Project Proposal

Pelumi Ajayi, Heather Seo, Qiyu Wang
Team Name: Infinity (∞)
Team 8

Business Problem

<u>• Discuss what is the exact business problem</u>

To minimize vacancy rates of the hotel.

<u>• Discuss how tackling that business problem with data science would add business value</u>

A typical approach hotels use to reduce vacancy rates is to target a certain predefined booking ratio. However, performance outcomes from this approach are inconsistent for high-demand and low-demand periods. By using a predictive model instead of simple descriptive statistics of historical data, we produce more precise estimates of cancellations by identifying the most influential factors driving these vacancies. With those refined estimates, we can decide on an overbook ratio accordingly and therefore reduce the vacancy rates and maximize the profit.

Modeling Ideas

<u>• What precisely is the data science problem? (e.g., classification vs regression)</u>

This is a classification problem. We want to predict the possibility of cancellation for hotel bookings based on numerous attributes related to booking such as time of the year, type of room, and occupant age.

<u>• Is the data science task a supervised or unsupervised task?</u>

This is a supervised classification task.

<u>• What is a data instance in your data set?</u>

A reservation for one room

<u>• What might be the target variable?</u>

The target variable is ‘is_canceled’, which is a historical record of whether a reservation was cancelled or not.

<u>• What features would be useful in predicting the target variable?</u>

By simply looking at the correlation matrix, the followings seem to be useful:

- The booking lead time which is the period of time between when a guest makes a reservation;
- If parking spaces are needed
- Any changes in booking happened
- Previous cancellations
- length of stay
- if any special requests were made
- Number of people
- Do they have children(or baby)
- if they’ve been a guest before
- If they use a agent

Data Details

• Give a short description of the data you are planning to use.

This data set contains booking information for a city hotel and a resort hotel, and includes information on each reservation made such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

• How have you obtained the data? What is the source of the data?

We obtained the data from Kaggle: https://www.kaggle.com/jessemostipak/hotel-booking-demand/tasks

• Please provide as much as you can from the following information:

o Number of examples?

119,390

o Number and types of variables?

There are 31 variables and they vary from categorical(polynomial and binary), and numerical

- Numeric: lead_time, stays in weekend nights
- Polynomial Categorical: meal, country, market segment
- Binary Categorical: Hotel type, cancellation, is repeated guest

o If you are planning on a classification problem – what is the ratio of the most prevalent class?

The prevalent class is “did not cancel reservation”, and that is 63% of total bookings.