Team 7: Team No Cap All Fax, Barbz Nation, Citizens of Ram, Wakanda Forever

Janvier Nshimyumukiza, Ashish Sangai, Sherraina Song, Wenqi Zhai

Introduction to Business Analytics

Business Overview

Team 7 will analyze hotel cancellation data. Booking cancellations can be a hassle for the hospitality industry impacting on demand-management decisions. Many hotel chains now implement rigid 24 to 72 hour cancellation policies and overbooking tactics. These two strategies do not always help in maintaining the hotel's reputation and repeat customers.

We want to demonstrate that by using machine learning and data mining, it is possible to build models to predict whether or not a customer will cancel their booking. By running the models daily against all reservations on-the-books, a hotel location can discover new information such as the number of room nights predicted to be canceled for each of the following days. Equipped with an accurate demand value, hotel managers can develop more effective overbooking and cancellation policies to accommodate their guests and maximize revenue.

Modeling Ideas

This is a classification problem and hence the data science task is supervised learning. But we will also use unsupervised learning to explore the data before employing supervised learning.

Each observation (instance) represents a hotel booking. The Observation contains the following information: hotel type, whether the reservation was canceled or not(is_canceled), lead time, arrival date year, arrival date month, arrival date week number, arrival date day of month, stays in weekend nights, stays in week nights, number of adults specified in the reservation, number of children, babies, meal type, country, market segment, distribution channel, is repeated guest, previous cancellations from that customer, previous bookings not canceled, reserved room type, assigned room type, booking changes, deposit type, agent, company, days in waiting list, customer type, adr, required car parking spaces, total of special requests, reservation status, and reservation status date.

We chose is_canceled (whether the booking was canceled or not) as our target variable. It is derived from the reservation status(check-out/canceled). Many variables can be used in predicting whether the booking was canceled or not. We will explore some variables like the lead time, hotel type, country of origin, repeat guest, previous cancellations etc.

Data Details

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

The dataset compares various booking information between two hotels: a city hotel and a resort hotel. It has multiple variables (32 columns) and we plan on using whether the booking was canceled or not as the target variable.

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

We obtained the data from Kaggle. The original data 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. Within the article, data were presented as two separate datasets: The first dataset is for resort hotel data which has 79,330 observations and the second is for city hotel data which has 40,060 observations. It contained 31 variables but the version provided on kaggle combines these datasets to form the version that we will use in this project. Thus, our dataset has 119,390 observations with 32 columns.

- **Load the data into RapidMiner/Python for EDA**

```
In [4]: df = pd.read_csv("./hotel_bookings.csv")
```

And we use info(), describe() to check the data structure and details of each variables, below are our discoveries:

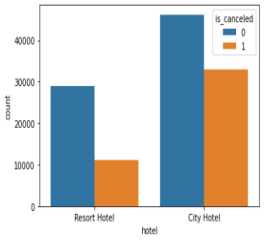
1) The dataset contains 119,390 rows and the number of the variables is 32
2) Variables with missing values: 4 variables namely country, agent, company, and children will require data cleansing
3) Types of variables: the datatype of variables are shown in the screenshot in python and the feature types are as follows:
    a) Categorical - hotel, is_canceled, customer_type, is_repeated_guest, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, agent, company, reservation_status
    b) Numerical - lead_time, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellations, booking_changes, previous_bookings_not_canceled, days_in_waiting_list, adr, required_car_parking_spaces, total_of_special_requests, arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month, reservation_status_date

```
In [7]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

```
In [8]: sns.countplot(data=df, x='hotel', hue='is_canceled')

resort_canceled = df[(df['hotel']=='Resort Hotel') & (df['is_canceled']==1)]
city_canceled = df[(df['hotel']=='City Hotel') & (df['is_canceled']==1)]
print('Cancelations in resort hotel= ', (len(resort_canceled))/(len(df[df['hotel']=='Resort Hotel'])))
print('Cancelations in city hotel= ', (len(city_canceled))/(len(df[df['hotel']=='City Hotel'])))

Cancelations in resort hotel=  0.27763354967548676
Cancelations in city hotel=  0.41726963317786464
```




```
In [9]: df['is_canceled'].value_counts(normalize=True)

Out[9]: 0    0.629584
        1    0.370416
        Name: is_canceled, dtype: float64
```

4) This is a binary classification problem, and the ratio of the prevalent class (is_canceled = 0) is 0.6296, which means the percentage for reservation not being canceled in the dataset is 62.96%. And we also explored the canceled (is_canceled =1) class in resort and city hotels and the cancellation rates are 27.76% and 41.72% respectively.