**Project Summary:**

|  |  |
| --- | --- |
| Batch Details | Bangalore, January 2022 |
| Team Members | 1. Shravya Pola.  2. Swapnil Sunil Loharkar.  3. Sanjay G K.  4. Nistha Kumari.  5. Raga Bhavana Dana. |
| Domain of Project | Hotel Industry |
| Proposed Project Title | Prediction of Booking Cancellation in Hotel Industry. |
| Group Number | 9 |
| Team Leader | Sanjay G K |
| Mentor Name | Vikash Chandra |

**Dataset name:** Hotel Booking Visualization dataset.

**Introduction to the problem/domain/background details:**

The domain been chosen for the capstone project is from service sector of hotel industry (Hotel Booking demand dataset). A hotel is an establishment that provides paid accommodation (lodging and usually meals), entertainment, generally for a short duration of stay. Hotels often provide a number of additional guest services, such as restaurants, bars, swimming pools, healthcare, retail shops; meeting, conference services and facilities, banquet halls, boardrooms; and space for private parties like birthdays, marriages, kitty parties, etc.

With increase in the spending capacity of the people, both leisure and business travel spending has seen YoY growth over the past 5 years. The hotel industry has been consistently growing with The Global Hotel Industry revenue. According to the Hospitality Global Market Report 2022 the global hospitality market is expected to grow from $ 3,952.87 billion in 2021 to $ 4,548.42 billion in 2022 at a compound annual growth rate (CAGR) of 15.1%.

There are various challenges that come in the way. Not all of them might be unique but they are certainly worth the notice. Some of the major factors affecting the business are high competition between hotels, improper marketing strategy, Hiring and retaining staff, marinating online reputation; Hotels are not data driven, ineffective pricing model, and customer dissatisfaction due to inability to satisfy their requests.

**Problem Statement:**

1. To be at top of the competition between other OTA agents and to attract the customers, the OTA platforms offer cancellation of suite/room booking at free of charge. It has been serious issues in respect to the hotel management that leads to the loss of revenue and loss of the customer at the same instance, final minute cancellation leads to lower of price to resell the room. Hence, with the acquired dataset likes to approach and understand to classify/predict how likely the hotel bookings will get cancelled by a customer to strategic better.
2. In recent years, the online booking platforms acts as the middle man between hotels and customers to book a room. As there are large active competitors are in this hotel industry, the competitors tend to offer various better marketing strategized deals in try influencing the customers decision while booking a room. With the available dataset will try identify the ideal time to book a hotel to get the effective prices based on the optimal length of the stay.
3. While booking of room/suite at an effective price for a customer is always been the major problem. So, predicting a particular market segment and customer type affecting the ticket cancellation will add up in increasing the revenue.

**Business problem/ Impact in business of your problem/Need for this study/Abstract (Executive summary):**

In recent years there has been a rapid increase in hotel cancellations. More and more frequently, guests tend to cancel their room reservation. In many cases the cancellation is free of charge up to 24 hours before arrival.

Cancellations often present a big challenge for hotel managers. Especially since the rise of online travel agents such as Booking.com, Expedia and Co. it became more and more common to enforce free cancellations up to 24 hours before arrival. On one hand, this policy brings a new dimension of flexibility to hotel guests, on the other hand, it means an increasing financial risk for hotels and difficulties while planning their occupancy rate. Due to these developments, it got more and more important to analyze cancellation behavior of hotel guests and find a pattern in order to create a forecast.

**What are the numbers telling us?**

**Duration of stay.**

In general, it can be said that the booked period of stay has an actual impact on the risk of cancellation. The following applies here: If a room is booked for only 1 or 2 nights, it is more likely to be cancelled then if the length of stay is 3 nights or longer.

**Time of booking.**

The time of booking also influences the cancellation behavior. The longer a room is booked in advance, the higher is the risk that it will be cancelled. The actual amount of cancellations from early bookers rises approximately 1 month before arrival.

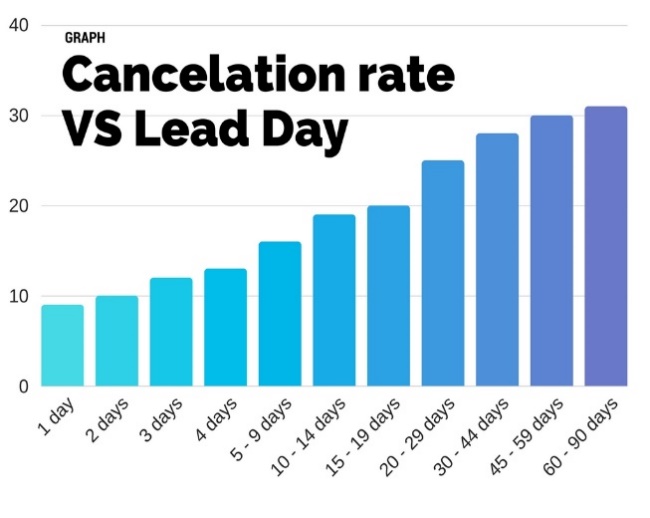
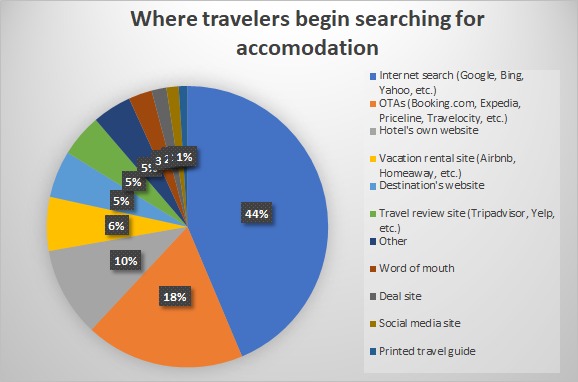
Contrary, the so-called last-minute bookings, which are made up to 10 days before arrival are less likely to be cancelled than the early ones.

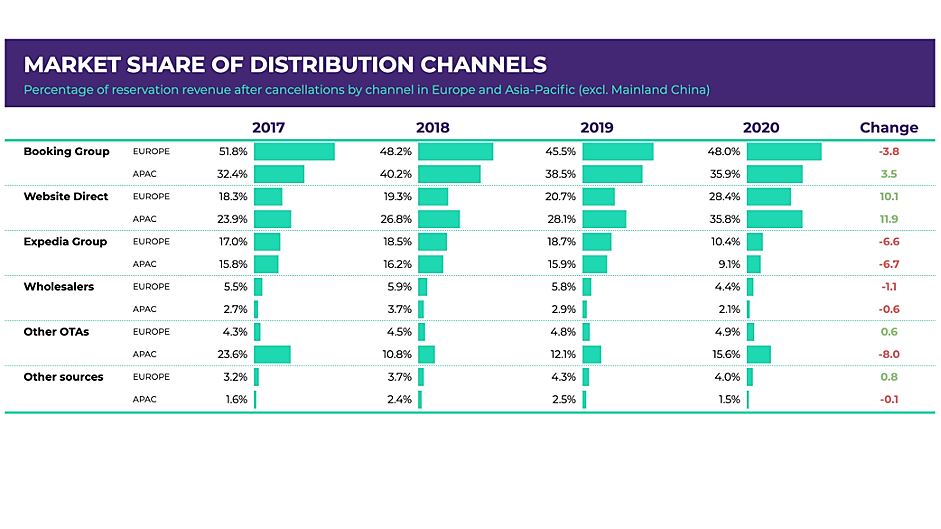
**Dealing with cancellations.**

Many hotels have come up with creative ways to counteract the trend of free cancellations up to 24 hours prior arrival. One Option is to offer several different rates and conditions at once. One of these is the Non-Refundable Rate. Although this rate is usually 10 to 20 percent cheaper than the hotel’s standard rate, it needs to be fully paid immediately and is not refundable.

Another focus is to generate more direct bookings via website. Not only does the hotelier avoid commissions from online travel agents like Booking.com who charge about 15%, but also it generates a stronger relationship between the customer and the hotel. Therefore, an attractive concept and benefits have to be developed in order to guide the guest from going through OTAs to booking directly on the hotel’s website.

It is important to regularly check the settings of the cancellation policies on the different distribution channels. Make sure that cancellation conditions on the hotel’s website do not put potential guests at a disadvantage. The hotel’s own cancellation policy should be at least as good or even better than those of online travel agencies.



**Variable identification:**

**Independent Variables:** There are 36 independent variables which are listed below:-

|  |  |
| --- | --- |
| 1. Hotel, | 1. Is\_Canceled, |
| 1. Lead\_Time, | 1. Arrival\_Date\_Year, |
| 1. Arrival\_Date\_Month, | 1. Arrival\_Date\_Week\_Number, |
| 1. Arrival\_Date\_Day\_Of\_Month, | 1. Stays\_In\_Weekend\_Nights, |
| 1. Stays\_In\_Week\_Nights, | 1. Adults, |
| 1. Children, | 1. Babies, |
| 1. Meal, | 1. Country, |
| 1. Market\_Segment, | 1. Distribution\_Channel, |
| 1. Is\_Repeated\_Guest, | 1. Previous\_Cancellations, |
| 1. Previous\_Bookings\_Not\_Canceled, | 1. Reserved\_Room\_Type, |
| 1. Assigned\_Room\_Type, | 1. Booking\_Changes, |
| 1. Deposit\_Type, | 1. Agent, |
| 1. Company, | 1. Days\_In\_Waiting\_List, |
| 1. Customer\_Type, | 1. Average Daily Rate (Adr,) |
| 1. Required\_Car\_Parking\_Spaces, | 1. Total\_Of\_Special\_Requests, |
| 1. Reservation\_Status, | 1. Reservation\_Status\_Date, |
| 1. Name, | 1. Email, |
| 1. Phone-Number, | 1. Credit\_Card, |

**Target variable:-**

1. is\_cancelled

**Variable information/Data description:**

|  |  |  |
| --- | --- | --- |
| VARIABLE | DATA TYPE | DESCRIPTION |
| 1. Hotel | Object | Hotel (H1 = Resort Hotel or H2 = City Hotel) |
| 1. Is\_Canceled | Int64 | If booking was cancelled then (1) otherwise (0) |
| 1. Lead\_Time | Int. | Number of days that passed between the entering date of the booking into the Property Management System (PMS) and the arrival date |
| 1. Arrival\_Date\_Year | Int64 | Year of arrival date |
| 1. Arrival\_Date\_Month | Object | Month of arrival date |
| 1. Arrival\_Date\_Week\_Number | Int64 | Week number of year of arrival date |
| 1. Arrival\_Date\_Day\_Of\_Month | Int64 | Day of arrival date |
| 1. Stays\_In\_Weekend\_Nights | Int64 | Number of Saturdays and Sunday’s guest stayed or booked to stay |
| 1. Stays\_In\_Week\_Nights | Int64 | Number of week nights(Monday to Friday) guests stayed or booked to stay |
| 1. Adults | Int64 | Number of adults |
| 1. Children | Float64 | Number of children |
| 1. Babies | Int64 | Number of babies |
| 1. Meal | Object | Types of meals:- 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) |
| 1. Country | Object | country of origin |
| 1. Market\_Segment | Object | Market segment designation ; through which hotels were booked.  “TA” means “Travel Agents” ; “TO” means “Tour Operators” ; ”GROUPS” means customers come in groups; “DIRECT” are those who came by their own; “CORPORATE” who got booked by their company; “COMPLEMENTARY” ; “AVITATION” and “UNDEFINED” |
| 1. Distribution\_Channel | Object | Booking distribution channel.  “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| 1. Is\_Repeated\_Guest | Int64 | If guest is repeated then(1) otherwise (0) |
| 1. Previous\_Cancellation | Int64 | Number of previous bookings that were cancelled by the customer prior to the current booking |
| 1. Previous\_Booking\_Not\_Canceled | Int64 | Number of previous booking not cancelled by the customer prior to the current booking |
| 1. Reserved\_Room\_Type | Object | Code of room type reserved. Code is presented instead of designation for anonymity reasons |
| 1. Assigned\_Room\_Type | Object | 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 |
| 1. Booking\_Changes | Int64 | Number of booking changes made till the cancellation |
| 1. Deposit\_Type | Object | Indication on if customer is making deposit to guarantee the booking. Assumed by 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. |
| 1. Agent | Float64 | ID of the travel agency that made the booking |
| 1. Company | Float64 | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons |
| 1. Days\_In\_Waiting\_List | Int64 | Number of days the booking was in waiting list before it was confirmed. |
| 1. Customer\_Type | Object | Type of customer booking can be assumed one of these categories:- 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 |
| 1. Average Daily Rate (Adr) | Float64 | Average Daily Rate:- defined by dividing the sum of all lodging transactions by the total number of staying nights. |
| 1. Required\_Car\_Parking\_Space | Int64 | Number of car parking space required by the customer |
| 1. Total\_Of\_Special\_Requests | Int64 | Number of special requests made by the customer |
| 1. Reservation\_Status | Object | Reservation status  assuming one of three categories:- Canceled- booking was cancelled; 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 |
| 1. Reservation\_Status\_Date | Object | Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel |
| 1. Name | Object | Name of the guest |
| 1. Email | Object | E-mail ID of the guest |
| 1. Phone\_Number | Object | Phone number of the guest |
| 1. Credit\_Card | Object | Credit-Card number of the guest |

**Future Work/Methodology (Details of algorithms):**

**Methodology to be followed:**

CRISP-DM which stands for Cross Industry Standard Process for Data Mining is a methodology created to help shape data mining projects. It describes the different phases/tasks involved in the project and provides an overview of data mining life cycle.

**1. Business Understanding -**  It focuses on determining the business requirements/objectives and

understanding what outcome to achieve. Also determine the business units being affected. Convert this business problem into a data mining problem and carve out an initial plan.

* Determine the business objectives: Understand what is needed to be accomplished for the customer.
* Assess situation: Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
* Determine data mining goals: Convert business problem to a data mining problem and recognize the data mining problem type such as classification, regression or clustering, etc.
* Produce a project plan: Devise a step-to-step plan for executing the project.

**2. Data understanding -** This phase starts with collecting the data and then examining the data for its surface properties like data format, number of records, etc. The next step is to better understand the data by understanding each attribute and perform basic statistics on them. Understand the relationship between different attributes. Determine the quality of data by checking the missing values, outliers, duplicates, etc.

* Collect initial data: Acquire the data and load it into the analysis tool to be used. Describe data: Examine the data and document its surface properties like data format, number of records, or field identities. Understand the meaning of each attribute and attribute value in business terms. For each attribute, compute basic statistics so as to get a higher-level understanding.
* Explore data: Find insights from the data. Query it, visualize it, and identify relationships among the data.
* Verify data quality: Identify special values, missing attributes and null data. Determine how clean/dirty is the data.

**3. Data preparation** -This stage, which is often referred to as data wrangling, has the objective to develop the final data set for EDA and modelling. Covers all activities to construct the final dataset from the initial raw data. Some of the tasks include table, record and attribute selection as well as transformation and cleaning of data for modelling tools.

* Select data: Determine which attributes/features will be used and document reasons for inclusion/exclusion.
* Clean data: Correct, impute and remove the improper data.
* Extract data: Derive new attributes from the existing ones
* Integrate data: Create features by combining data from multiple sources.
* Format data: Re-format data as necessary. For example, convert string values to numeric values so as to perform mathematical operations.

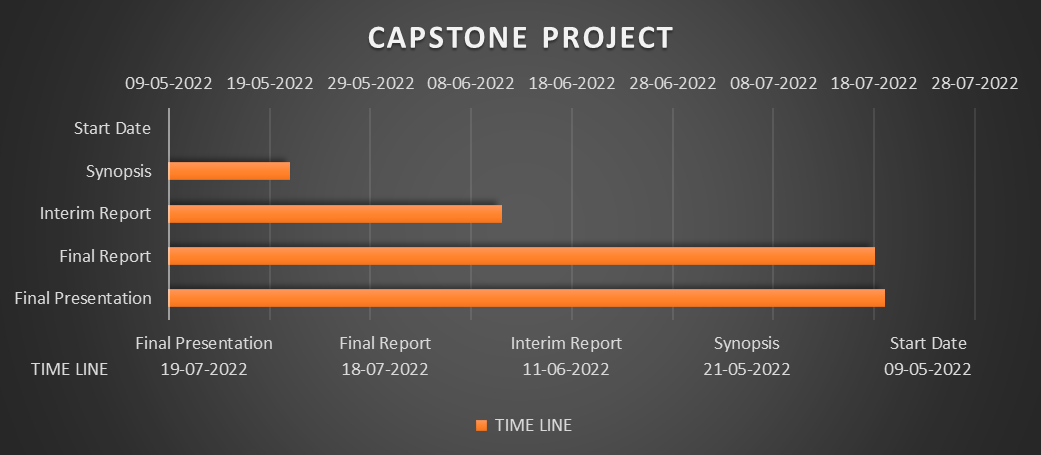
**4. Modeling -** In this stage we build and assess different models built using various techniques from the training dataset.

* Select modelling technique: Determine the algorithms to be used to model the data based on the business requirement.
* Generate test design: In order to build and test the model, we need to divide the dataset into training and testing data set. In this step we divide the data into train and test data set.
* Build model: Based on the modelling technique selected, build the model on the input data set.
* Assess model: Compare the results of different models based on confusion matrix. The outcome of this step frequently leads to model tuning iterations until the best model is found.

**5. Evaluation -** Evaluate the models and review the steps executed to construct the model to be certain it properly achieves the business objectives.

* Evaluate results: Understand the data mining results and check how impactful they are in achieving the data mining goal. Select appropriate model based on confusion matrix.
* Review process: Review the work accomplished and make sure that nothing was overlooked and all steps were properly executed. Summarize the findings and correct anything if needed.
* Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

**Timeline Chart (Weekly Plan):**

****

**References (Dataset Source/Journals/Articles):**

* <https://www.kaggle.com/code/touba7/hotel-booking-visualization/data>
* <https://pbiecek.github.io/xai_stories/story-hotel-booking.html>
* <http://bright-journal.org/Journal/index.php/JADS/article/view/20/7>
* <https://opus.govst.edu/cgi/viewcontent.cgi?article=1199&context=capstones>
* <https://www.researchgate.net/publication/350699766_Performance_Analysis_of_Machine_Learning_Techniques_to_Predict_Hotel_booking_Cancellations_in_Hospitality_Industry>
* Antonio, N, et al. 2019. An Automated Machine Learning Based Decision Support System to Predict Hotel Booking Cancellations. Data Science Journal, 18: 32, pp. 1–20. DOI: <https://doi.org/10.5334/dsj-2019-032>

**Declaration:** This is to declare that the dataset that we are using for our capstone project does not have any relevant legality associated to it and can be used to showcase the work we do on it as a presentation in Great Learning.