Theodore Koby-Hercsky

12/01/2022

DSC680-T301 2233-1

Professor Catie Williams

Milestone 1 - Proposal

Topic - Describe and name your project in 1-2 sentences max

Demand for hotel bookings and how one can predict guest cancellations due to different variables such as room rates, busy months, and length of visits.

Business Problem - Describe the business problem your project is trying to solve and/or the research questions you will explore

The research question I would like to explore is the demand for hotels depending on the time of year. As this will allow customers to determine if a busy season that has an increase in price and occupancy is better than a slower season with less occupancy and lower rates. By researching into these variables and creating prediction models' businesses can use this to solve problems such as increasing occupancy during slower seasons and how to bounce back from cancellations.

Datasets - where are you getting your data? Describe the data that you will use to solve the problem

Data set: Hotel booking demand | Kaggle

The data set that I will be utilizing contains data for a resort hotel and a city hotel. The data set also includes 31 other variables that can be seen below with a brief description of each.

- Is_Canceled – Indicates if the booking has been canceled or not with a (1) for cancelled and (0) if not.
- Lead_Time – Indicates the number of days that elapsed between the entering date of the booking into the PMS and arrival date.
- Arrival_Date_Year – Indicates the arrival date
- Arrival_Date_Month – Indicates the month of the arrival date.
- Arrival_Date_Week_Month – Indicates the week number of year for the arrival date
- Arrival_Date_Day_Of_Month – Indicates the day of arrival
- Stays_In_Weekend_Nights – Indicates the number of weekend nights such as Saturday and Sunday the guest will be staying or booked to stay at the hotel
- Stays_In_Week_Nights – Indicates the number of week nights such as Monday through Friday the guests will be staying or have booked to stay at the hotel.
- Adults – Indicates the number of adults
- Children - Indicates the number of children
- Babies - Indicates the number of babies
- Meal – Indicates the type of meal booked. These categories include standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board

(breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)

- Country – Indicates the country of origin
- Market_Segment – Indicates the market segment designation such as the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
- Distribution_Channel – Indicates the booking distribution channel such as the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
- Is_Repeated_Guest – Indicates the value indicating if the booking name was from a repeated guest is (1) and (0) if not
- Previous_Cancellations – Indicates the number of previous bookings that were cancelled by the customer prior to the current booking.
- Previous_Bookings_Not_Cancelled – Indicates the Number of previous bookings not cancelled by the customer prior to the current booking.
- Reserved_Room_Type – Indicates the Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- Assigned_Room_Type – Indicates the 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 (such as overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
- Booking_Changes – Indicates the 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 – Indicates the 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 – Indicates the ID of the travel agency that made the booking.
- Company – Indicates the 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 – Indicates the Number of days the booking was in the waiting list before it was confirmed to the customer.
- Customer_Type – Indicates the 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 – Indicates the Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

- Required_Car_Parking_Spaces – Indicates the Number of car parking spaces required by the customer
- Total_Of_Special_Requests – Indicates the Number of special requests made by the customer (such as twin bed or high floor)
- Reservation_Status – Indicates the 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 – Indicates the Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when the booking was canceled or when did the customer checked-out of the hotel.

Methods - What analysis methods will you use to complete this project? Note: this is just a proposal, your project can adapt as you work on it

This project will require exploratory data analysis to get a view at what the data holds and what improvements I can make to better utilize the data set. I will also create different predictive models that focus on cancellations from the data set. One model I plan on using will be the Random Forest Classifier which is a supervised machine learning algorithm that is used for regression analysis and classification. This model will use a randomly selected subset of the training data to create a set of decision trees that will collect votes from trees to determine a prediction.

Another model that will be used is the Gradient Boosting Classifier which is used to improve a prediction throughout each predecessor which will reduce errors and increase the prediction. The model will use the log of the odds from the targeted feature that will allow for the creation of the initial prediction from the data. A Gradient Boosting Classifier will allow viewers to see the accuracy of the model that includes the precision, recall, f1-score, and support.

The K-Nearest Neighbors or KNN will be another model that will be utilized in this project. This model is a machine learning supervised algorithm that is very user friendly and a go to model when working with predictions. The KNN is mainly used for regression and classification needs but can also be utilized for data sets with missing values. The algorithm takes observations that are closest to a given data variable is seen as the most similar. Which allows the individual the capability to select the number of nearby observations by choosing K to use in the algorithm.

These are just three models I plan on trying to utilize for my project but will also pursue other models such as but not limited to the Decision Tree, XgBoost Classifier, and more. I would also like to focus on EDA and different visualizations to help viewers see the picture behind different questions they may have.

Ethical Considerations - What are some potential ethical concerns of this topic or analyzing the data?

Anonymity is one ethical issue that this data set has as the user does not know the identities of the hotel reservations. The personal data of the hotel guests was not provided. Being that the data set provided did not include any person information I do not see any ethical issues with voluntary participation as no personal information was used. While informed consent would also not be an ethical issue as no personal information was used.

Challenges/Issues - What are some issues and challenges do you think you might face?

One challenge I could face is pinpointing the cause of cancellations and where these cancellations are coming from due to no personal information this might be a challenge. Another challenge I might face is missing data I will have to perform some EDA to determine what data is missing if any and determine the best course of action.

References - What sources will you use to validate your results and support your project topic?

- I will use a Random Forest Classifier article to gain a deeper understanding of the model I will be creating.
    - Random Forest Classifier using Scikit-learn - GeeksforGeeks
- A Gradient Boosting Classification article from Towards Data science will be utilized to get a better understanding of the model and how I can use it to my advantage.
    - Gradient Boosting Classification explained through Python | by Vagif Aliyev | Towards Data Science
- The K-Nearest Neighbor article will be utilized to create a KNN algorithm with our data and understand how it is used.
    - Python Machine Learning - K-nearest neighbors (KNN) (w3schools.com)
- Ethical Consideration article used to review potential ethical concerns within the data set.
    - Ethical Considerations in Research | Types & Examples (scribbr.com)