Predicting Hotel Booking Cancellation

Waad Alotaibi

Abstract

The increasing number of hotel booking cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. The goal of this project is to use classification models to predict the hotel booking cancellation. I worked with the data provided to build a predictive model that can predict which booking is going to be canceled in advance and help in formulating profitable policies for cancellations and refunds.

Design

This project originates for predicting hotel booking cancellation. the 'Customers Booking Details' dataset is used for this project. The prediction was checked through four different Machine learning models Random Forest, Decision Tree, Support Vector Machine and k-nearest neighbors.

Data

For all the experiments and evaluations performed in this project I have used the Customers' Booking Details dataset. It used for supervised binary classification tasks. Total number of features in the dataset is 5499 distributing 18 attributes as some features input to the model and single attribute as output label.

Algorithms

Firstly, I checked the insgihts of dataset by perform the Exploratory Data Analysis (EDA) and then performed the preprocessing to get the most relevant features for prediction. After all the anaylises, I check the prediction through four different models Random Forest, Decision Tree, Support Vector Machine and k-nearest neighbors. Then compare all the models through different performance metrics eaccuracy, precision, recall and f1-score and use the model which gives best results as compared to other models.

1. Exploratory Data Analysis (EDA)

Exploratory data analysis is performed to gain different useful information and hidden insights from dataset. In this section different statistical techniques have been used to gain insights and then being visualized into appropriate charts and plots. A few questions have been mentioned below which help approach the analysis in the right manner and generate insights from the data.

1. What are the busiest months in the hotel?
2. Which market segment do most of the guests come from?
3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

4. What percentage of bookings are canceled?
5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

2. Feature Selection

I removed the features which has high one category value percentage greater than 90% because if one feature has 90% values belong to single value so our model will always predict the high category values. Features are: no_of_childeren, no_of_previous_booking_not_cancelled, no_of_previous_cancellations, reprepeated_guest, required_car_parking_space.

3. Label Encoding and Feature scaling

In our dataset, there are many columns are categorical variables. So, we transform non-numerical labels to numerical labels since we would only need numbers in the equations. Then, we limit the range of variables so that we can compare them on common grounds. We also standardize our data that will be better for model prediction.

4. Splitting data

Every dataset for Machine Learning model must be split into two separate sets – training set and test set. We split the dataset to 80% for training and 20% for testing and we can see the shape of training (4399, 9) and (1100, 9) for testing.

Models

1. k-nearest neighbours

Accuracy on validation set: 0.8173
Precision on validation set: 0.8535
Recall on validation set: 0.8620
F1_Score on validation set: 0.8577

2. Support Vector Machine

Accuracy on validation set: 0.8282
Precision on validation set: 0.8589
Recall on validation set: 0.8748
F1_Score on validation set: 0.8668

3. Random Forest

Accuracy on validation set: 0.8427
Precision on validation set: 0.8743
Recall on validation set: 0.8805
F1_Score on validation set: 0.8774

4. Decision tree

Accuracy on validation set: 0.8000
Precision on validation set: 0.8567
Recall on validation set: 0.8250
F1_Score on validation set: 0.8406

Model Evaluation and Selection

Evaluation Metrics

State of the art evaluation metrics for supervised binary classification problems are applied : accuracy ,precision, recall and f1-score to measure each model performance.

Results

This section delves into the performance of all machine learning models that used on our dataset and conducts a systematic comparative analysis to determine which model is the best.

the table below shows result of all models on testing data. It clear that Random Forest attains the highest Accuracy (0.84) on testing data and high other metrics as compare to others.

Metrics	decision tree	k-nearest neighbours	support vector machine	random forest
accuracy	0.81	0.817273	0.828182	0.847273
precision	0.863235	0.853521	0.858939	0.881598
recall	0.834993	0.86202	0.874822	0.87909
f1-score	0.848879	0.857749	0.866808	0.880342

Conclusion

We check the results through all model. So, we will use Random Forest model for prediction because it giving good score as compared to other model.

Tools

- ✓ Numpy and Pandas for data manipulation
- ✓ Scikit-learn for modeling
- ✓ Matplotlib and Seaborn for plotting
- ✓ Tableau for visualizations

Communication

In addition to the slides and visuals of Predicting Hotel Booking Cancellation project will be embedded on my personal GitHub.