Maximising Revenue for Hotels Through Supervised Learning

BT2101 Final Project - Group 28

Lew Kee Siong Lionel (A0185418X)

e0318709@u.nus.edu

Edward Lee (A0183660A)

edwardlee@u.nus.edu

1. Introduction

Using a dataset which contains booking information from two different hotels between July 2015 to August 2017 in Portugal, we aim to create supervised learning algorithms to increase the revenue for hotels, which is traditionally influenced by either the average daily rate, revenue per available room, or occupancy rate.

1.1. Problem

The occupancy rate of hotel rooms around the world have been steadily rising since 2009, as seen in Figure 1. However, according to Statista, the occupancy rate of hotel rooms in Portugal have shown lacklustre performance year on year, having a substandard rate of only a 56.98% in 2018 [1], which is about 25% lower compared to the average of other European countries in the same year.

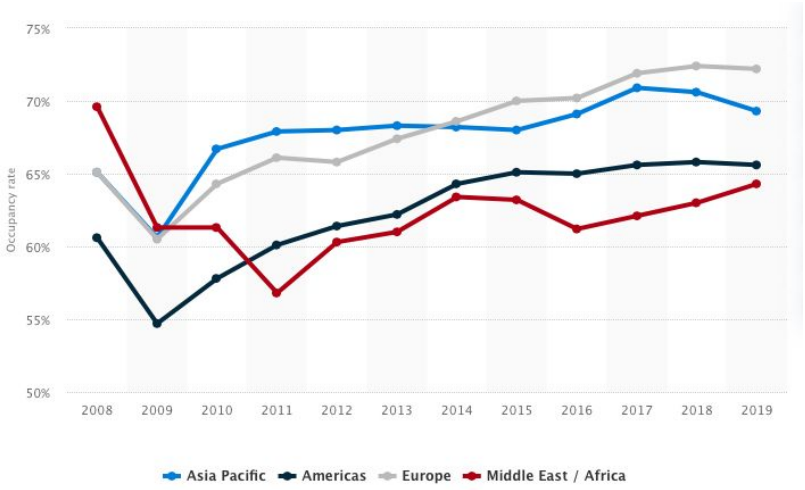

Figure 1: Hotel room occupancy rates by region

Furthermore, the global economy is undergoing a major shift in decision-making and buying power, hence changing the hospitality industry. Since the founding of Airbnb, the hotel industry has been dramatically disrupted, with every 1% increase in the number of Airbnb properties decreasing the average revenue per room by 0.2% for hotels[2].

[1] Hotel bedroom occupancy rates Portugal 2012-2018. (2020). Retrieved 21 April 2020, from https://www.statista.com/statistics/778245/hotel-bedroom-occupancy-rate-portugal/

[2] Dogru, T. (2019). Hotels Should Continue to Fear Airbnb - CityLab. Retrieved 21 April 2020, from https://www.citylab.com/life/2019/05/heres-how-much-airbnb-is-lowering-hotel-prices-and-occupancy/590485/

1.2. Objective

We hope to increase the revenue for the two hotels, “City Hotel” and “Resort Hotel” by increasing the hotel occupancy rate through the ability to accurately predict bookings that end up being cancelled.

The steps taken in this report are as follows:

1. Exploratory data analysis, including checking for anomalies, missing data and cleaning the data
2. Performing statistical analysis and data visualizations
3. Formulating hypothesis to support our case
4. Building suitable models for our datasets
5. Selecting the most suitable model and hypothesis
6. Improving upon and finalizing the model to meet our objective

2. Data Exploration

2.1. Variables of interest

1. Hotel (H1= Resort Hotel, H2 = City Hotel)
2. **Is_canceled** (Value indicated if the booking was canceled (1) or not (0))
3. Lead_time (Number of days that elapsed between the entering date of the booking into the PMS and the arrival date)
4. Arrival_date_year (Year of arrival date)
5. Arrival_date_month (Month of arrival date)
6. Arrival_date_week_number (Week number of year for arrival date)
7. Arrival_date_day_of_month (Day of arrival date)
8. Stays_in_weekend_nights (Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel)
9. Stays_in_week_nights (Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel)
10. Market_segment (Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”)
11. Deposit_type (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.)

12. ADR (Average Daily Rate)
13. country_PRT (1 indicates that the guest is from Portugal and 0 otherwise)

2.2. Analysis

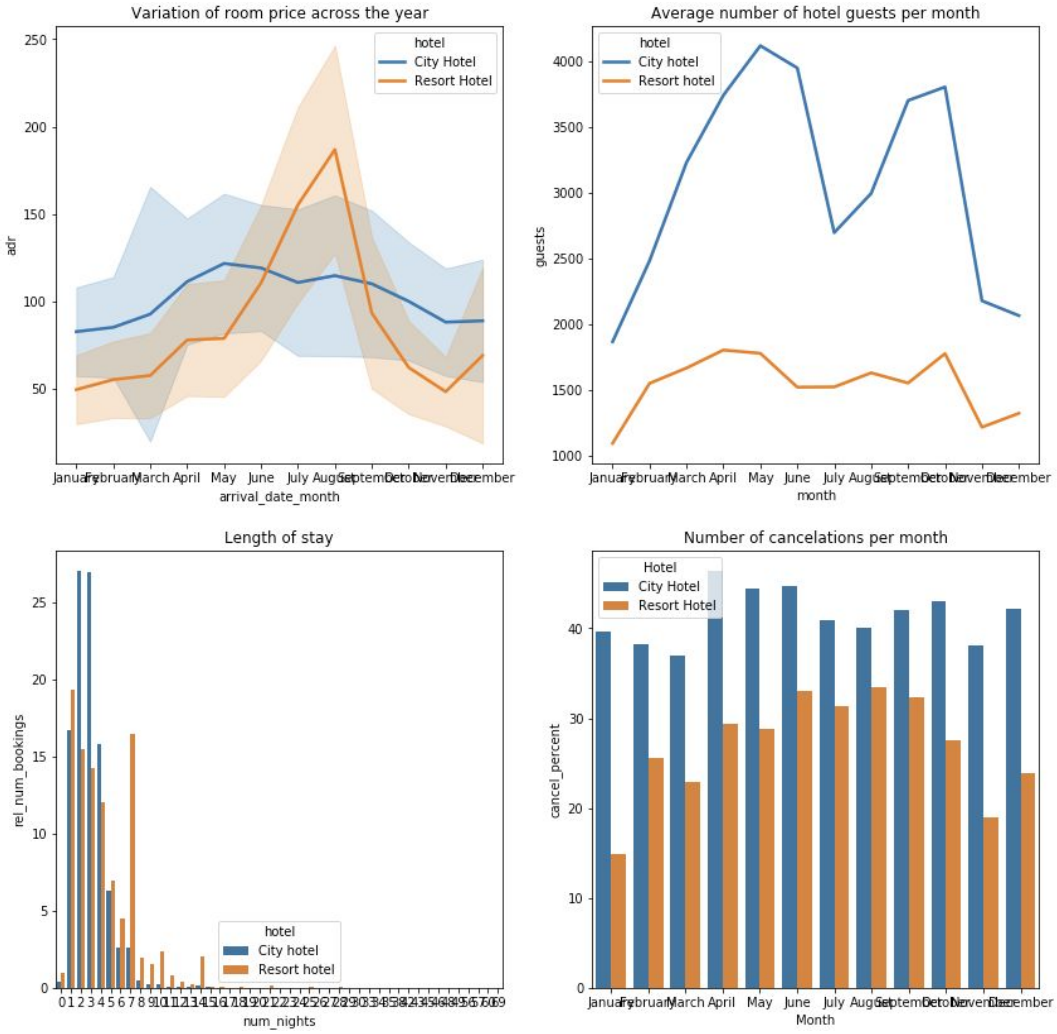

Figure 2: Charts on price, number of guests, length of stay, and cancellations

From our analysis, it can be seen that "City Hotel" has a higher average of human traffic, room prices, and room cancellation rate as compared to "Resort Hotel". Furthermore, bookings in "City Hotel" tend to have shorter durations while bookings in "Resort Hotel" tends to have longer durations. We have identified that the hotels are losing a significant amount of revenue due to the cancellation of bookings by hotel guests. Based on the data, we have found that the two hotels have lost over $4,641,942 in revenue over 2015-2017 purely due to customers who cancelled their bookings.

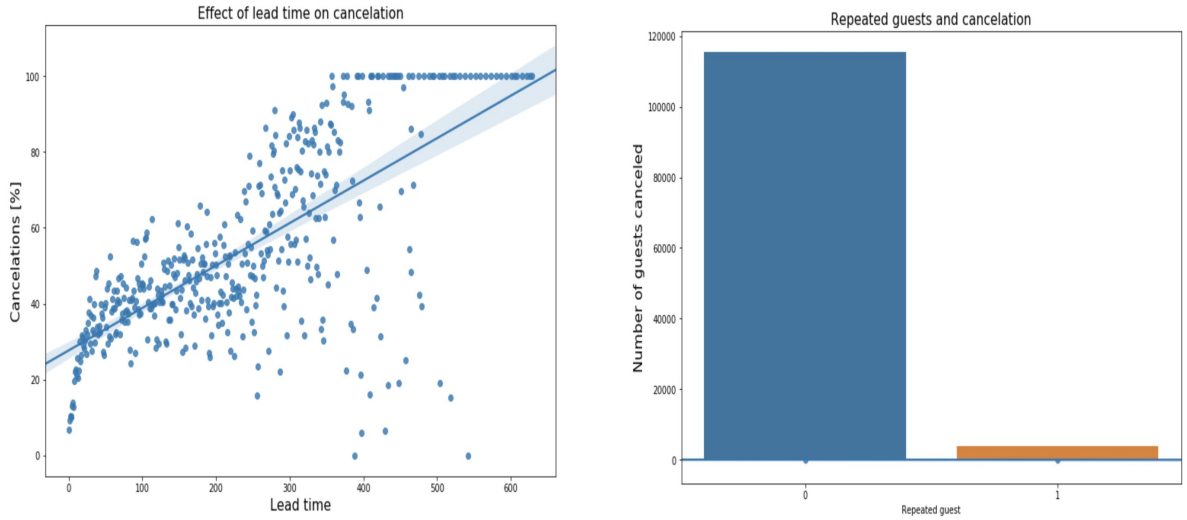

Figure 3: Relationship between duration of stay (lead_time/is_repeat_guest) and cancelation

Lead time appears to have a strong positive correlation with cancellations and non-repeat guests have a higher frequency of cancellations. This means that the longer the period between making the booking and the actual date of stay, and if the guest is not a former customer, then the likelier the booking gets cancelled.

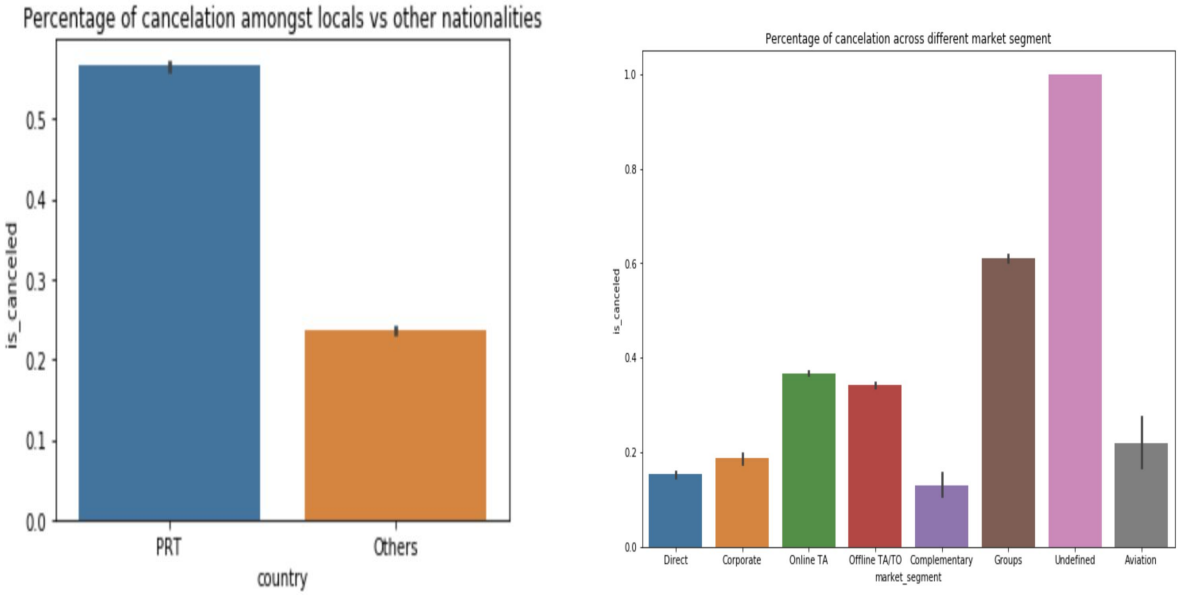

Figure 4: Relationship between background (country_PRT/market_segment) and cancelation

The customer's background and their market segment seem to affect cancellation rates as well where local Portuguese bookings seem to have a significantly higher rate of cancellation, almost double that of non-local bookings. The different categories of market segments also have a very distinct difference in cancellation rates, with the highest being "Undefined" and "Groups" and the lowest being "Direct" and "Complementary".

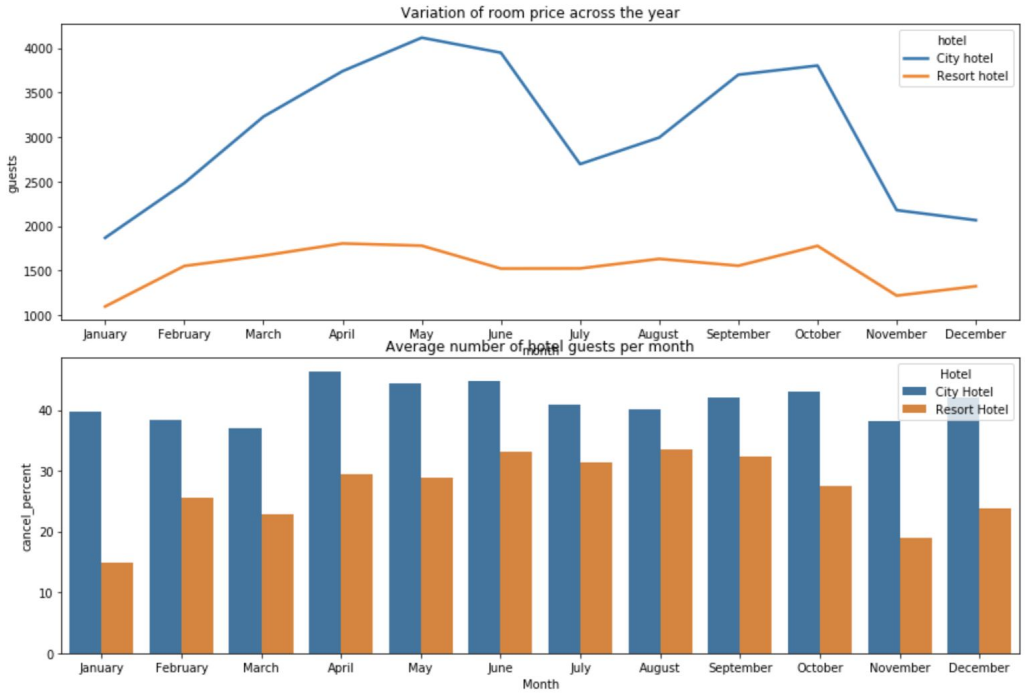

Figure 5: Relationship between seasonality (hotel/month) and cancelation

Prices and cancellations seem to be affected by seasonality as there appears to be a different pattern in room prices for both hotel types throughout the year, especially during the holidays season (May to August). Room prices seem to be fluctuating the most for "City Hotel", while the prices seem to be relatively stable for "Resort Hotel". The cancellation rates for "City Hotel" seem to be relatively constant throughout the year, while the cancellation rates distribution for "Resort Hotel" seem to take on somewhat of an inverted-U shape.

Through our exploratory data analysis, we have identified several interesting relationships between booking cancellation and customer's background, seasonality and duration of stay in the hotel (Figures 3, 4, 5). As such we have formulated 3 different hypotheses which we will be investigating in this report.

2.3. Hypotheses

1. Customers who book for a **short stay** (less than 4 nights) are more likely to cancel their booking.
2. **Local** customers (Portugese) and from the **"Groups" market segment** are more likely to cancel their booking.
3. Booking cancellation rate is higher during **months** where the hotel is **less busy**

4. **Local** customers who book for a **short stay** and from the **"Groups" market segment** are more likely to cancel their booking.

3. Methodology

It is a binary classification problem with two possible discrete outcomes - Hotel guests that cancel their bookings are classified as "1" under the cancelled column, and "0" otherwise. In order to test out our hypotheses, we have created a number of different supervised learning models namely the logistic regression and random forest classification models. We will also be looking at 2 key metrics - the **accuracy** rate and the **sensitivity** rate of the model.

3.1. Binary Logistic Regression

Binary Logistic regression mainly measures the relationship between the dependent variable("is_canceled") and the independent variables by estimating the probabilities using its underlying logistic functions. Therefore the predicted probability values can be transformed into binary values in order to make the prediction "Canceled" or "Not canceled"[3].

Mathematically, let Y be the binary outcome variable indicating booking cancelation, where 1 is canceled and 0 otherwise and p be the probability of Y to be 1. Let $X_1,...,X_k$ be a set of predictor variables. Then the logistic regression of Y on $X_1,...,X_k$ gives us the coefficient estimates $\beta 0,\beta 1,\cdots,\beta k$

$$\text{logit}(p)=\log\left(\frac{1}{1-p}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_1 x_k$$

Each estimated coefficient $\beta_0, \beta_{1,\ldots,}\beta_1$ is the expected change in the log odds of a booking cancellation for a unit increase in the corresponding predictor variables holding the other predictor variables constant. The task of the logistic function, also called the sigmoid function, is to transform probabilities into binary values to make the prediction.

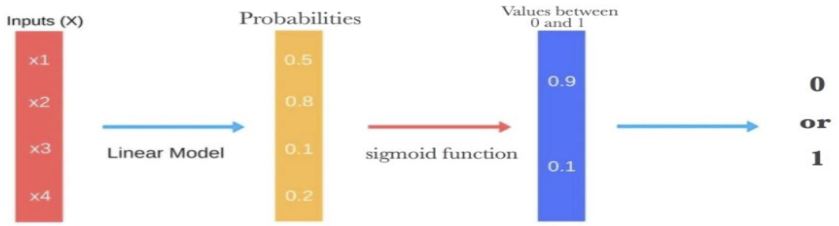

Figure 6: The process of a logistic regression

The sigmoid function is an S-shaped curve that maps inputs to values ranging from 0 to 1. The threshold value will then classify these ranges of values into either 0 or 1. The figure below illustrates the sigmoid function

[3] Logistic Regression — Detailed Overview. (2018). Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

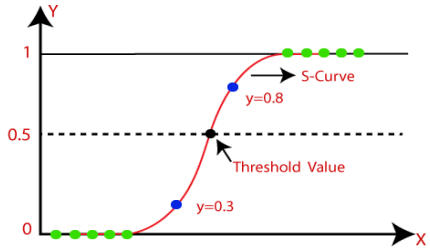

Figure 7: Sigmoid function curve

In order to predict if a booking is cancelled, if the predicted probability of the input falls above the threshold value, we would predict that the booking is canceled and vice versa. In this report, we shall set the threshold value to 0.5.

3.1.1. Advantages of using Logistic Regression

Since we are focused on binary classification, logistic regression would be a suitable tool to use. Furthermore, logistic regression is highly efficient and does not require very high computational power. As such, we would use the accuracy of the logistic regression model as well as the No Information Rate to evaluate the performance the predictive models used.

3.2. Random Forest Classification Model

Random Forest is one of the most popular and most powerful machine learning algorithms. It is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging. The "forest" it builds, is an ensemble of decision trees with the general idea that a combination of learning models improves the overall result.

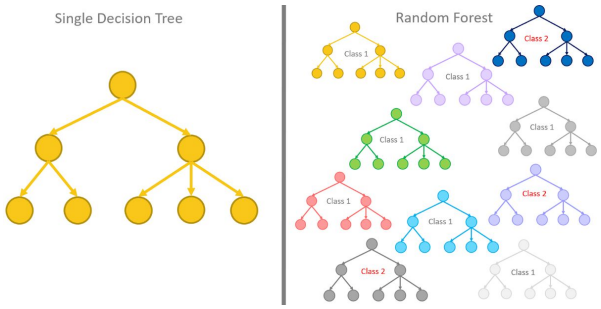

Figure 7: Comparison between a decision tree and a random forest model

The algorithm used to train each decision tree is called CART - Classification And Regression Tree. It seeks the best feature–value pair of variables within the training set to create nodes and branches. After each split, this task is performed recursively until the maximum depth of the tree

is reached or an optimal tree is found. Depending on the task, the algorithm may use a different metric such as the Gini impurity, information gain or mean square error, to measure the quality of the split. [4]

3.2.1. Advantages of using Random Forest

Random Forest models are also suitable to use to test our hypotheses since they are flexible enough to be used for both classification and regression problems. It has a very strong and accurate predictive performance, and also makes it incredibly easy to measure the relative importance of each feature to the model, which will enable us to make quick adjustments.

[4] Random Forests®, Explained. (2017). Retrieved from https://www.kdnuggets.com/2017/10/random-forests-explained.html

4. Hypotheses Evaluation

4.1. Hypothesis 1: Short term bookings are more likely to cancel their booking

Cancelled = $\beta_1 + \beta_2$ * lead_time (X_1) + β_3 * is_repeated_guest (X_2) + β_4 * total_stay (X_3)

4.1.1. Logit Model

```
Call:
glm(formula = cancelled ~ lead_time + repeatguest + total_stay,
    family = binomial(link = "logit"), data = trainData.1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5768  -0.8823  -0.7656   1.2277   2.3141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.049e+00  1.448e-02 -72.447   <2e-16 ***
lead_time    5.878e-03  7.524e-05  78.122   <2e-16 ***
repeatguest -9.311e-01  5.733e-02 -16.240   <2e-16 ***
total_stay  -2.847e-02  3.047e-03  -9.345   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110115  on 83572  degrees of freedom
Residual deviance: 102498  on 83569  degrees of freedom
AIC: 102506

Number of Fisher Scoring iterations: 4
```

Figure 8: Logit model for hypothesis 1

Looking at the summary output of the logistic regression model, we can see that the 3 variables used are highly significant even at the 0.1% level of significance. In a logistic regression, the coefficient estimate corresponds to the logarithm of the odd ratio between the dependent variable and the independent variable. The variables X_1, X_2 and X_3 have a coefficient estimate of 0.005878, -0.9311 and -0.02847respectively. For X_1 the positive coefficient estimate implies that the longer the duration from the day the booking is made till the check in day, the higher the probability of booking cancellation. Looking at the odds ratio, we can infer for the regression results that one unit increase in X_1 will increase the odds of a cancelled booking by exp($5.878e^{-03}$) 1.004 times. By applying the same method, we can gather that a unit increase in X_2 reduces the odds of cancellation by 0.3941 times and a unit increase in X_3 reduces the odds of cancellation by 0.9719 times.

Confusion matrix

	Actual		
Predicted		No	Yes
	No	19926	9647
	Yes	2564	3680

This gives us a model accuracy of 65.9%. With sensitivity of 27.6%, specificity of 88.6% and positive predictive value of 58.9%. This implies that out of the customers that were predicted to cancel their booking, only 58.9% of them actually did. Out of all the customers that canceled, only 27.6% was predicted accurately by our model but out of all the customers who did not cancel, 88.6% was predicted accurately by our model. As seen in the regression output,with a longer duration of stay as well as if the guest is a repeated guest, the probability of a booking cancellation is decreased.

4.1.2. Random Forest Model

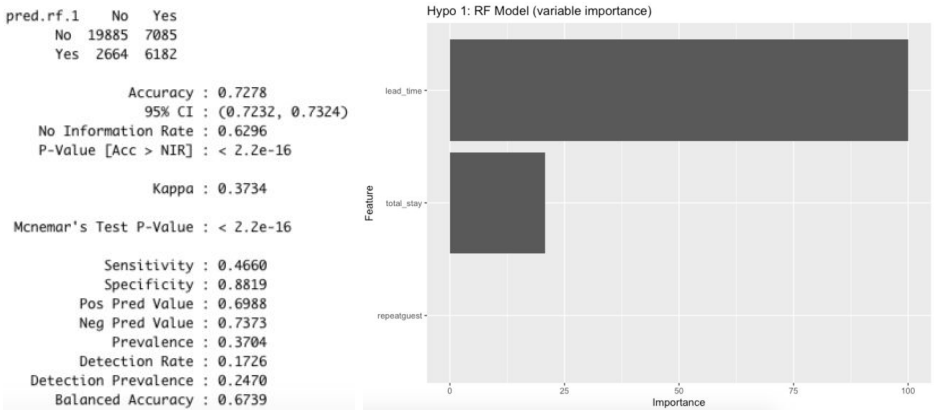

Figure 9: Random forest model and feature importance for hypothesis 1

Using a 5-fold cross-validation Random Forest classification tree model created from the CARET package in R, we are able to obtain a model that predicts the cancellation of our test data with a 72.8% accuracy rate. It also has a 46.6% sensitivity rate, a 88.2% specificity rate, and a positive predictive value of 69.8%. When evaluating the importance of the variables used, "is_repeated_guest" showed to be of low importance. It is however still included since the created model showed that it is the most accurate when all the IVs are included.

4.2. Hypothesis 2: Local guests from the "Groups" market segment are more likely to cancel their bookings

Cancelled = $\beta_1 + \beta_2$ * country_PRT (X_1) + β_3 * market_segment_Online TA (X_2)+ β_4 * market_segment_Offline TA/TO (X_3) + β_5 * market_segment_Groups (X_4) + β_6 * market_segment_Direct (X_5) + β_7* market_segment_Corporate (X_6)

4.2.1. Logit Model

```
> summary(glmModel)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6304  -0.8076  -0.5895   0.8551   2.5780

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -3.28623    0.10865 -30.246  < 2e-16 ***
country_PRT                       1.77097    0.01831  96.697  < 2e-16 ***
`\\`market_segment_Online TA\\``  2.33322    0.10866  21.473  < 2e-16 ***
`\\`market_segment_Offline TA/TO\\`` 1.62441  0.10884  14.925  < 2e-16 ***
market_segment_Groups             2.53688    0.10904  23.266  < 2e-16 ***
market_segment_Direct             0.51827    0.11160   4.644 3.42e-06 ***
market_segment_Corporate          0.30904    0.11548   2.676  0.00745 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110180  on 83573  degrees of freedom
Residual deviance:  93593  on 83567  degrees of freedom
AIC: 93607

Number of Fisher Scoring iterations: 4
```

Figure 10: Logit model for hypothesis 2

Looking at the summary output of the logistic regression model, we can see that all the variables used are highly significant even at the 0.1% level of significance except X_6 which is significant at the 1% level of significance . The variables X_1, X_2, X_3, X_4, X_5 and X_6 have a coefficient estimate of 1.771 , 2.333, 1.624, 2.536, 0.5183 and 0.3090 respectively. As explained previously, we can predict that a local guest will increase the odds of cancellation by 5.87 times, a hotel guest from the online tour agent, offline tour agent, groups, direct and corporate market segment have an increased odds in cancelation of 10.308, 5.0754, 12.64, 1.679 and 1.362 times respectively.

Confusion matrix

	Actual		
Predicted		No	Yes
	No	18814	5712
	Yes	3735	7555

This gives us a module accuracy of 73.6%. With sensitivity of 57.0%, specificity of 83.44% and positive predictive value of 67.0%. This implies that out of the customers that were predicted to cancel their booking, only 67.0% of them actually did. Out of all the customers that canceled, only 57.0% was predicted accurately by our model but out of all the customers who did not cancel, 73.6% was predicted accurately by our model.

4.2.2. Random Forest Model

```
Confusion Matrix and Statistics


pred.randomForest    No   Yes
              No  20660  7237
              Yes  1889  6030

               Accuracy : 0.7452
                 95% CI : (0.7407, 0.7497)
    No Information Rate : 0.6296
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4043

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4545
            Specificity : 0.9162
         Pos Pred Value : 0.7615
         Neg Pred Value : 0.7406
             Prevalence : 0.3704
         Detection Rate : 0.1684
   Detection Prevalence : 0.2211
      Balanced Accuracy : 0.6854

       'Positive' Class : Yes
```

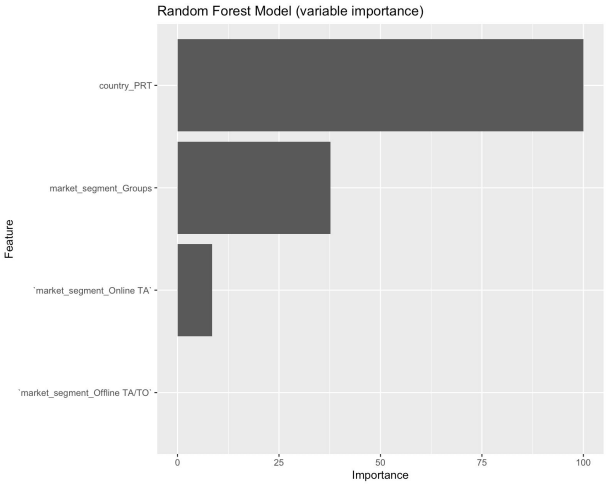

Figure 11: Random forest model and feature importance for hypothesis 2

Using a 5-fold cross-validation Random Forest classification tree model, we are able to obtain a model that predicts the cancellation of our test data with a 74.5% accuracy rate. It also has a 45.4% sensitivity rate, a 91.6% specificity rate, and a positive predictive value of 76%. When evaluating the feature importance of the variables used, "market_segment_Offline TA/TO" showed to be of low importance. It is however still included since the created model showed that it is the most accurate when all the IVs are included.

4.3. Hypothesis 3: Booking cancellation rate is higher during months where the hotel is less busy

Canceled = $\beta_0 + \beta_1$ * hotel (X_1) + β_2 * dayofmonth (X_2) + β_3 * weeknumber (X_3) + β_4 * month (X_5) + β_5 * year (X_6)

4.3.1. Logit Model

```
Call:
glm(formula = cancelled ~ hotel + dayofmonth + weeknumber + month +
    year, family = binomial(link = "logit"), data = trainData.1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2174  -1.0027  -0.8047   1.2985   1.8500

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       102.239687  28.938143   3.533 0.000411 ***
hotelResort Hotel  -0.616562   0.015991 -38.557  < 2e-16 ***
dayofmonth          0.045639   0.002777  16.433  < 2e-16 ***
weeknumber         -0.333901   0.018454 -18.094  < 2e-16 ***
monthAugust         5.712332   0.321922  17.744  < 2e-16 ***
monthDecember      11.464708   0.643546  17.815  < 2e-16 ***
monthFebruary      -3.120906   0.161155 -19.366  < 2e-16 ***
monthJanuary       -4.744006   0.242553 -19.559  < 2e-16 ***
monthJuly           4.223016   0.240616  17.551  < 2e-16 ***
monthJune           2.901055   0.164440  17.642  < 2e-16 ***
monthMarch         -1.867866   0.088816 -21.031  < 2e-16 ***
monthMay            1.367242   0.086127  15.875  < 2e-16 ***
monthNovember       9.868946   0.566563  17.419  < 2e-16 ***
monthOctober        8.689446   0.483334  17.978  < 2e-16 ***
monthSeptember      7.249074   0.403862  17.949  < 2e-16 ***
year               -0.048513   0.014292  -3.394 0.000688 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

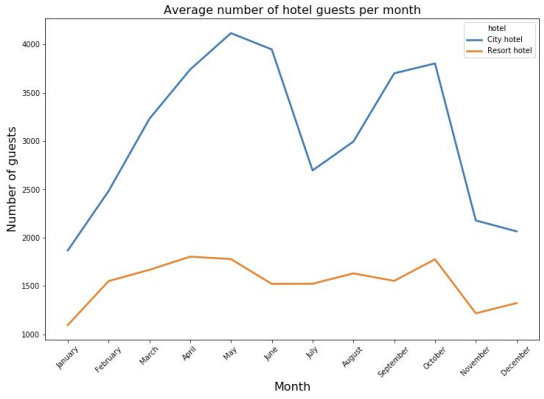

Figure 12 and 13: Logit model for hypothesis 3 and customer traffic by month

By referring to Figure 13, we can see that the months with fewer guests on average are January, July, November and December, while the busier months are May and october. Looking at the summary output of the regression analysis, there seems to be no observable pattern in the probability of booking cancellation across the months. coefficient estimates for the month of January, July, November and December are -4.744,4.223,9.868,11.464 respectively while that of the months May and October are 1.36,8.68 respectively. We can only observe that during the month of December, one of the months with fewer guests, the increase in odds of booking cancelation is the highest and in May, the increase in odds of booking cancelation is the lowest.

Confusion matrix

	Actual		
Predicted		No	Yes
	No	22142	12953
	Yes	348	374

This gives us a module accuracy of 62.9%. With sensitivity of 2.806%, specificity of 98.45% and positive predictive value of 51.8%. This implies that out of the customers that were predicted to cancel their booking, only 51.8% of them actually did. Out of all the customers that canceled, only 2.806% was predicted accurately by our model but out of all the customers who did not cancel, 98.45% was predicted accurately by our model. With such a poor sensitivity score, we concluded that we can reject this hypothesis for the prediction of booking cancelations.

4.3.2. Random Forest Model

After tuning, the new model is: Canceled = $\beta_1 + \beta_2$ * hotel + β_3 * dayofmonth + β_4 * weeknumber

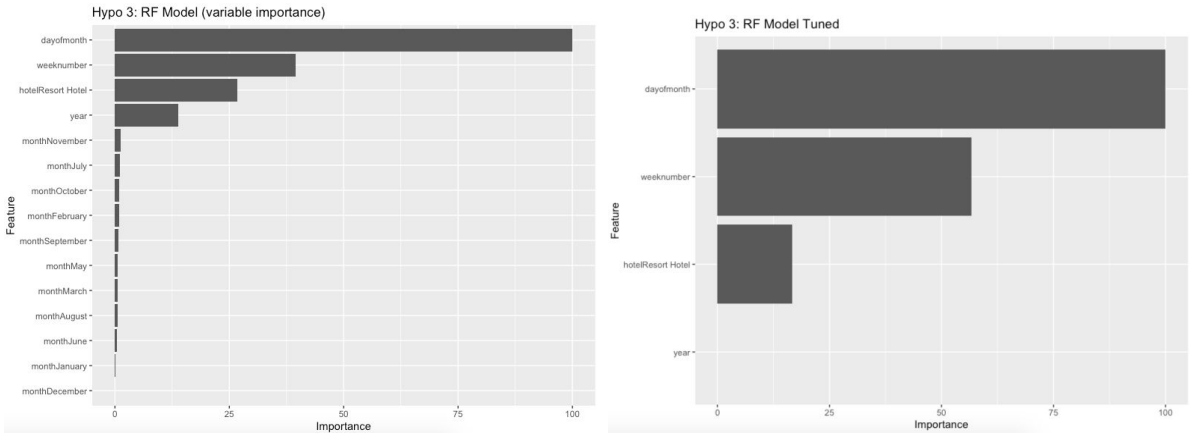

Figure 14:Feature importance for hypothesis 3

Upon evaluating the feature importance of the 5-fold cross-validation Random Forest classification tree model that we initially created, we found that the one-hot encoded splits of the "Months" variable was of low importance and made the model less accurate. We then recreated a similar model without it, which gave us a 66% accuracy rate. It also has a 24.5% sensitivity rate, a 90.4% specificity rate, and a positive predictive value of 60.1%.

```
Confusion Matrix and Statistics


pred.rf.3.new    No   Yes
          No  20389 10008
          Yes  2160  3259

               Accuracy : 0.6603
                 95% CI : (0.6553, 0.6652)
    No Information Rate : 0.6296
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1706

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.24565
            Specificity : 0.90421
         Pos Pred Value : 0.60140
         Neg Pred Value : 0.67076
             Prevalence : 0.37042
         Detection Rate : 0.09099
   Detection Prevalence : 0.15130
      Balanced Accuracy : 0.57493

       'Positive' Class : Yes
```

Figure 15: Random forest model for hypothesis 3

4.4. Hypothesis 4: Local guests from the "Groups" market segment with short term bookings are more likely to cancel their booking.

Cancelled = β_0 + β_1 * length_of_stay + β_2 * lead_time + β_3 * is_repeated_guest + β_4 * country_PRT + β_5 * `market_segment_Online TA` + β_6 * `market_segment_Offline TA/TO` + β_7 * market_segment_Groups + β_8 * market_segment_Direct + β_9 * market_segment_Corporate

To prevent overfitting of the model, we decided to trim the model by only selecting the top few variables in terms of their level of significance. Hence after tuning, our revised model is as follow,

Cancelled = β_0 + β_1 * length_of_stay (X_1) + β_2 * lead_time(X_2) + β_3 * is_repeated_guest (X_3) + β_4 * country_PRT (X_4) + β_5 * `market_segment_Online TA` (X_5) + β_6 `market_segment_Offline TA/TO` (X_6) + β_7 market_segment_Groups (X_7)

4.4.1 Logit Model

```
> summary(glmModel2)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4480  -0.7785  -0.5279   0.8950   2.9476

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -3.166e+00  3.213e-02 -98.537   <2e-16 ***
LOS                                2.959e-02  3.330e-03   8.885   <2e-16 ***
lead_time                          5.560e-03  9.057e-05  61.388   <2e-16 ***
is_repeated_guest                 -1.170e+00  6.099e-02 -19.185   <2e-16 ***
country_PRT                        1.949e+00  1.954e-02  99.722   <2e-16 ***
`\\`market_segment_Online TA\\``   1.597e+00  2.951e-02  54.123   <2e-16 ***
`\\`market_segment_Offline TA/TO\\`` 5.083e-01  3.227e-02  15.749   <2e-16 ***
market_segment_Groups              1.257e+00  3.364e-02  37.372   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110180  on 83573  degrees of freedom
Residual deviance:  88494  on 83566  degrees of freedom
AIC: 88510

Number of Fisher Scoring iterations: 4
```

Figure 16: Logit model for hypothesis 4

Looking at the summary output of the logistic regression model, we can see that all the variables used are highly significant even at the 0.1% level of significance. The variables X_1, X_2, X_3, X_4, X_5, X_6 and X_7 have a coefficient estimate of 0.02959 , 0.00560, -1.170, 1.949, 1.597, 0.05083 and 1.257 respectively.

We can predict that a unit increase in X_1 will increase the odds of cancellation by 1.03 times, a unit increase in X_2 will increase the odds of cancellation by 1.0056 times and a unit increase in X_6 will decrease the odds of cancellation by 1.052 times. It is interesting to note that even though the increase in odds associated with these variables are small, it is still considered as a significant variable. We can infer that although the effect size of this variable is minute, there is still a strong relationship between the underlying distribution of independent variables and the dependent variable. Furthermore, this may be an indication that there is an interaction between these variables. A unit increase in X_3, will decrease the odds of cancellation by 0.310 times, a unit increase in X_4, X_5 and X_7 will increase the odds of cancellation by 7.021, 4.938 and 3.514 times respectively.

Confusion matrix

	Actual		
Predicted		No	Yes
	No	19193	6297
	Yes	3356	6970

This gives us a module accuracy of 73.05%. With sensitivity of 52.54%, specificity of 85.12% and positive predictive value of 67.50%. This implies that out of the customers that were predicted to cancel their booking, only 67.50% of them actually did. Out of all the customers that canceled, only 52.54% was predicted accurately by our model but out of all the customers who did not cancel, 85.12% was predicted accurately by our model. We then tried to use a seperate model which did not include variables X1, X2 and X6 since their effect size in the model is small. This gave us a model with a slightly lower accuracy rate of 72.05%, sensitivity rate of 50.92% and specificity rate of 84.5%.

4.4.2. Random Forest Model

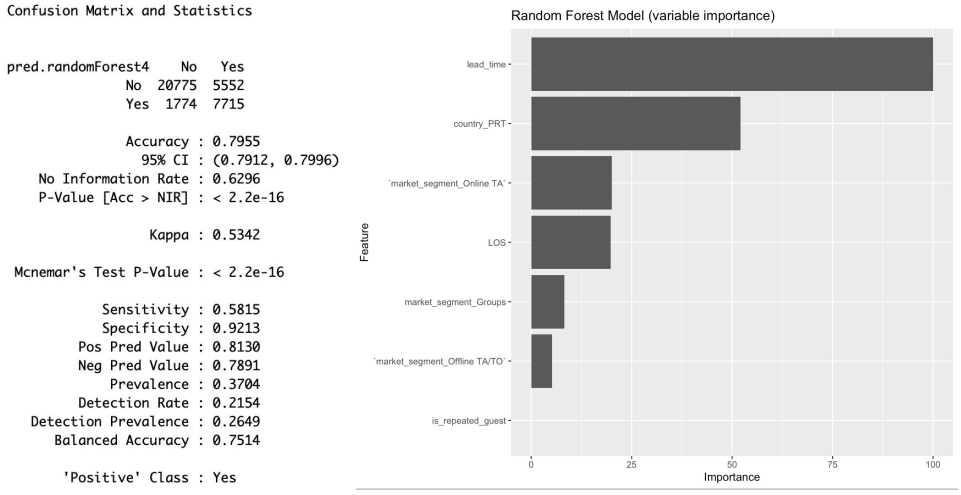

Figure 17: Random forest model and feature importance for hypothesis 4

Using a 5-fold cross-validation Random Forest classification tree model created from the CARET package in R, we are able to obtain a model that predicts the cancellation of our test data with a 79.6% accuracy rate. It also has a 58.2% sensitivity rate, a 92.1% specificity rate, and a positive predictive value of 81%. When evaluating the importance of the variables used, “is_repeated_guest” showed to be of low importance. It is however still included since the created model showed that it is the most accurate when all the IVs are included.

4.5. Findings and conclusion

The metric used by our group to evaluate our prediction models would be model accuracy and sensitivity rate. Since our objective is to increase hotel revenue through the prediction of cancellation, our model must not only be overall accurate, but more importantly, have a strong ability to detect **true positives**.

A model's accuracy is measured by its **sensitivity** and **specificity** rate, which in a nutshell, is the true positive rate and the true negative rate respectively. We should always aim to get both sensitivity and specificity as high as possible but depending on the problem at hand, one measure may be more important than the other.

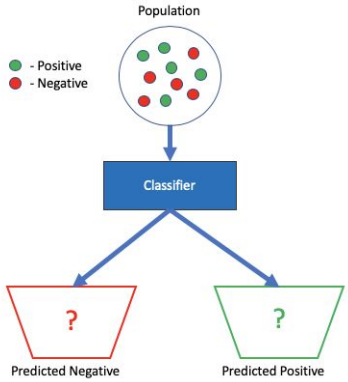

Figure 18: Breakdown of predictions in any given model

The benchmark accuracy rate we use would be to have 10% higher accuracy than the "No Information Rate" (NIR). For our dataset, the No Information Rate is 62.9%, this implies that the expected probability of a booking cancelation is 62.9% given no information other than the distribution of the target variable, "is_canceled". [5] Thus, we would only select models which outperform 72.9% in terms of accuracy.

	Accuracy (%)		Sensitivity (%)	
	Random Forest	Logistic Regression	Random Forest	Logistic Regression
Hypothesis 1	72.8	65.9	46.6	27.6
Hypothesis 2	74.5	73.6	45.5	57.0
Hypothesis 3	66.0	62.9	24.5	2.8
Hypothesis 4	79.6	73.05	58.2	52.5

Since the features in each model were selected based on the hypothesis we wanted to test, we can assume that the better the performance of the predictive models, the more plausible the hypothesis to be significant for our business problem. As such, we can see that the prediction models built for hypothesis 2 and hypothesis 4 prove to have stronger predictive power

[5] Tutorial: How to Assess Model Accuracy. (n.d.). Retrieved from https://www.hranalytics101.com/how-to-assess-model-accuracy-the-basics/#confusion-matrix-and-the-no-information-rate

compared to the others. We can conclude that whether a guest is local, the duration of stay and the market segment has a significant effect on booking cancelation.

Since the models built for hypothesis 3 performed poorly and did not deviate much from the NIR, our group has decided to reject hypothesis 3 and conclude that seasonality has little effect on booking cancelation.

5. Improvements

According to the models that we have created for our various hypotheses, we will be focusing on the fourth hypothesis, claiming that local customers who book for a short stay and from the "Groups" market segment are more likely to cancel their booking. Using our metric to evaluate the various models we have created, the Random Forest classification model for hypothesis 5 has the highest accuracy rate of 79.6% and sensitivity rate of 58.15%. This means that this hypothesis may be the most helpful in helping the hotel to successfully predict booking cancellations.

In the Random Forest model case, most of the model's high accuracy is attributed to its high specificity rate of 92.1% as compared to its mediocre sensitivity rate of 58.2%. This means that the model's accuracy is mainly derived from its ability to correctly predict non-cancellations. However, in the context of our business problem, the **sensitivity rate is of greater significance** in order to maximise profits through the ability to successfully predict booking cancellations. Thus, we will be looking at ways to improve the sensitivity of our final selected model.

5.1. Potential Model Improvements

Adopting the same dependent and independent variables equation from the selected hypothesis - **Cancelled = β_1 + β_2 * lead_time (X1) + β_3 * country_PRT (X2) + β_4 * is_repeated_guest (X3) + β_5 * total_stay (X4) + β_6 * market_segment (X5),** we continue to explore various other methods and algorithms to improve our sensitivity rate and our overall predictive performance.

Boosting is an ensemble technique that reduces variance and bias by using multiple models and training each subsequent model through learning from previous errors. There are many implementations of boosting algorithms such as Adaboost, Boosted Logit Regression, XGBoost, and Gradient Boosting[6].

AdaBoost was selected as the algorithm of choice as it is very effective in boosting the performance of decision trees on binary classification problems and have been referred to as

[6] Cruz, R. (2017). Why do we need XGBoost and Random Forest?. Retrieved 21 April 2020, from https://datascience.stackexchange.com/questions/23789/why-do-we-need-xgboost-and-random-forest

"discrete AdaBoost" since it's often used for classification rather than regression. It boosts performance by combining multiple weak learners into a single strong learner which allows us to better capture nonlinear relationships and improve our prediction accuracy. This will hopefully help us improve the sensitivity rate of our model, making it suitable to predict actual cancellations.

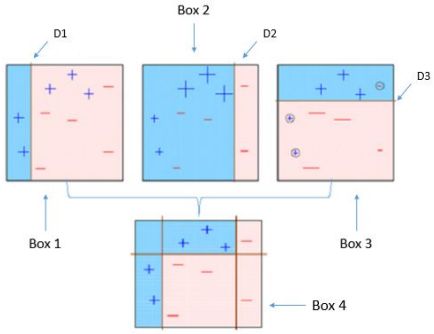

Figure 19: Adaboost algorithm learning

For comparison, we have also created a SVM model using the polynomial kernel, and a boosted logit regression model to be put side-by-side with our random forest model and adaboost model.

	SVM Polynomial	Adaboost	Boosted Logistic Regression	Random Forest
Accuracy	71.6%	**79.7%**	70.4%	79.6%
Sensitivity	42.9%	**83.1%**	46.0%	58.2%
Specificity	88.4%	77.8%	84.8%	**92.1%**

5.2. Final Model

```
Confusion Matrix and Statistics


pred.stackedada    No   Yes
            No  17537  2248
            Yes  5012 11019

               Accuracy : 0.7973
                 95% CI : (0.7931, 0.8015)
    No Information Rate : 0.6296
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5833

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8306
            Specificity : 0.7777
         Pos Pred Value : 0.6874
         Neg Pred Value : 0.8864
             Prevalence : 0.3704
         Detection Rate : 0.3077
   Detection Prevalence : 0.4476
      Balanced Accuracy : 0.8041

       'Positive' Class : Yes
```

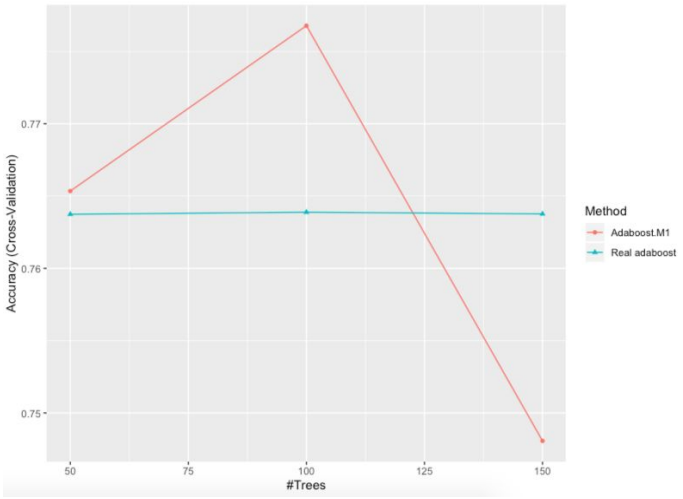

Figure 20: Adaboost model and number of trees used for maximum accuracy

The **Adaboost model** provides us with not only the highest overall **accuracy at 79.7%**, but also an incredible **sensitivity rate of 83.1%** which miles ahead of all other models including the previous best performing random forest model, and an improved positive predictive value of 68.7%. Therefore, the Adaboost model will be selected as the model of choice as it is a good fit for enabling hotels to meet the objective of increasing their revenue through **improving their hotel occupancy rates** since they will be able to predict true cancellations more accurately. Hotels will then be able to take necessary measures and set up arrangements to utilise and monetize the empty rooms. These measures could include heavily discounting the rates charged and providing free upgrades in order to utilize as many rooms as possible.

For the City Hotel and Resort Hotel, some measures that could be put in place would be to allow overbooking of hotel rooms, especially during peak periods with some reserve rooms in place. Assuming a total of 100 hotel rooms are predicted to cancel their bookings. With a sensitivity rate of 83.1%, the hotel could allow overbooking up till 83.1% or 83 rooms. With a positive predicted value of 68.7%, about 31 rooms could be set aside as reserve rooms or a contingency plan can be implemented for these people who were predicted to cancel their booking but actually did not.