DePaul University
Jarvis College of Computing Digital Media

DSC 478 Final Project Executive Summary

Project Title:

Hotel Booking (A Data Analysis Project)

Team Members:

Bramhashree Manoharan, Hoda Masteri, Goutham Selvakumar

Fall 2022

The Project goal:

The goal of our project is to predict whether a hotel reservation made by a customer would lead to revenue loss or gain, and in general, what are the characteristics of hotel reservations that can be considered profitable for the hotel and what are the characteristics of the reservations that would result in loss of revenue.

Methods used:

To predict whether a hotel reservation would result in profit or revenue loss for a hotel, we needed an appropriate target or label for our dataset. None of the single columns of the original dataset provides information on revenue, so we created a revenue column to be used as our target column for classification tasks. The steps of creating the revenue target column are explained in depth in the report. To predict profit or loss for each reservation, we utilized different classification algorithms from Scikit-Learn. We created KNN, Decision Tree, SVM, Naïve Bayes, LDA, and Random Forest classifiers and tuned their hyperparameters using grid search. We then evaluated them by reviewing the train, test, and cross-validation accuracies and the confusion matrices. What we were expecting was that for a good model, the accuracies should be high, the gap between the train and test accuracies should be low (no overfitting; a more generalizable model), and the confusion matrices should have very few off-diagonal non-zero elements (low misclassification). The results of comparisons between these classifiers are provided in the conclusion section. We also performed unsupervised knowledge discovery to divide the dataset instances into clusters for pattern recognition. We repeated the clustering with different values for the parameter k (number of clusters) and chose k=2 based on the scores and the ease of detecting patterns. Even though the Silhouette mean, and completeness and homogeneity scores were not high, we managed to discover that one of the cluster centroids had characteristics of a profitable reservation and the other cluster centroid indicated that it mostly includes the non-profitable reservations.

Conclusions:

Below are the Overall Accuracies for each classifier:

- Random Forest: 0.85 (+/- 0.02)
- Decision Trees: 0.82 (+/- 0.02)
- KNN: 0.80(+/- 0.01)
- SVM: 0.79 (+/- 0.01)
- LDA: 0.78 (+/- 0.02)
- Naïve Bayes: 0.53 (+/- 0.03)

Based on our analysis on each classifier, we conclude that Random Forest, being an ensemble model, outperformed the simple classifiers with an overall accuracy of 85%.

In addition, the results, and observations from the unsupervised knowledge discovery (clustering) helped us in pattern recognition in the sense that the findings were in agreement with the information we had previously obtained from exploratory data analysis and visualizations, and even clustering provided more information about characteristics of each class beyond what the EDA and visualizations had provided us.