Capstone Project Proposal

Domain Background

Have you ever book a hotel room but canceled it later? Did you hear the complaint from a hotel manager when he/she has prepared everything but the visiter was absent? Is there a way to predict whether the customer is going to cancel the reseration? This research is going to use machine learning to solve this realistic issue.

Problem Statement

I will use the collected data set which contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, among other things to train a model. The goal is to predict whether the custmer may cancel the reservation by using some relavent variables with machine learning.

Datasets and Inputs

The data set was generated from https://www.kaggle.com/jessemostipak/hotel-booking-(https://www.kaggle.com/jessemostipak/hotel-booking-) demand (https://www.kaggle.com/jessemostipak/hotel-booking-demand (https://www.kaggle.com/jessemostipak/hotel-booking-demand)). It is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The target variable is *is_canceled*. I have chosen some varaibles that may be relevent with *is_canceled* and will deal with the other variables to select the ones that have the largest affect.

“hotel”: Resort Hotel/City Hotel

“is_canceled”: Value indicating if the booking was canceled (1) or not (0)

“lead_time”: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

“arrival_date_week_number”: Week number of year for arrival date

“stays_in_weekend_nights”: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

“stays_in_week_nights”: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

“adults”: Number of adults

“children”: Number of children

“babies”: Number of babies

“meal”: Type of meal booked

“reserved_room_type”: Code of room type reserved

“customer_type”: Type of booking

“adr”: Average Daily Rate

“total_of_special_requests”: Number of special requests made by the customer

Solution Statement

I first select the variables that are most relevent with *is_canceled*. Then, split the data into training set and test set in order to train the model and test the accuract. I will try several kinds of models in SKlearn API and find the best one to fit.

Benchmark Model

I have made a naive Bayes classifier assuming the predictors have independent normal distributions. In this way, the naive Bayes classifier can predict a not canceled booking with high accuracy, though it cannot predict a canceled booking well, with accuracy 94.7% and 15.6% respectively.

Evaluation Metrics

When the predicted results are shown, I will make a *is_canceled-predicted_canceled* table to see the accuracy of the prediction on test data. Due to the previous attempt, I will consider the canceled room and not canceled room seperately.

Project Design

Just like in the Plagiarism-Detector project, there are 2 main steps: variable selection and model construction.

First, transfrom all the variables to numbers. Then, try to find some highly correlated variables. Next, train the model with different types and find the one with highest accuracy. Finall, deal with the variables to see if deleting or combining some of them may make the prediction better.

That is all!

In [ ]: