



**IST 687: Introduction to Data Science**

**Project Group 2**

*Hotel Booking Cancellation Analysis*

1. Anjali Shahi
2. Pranali Shenvi
3. Kirat Saran
4. Anusha Ramprasad
5. Jacob Murrer

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| Sr. No. | Title | Page No. |
| 1. | Case Introduction | 3 |
| 2. | Summary of the Dataset | 3 |
| 3. | In-depth Dataset Information | 3 |
| 4. | Data Exploration | 5 |
| 5. | Data Cleansing | 6 |
| 6. | Data Transformation | 8 |
| 7. | Data Visualization and Analysis | 9 |
| 8. | Data Modelling | 17 |
|  | 1. Linear Modelling | 17 |
|  | 1. Support Vector Machines | 19 |
|  | 1. c. Association Rule Mining | 19 |
| 9. | Business Questions | 21 |
| 10. | Insights and Recommendations | 22 |

1. **Case Introduction**

We have been provided with the data which contains real - life hotel stay data, with each row representing a hotel booking.  Our goal of this project is to understand some key drivers for why people cancel hotel reservations and better predict who will cancel. The problem statement depicts that the hotel faces some cancellations, and we are going to give some insights to prevent the cancellations and try to predict the future cancellations. We will be providing descriptive visualizations and graphs to support our analysis of prediction. Furthermore, we will be applying three of the modelling techniques; Linear Modelling, Association rules Mining and Support Vector Machines to find the statistical significance between the variables and finally give out recommendations that answer the key business questions. We will also find and use the NPS (Net Promoter Score) using the data attributes to find out how likely a customer will recommend the hotel for stay.

1. **Summary of the Dataset**

The dataset consists of a real - life hotel stay data with 40,060 rows and 20 columns.  The attributes include personal information attributes such as Number of Adults, children, babies, and travel attributes like the Country. It also includes hotel attributes like Previous cancellations, the room assigned, meal given, and the weekend or weekday stay.

We need to find the factors which affect the hotel cancellation and then provide actionable insights to reduce the cancellations and retain the customers. Thus, using the data transformation after cleaning the data and using some modelling techniques we can find the statistically significant attributes and frame effective recommendations which would help determine the reason for the cancellations and prevent them. The visualizations would also help give a descriptive view and finally we can answer the business questions.

1. **In-depth Dataset Information**

As provided in the PDF for the information of the data we can see that these are the following columns in the dataset and their information:

* IsCanceled: Categorical Value indicating if the booking was cancelled or not (1 for cancelled booking and 0 for the bookings not cancelled)

* LeadTime: Integer, Number of days that elapsed between the entering date of the booking into and the arrival date

* StaysInWeekendNights: Integer, Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

* StaysInWeekNights: Integer, Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

* Adults: Integer, Number of adults

* Children: Integer, Number of children

* Babies: Integer, Number of babies

* Meal: Categorical, Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC - no meal package; BB - Bed and Breakfast; HB - Half board (breakfast and one other meal - usually dinner); FB - -Full board (breakfast, lunch and dinner)

* Country: Categorical, Country of origin. Categories are represented in the ISO 3155 - 3:2013 format

* MarketSegment: Categorical, Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

* IsRepeatedGuest: Categorical, Value indicating if the booking name was from a repeated guest (1) or not (0)
* PreviousCancellations: Integer, Number of previous bookings that were cancelled by the customer prior to the current booking

* PreviousBookingsNotCanceled: Integoer, Number of previous bookings not cancelled by the customer prior to the current booking

* ReservedRoomType: Categorical, Code of room type preserved. Code is presented instead of designation for anonymity reasons

* AssignedRoomType: Categorical, Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved froom type due to hotel operation reasons (e.g., overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
* BookingChanges: Integer, 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

* DepositType: Categorical, 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

* CustomerType: Categorical, 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 a 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

* RequiredCarParkingSpaces: Integer, Number of car parking spaces required by the customer

* TotalOfSpecialRequests: Integer, Number of special requests made by the customer (e.g., twin bed or high floor)

1. **Data Exploration**

For the analysis of the dataset given, there are a few steps to be followed. First, we will load the data into the appropriate data frame, then we will cleanse the data so that we are left with the required data to work on after which we will work on some visualizations to understand the data and then use prediction modelling to come to conclusions regarding the data.

First, we are using the appropriate libraries and loading the data into a data frame as shown below:

Graphical user interface, text, application

Description automatically generated

Fig 1: Data Exploration

After loading the input file and exploring it we can see the data in each of the columns and understand their data type.

1. **Data Cleansing**

Data cleansing is an important part as it helps us get rid of unwanted or faulty data and work with only the required data which is of relevance.

First, we are creating a table for each of the numeric forms to understand the data.

Text

Description automatically generated

Fig 2: Creating tables for attributes

After checking for the NA values, we can see that none of the above numeric fields have any NA values present.

We can see that there are some NULL values in the attribute Country. These account for 464 rows. We are removing these rows as we cannot plot maps when we have NULL values in the column for the country.

Text

Description automatically generated with medium confidence

Fig 3: Removing the null valued rows in Country

To standardize the dataset values, we are replacing the rows which contain “Undefined” values for Meal and change that with “SC” as they mean the same thing that is no meal package according to the attribute details in the dataset.

Text

Description automatically generated

Fig 4: Making changes in the values of Meal column (from Undefined to SC)

The next thing we did is that we changed the column name from LeadTime to DaysSinceBooking for easier understanding.

A picture containing diagram

Description automatically generated

Fig 5: Renaming the column LeadTime

Further, we will be filtering out those rows from the dataset where the number of StaysInWeekendNights and StaysInWeekNights are zero. We do this because, most hotels have a nightly pay rate so a hotel booking that does not have a single night stay can be considered as dummy data

Text

Description automatically generated

The final cleaned data has 39209 observations and 21 attributes.

1. **Data Transformation**

Changing country code for ‘CN’ to ‘CHN’ inorder to plot the map. A new column called, modifiedCountryCode is added to implement the same

Text

Description automatically generated

We have also created a new column called Detractor based on another attribute:

PreviousBookingCancelations. The column stores TRUE value if the number of previous cancellations is greater than 12.

Graphical user interface, text

Description automatically generated

1. **Data Visualization**

As we investigate developing an analysis and using various regression models and techniques, we first need to gain some general insights into our data.

To start, we create a bar graph that looks at the percentage of canceled vs non-canceled bookings.

Chart, square

Description automatically generated

As you can see from the graph, 30% of all bookings in the data set were canceled. Our job now is to figure out the various characteristics that make up this 30%.

As we look deeper into our dataset, we create visualizations regarding cancelations and various factors that make up the booking. One of these being a bar graph associating the room type and percentage of bookings.

Chart, bar chart

Description automatically generated

From this visualization we can see that room type A was the most assigned room type followed by D, E, and G respectively. This is important to understand as we move deeper into our analysis as these rooms have much more significance relative to the other room types.

Our next insight was gathered through examining how cancelations are distributed through market types.

Chart, bar chart

Description automatically generated

From this we can interpret that about 50% of customers who are using an online travel agent end up canceling. We also see that a good number of cancelations are coming from the offline travel agent market type. Moreover, we can see that a very large percentage of group bookings are resulting in cancelations. These are areas that will be important to include in our analysis models later.

We then look at the distribution of cancelations through market types based on whether the booking is a weekend or weekday.

Chart, bar chart

Description automatically generated

Following the same trend seen previously, the Groups, Online TA, and Offline TA market types are showing the most cancelations. We can also determine that weeknights are being canceled the most and that this is something to improve on.

Continuing our visualization of the data we look at the cancellations based on the selected meal type.

Chart, waterfall chart

Description automatically generated

Here we can see that both the bed and breakfast and half-board meal types are showing many cancelations and are an area of importance when looking to decrease the cancelation rate.

Combining both the meal type and time of booking we obtain this graph.

This is following the same trend as before where weeknights are showing the highest cancelations, even when looking at the individual meal types.

Here we examine how cancelations are distributed through deposit type.

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

From this we can determine that non-deposit types are seeing the most cancelations and will be an important factor in our analysis. We can also see a large number of cancelations in the non-refund market type, however as the company is still receiving a profit off this, we do not deem this to be important.

Following the same trend as before, we examine how cancelations are distributed through deposit type based on the time of booking.

Chart, bar chart

Description automatically generated

Reinforcing our earlier findings, cancelations are much more likely to occur when the booking is on a weeknight.

Now we look at how cancelations are affected when the time since booking changes.

Here we can see that when the amount of time since booking increases, the more likely the booking is to be canceled.

Chart, box and whisker chart

Description automatically generated

Our final visualization is looking at the geographical distribution of cancelations for this hotel.

A picture containing whiteboard

Description automatically generated

From this we can see that Portugal is showing the highest number of cancelations and may be an important factor to examine later.

**8. Analysis: Modeling techniques:**

* 1. **Logistic Regression:**

All these visualizations have been valuable in giving us insights as to what factors are important to include in our various models. We have already seen some trends and relationships that should be reinforced by the output of our models.

For our logistic modeling our first model was to use all the variables in efforts to narrow down what variables might be the most significant in predicting what influences hotel cancelations.

The IsCanceled variable is derived from multiple variables, it is our dependent variable in our linear model analysis. The key measurements we will pay attention to in this analysis will be the Residual Standard Error, Multiple R-Squared value, and the p-value statistic based on the level of significance.

Graphical user interface, text, email

Description automatically generated

From this we can see the most significant variables. The list of all variables the analysis was performed on is much too large to copy the whole things, so we selected the top few to show.

In the output, of the regression model we observe that different categories in categorical dependent variables have been listed out separately and may or may not be significant. Further, if an attribute has 5 categories, then only 4 categories will be listed out as one is included in the intercept itself.

From the result of the model, we obtain a initial list of significant variables that we can further narrow done by applying additional models

* 1. **SVM:**

The support vector machine is used to find the accuracy of the prediction of a particular model. Here, we split the sampled dataset into 2 parts, training and testing and the train data was used on a model and the test data was used to determine its accuracy of prediction.

Text

Description automatically generated

Chart, pie chart

Description automatically generated

* 1. **Association Rules Mining:**

Association rules are created using the apriori function. It is useful in finding variables that often occur together. It is created with a left-hand side that are often together with one variable of the right-hand side.

We converted all integer variables that we plan to use to categorical variables. We then created a data frame of the key variables to be able to perform an apriori analysis on the new data set.

These are some of the rules obtained from the model -

Graphical user interface, text, application

Description automatically generated

From the output we observe that there are certain variables that are recurring a number of times like: DepositType=Non-Refund, RequiredCarParkingSpaces=0.

We can conclude that these are some of the most important contributing factors in hotel cancellations.

Chart, line chart

Description automatically generated

This graph is a visualization of the rules obtained from association rules mining. As we can see from the graph, DepositType=non-Refund is the strongest influencer to hotel cancellations.

* + 1. **Business Questions:**

These are some of the business questions we had aimed to answer through our visualizations and modelling:

* Does opting for RequiredParkingSpace while making hotel reservations impact cancellations?

Yes, as seen in the visualization as well as backed up by the rules generated through association rule mining. We see those bookings where the RequiredParkingSpace is zero are more likely to cancel.

* Is it possible to associate meal plans and StaysinWeeknights with possible hotel booking cancellations?

Yes, BB and FB are two meal plans that show a trend of high cancellation along with StaysinWeeknights. However, since the number of instances for FB are less.

* Which market segment influences the cancellations the most?

Bookings made through OnlineTA have the highest number of cancellations.

* Does DepositType affect cancellations?

Bookings made through; Non-Refundable deposits have the most cancellations.

* Are customers making cancellations based on the rooms they were assigned?

We observe that bookings where room type A was assigned has the highest cancellations however, since we were not given enough information regarding the room types in the dataset no interpretation can be made.

**10. Insights and Recommendations**

* 1. *Guest Retention*

Insight: There are a lot of cancellations which tend to happen in BB and HB meals. This trend can be substantially seen in the weekend night stay bookings.

Recommendation: To pull those customers from not cancelling on weekend nights and the meal types, better deals like loyalty programs should be offered on BB and HB and on weekend night stays.

* 1. *Payment Structure*

Insight: The payment type ‘No Deposit’ tends to experience a greater number of cancellations compared to other payment types. Also, Non-refundable payment types tend to cancel more than Refundable payment types.

Recommendation: The hotel management should keep a nominal amount as minimum deposit in order to get rid of the category of ‘No Deposit’ thus reducing the cancellations.

* 1. *Market Segment*

Insight: Out of all the categories of market segments, the Group category experiences a rather large number of cancellations. It can also be minutely observed in the complimentary category. But, as it contributes to a very small part in business, we will not be. Considering that.

Recommendation: As the people are already staying as a group, the hotel can suggest getting a bigger group to avail more discounts. For the company, it would still get them revenue as many people are booking and even after giving some discount, they will still profit.