Analytics Startup Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

Project	*Hotel Booking Cancellation Prediction*
Requestor	*Centennial College* *Bilal Hasanzadah*
Date of Request	*August 15, 2024*
Target Quarter for Delivery	*Q4 2024*
Epic Link(s)	• *https://www.kaggle.com/datasets/thedevastator/hotel-bookings-analysis/data* • https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/ • https://revenue-hub.com/three-most-common-trends-impacting-cancellation-rates/#:~:text=2022%20state%20of%20hotel%20cancellations,seeing%20cancellations%20occur%20the%20most.
Business Impact	*This model will help hotel management optimize their operations by anticipating potential booking cancellation what will improve their operational efficiency and revenue optimization.*

1.0 Business Opportunity Brief

Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment.

Business Problem Statement

High cancellation rates can disrupt inventory management, staffing schedules, and financial forecasting, ultimately impacting the revenue. Understanding the factors contributing to booking cancellations and predicting which bookings are likely to be cancelled is crucial for hotel managers to implement proactive measures to mitigate these issues.

Opportunity

This project presents a valuable opportunity to leverage data analytics and machine learning to predict hotel booking cancellations. By developing an accurate predictive model, the hotel can gain insights into the key drivers of cancellations, enabling targeted interventions and strategies to optimize revenue management, adjust staffing, and create customer retention strategies.

1.1 Supporting Insights

Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally?

- Nearly 40% of hotel bookings were cancelled before arrival in 2018 in Europe.
- Bookings with lead times longer than 60 days have a 65% higher likelihood of being canceled.
- In 2018, Online Travel Agencies (OTA) accounted for 71% of the online distribution for independent hotels in Europe.
- Couples have a higher rate of 66.7% compared to 10% for family bookings.

1.2 Project Gains

Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI's.

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.

- By anticipating cancellations, the hotel can optimize its pricing and inventory management strategies, ensuring that rooms are sold to committed guests and reducing the financial impact of cancellations.
- Identifying customers likely to cancel allows the hotel to engage with then proactively through personalized offers or incentives, improving customer satisfaction and loyalty.
- Predicting cancellations enables better planning for staffing, housekeeping, and other operational resources, leading to cost savings and improved service quality.
- Data-driven insights into booking patterns and cancellation trends help hotel managers make informed decisions about marketing strategies, promotions, and customer engagement initiatives.

2.0 Analytics Objective

List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.

Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed.

Key questions:

- Which factors are the most significant predictors of booking cancellations?
- How accurately can we predict cancellations based on historical data?
- What strategies can be implemented to mitigate the risk of cancellations?
- Are all the columns important to predict hotel cancellations? If not, which columns could we drop?
- What is the relationship between lead time and cancellation rates?
- How do different customer demographics influence cancellation rates?
- What is the impact of the distribution channel (i.e. travel agency, or direct) on the likelihood of cancellation?
- How do previous booking behaviors (e.g., previous_cancellations, is_repeated_guest, booking_changes) affect the probability of cancellation?
- What role does the type of hotel (city hotel vs. resort hotel) play in the cancellation rates?
- What is the role of average daily rate in the cancellation rate?
- Are there specific periods (e.g., holidays or weekends) with higher cancellation rates?

Assumptions:

- TBD

Hypothesis:

- Bookings with longer lead times (more than 30 days) are more likely to be canceled compared to bookings made closer to the arrival date.
- Bookings made through online travel agencies (OTAs) may have a higher likelihood of cancellation compared to direct bookings.
- Customers with a history of cancellation or multiple changes are more likely to cancel their current bookings.

- City hotels have a higher cancellation rate.
- Groups without children are more likely to cancel their bookings.
- Higher average daily rate bookings are less likely to be canceled.

2.1 Other related questions and Assumptions:

List any assumptions that may affect the analysis

- How do external factors (e.g., economic conditions, weather events) impact booking cancellations?
- What are the financial implications of booking cancellations?
- Can customer engagement and communication strategies reduce cancellations?

2.2 Success measures/metrics

What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. *These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.*

Business impact metrics

- **Reduction in cancellation rates:** measure the decrease in cancellation rates after implementing the model.
- **Revenue increases:** compare the revenue before and after model implementation.
- **Booking retention rates:** calculate the percentage of bookings retained after implementing strategies based on model predictions.
- **Customer satisfaction:** analyze customer feedback and reviews after implement the measures taken based on model predictions.

Model performance metrics

- Model accuracy
- Precision: ratio of true positive predictions to the total of positive predictions. This helps to minimize false positives.
- Confusion Matrix

2.3 Methodology and Approach

Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.

Type of Analysis: logistic regression, decision tree, random forest.

Methodology:

We will start by identifying all hotel bookings recorded in the dataset, focusing on their respective cancellation statuses. The response variable will be defined as 1 if the booking was cancelled and 0 otherwise.

First, we will start with an exploratory data analysis (EDA) to understand the distributions, trends, and relationships within the dataset. This includes identifying missing values, outliers, and any data inconsistencies. Next, we will perform feature engineering including normalization, scaling, or transformation if necessary.

After preprocessing the data, we will split the dataset intro training and test sets to ensure the robust model evaluation. We will build several machine learning models, including decision tree, logistic regression, and random forest, to predict the likelihood of booking cancellations. We will evaluate the performance of the model focusing on metrics such as accuracy and precision.

Output: the output will be a set of insights, rules, and strategic recommendations to help the hotel management predict cancellations and take proactive measures.

3.0 Population, Variable Selection, considerations

Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today.

Audience/population selection: the audience for this analysis is the hotel industry and the population includes both city hotels and resort hotels. The dataset contains 119,390 records and 33 variables before preprocessing process.

Observation window: the observation window includes records from 2015 to 2017.

Inclusions: all relevant features that could potentially influence booking cancellation will be included in the analysis. These features include:

Column name	Description
hotel	Indicates the type of hotel (resort or city)
is_canceled	Indicates whether the boking was cancelled or not (0=not cancelled, 1=cancelled)
lead_time	Represents the number of days between the booking date and the arrival date
arrival_date	Denotes the arrival date. **This variable is not part of the original dataset; however, it is the result of aggregate the columns:** arrival_date_year, arrival_date_day_of_month, arrival_date_day_of_month
stays_in_weekend_nights	Indicates how many nights (Saturday or Sunday) guests stayed or booked to stay at a hotel during weekends

stays_in_week_nights	Represents how many weeknights (Monday to Friday) guests stayed or booked to stay at a hotel during weekdays.
adults	Indicates the number of adults included in each booking.
children	Indicates the number of children included in each booking.
babies	Indicates the number of babies included in each booking.
meal	Describes what type of meal was booked (Breakfast only, Half board, Full board, or Undefined/SC – no meal package)
country	Denotes the country-of-origin for each guest who made a reservation.
market_segment	Shows various market segments that individuals belong to when making reservations (e.g., Online Travel Agents, Offline Travel Agents, Corporate clients).
distribution_channel	Specifies different channels through which bookings were made (e.g., online travel agencies, direct bookings with hotels/tour operators, corporate arrangements).
is_repeated_guest	Indicates whether the guest is a repeated visitor (0=not repeated guest, 1=repeated guest).
previous_cancellations	Represents the number of times guests previously canceled their bookings.
previous_bookings_not_canceled	Denotes the count of previous bookings made by guests that were not canceled.
booking_changes	Represents the number of changes made to the booking.
deposit_type	Indicates the type of deposit made for the booking.
days_in_waiting_list	Represents the number of days the booking was on the waiting list before being confirmed.
customer_type	Indicates the type of customer (e.g., transient, contract, group, or other).
Average daily rate	Represents the average daily rate (price per room) for the booking.
required_car_parking_spaces	Indicates the number of car parking spaces required by the guest.
total_of_special_requests	Represents the total number of special requests made by the guest (e.g., extra bed, room amenities).
reservation_status_date	Represents the date on which the reservation status was last updated.

Exclusions: irrelevant or redundant features will be excluded from the analysis. These might include:

Column name	Description	Reason to exclude
Index	Identifier for each row	Irrelevant
agent	Represents the ID of the travel agency that made the booking.	Specific. Not information about its values
company	Represents the ID of the company that made the booking.	Specific. Not information about its values.
arrival_date_year	Denotes the year of the arrival date	Redundant variables with the new created variable called arrival_date.
arrival_date_month	Indicates the month of the arrival date	
arrival_date_day_of_month	Represents a specific day of arrival within a month	
arrival_date_week_number	Specifies the week number in which guests arrived at the hotel	Redundant, with the variable arrival_date is possible to now the week number.
reservation_status	Indicates the status of the reservation (e.g., canceled, checked-in, no-show).	Redundant with our target variable. Canceled and No-Show are counted as 1 (Cancelled) and Check-in is counted as 0 (Not Cancelled)
reserved_room_type	Identifies the type of room initially reserved.	Code of room type reserved, values from A to P. Due to anonymity reason, we do not have the meaning of each letter.
assigned_room_type	Identifies the type of room that was assigned to guests.	Code of room type reserved, values from A to P. Due to anonymity reason, we do not have the meaning of each letter.

Data Sources: *https://www.kaggle.com/datasets/thedevastator/hotel-bookings-analysis/data*

Audience Level: hotel industry

Variable Selection: all the variables mentioned above in the inclusion section.

Derived Variables:

- **total_guests:** sum of adult, children and babies per booking.
- **arrival_date:** aggregate the variables arrival_date_year, arrival_date_month, and arrival_date_day_of_month.

Assumptions and data limitations:

- The impact of external factors such as economic conditions or travel restrictions are not included in the model but could affect booking patterns and cancellations.

4.0 Dependencies and Risks

Identification of key factors that may influence the outcome of the project and likelihood of it happening:

Risk	Likelihood (based on historical data)	Delay (based on historical data)	Impact
Class Imbalance	*High*	*Low*	*Imbalance classes (cancellations vs. non-cancellations) can lead a model that is biased towards the majority class.*
Feature selection and engineering	*Medium*	*Medium*	*Incorrect feature selection or engineering may lead to poor model performance.*
Changes in booking patterns	*Low*	*Low*	*Sudden changes in bookings behaviors due to external events (e.g., pandemics, economic shifts) may affect the model performance.*

5.0 Deliverable Timelines

List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process.

Item	Major Events / Milestones	Description	Scope	Days	Date
1.	Assessment / Triage	Initial meeting with the advisor.	Define problem statement, business opportunity. Analytics Plan is documented.	5	07/15/2024
2.	Data Exploration & Analysis • Issues with duplicates • Issues with	Exploring and preprocessing data.	Conduct EDA to understand data distribution, identifying trends, and relationships between variables. Visualize key features using	5	07/18/2024

	Spend data		boxplots, scatterplots, heat map, and others.		
3.	Modeling	Develop and evaluation of predictive models.	Develop logistic regression, decision tree, and random forest to forecast booking cancellations. Perform cross-validation and hyperparameter tuning.	15	07/31/2024
4.	Governance and documentation	Ensuring proper documentation and compliance with data governance.	Prepare detailed documentation for the project, including data sources, preprocessing steps, model development, and evaluation processes. Creäte paper and presentation.	5	08/05/2024
5.	Final Presentation	Stakeholder presentation.	Conduct a final presentation that showcase the development and result of the project.	5	08/15/2024
6.	Portfolio	Update of the portfolio.	Include all the project deliverables into the portfolio.	2	08/16/2024