors by the ability to do so without charge or preferable at a low cost, but it is a less desirable and potentially revenue-decreasing element for hotels to cope with.

The objective of this project is to predict whether the consumer will keep their reservation or cancel it?

Dataset-

- The dataset was obtained from Kaggle and can be found in the Hotel_Reservations.csv file, and the data dictionary is provided in HotelReservations_dictionary.csv file.
- It includes 18 predictors and 1 response variable.
- Tools used – Python (Jupyter Notebook)

Approach to Analysis-

1. **Explore the Data Dictionary**

   This step includes exploring the given data dictionary to understand the definition of different variables in the dataset and differentiate between predictors and response variables.
2. **Data exploration and Preprocessing**

   Exploratory data analysis is done to understand the variables, the relationship between them and to clean the dataset. This comprises of following steps:
   a. Checking number of rows, columns, data types, unique values for each variable, finding outliers etc.

b. Determining the relationship between variables
c. Predictor selection
d. Dealing with null values
e. Dealing with categorical predictors
f. Scaling predictors

3. **Model selection and building**

   Since the response variable is binary (Booking canceled or not), this is a case of Classification. In this project, the below two classification models have been used:

   i. k-NN
   ii. Random forest classifier

4. **Model performance evaluation and comparison**

   The performance of two models are compared through F1 scores to determine which one is more accurate in predicting the response variable.

Conclusion –

- F1 scores are the harmonic mean between Precision and Recall. It ranges between 0 to 1, higher the score, better the predictive performance of the model.
- Both the models are just fit as the F1 scores for the training and test dataset is very close. So, there is no case of underfitting or overfitting.
- **F1 scores from the k-NN model (75%) is lower than the Random Forest classifier (85%).**
- Thys, the latter should be used for further testing and deployment for predicting which customers are more likely to cancel their bookings.