Analytic Plan

DSE6211

Adeline Casali

*Instructions:*

*The first step in the machine learning process is to carefully consider the objectives (i.e., the business needs) in the context of the available data, appropriate and applicable machine learning methods, as well as the expected analytic and informational outcomes. Consequently, the Analytic Plan is a detailed and specific outline addressing at least the following:*

* *Using the business need and project data available in Canvas (on the "Course Project" page of the "Start Here" module), propose a supervised classification problem to address the business need.*
  + *What is the label (i.e., the target or dependent variable) for the supervised classification problem?*
  + *What data processing is needed and how will it be performed? Note: all variables should be included in the analysis unless a reason is given for exclusion.*
  + *What features will be initially included?*
* *What are the expected analytic and informational outcomes to be produced? How will the model be used in practice?*

*The Data*

***For the project in this course, address the following business need of the customer using the provided data set.***

***Customer:****ABC Hotels.*

***Business Need****: ABC Hotels would like to identify bookings that have a high risk of cancellation. The risk of cancellation should be a value between 0 and 1, so it can be interpreted as the probability of cancellation. With this capability, hotel management can target bookings that have a high risk (i.e., probability) of cancellation with additional advertisements and/or offers in an effort to prevent them from being cancelled.*

***Data****: ABC Hotels has provided a data set containing over 35,000 bookings for which it is known whether or not the booking was cancelled. Students are required to use this data set provided in the zipped folder below.*

**Objectives**

* Address business needs.
* Build a model capable of determining the probability of cancellation for future customers using data from 35,000 past customers.
* Use this information to target bookings.
* Brief description of machine learning methods utilized.
* Model over time to state increasing over time, include figure.

Hotel booking cancellation rates in the US have increased a whopping 33% from 2019-2022, resulting in a total of 20% of all bookings canceled in 2022 (SHR Group, 2023). This project aims to develop a model utilizing neural networks to predict the probability of cancellation for every new booking. ABC Hotels has provided data on 35,000 past customers' booking and cancellation behavior to fit the model. Within this data, cancellation rates have varied over time but overall have risen from 14.8% in 2017 to 36.7% in 2018 (Figure 1). This model will allow ABC Hotels to target bookings with a high risk of cancellation with additional advertisements, offers, and discounts in an effort to retain them as customers. Furthermore, this model will reduce uncertainty in demand management decisions and the need for restrictive cancellation policies, often resulting in negative social impacts (Antonio et al., 2019).

A graph showing the growth of a number of years

Description automatically generated

*Figure 1: Cancellation rates over time, based on ABC Hotels data*

**Data**

* Describe the data: 35,000 observations, 17 columns.
* Define label and features utilized (and not – booking id not needed, justify).
* Data preprocessing – encoding categorical variables into integers, separating dates into month, day of month, and day of week.
* Usual data preprocessing - correlation, missing values, outliers.
* X number of bookings were cancellations.

The data provided from ABC Hotels contains 35,000 observations (bookings) over 17 variables (columns). Within the data, 11,878 bookings were cancellations, and 24,360 number were not cancellations, for a total cancellation rate of 32.8%. The label, also known as target or dependent variable, investigated in this scenario is the booking status feature, which states whether or not the reservation was canceled. The other features being used to fit the model are the number of children and adults, the number of weeknights and weekend nights, the type of meal plan, whether they required a car parking space, the room type, lead time, arrival date, market segment type, whether they were a repeat guest, the number of previous cancellations and bookings not canceled, the average price per room and the number of special requests. The categorical variables such as meal plan, room type, market segment type, and booking status will be encoded into integers to prepare this data for use in the model. Additionally, dates will be separated into the month, day of the month, and day of the week to extract more information. Finally, the data will be explored to identify any outliers or missing values as well as any correlations between variables that might cause interference in the model.

**Outcomes and Impact**

In practice, information for new bookings will be fed into the model to calculate the probability of the customers canceling their booking. The model will produce the probability of cancellation for new bookings as a value between 0 and 1. In turn, this information can be passed on to marketing teams and used to provide customers with a high risk of cancellation with additional offers or advertisements in an effort to prevent their cancellation. As cancellations result in a large loss of income to hotel businesses, this investment will result in a reduction in the loss of income suffered from cancellations for ABC Hotels.

**References**