**Final Project Report**

Kangsheng Pu

December 5th, 2021

**Introduction**

This report’s main goal is to discover the key factors of people canceling their hotel reservations. The factors could help hotel managers to predict customers who are more likely to cancel the booking so that hotel managers might allocate those resources in a more efficient approach.

**Dataset Description & Preview of Database**

This section discusses the raw variables of the database and cleanup process of the dataset. The section will also discuss some suspected factors that might influence the cancellation rate of hotel booking according to the cleaned-up variable. However, the report will still conduct an exploratory data analysis in case of unexpected factors.

1. **Variable Transformation**

IsCanceled, the variable that identifies whether the unit of observation canceled their hotel bool, is the target dependent variable we need to predict this semester. Except this variable, there are 19 variables that could be used to predict the dependent variable.

As the data set contains unprocessed factor variables without dummy variables, and they contain more complex information and needs data cleaning, they will be the first to discuss.

* **Country**: Shows the user’s country. Although being straightforward, this variable might be the most complex factor variable in the whole database. First of all, there are numerous variables that only occupy one to two units of observations in the presentation. It is unreasonable to use these barren amounts of data to represent the customers from the country. To solve the problem, adding more samples will be a good choice, while this method is unavailable at this moment. Instead, it might be more sensible to focus on certain countries with rich data size to see what the changes in the factors of hotel cancellations might be. However, because of the time limits, this report will drop the country variable to
* **Meal**: According to the project description, it represents the type of meal the customer booked. The values have an ordinal trend: SC(no meal package)- BB(breakfast – one meal) – HB(two meals) – FB(three meals). Thus this variable could be transferred with numerical value according to the meals booked: 0[SC] – 1[BB] – 2[HB] – 3[FB].
* **Reserved Room Type** – Assigned Room Type: Those two variables separately show the customers’ reserved room type and assigned room types. Although being nominal categorical variables, the two variables might derive into a new Boolean variable showing whether the customers received their assigned room types. The new Boolean variable is called **AssignedReservedRoom**, 1 stands for the customer is assigned with reserved room type(true), while 0 stands for the opposite(false).
* **Deposit Type**: This variable has three values: No deposit, refundable, and nonrefundable. They also have a slightly ordinal trend, as three values start from free book to refundable booking, and eventually nonrefundable booking. Thus, this variable will also set to ordinal values:
  + No Deposit: 0
  + refundable: 1
  + nonrefundable: 2

One numerical variable also need transformation:

* **Previous cancellations & previous booking not cancelled**: those two variables could derive a new variable called the cancellation rate(**cancellationRate**). It will calculate the chance of user cancelled their booking, and the formula would be previous cancellation / (previous cancellation + previous not cancelled). For new customers who never cancelled or booked, their rate will be set at 0.5(50%).

Finally, the list below is the candidates variables that will be used in the prediction model:

**StaysInWeekendNights**: the number of weekend nights stays of the booking.

**StaysInWeekNights**: the number of weekday nights stays of the booking:

**Adults**: The number of adults, minimum 1

**Children**: The number of children

**Babies**: The number of babies:

**Meal**: Ordinal Categorical Value, the meal plans for the booking. [0-3] represents number of meals booked

**IsRepeatedGuest**: Factor value that represents whether the customer is a repeated guest.

**BookingChanges**: The number of changes for the booking.

**DepositType**: Ordinal Categorical Value, the deposit type user used for booking the hotel. 0 – No deposit, 1 – Refundable deposit, 2 – Non refundable deposit.

**CustomerType**: Nominal Categorical Value, the type of booking. (Contract, Group, Transient, Transient-Party)

**RequiredCarParkingSpaces**: the number of parking spaces customer required.

**TotalOfSpecialRequests:** The number of special request for the users booked.

**AssignedReservedRoom**: Factor(Boolean), checks whether the user have the requested room assigned.

**cancellationRate:** The percentage of user previously cancelled the booking. If user never cancelled or booked any hotels, the value is set to default value 50%.

**PreviousCancellations:** The number of times usercanceled his/her/their hotel booking

**PreviousBookingsNotCanceled:** The number of times user booked and did not cancel the hotel booking.

**MarketSegment:** Nominal factors, Market segment designation. The factors are: Complementary, Corporate, Direct, Groups, Offline TA/TO,Online TA.

**Data Cleaning**

**Invalid Values:**

Although every value in the dataset is not null, there are several variables containing invalid values, and requires discussion in this section.

* **Stays in weekend nights + Stays in weeknights**: The sum of the two values means the number of nights stays booked. Those uncertain values will be removed for the primary analysis.
* **Meal**: After completing the variable transformation, there are still 1160 units of observation having undefined meal plans. Although possessing a relatively portion of larger dataset, those undefined value will be remove primarily from the dataset.
* **Adults**: There must be at least one adult booking for the room. However, there are 10 rows that does not have any adults, thus removed.

**Outliers:**

Outliers will influence the prediction model negatively. According the graphs, some outliers are discovered and requires removal or winsorization. In this example, the author applied z-score identification, and removes all the values if their z-score exceeds 3. Those are the variables that are processed under this method: "Lead Time", "Stays In Weekend Nights", "Stays In Week Nights", "Adults", "Previous Cancellations", "Previous Bookings Not Canceled", "Booking Changes", "Required Car Parking Spaces", "Total Of Special Requests". The boxplot set is shown as below:

Diagram

Description automatically generated

It is not hard to discover that the graphs are extremely skewed and requires outliers removal. The boxplot after the outliers removal is shown below:

Diagram

Description automatically generated

Although many graphs are still seems skewed, while the range have been dramatically decreased, which is a good sign of removing outliers.

Variable “Babies” and “Children” did not apply this method as after z-score identification, Column “Babies” only have value 0 and Column “Children” only have value (0, 1). Thus the author manually removed several rows for

After data cleaning, the final data size is 35689.

**Exploratory Data Analysis**

This session will illustrate three models that were applied to predict whether the hotel booking will cancel. All the training data and randomly selected and occupies 80% of the original data, and the test data possess 20% of the original data.

**Logistic Regression Model**

Logistic regression is a prediction model that is used to predict categorical values,

which is suitable for our dataset. To select the variables that will be significant for the logistic model, the author decided to conduct a stepwise removal logistic regression. By looping the process of comparing the previous model and the next model, the author would be able to discover the best model or significant predictors of predicting hotel cancellations. The graph below shows the result of first logistic regression by putting all candidates variables as the explanatory variables:

Table

Description automatically generated

It is clear that there are many insignificant predictors in the model, according to the p values revealed in the summary image above. According to the p-values, a list called “Failed List” is generated for the stepwise regression looping. The failed list contains all the variables that its p-value exceeds α, 0.05. By removing each variable in the failed list and check the regression result each time, the author will remove all the insignificant variable predictors. The variables in the failed list are: "StaysInWeekendNights", "StaysInWeekNights", "Meal", "MarketSegment", "Babies", "RequiredCarParkingSpaces", “PreviousBookingsNotCanceled", "CustomerType".

Upon removal, the result of final logistic regression is shown below:

Table

Description automatically generated

All the variables are statistically significant, which means that this might be the ultimate logistic regression model for the prediction. These final predictor variables are: Adults, Children, IsRepeatedGuest, DepositType, TotalOfSpecialRequests, AssignedReservedRoom, cancellationRate, and PreviousCancellations.

To double check the final result of the variables. The author applied Chi-Squared test to check whether all those variables are dependent on the variable”. Thus, setting up null hypothesis and alternative hypothesis within a loop for all those final predictors:

Null Hypothesis: One or more of the final predictors are independent to the cancellation of the hotel booking.

Alternative Hypothesis: All final predictors are dependent to the cancellation of the hotel booking.

Text

Description automatically generatedText

Description automatically generatedThe results are shown below in two images. By setting α = 0.05 and according to the p-values, the author is able to reject the null hypothesis and concluded that all the final predictors variables the author selected are dependent on the cancellations of the hotel bookings.

**Interpretation**

Transferring all the variables into x:

Adults – x1

Children - x2

IsRepeatedGuest – x3

BookingChanges – x4

DepositType\_Linear – x5

DepositType\_Quadratic – x6

TotalOfSpecialRequests – x7

AssignedReservedRoom – x8

cancellationRate – x9

PreviousCancellations – x10

The final formula would be:

The following would be the interpretation of each variables:

Adults: Controlling all variables constant, one number increase of adult will increase the logit by 0.276, which means that by having one more adult booking the hotel, the possibility of cancellation rate increase by 31.7%.

Children: Ceteris Paribus, one more child in the booking group will increase the logit by 48.6%

IsRepeatedGuest: Ceteris Paribus, repeated guests are 72.7% less likely to cancel the hotel booking.

DepositType: Ceteris Paribus, as the deposit type becomes stricter, the user will be 12 times more likely to cancel the next reservation.

TotalOfSpecialRequests: Ceteris Paribus, 1 case increase of special requests will decrease the possibility of cancellation by 17.3%.

AssignedReservedRoom: Ceteris Paribus, the user with the same assigned room type and reserved room type is 7.1 times more likely to cancel the booking.

cancellationRate: Cetris Paribus, by increasing ten percent of cancellation rate, the customer will be 58.1% more possible to cancel their next booking.

PreviousCancellations: Cetris Paribus, 1 number increase of previous cancellations is 6.43 times more possible to cancel the hotel booking.

**Accuracy**

By looping the training and predictions 30 times, the average accuracy of the model is around 72.8%, which is the lowest among the three models.

**Discussion:**

Among all those variables, cancellation rates, previous cancellations, whether the customer is a repeated guest and number of Special requests are the relatively straightforward. The higher the cancellation rates and the number of cancellations, the higher possibility that the customer will cancel the booking again. Repeated guest means that the customer is satisfied with the service and location of the hotel, which is reasonable. Approved special requests means that the hotel’s service is unique to the customer, thus the customer will stick more to the hotel.

However, there are also some counter-intuitive results. Some of them might be explainable. For example, users with different unexpected room types are less likely to cancel the room. It is possible that the hotel is accepting large numbers of customers, and the demand for the room is extremely high. Also, if the customer is not satisfied with the room type change, they will just not book the hotel instead of cancelling them. This might be an issue of survivorship bias. However, the results like “some high more adults and children are more likely to cancel the reservation” and “stricter deposit type will lead to higher cancellation possibility” are hard to interpret and discuss.

**Restrictions**

Unfortunately, this model does not seem to be a good prediction model and removing the insignificant variables according to P-values directly might be the reason. Instead, it might be a better approach to conduct chi-square test first and then extract the variables and conduct stepwise logistic regression. However, the author is not able to complete this approach due to time limits.

What’s worse, as the research became deeper, the author discovered that instead of using logistic regression model, an ordinal logistic regression model would be more appropriate for the prediction. This late discovery means that it is highly possible that this model might get wasted. In spite of this, the prediction model still has its usage: this model gives the author directions what variables are dependent to the user cancelling the hotel booking

**SVM Model**

The second prediction model would be SVM model. One critical issue of SVM model is to discover the best cost of the variable. Also, to prevent overfitting, using cross-validations data analysis might be a better choice disregarding the time cost of training the model. The author selected 5 fold cross validation and will repeat three times.

Chart, line chart

Description automatically generatedText

Description automatically generated

**SVM Model - Discussion**

Unlike logistic regression, the SVM model is hard to determine whether a variable will positively or negatively influence the model. However, SVM model could still produce the importance of the variables, and the result is shown below:

Graphical user interface, application, table

Description automatically generated

According to the chart above, we discovered that whether the customer got the reserved room type has the highest importance, followed by customers’ number of required car parking spaces, market segment, and stays in week nights.

One interesting variable that SVM mentioned would be required car parking spaces and market segments, as both were not mentioned by the logistic regression. No parking space required might means that the customers are harder to move, and it will be harder for them to find a new hotel around the area.

**SVM Model - Accuracy**

The accuracy of SVM model is 81.15%, which is the highest among the three prediction models. The accuracy of the model peaked at cost equals 2.

A screenshot of a computer

Description automatically generated with low confidence

**SVM Model - Restrictions**

As the model has a high accuracy model, it is significant to check whether the model is overfitting the dataset. The p-value is also small enough to conclude that the this could be a statistically essential model. However, the specificity is low and at around 0.47. This figure means that 53% of the customers who won’t cancel the booking will be incorrectly regarded as they will cancel the booking. It indicates that the data still need improvement by removing overfitting factors.

**TreeBag Model(Decision Tree)**

The final model would be TreeBag Model, and this is the model author is not familiar with, as the author does not know skills of overfit control. The few possible methods would be having more samples and conduct data cleaning by removing more outliers, which is already unavailable at this state.

The result decision tree graph is shown as below:

Diagram

Description automatically generated

**TreeBag Model -Accuracy**

The accuracy of Treebag model is 79.4%, which is mediocre among the three. However, as it does not contain overfit control methods, the accuracy of this model remains questionable.

**TreeBag Model -Discussion**

In the decision tree, the included variables includes Deposit Type, Required car parking space, whether user get the request room, market segment, previous Cancellation rate, number of special requests, and customer group.

Required Parking space, assigned reserved room, cancellation rate and total special request are the variables that are similar to the SVM models, and most of them gave accurate and satisfactory predictions.

Although the deposit type is the first leaf, it incorrectly predicted customers whose deposit type is nonrefundable. Also, the customer type had a poor performance for predicting customers who are transient or under contract.

**TreeBag Model -Restrictions**

Same as the SVM model, the low specificity means that the model is not good at identifying users who won’t cancel the booking. This result means that the model also have overfitting issues.

A screenshot of a computer

Description automatically generated with low confidence

**Model Selection**

As the author discovered that it is incorrect to apply logistic regression for data analysis, it turns out SVM model might be the best model to choose at hand. Even so, the low specificity still indicated that the model overfits for the data.

**Recommendations**

Although the model buildings of the whole project weren’t ideal, the author still got some useful insights. Those factors may not help users to predict but will give directions to the hotel managers to some extent.

Cancellation Rate, Deposit type, total of special requests and whether user get the reserved room type are the factor that all of the models indicated as significant. However, for deposit types, logistic regression provided counterintuitive results and tree models gave wrong predictions.

Required car parking space is a variable that both SVM model and treebag model mentioned as significant, and after discussion, the author consider it as a significant factor since customers with no cars are harder to look for another hotel within similar range than customers with a car.

**Future possible work and Reflection**

There are still numerous tasks that requires improvement in the future:

For Model improvement of regression, it is clear that applying ordinal logistic model will be more meaningful and might produce a better model.

For SVM model, putting more time and creating a multi-fold model might help improve overfit control. More detailed data cleaning might also improve the situation.

Also, the author removed country for the primary analysis. It might be useful to filter out countries with less customers and check other countries’ customer cancellation patterns.

The author considered the final project as a learning process. During this project, the author learned numerous skills and model building techniques. Even so, there are still a lot of machine learning skills the author did not study. Studying them and returning to the project will have a better improvement on the model building and prediction.