Using API with Logistic Regression Model to Predict Hotel Reservation Cancellation by Detecting the Cancellation Factors

Sultan Almotiri[1], Nouf Alosaimi[2], Bayan Abdullah[3],
College of Computer and Information Systems
Umm Al-Qura Univeristy

***Abstract*—The aim of establishing hotels is to provide a service activity to its customers with the aim of making a profit. So, for that the cancellations are a key perspective of inn income administration since their effectiveness on room reservation systems. Cancelling the reservation eliminates the outcome. Many expected factors affect this problem. By knowing these factors, the hotel management can make a suitable cancellation policy. This project aims to create an API that can provide a function to predict if a reservation is most likely to cancel or not. That API can integrate with the hotel management systems to evaluate each reservation process with the same parameters. To do this, the study starts by defining the factors using Chi test, correlation to find the effective variables, and coefficient of the variables in the linear regression. And the results that have been found for the factors are: is_repeated_guest, previous_cancellations, previous_bookings_not_cancelled, required_car_parking_spaces, and deposit_type. For API function, the intercept and coefficients have been used from the logistics regression model to create a scoring function. Scoring function can be calculated by the sum of the factors multiplied by their coefficients in addition to the intercept. This score is to be evaluated as a probability later using the logistic function.**

***Keywords*—Prediction; API; factors; logistics regression**

I. Introduction

The tourism sector in any region is based on a several of constituents and pillars, among them are the availability of tourist facilities and number and quality of services provided by those facilities. Perhaps the most prominent of these facilities, and the most important, is the hotel sector. In numerous tourism goals, the inn industry is regularly the major financial sector [1]. Hotels are defined as those that provide places for accommodation, meals, and other services for visitors in general. Hence, the presence of hotels in countries has become an indispensable thing.Committing room numbers to the diverse reservations could be a day by day action in inns [2]. In hotels, the revenue Management employments data-driven modelling and optimization strategies to choose What, when, whom to offer, and for which cost, in arrange to extend revenue and benefit [3].

The hotel industry is considered as a basic pillar of tourism because it provides services in the field of economy, because of the means of transportation it pumps and because it is a means of obtaining aid for comprehensive development plans in the world. The aim of establishing hotels is to provide a service activity to its customers with the aim of making a profit. Consequently, the cancellations are a key perspective of inn income administration since their effectiveness on room reservation systems. By means of the e-commerce increases the chance of cancellation since client can effortlessly compare conditions among lodgings and can cancel his/her reservation [4].

Moreover, Lodging cancellation can cause numerous issues on the hospitality industry in numerous perspectives such as affecting hotel's reputation, since clients can be influenced by others' criticism about the inn [5], [6]. A lodging booking cancellation arrangement can depend on a few variables, such as the rate of the booking and the date of check-in. Cancelled by certain date approach, this type of cancellation approach gives travelers the choice to cancel inn reservations free of charge up until a certain date. Once this date passes, inn cancellation approaches can either: Charge a standard cancellation fee require the complete installment for the reservation. One-night penalty policy, for this sort of lodging room cancellation arrangement, inn charges a cancellation expense proportionate to one night's remain at the inn. The prepaid, nonrefundable hotel reservation policy, this sort of inn booking requires installment for the aggregate of the reservation at the time of booking and is nonrefundable. In truth, exceptionally small is known almost the reasons that lead clients to cancel, or how it can be maintained a strategic distance from.

It is more than common in these days to employ computer programs and services to support management and decision taking. The contemporary algorithms and artificial intelligence (AI) techniques encouraged the involvement of automated decision support systems in every business-oriented information system, especially ERP programs, to analyze problems, detect factors, propose solutions, and help in improving or building new strategies using data-driven approaches.

One of the applications of these services is hotel management application. It started as invoices application which developed to include keeping the reservation information. Later, it added automated reservation services. After many reservations, the data can be analyzed to detect reservation problems' factors and expect results.

Cancelling reservation is one of the most common problems in hotel reservation services. Many expected factors affect this problem like the reservation location, the hotel rate, the price offer, and most importantly, the history of the client and his solemnity. The modern techniques can elicit the question of what these factors are based on data-driven analysis and data mining in the reservation records.

The current state of hotel management and reservation systems is interrelated and cooperative. Almost all reservation systems provide API for the third-party reservation systems which promote the offers and spread the popularity of the hotel. This means we have three main stakeholders in the reservation process: the hotel management, the external reservation promotional, and the client. Although the process is automated, it can be improved by implementing data mining and AI techniques to solve the major problems. We still do not have a concrete API that can evaluate each reservation process to control that process accordingly in the matter of price, payments, and priority.

This study aim to provide a technique that can detect the reasons behind reservation cancelling. These reasons are to be evaluated and weighted where they can be used later to evaluate reservations with similar parameter values. This technique is to be shaped in a ready to implement model that specify the inputs, processes, and output of factor detecting processes. The outputs of this technique is a weighted factor rule that will be added to the factors knowledge base to be checked in later reservations.

A. The Motivation

The importance of the study is derived from its function. Reservation imply the fact that the asset will not be usable until that reservation is over which prevent any exploit of that asset. This is acceptable as the expected outcome of that reservation is granted. However, cancelling the reservation prevents the usability of that asset and eliminate the outcome. In hotels, cancelling reservation have a higher impact as it eliminates the main purpose of the business which is renting the rooms. While there are many consequences for reservation canceling, it is better to avoid that cancel in the first place. We can avoid that cancel, or at least measure the risk of cancellation, by knowing the effective factors that results in cancellation. As knowing hotel cancellation factors can be imperative to make strides overbooking and cancellation approaches, which are two exceptionally critical themes in RM investigate [7].

B. The Contribution

This study aim to create an API that can provide a function to predict if a reservation is most likely to cancel or not. To do this, the study starts to define the factors. And the results that have been found of the factors are is repeated guest, previous cancellations, previous bookings not cancelled, required car parking spaces,and deposit type. The model using these factors show precision of 0.86 while the one that uses all the variables shows an average precision of 0.82.

C. The Paper Structure

First, the paper starts with introduction that introduce the problem and the background. Then, related work that present the current work. After that, the methodology followed by the results discussion. Finally, the conclusion for the paper.

II. Background and Related Work

A. Hotel Cancellation Policies

As is known, all the hotels establish a price and conditions or policies for booking. Some of lodgings make the cancellation access tough by limiting the free cancellation windows and/or by setting higher cancellation punishments. Some hotels put some cancellation fees and consider it as an other revenue [8]. In expansion, cancellation arrangements are outlined to influence travelers' booking behaviors in a way that's more alluring or profitable to inns. Furthermore, some hotels use debit or credit card which tends to decrease the possibility of no appears [9].

There are many indications that the searchers for a hotel continues to search for a suitable hotel even after they have booked a hotel in order to search for another hotel that has better policies and a lower price. In the event that a hotel is found that has suitable policies and has a lower price, the researcher cancels the reservation in the previous hotel and rebook the new one. Cancellation approaches exist for the reason to avoid such "switching" behavior [10].

In [11] inspected the effect of diverse sorts of cancellation confinements on lodging guests' reservation choices. Particularly, the consider explored how cancellation due dates and costs impact the vital booking behavior of deal-seeking travelers, advanced-booking. They found out the cancellation due date influenced participants' behavior whereas the estimate of the cancellation expense had no measurably critical affect. In expansion, there was no noteworthy difference between no cancellation approach and cancellation due date.

This considered in [12] that gives experimental prove on how lodging cancellation approaches are changing in later a long time. The discoveries demonstrate that whereas lodgings are testing with stricter cancellation windows, their cancellation punishments don't appear to end up stricter. There discoveries indicate that in spite of the fact that US inns have as of late fixed their free cancellation windows, there's no clear sign that the cancellation punishments have expanded. They found that together with the tightening of cancellation windows, there appears to be a move toward "standardizing" the cancellation punishments.

Indicated a cancellation policy means choosing the terms and conditions that allow booking to be cancelled, and penalization to the clients who cancel a booking or don't appear up at the concurred date. Such penalization is known to have a non-negligible part in income administration of inn chains. In [13] stated that inns that set extensive penalties for cancellation, especially with it gets closer to the arranged entry date point to play down revenue losses. For occurrence, it has become necessary to request reservations with credit card sponsorship, so that in the event that the customer does not show up, a fee for at least one night will be charged.

Lodging cancellation approaches not only affect Income but it also influences consumer behaviors. This consider [14] confirms the fact that big cities have more Free Cancellation Approaches. The reason is twofold. One reason is due to Competition within the markets in these cities. The other reason dues to the fact that small towns have few high-end hotels Which has permissive policies.

The discovery [14] shows that high-end inns tend to have less-strict approaches because maintaining great relationships with their customers is one of the goals of most high-end hotels. A later industry report demonstrates that the cancellation arrangement is conditional based on diverse variables

such as, the larger part of the sort of clients, the area, and the economy situation. For example, amid the subsidence, the tradition lodgings have tighter cancellation arrangements.

Utilizing the proposition-based theorizing within the setting of cancellation approaches, this study provides a few suggestions that seem have wide suggestions for future investigate. In [15], they introduce a theory about hotel cancellation policies that can Future research examine the effect of these policies. One of the policies that they introduce if the tightening cancellation policy might flag solid demand because the inn anticipates to exchange the room effectively notwithstanding of the restrictive policy or a strict cancellation approach might flag frail request because the lodging does not anticipate to exchange the room, and so endeavors to recoup its losses.

B. Predict Hotel Cancellation using Machine Learning

Inn income administration employments machine learning. The lodging industry endures 20% of income due to cancellation. In Addition, the administration cannot gadget strict arrangements of non-cancellation, as the guests/customers might switch to another inn. The machine learning-based estimating employments a neural network [16].

In [17] the authors proposed a forecasting model that forecast the cancellation in hotels. The forecasting model that they have been used was based on artificial intelligence by using personal name records (PNR). This investigate has been created utilizing genuine booking records provided by a lodging accomplice found in Gran Canaria (Spain) with the point of forecasting future cancellations.There approach for forecasting the cancellation in hotels was using 13 independent variables which is a decreased number in comparison with related investigate. For the result, they used three algorithms which were Random forest with 80% of accuracy, Support vector Machine with 75% and GA for ANN with 79%.

As the cancellation effects the reservation system and the income. In [18] they used a Personal Name Records (PNR) for predict hotel cancellation. Their approach differ from other research by aiming to distinguish those people likely to create cancellations in a short-horizon of time through PNR. Their promising comes about have been accomplished with 80% accuracy for cancellations made 7 days in progress. They used a PNR data from inn found in Gran Canaria (Spain). There method gave accuracy 73% for C5.0, SVM 71% ANN 69% Boosting ensemble 80%. This approach endeavors to forecast individual cancellations likely to be made exceptionally near to the section day, from 4 to 7 days in progress, and which can be considered “critical cancellations”.

It is conceivable to construct models to anticipate bookings cancellation probability. These models has never been evaluated in a real environment. Consequently, in this paper [19] the author designed a prototype of the prediction model and transferred it to two hotels. First, they developed a trained model that learns from previous reservations to see a pattern of cancellations over time. From a commercial angle, the model demonstrated its applicability, being above 84% in accuracy, 82% in precision. The framework allowed the residences to predict their net demand, thus making better options for which reservations should be acknowledged and rejected, what construction costs would be, and how many rooms to sell.

As well as, reservation cancellations within the hospitality industry not as it were creating income misfortune and influence estimating and inventory allotment choices. Moreover, in overbooking circumstances, have the potential to influence the hotel’s online social notoriety. In [20] authors used data from four resort hotel Property Management Systems (PMS) to predict hotel cancellation. The creators reach the accuracy of 90%.

Furthermore, to overcome the negative affect caused by overbooking and the usage of unbending cancellation approaches to manage with booking cancellations. This paper [21] the author used a real data from four resort hotels of the Algarve, Portugal. The author used Boosted Decision Tree, Decision Forest, Decision Jungle, Locally Deep Support Vector Machine, and Neural Network models to predict hotels reservations cancellation. the author got reached accuracy values above 90%, while models of H2 and H3 reaching 98.6% and 97.4%, respectively.

The online booking system in hotels is one of the most important alluring arrangements within the hospitality industry. Cancellation of inn bookings or reservations through the online framework is as of now one of the issues within the inn administration framework. In [22] the authors used a data set from the paper ”Hotel Booking Demand Datasets”. That been used XGBoost, Catboost, Light Gradient Boosting Machine, and Random Forest. By using CRISP-DM framework, they got about 0.8725 of accuracy in random forest.

Supervised anomaly detection concept is using to predict hotel cancellation. In [23] the authors used a few classification models such as Decision Tree, Gradient Boost, and XGBoost. Moreover, they reached high accuracy which is 86%. Their result appears that in their circumstance, the decision tree calculation, utilizing data sets with the accurately characterized properties, may be an incredible procedure for making prescient models for booking cancellations. These discoveries too affirm Chiang (2007) articulation ”as modern trade models keep on rising, the ancient estimating strategies that worked well some time recently may not work well within the future. Confronting these challenges, analysts ought to proceed to create unused and superior estimating methods”.

In this paper [24], they indicated the factors that influence the cancellation behavior. The factors that they found out are lead time, country, and season. The timing of booking is vital, with early bookings showing an altogether higher cancellation likelihood. Moreover, they found out that those who booked through agencies are more likely to not cancel their cancellation while those who booked through online or offline are more likely to cancel. The lead time and country effect the reservations which made through online more than offline. The cancellation probability indicated using cluster adjusted standard errors.

In [25] they use the data from eight hotels with few data from other sources (weather, holidays, events, social reputation, and online prices/inventory). Also, they developed a model to predict reservation cancellation in hotels. Moreover, they indicated cancellation drivers. For the drivers, they found out the most important features are country, deposit type, adults, Stays at Weekend Nights, Distribution Channel and Agent. They got 86% of accuracy for XGBoost model.

We still do not have a concrete API that can evaluate each reservation process to control that process accordingly in the matter of price, payments, and priority. This study aim to provide a technique that can detect the reasons behind reservation cancelling. These reasons are to be evaluated and weighted where they can be used later to evaluate reservations with similar parameter values. This technique is to be shaped in a ready to implement model that specify the inputs, processes, and output of factor detecting processes. The outputs of this technique is a weighted factor rule that will be added to the factors knowledge base to be checked in later reservations. In this study, the factors that influence the cancellation have been found are: is repeated guest, previous cancellations, previous bookings not cancelled, required car parking spaces,and deposit type. These factors evaluated using logistics regression. Later, these factors used in API function. Table I indicates the methodology and the results of related work.

C. Logistics Regression

Logistics regression is a statistical model utilizes to Logistic equation conditional probability. In addition, it uses to model a binary dependent variable. Logistic regression is used to predict the likelihood of an event occurring with additional knowledge of the values of variables that can be explained or for that event. This modeling is widely used in commercial and commercial transactions and is one of the modeling methods most applied in the field of learning, as it is classified within the methods of supervised learning.

In order to know the effectiveness of internal control systems to increase the revenue of hotels [26].As to know the relationship between two variable, they used logistics regression. that been found out some internal control components have positive effect on revenue in hotels such as controlling activity, information, and communication and observing action have positive and critical impact in determining the results. Moreover, checking action contains a more prominent impact on the outcomes than data communication and control action.

Moreover, to know how auxiliary and organizational components impact hotel's probability of creating service/product, process, organizational and promoting advancements. In [27], the authors used responses from 174 hotels Chi-square test was utilized to test the primary theory. An arrangement of different logistic regression investigations was performed to test the connections between the independent variables. This paper gives bits of knowledge almost the nature and degree of advancements within the inn sector. In spite of the fact that generally considered inflexible and non-innovative, around half of the responding lodgings made at slightest one sort of improvement. Most common are service/product and displaying advancements. A hotel's probability of moving forward depends to a incredible degree on fundamental flexibility (non-chain), having an unequivocal advancement strategy.

TABLE I. SUMMARY OF RELATED WORK

Citation	Methodology	Result
[17]	SVM and random forest	predict hotel cancellation with accuracy 80%
[18]	SVM, ANN,and Boosting ensemble	forecast individual cancellations likely to be made exceptionally near to the section day, from 4 day to 7 days in progress
[24]	probit model with cluster adjusted standard errors	The factors that they found out are lead time, country, and season
[19]	they designed a prototype of a prediction model and conveyed in with two hotels	the model illustrated its viability, with comes about surpassing 84% in accuracy
[25]	they indicated cancellation drivers. For the drivers, they found out the most important features are country, deposit type, adults,Stays At Weekend Nights,Distribution Channel and Agent	They got 86% of accuracy for XGBoost model
[20]	they used data from four resort hotel Property Management Systems (PMS) to predict hotel cancellation	it is possible to construct models for anticipating booking cancellations with exactness comes about in accuracy of 90%
[21]	they used Boosted Decision Tree, Decision Forest, Decision Jungle, Locally Deep Support Vector Machine, and Neural Network models	they got reached accuracy values above 90%, while models of H2 and H3 reaching 98.6% and 97.4%, respectively
[22]	They used XGBoost,Catboost,Light Gradient Boosting Machine, and Random Forest	they got about 0.8725 of accuracy in random forest
[23]	they used a few classification models such as Decision Tree, Gradient Boost, and XGBoost	they reached high accuracy which is 86%

III. DATA SET

The dataset used for this project is downloaded from Kaggle website. it contains a variety of hotel reservations information collected based on some potential factors. This data set contains booking information for city and resort hotels. These information such as the date of the reservation, length of stay, number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data.

The dataset contains 32 columns described in their original project[1] as the following:

- hotel: Hotel (H1 = Resort Hotel or H2 = City Hotel)
- is_cancelled: Value indicating if the booking was cancelled (1) or not (0)
- lead_time: Number of days that between the entering

[1] https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md

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_weeknights: 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_cancelled: 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: Cancelled – booking was cancelled 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 Reservation Status to understand when the booking was cancelled or when did the customer checked-out of the hotel.

The second column is the dependent variable in this study, while the other columns represents the independent variables as potential factors.

A. Data Wrangling

Start with 31 proposed variables.These included timestamps, details, and status data. For the timestamps, It can be decided what kind of time variables that should be accepted by understanding the logical scope of the problem. In other words, how would time affect the hotel reservation? and what timestamp are we talking about? If that has been considered, there might be some yearly routine that during some period of the year there would be many cancellations, then it does not need to know which year exactly that has been talking about. In addition, this cannot be observed on daily bases. It is convenient to consider the period on weekly and monthly basis. For this, other timestamps can be drooped.

Now, by talking about dates, considering the arriving date in the reservation, not the reservation date, leaving date, nor status date. The other timestamps might be eliminated too. The details of the reservation include information about the hotel, agency, travelers, reservation type, average daily rate, and meals. Any variable considering this reservation is considered as a potential factor. The status is what we are trying to predict and analyze to find its factor, namely, the calculation status. For this, we can consider all other status as not cancelled. We do not need more details about the status. Based on the criteria above, we will drop the following columns:

- lead_time: Unneeded timestamp
- arrival_date_year: Unneeded timestamp
- arrival_date_day_of_month: Unneeded timestamp
- reservation_status: Unneeded status details
- reservation_status_date: Unneeded timestamp

There are missing data in the children field, country field, agent field, and company field. From the data description in the introduction it can be found that it might be gotten the types that needed from the market segment, distribution channel, and customer type and disregard the agent and company columns. So, these two are just dropped. For the children field and country field, it might been filling the missed data with 0 and 'Not' as mark that it was not inserted.

After cleaning, the dataset contains 33,775 duplicates which is more 30% of the dataset. However, these duplicates are not real duplicates. They might be different in term of year, day, or other removed fields. For this reason, no duplicates might be eliminated.

IV. Methodology

In order to answer the question of this study, the study starting by making an exploratory data analysis . Then, creating a list of the most effective factors. Finally, evaluating what the study has done by implementing prediction model based on the factors that will be found. The EDA includes having an insight of the statistics of the dataset. This insight will help in detecting if any bias occurs. After that the study will evaluate each factor using chi-square test of independence. To create the list of the effective factors, it will calculate the correlation coefficient between all the variables to understand the relationship between the variables. This process might propose eliminating some unrelated variables. Finally, the remaining factors will be used in creating prediction models, and compare these prediction models with same models which included all the eliminated data. Fig. 1 shows the stages for the methodology.

Fig. 1. Stages.

A. Data Analysis

The dataset has total bookings cancelled about 44,224 (37%). Bookings cancelled in resort hotel are 11,122 (28%). While bookings cancelled in a City hotel are 33,102 (42%). We can see the cancellation in the dataset more than 60% as Fig. 2 shows:

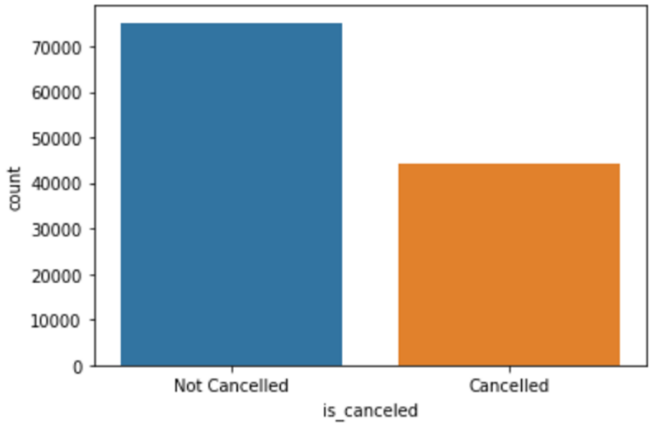

Fig. 2. Bookings Cancellation.

The question of this study is what are the main factors of hotel reservation cancelling? From this question it can be hypothetically derived a sub question for each of the other parameters. However, it can be disregarded some parameters logically or practically during data wrangling. To investigate the dataset statistics,the study starts with exploring the numerical variables. Starting with the numerical variables, the following Fig. 3 the histogram that has been found: There

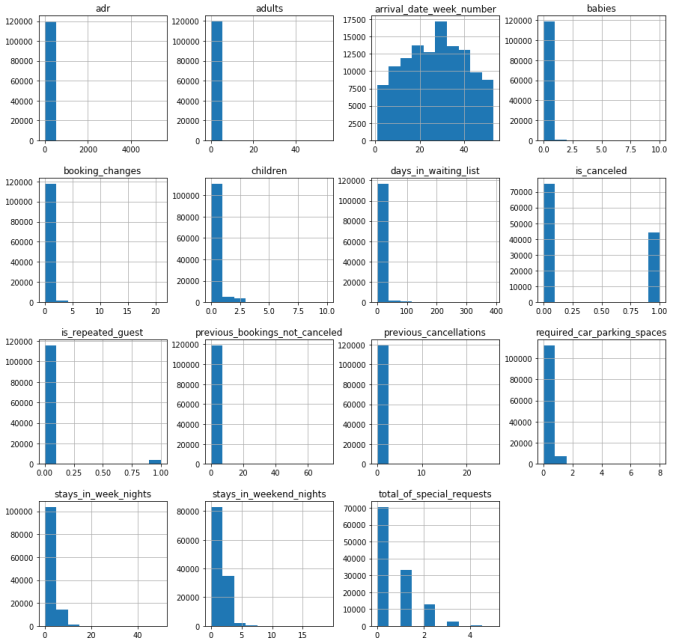

Fig. 3. Histogram of the Variable.

are 15 numerical variables including two of them are Boolean but treated as numerical which are: is_cancelled variable and

is_repeated_guest variable. Numerically, it seems that there are a lot of outliers. Logically, it is understandable that most of the reservation would have adults less than ten for example. it would not consider a reservation for group of 50 is an outlier. This also applies to the average daily rate, babies, and car parking spaces. However, for fields like the booking_changes, previous_booking_not_cancelled, previous_cancellations, stays_in_weeknights, and stays_in_weekend_nights, the concept is the important but not the quantity. This means we might look on the impact of changing booking on the reservation cancellation. This impact would be effective from the first and second change. No need for to look at those of 20 changes. It may combine all more than 5 changes as 5. In other words, In other words, the study will decrease the values more than the outlier threshold to the outlier threshold values in these fields. So, by calculating the Outliers based on quartile which is:

interquartile Range

IQR = Q3 - Q1

threshold range

threshold_range = IQR$\times$1.5.

It can be seen from Table II the 1.5-based outlier threshold for most of the field is 0. This makes infeasible to use it. According to these results, it can be only used the threshold for stays_in_weekend_nights and stays_in_weeknights fields. For these two fields, we will reduce all outliers to the threshold.

TABLE II. CALCULATING OUTLIERS

Variables	Q1	Q3	IQR	Threshold range
arrival_date_week_number	16.00	38.0	22.00	33.000
stays_in_weekend_nights	0.00	2.0	2.00	3.000
stays_in_weeknights	1.00	3.0	2.00	3.000
adults	2.00	2.0	0.00	0.000
children	0.00	0.0	0.00	0.000
babies	0.00	0.0	0.00	0.000
is_repeated_guest	0.00	0.0	0.00	0.000
previous_cancellations	0.00	0.0	0.00	0.000
previous_bookings_not_cancelled	0.00	0.0	0.00	0.000
booking_changes	0.00	0.0	0.00	0.000
days_in_waiting_list	0.00	0.0	0.00	0.000
ADR	69.29	126.0	56.71	85.065
required_car_parking_spaces	0.00	0.0	0.00	0.000
total_of_special_requests	0.00	1.0	1.00	1.500

B. Factor Extraction

To extract the factors, the study uses chi-square test of independence. For nominated variables, correlation has been used to find the effective variables. After that another test has been added based on the coefficient of the variables in the linear regression.

C. Factors Evaluation

To evaluate the factors, two logistic regression models have been used. One to predict the cancellation based on all variables in the dataset. The other one predicts the cancellation based on the nominated factors. The factorizing process is considered efficient if the precision of the factors model revealed better performance than the all variables model.

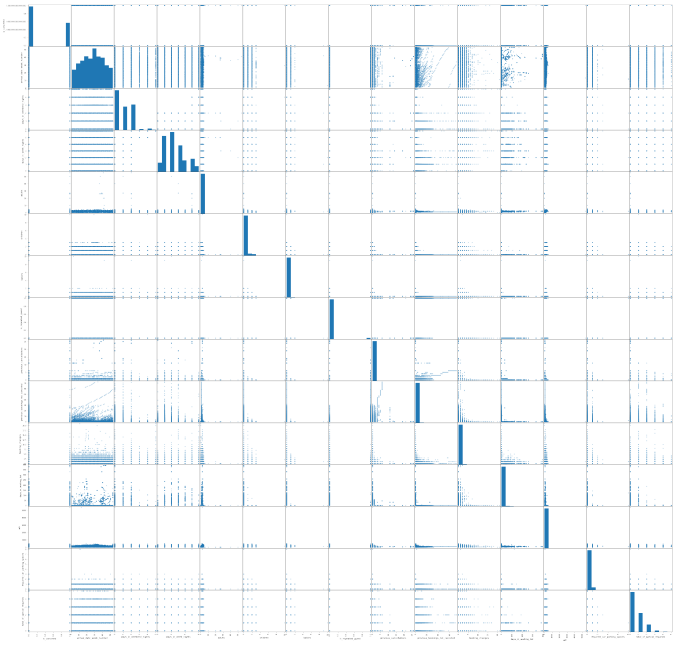

Fig. 4. Scatter Plot Matrix.

D. Scoring Function

by assuming that the cancellation process follows a linear equation, the intercept and coefficients have been used from the prediction model to create a scoring function that can calculate the sum of the factors multiplied by their coefficients in addition to the intercept. This score is to be evaluated as a probability later using the logistic function:

$$P(x) = \frac{e^x}{e^x + 1} \quad (1)$$

The result will give the probability of cancellation. If the probability is more than 0.5, the reservation process is more likely to cancel.

E. API Function

The API function is the last function that will use the scouring function to predict if a reservation would likely cancel or not. Its definition would be Boolean cancel (factorsArray). It will return true if the factors indicate probability of cancellations and false otherwise.

V. RESULTS DISCUSSION

For detecting the factors, it can be tried the scatter plot matrix which is used for seeing the connections between combinations of factors. Each scatter plot within the matrix visualizes the relationship between a combine of factors, permitting numerous relationships to be investigated in one chart. Therefore, in this project scatter plot matrix in Fig. 4 has been tried to have an insight about the relationship between the variables:

The scatter plot matrix initially suggests that there might be a relationship between the cancellation and adults' number, babies number. also, it seems that when there are previous cancellations, it is more likely it will cancel, while it might

not cancel with previous uncancelled reservations. There is also relationship with car parking and booking changes.

The study starts by initial evaluation for the factors using chi-square test of independence. The chi test is using for indicating if there is a relationship between tow variables based on the null hypothesis and alternative hypothesis. Null hypothesis assumes that there is no relationship between the two variables while the alternative hypothesis assume there is a relationship between the two variables. So, when the result from chi test less than alpha = 0.05, null hypothesis will be rejected and accept an alternative hypothesis. there are 10 categorical variables. To deal with them, they have been encoded into numerical values. The test shows the following results in Table III:

TABLE III. CHI TEST RESULT

Variable	Chi test result
0	2.78263008e-02
1	2.29104890e-02
2	1.06040302e-01
3	4.42317339e-01
4	3.20597577e-02
5	5.69348351e-02
6	2.22306546e-02
7	1.84643301e-02
8	1.68491108e-02
9	2.60087415e-02
10	1.45654133e-01
11	5.79706198e-02
12	4.51720898e-02
13	1.72514103e-02
14	1.79325050e-02
15	1.90153390e-02
16	4.20164259e-02
17	4.51832920e-02
18	5.28838342e-02
19	2.08257569e-02
20	4.67504818e-02
21	2.45671565e-02
22	2.36476367e+01
23	1.77620913e-02
24	2.77362630e-02

The variable number 1 is the 'is cancelled' field itself. In general, the test results indicate there is no significant dependency between arrival date month, arrival date week number, stays in week nights, country, market segment, booking changes, and ADR with is cancel variable.

Numerically, the study need to find the correlation between the numerical variables and the dependent variables. Correlation coefficient could be a factual degree of the quality of the relationship between the relative developments of two factors. In the following Table IV the correlation for numerical variable:

The correlation analysis reveals that there is significant correlation at alpha = 0.05 with the following columns: adults, is_repeated_guest, previous_cancellations, previous_bookings_not_cancelled, booking_changes, days_in_waiting_list, required_car_parking_spaces, and total_of_special_requests. At alpha = 0.01, stays_in_weeknights, babies, and ADR are added. The correlation results for the categorical variables in the following Table V:

Based on these results, we will consider the following variables as factors of reservation cancelling: adults, is repeated guest, previous cancellations, previous bookings not cancelled, booking changes, days in waiting list, required car parking spaces, total of special requests, hotel, country, market segment, distribution channel, assigned room type, deposit type, customer type.

TABLE IV. CORRELATION FOR NUMERICAL VARIABLES

Variables	Result
arrival_date_week_number	0.008148065395052901
stays_in_weekend_nights	-0.0017910780782611744
stays_in_weeknights	0.024764629045872715
adults	0.06001721283956815
children	0.005036254836439323
babies	-0.03249108920833264
is_repeated_guest	-0.0847934183570878
previous_cancellations	0.11013280822284255
previous_bookings_not_cancelled	-0.057357723165947075
booking_changes	-0.14438099106132224
days_in_waiting_list	0.054185824117780376
ADR	0.047556597880386124
required_car_parking_spaces	-0.1954978174945085
total_of_special_requests	-0.2346577739690198

TABLE V. CORRELATION FOR CATEGORICAL VARIABLES

Variables	Result
Hotel	0.13653126949161642
arrival_date_month	0.011822120071305441
meal	-0.01767760995132292
country	-0.10044912870002404
market_segment	0.23833549336078935
distribution_channel	0.1697270301121236
reserved_room_type	-0.04397743747112238
assigned_room_type	-0.1252105041585338
deposit_type	0.4804339866053031
customer_type	-0.13581931980513778

A. Evaluation

The precision defines as the number of true positives divided by the number of true positives plus the number of false positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

For evaluation, two prediction model are created. The one that uses all the variables shows an average precision of 0.82 while the model used the factors shows average precision of 0.85.By exploring the coefficients in the following Table VI and eliminating the insignificant effective variables:

We remain with the following five factors: is_repeated_guest, previous_cancellations, previous_bookings_not_cancelled, required_car_parking_spaces, and deposit_type. The model using these factors show precision of 0.86. The factors are used to create a function that can evaluate these factors and predict if the reservation is more likely to cancel or not. To create this function a scoring function can be started. The intercept and coefficients are rounded to 2 decimal level and used in an equation model. The coefficient values are -0.84, 2.21, -0.76, -6.3, and 2.66. The score is then processed in a logistic function.

Based on the result of the logistic function, the cancel evaluation function returns true if the probability is more 0.5. Otherwise, it returns false. The following Table VII and Table

TABLE VI. COEFFICIENT FOR THE FACTORS

Factors	Coefficient
adults	-0.01
is_repeated_guest	-0.75
previous_cancellations	2.42
previous_bookings_not_cancelled	-0.76
booking_changes	-0.29
days_in_waiting_list	0.0
required_car_parking_spaces	-6.11
total_of_special_requests	-0.37
hotel	0.03
country	-0.0
market_segment	-0.15
distribution_channel	0.6
assigned_room_type	-0.04
deposit_type	2.61
customer_type	-0.37

VIII shows the result from the API function. Based on the result in Table VIII, it seems that there is a probability of cancelling the reservation about 83%. So, the cancel evaluation function gives a true for this reservation.

TABLE VII. EXPERIMENT 1 AFTER USING API

Scoring	7.157
API function result	0.00077
Cancel evaluation function	False

TABLE VIII. EXPERIMENT 2 AFTER USING API

Scoring	1.632
API function result	0.83
Cancel evaluation function	True

Finally, the cancel function can be interfaced in the requested form:
Boolean cancel (factorsArray)

VI. CONCLUSION AND FUTURE WORK

For a future work, it can be considered a data set from hotels in Saudi Arabia. Moreover, API function can be used with the data set that have the same parameters to evaluate it. Because of the unavailability of many data sets that can be used, only this data set was used in this study. Based on the results, the hotel management can consider a good cancellation policy. Furthermore, they can adjust their cancellation policy based on the proprieties that the API function has.

REFERENCES

[1] A. Serra-Cantallops, D. D. Peña-Miranda, J. Ramón-Cardona, and O. Martorell-Cunill, "Progress in research on csr and the hotel industry (2006-2015)," *Cornell Hospitality Quarterly*, vol. 59, no. 1, pp. 15–38, 2018.

[2] R. Battiti, M. Brunato, and F. Battiti, "Roomtetris: an optimal procedure for committing rooms to reservations in hotels," *Journal of Hospitality and Tourism Technology*, 2020.

[3] M. Brunato and R. Battiti, "Combining intelligent heuristics with simulators in hotel revenue management," *Annals of mathematics and artificial intelligence*, vol. 88, no. 1, pp. 71–90, 2020.

[4] T. Koide and H. Ishii, "The hotel yield management with two types of room prices, overbooking and cancellations," *International journal of production economics*, vol. 93, pp. 417–428, 2005.

[5] M. Gellerstedt and T. Arvemo, "The impact of word of mouth when booking a hotel: could a good friend's opinion outweigh the online majority?," *Information Technology & Tourism*, vol. 21, no. 3, pp. 289–311, 2019.

[6] S. Park, Y. Yin, and B.-G. Son, "Understanding of online hotel booking process: A multiple method approach," *Journal of Vacation Marketing*, vol. 25, no. 3, pp. 334–348, 2019.

[7] N. António, "Predictive models of hotel booking cancellation: a semi-automated analysis of the literature," *Tourism & Management Studies*, vol. 15, no. 1, pp. 7–21, 2019.

[8] T. A. Maier and C. Roberts Ph D, "Exploratory analysis of 'other revenue'impact on full and limited service hotel noi," *Perspectives in Asian Leisure and Tourism*, vol. 3, no. 1, p. 4, 2018.

[9] C. Chen, "Cancellation policies in the hotel, airline and restaurant industries," *Journal of Revenue and Pricing Management*, vol. 15, no. 3, pp. 270–275, 2016.

[10] F. DeKay, B. Yates, and R. S. Toh, "Non-performance penalties in the hotel industry," *International Journal of Hospitality Management*, vol. 23, no. 3, pp. 273–286, 2004.

[11] C.-C. Chen, Z. Schwartz, and P. Vargas, "The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers," *International Journal of Hospitality Management*, vol. 30, no. 1, pp. 129–135, 2011.

[12] A. Riasi, Z. Schwartz, and C.-C. Chen, "A paradigm shift in revenue management? the new landscape of hotel cancellation policies," *Journal of Revenue and Pricing Management*, vol. 18, no. 6, pp. 434–440, 2019.

[13] B. Benítez-Aurioles, "Why are flexible booking policies priced negatively?," *Tourism Management*, vol. 67, pp. 312–325, 2018.

[14] C.-C. Chen and K. L. Xie, "Differentiation of cancellation policies in the us hotel industry," *International Journal of Hospitality Management*, vol. 34, pp. 66–72, 2013.

[15] A. Riasi, Z. Schwartz, and C.-C. Chen, "A proposition-based theorizing approach to hotel cancellation practices research," *International Journal of Contemporary Hospitality Management*, 2018.

[16] E. Alotaibi, "Application of machine learning in the hotel industry: A critical review," *Journal of Association of Arab Universities for Tourism and Hospitality*, vol. 18, no. 3, pp. 78–96, 2020.

[17] A. J. Sánchez-Medina, C. Eleazar, *et al.*, "Using machine learning and big data for efficient forecasting of hotel booking cancellations," *International Journal of Hospitality Management*, vol. 89, p. 102546, 2020.

[18] E. C. Sánchez, A. J. Sánchez-Medina, and M. Pellejero, "Identifying critical hotel cancellations using artificial intelligence," *Tourism Management Perspectives*, vol. 35, p. 100718, 2020.

[19] N. Antonio, A. de Almeida, and L. Nunes, "An automated machine learning based decision support system to predict hotel booking cancellations," *An automated machine learning based decision support system to predict hotel booking cancellations*, no. 1, pp. 1–20, 2019.

[20] N. António, A. de Almeida, and L. M. Nunes, "Using data science to predict hotel booking cancellations," in *Handbook of research on holistic optimization techniques in the hospitality, tourism, and travel industry*, pp. 141–167, IGI Global, 2017.

[21] N. Antonio, A. De Almeida, and L. Nunes, "Predicting hotel booking cancellations to decrease uncertainty and increase revenue," *Tourism & Management Studies*, vol. 13, no. 2, pp. 25–39, 2017.

[22] Z. A. Andriawan, S. R. Purnama, A. S. Darmawan, A. Wibowo, A. Sugiharto, F. Wijayanto, *et al.*, "Prediction of hotel booking cancellation using crisp-dm," in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 1–6, IEEE, 2020.

[23] C. Timamopoulos, "Anomaly detection: Predicting hotel booking cancellations," 2020.

[24] M. Falk and M. Vieru, "Modelling the cancellation behaviour of hotel guests," *International Journal of Contemporary Hospitality Management*, 2018.

[25] N. Antonio, A. de Almeida, and L. Nunes, "Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior," *Cornell Hospitality Quarterly*, vol. 60, no. 4, pp. 298–319, 2019.

[26] M. Yemer, "The effect of internal controls systems on hotels revenue. a case of hotels in bahir dar and gondar cities," *Arabian Journal of Business and Management Review (Oman Chapter)*, vol. 6, no. 6, p. 19, 2017.

[27] W. Wikhamn, J. Armbrecht, and B. R. Wikhamn, "Innovation in swedish hotels," *International Journal of Contemporary Hospitality Management*, 2018.