AMIRKABIR UNIVERSITY OF TECHNOLOGY

DEPARTMENT OF COMPUTER ENGINEERING AND IT

Principles of Data mining

Final Project

Kian Eliasi

MohammadMahdi Heydari

Armin Kazemi

Saeedeh Sadeghpour

Under the supervision of: Dr. E. Nazerfard

Spring 2020

1 Task

Have you ever wondered when the best time of year, to book a hotel room is? Or the optimal length of stay to get the best daily rate? What if you want to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? In this project, you will treat a hotel booking dataset that can help you explore those questions. More specifically, this is a supervised learning project in which you will find a way to predict the chance of cancellation for a particular hotel booking record.

2 Implementation Details

You can find the main page for this project in Kaggle. Kaggle is a website that supports an online judge for data mining contests. By registering in the website you will be able to access the dataset, read the instructions and discussions about the problem. You can also access the dataset here.

- Dataset description
  - hotel bookings.csv: This file contains a table with 119390 rows and 32 columns that is all the data you are provided. You have to split the dataset to train set and test set yourself.
- Data fields: a brief explanation of what each column means.
  - hotel: Type of the hotel (H1 = Resort Hotel or H2 = 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 (Property Management System) and the arrival date.
  - arrival_date_year: Year of arrival date.

- arrival_date_month: Month of arrival date.
- arrival_date_week_number:Week number of year for arrival date.
- arrival_date_day_of_month: Day of 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. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal.
- country: Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- market_segment: Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators".
- distribution_channel: Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators".
- is_repeated_guest: Value indicating if the booking name was from a repeated guest (1) or not (0).
- previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking.
- previous_bookings_not_canceled: Number of previous bookings not cancelled by the customer prior to the current booking.
- reserved_room_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- assigned_room_type: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
- booking_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
- deposit_type: Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
- agent: ID of the travel agency that made the booking.
- company: ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons.

- days_in_waiting_list: Number of days the booking was in the waiting list before it was confirmed to the customer.
- customer_type: Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking.
- adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.
- required_car_parking_spaces: Number of car parking spaces required by the customer.
- total_of_special_requests: Number of special requests made by the customer (e.g. twin bed or high floor).
- reservation_status: Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why.
- reservation_status_date: Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel.

As your first step into finding a solution for a datamining problem, You should explore into the dataset which is formally called EDA. You have to determine which features of the dataset are more informative. You may also need to create new features out of the existing features. For example, you can sum values of two features divided by a third feature and make a new feature that would be more helpful for this classification task. Then you have to take at least two methods, compare the results of them and explain the reasons that have possibly caused the difference in the performance of your machine learning methods. it is highly recommended that you implement one ensemble method of your own choice and compare it with a neural network. you should report the performance metrics both on the training set and test set. Your code must be implemented with python programming language and its related libraries like Scikit-learn, Tensorflow, Keras, etc. You can find initial ideas online but they should be fully elaborated on, in your report.

3 Caution

- Report is an important part of your grade. So please write it completely, carefully ,and explain your analysis. Your report is only accepted in ’pdf’ format. Please put it in a folder named ”report”. (There is no force on the language of the report)

- Your codes should be written in python. Put them in ”supporting material” folder. They can either be ’py’ or ’ipynb’ files.

- The deadline for this project is 1399/05/25, 11:55 PM. you will lose 10 percent of your grade for each day of delay.

- Put all your folders and files in a ’zip’ file and upload it on moodle.

- note that this is much more like a data mining contest. i.e. building a specific model or method is not obligatory. (but please pay attention that you have to explain all the processes you have made to your data, building your model, evaluation metrics, etc.)