**Hotel Booking Cancellation modelling and Number of Guests’ FORECAST**

**Project Summary Document**

Annu Priya (apriya3@buffalo.edu)

**Problem Statement**

* The hospitality industry often fails to better analyze their bookings, cancellations, etc. as the industry is quite volatile.
* A lot of dynamic factors impact their business continuity.
* The inspiration for this project came in as we often find ourselves intrigued by the below questions:
  + Grounds on which a hotel is likely to receive a cancellation.
  + What would be the expected footfall of the customers in future?

**Data Source**

* The dataset has been sourced from www.kaggle.com, from which any personal identifying data have been removed.
* The dataset, however, 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***.
* Link to the dataset: <https://www.kaggle.com/jessemostipak/hotel-booking-demand>.
* It contains over 100k booking information records having the below 32 features(Nuno Antonio):
  + Hotel: Resort Hotel or 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\_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 package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
  + 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 the booking was canceled or when did the customer checked-out of the hotel

**Platform and Packages Used**

The project was developed on Jupyter Notebook and the below packages:

* datetime
* numpy
* pandas
* seaborn
* matplotlib
* auto\_arima
* statsmodels
* sklearn
* math
* warnings

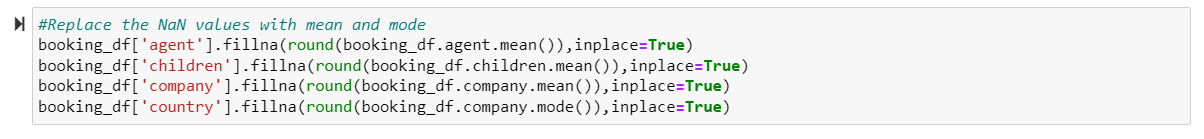
**Data preparation**

* Data Cleaning
  + Counting all the NaN values in the columns of the dataset: isna()
    - Company, agent, country and children’s columns had NaN values

Table

Description automatically generated

* + Filling NaN values in the dataset with appropriate values: fillna()
    - The NaN values in the above columns were replaced by their respective mean and mode values



* + Dropping irrelevant columns: drop()
    - We dropped certain columns from the dataframe as they were not participating in the modeling as per the correlation.



* Data Preprocessing
  + Label Encoder: LabelEncoder()
    - We encoded the categorical variables required to execute the correlation

Table

Description automatically generated

* Data Preprocessing
  + Scaling: StandardScaler()
    - Scaling was also performed on the data before building the model



* Data Visualization
  + Correlation

Chart, scatter chart

Description automatically generated

What percent of the bookings were cancelled?

Chart, bar chart

Description automatically generated

Booking ratio between the resort and city hotels:

Chart, bar chart

Description automatically generated

Duration of stay for the customers.

Chart, line chart

Description automatically generated

* Data Partition:
  + For the modeling, we split the dataset into train and test using train\_test\_split().
  + 75% data were used to train the model and the rest 25% were used to test it.

Graphical user interface, application

Description automatically generated with medium confidence

**Modelling**

**All the modelling were done on Jupyter Notebook**.

Below were the inputs and output for the model:

* Inputs: Columns Predictors from the dataset
* Output: is\_canceled – whether the booking was canceled

The models were developed using the below two methods:

* + Logistic Regression

Graphical user interface, text, application, email

Description automatically generated

* + Random Forest

Text

Description automatically generated with medium confidence

No. of guests' prediction was forecasted by using time series ARIMA model.

Inputs to time series forecasting ARIMA model:

* + Total Guests = No. of adults + No. of children
  + Arrival Date = MM/DD/YYY
  + 2 separate dataframes for City and Resort Hotels
  + Time duration: July 2015 till Aug 2017 (weekly data used - 113 weeks)
  + Training set: 75% and Test set: 25%
  + Order of ARIMA model from auto\_arima(): p (AR), d (I) and q (MA)

Auto ARIMA execution to get p, d and q values for City Hotel:

Text

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

Auto ARIMA execution to get p, d and q values for Resort Hotel:

Text

Description automatically generated

Table

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* ARIMA Model for to predict the n. of guests in City and Resort Hotels:

Graphical user interface, text

Description automatically generated

**Evaluation**

* Accuracy and RMSE Values:
  + Logistic Regression
    - Accuracy: 76.75%
    - RMSE: 0.4821

Graphical user interface, text, application, email

Description automatically generated

* + Random Forest
    - Accuracy: 82.09%
    - RMSE: 0.4232

Text

Description automatically generated with medium confidence

* Accuracy comparison plot

Chart, bar chart

Description automatically generated

* Confusion Matrices:
  + Logistic Regression

Chart

Description automatically generated

* + Random Forest

Chart

Description automatically generated

* Number of guests’ predictions in the future for Resort Hotel and City Hotel:
  + - City Hotel –

Graphical user interface, chart, line chart

Description automatically generated

* + - MAPE Value: 0.266247

Graphical user interface, text, application

Description automatically generated

* + - Resort Hotel –

Graphical user interface, chart

Description automatically generated

* + - MAPE Value: 0.219413

Graphical user interface, text, application, email

Description automatically generated

* Some of the important factors effecting the booking cancellations are:
  + lead\_time: If there is a time gap between the booking and the actual check-in date, customer might venture out seeking more lucrative hotel deals that are being offered by the competitors of the hotel where he had done their bookings, which leads to the cancellations of current booking.
  + previous\_cancellations: If a customer has a history of booking cancellations, chances are there, that they might cancel the current booking as well.
  + required\_car\_parking\_spaces: Insufficient parking spaces lead to cancellation
  + deposit\_type: In case, the deposit is refundable or if the customer has made no deposits, the number of cancellations is more.

**Recommendations**

* As per the observation and evaluation we infer that the management of the hotels should strive towards raising the customer satisfaction in order to minimize the cancellations.
* The project would help the hospitality industries to grow their business by providing useful insights so that they can:
  + Better plan their business
  + Offer personalized experience to the customers
  + Offer competitive rates and deals
  + Help in marketing
* The hotel management should address the cancellations factors by offering competitive rates and improved facilities such as: sufficient parking lots, several complimentary special services, etc.
* The management should also plan to introduce a feedback survey for all the customers who have previously cancelled their bookings in the quest to improve their services and make the customers’ experiences enriching.
* The number of guests’ forecasting piece of the project should be utilized to ready the hotel inventory for the anticipated footfall of the guests.

**References**

* Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.
* <https://www.geeksforgeeks.org/python-arima-model-for-time-series-forecasting/>
* <https://towardsdatascience.com/time-series-forecasting-using-auto-arima-in-python-bb83e49210cd>
* <https://vitalflux.com/labelencoder-example-single-multiple-columns/>