Steven Simonsen

DSE6211

9/08/2024

Course Project – Analytic Plan

The purpose of this document is to provide a detailed analytic plan surrounding the company ABC Hotels. The label for this supervised classification problem is to determine the bookings that have a high risk of cancellation. In the creation of this machine learning model, a probability value closer to 1 will indicate a higher risk for cancellation, and a value closer to 0 will indicate a lower risk for cancellation.

There are a total of 17 predictors within the supervised classification dataset, not including the label, which is titled as the variable booking\_status. To begin data processing, an analysis will be conducted to determine whether specific variables should be excluded from the model. Initially, it appears that the Booking\_ID variable should be excluded from the model. This is a unique ID for each of the 35,000 records contained within the data and does not serve relevance as to whether a booking is at high risk for cancellation. Second, it appears that some variables are numeric, while others are in categorical format. The categorical data, such as the type\_of\_meal\_plan, will need to have one-hot encoding applied, so that all inputs to the neural network are numeric. Since Keras and TensorFlow will be used to create the model in this exercise, all data will need to be converted into the form of arrays.

As previously noted, all predictors aside from the Booking\_ID will be included as predictors within the model. Some of the features resulting from data engineering will include the extraction of the day, month, and/or year from the arrival\_date variable. Additionally, experimentation will be necessary to determine which, if not all, of these date-based features from this variable will best contribute to the model’s overall performance. The lead\_time variable may also provide valuable features. At a minimum, this variable, along with the avg\_price\_per\_room variable, will be scaled to align more closely with the rest of the dataset. The lead\_time may also prove useful in combination with the arrival\_date to create a time-series based variable that could contribute to interactions and relationships within the data.

The main analytic outcome of this machine learning project will be the generation of a value between 0 and 1, to be interpreted as the probability of cancellation. Additionally, the performance of the model can be evaluated with various metrics such as roc\_auc and accuracy. From an informational perspective, class-based predictions will be made to determine whether a booking will be cancelled or not cancelled. This will assist the ABC Hotels company in their efforts to retain high-risk customers through this use of targeted advertising and other efforts to ultimately lead to a lesser cancellation rate. This, in turn, will result in higher profits for the ABC Hotels company.