5.3 Assignment: Create Optimal Hotel Recommendations

Soukhna Wade

Bellevue University

Top of Form

**Assignment 5.3 – Summary**

**DSC630**

After completing the Exploratory Data Analysis (EDA), we got the idea to select the following algorithms based on the understanding of the datasets. In the following you will find the selected algorithms:

First, the Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors. SVM can do both classification and regression. The clusters are multi-level (100) and used non-linear SVM. Non-linear SVM means that the boundary that the algorithm calculates doesn't have to be a straight line. The advantage is that we can capture much more complex relationships between the data points without having to perform difficult transformations on our own. The downside is that the training time is much longer as it's much more computationally intensive.

Using SVM, help to achieve the highest cross-validation score.

Second, using the Naive Bayes classifier, which assumes that the presence (or absence) of a feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Naive Bayes uses a similar method to predict the probability of different classes based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes. But it has the worst performance of the four models. Therefore, this classifier is not recommended for the problem at hand.

Third, Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. The hotel falls in a specific cluster (yes/no) based on the chosen features. Logistic Regression was close to the performance of SVM but slightly worse.

Fourth, K nearest neighbors is a simple algorithm that stores all available cases and classify new cases based on a similarity measure (e.g., distance functions). It has been used in statistical estimation and pattern recognition already at the beginning of the 1970s as a non-parametric technique. The idea of KNN is to teach the model which users (with other similar characteristics) chose which hotel cluster and predict future cluster assignment based on that learning.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

KNN performed very similar to Logistic Regression for the model in question.