INTRODUCTION

The project refers to a classification problem.The data refer to a random sample of room bookings in some hotel and whether the booking has finally cancelled or not,the purpose of the project is to create a model to classify whether a booking will be canceled or not.In total the dataset has 2000 observation and 16 variables,which are the follow

VARIABLE	TYPE
Booking_ID--Unique identifier for each booking	Character
number of adults--Number of adults included in the booking	numeric
number of children--Number of adults included in the booking	numeric
number of weekend nights--Number of weekend nights included in the booking	numeric
number of week nights-- Number of week nights included in the booking	numeric
type of meal--Type of meal included in the booking	factor(meal plan 1,meal plan 2,meal plan 3)
car parking space--Indicates whether a car parking space was requested or included in the booking	numeric
room type--Type of room booked	factor(room type 1,room type 2,room type 3 Room type 4 ,room type 5,room type 6 ,room type 7)
lead time--Number of days between the booking date and the arrival date	numeric
market segment type--Type of market segment associated with the booking	Factor(Aviation,complementary,corporate,offline,online)
repeated--Indicates whether the booking is a repeat booking	Numeric
P-C--Number of previous bookings that were canceled by the customer prior to the current booking	Numeric
P-not-C--Number of previous bookings not canceled by the customer prior to the current booking	Numeric
average price--Average price associated with the booking	Numeric
special requests--Number of special requests made by the guest	Numeric
date of reservation--Date of the reservation	Date
booking status--Status of the booking canceled or not canceled	Factor(not_canceled=1,canceled=0)

DESCRIBE STATISTICS

In this section it will be presented some describe statistics and plots about the dataset so to understand the data.Before presented,to mention that the variable booking_id removed because it does not play a role about the booking_status.Also the variable date removed because it has non-sense values that dont contribute to evaluate the model and to extract usefull infomation.

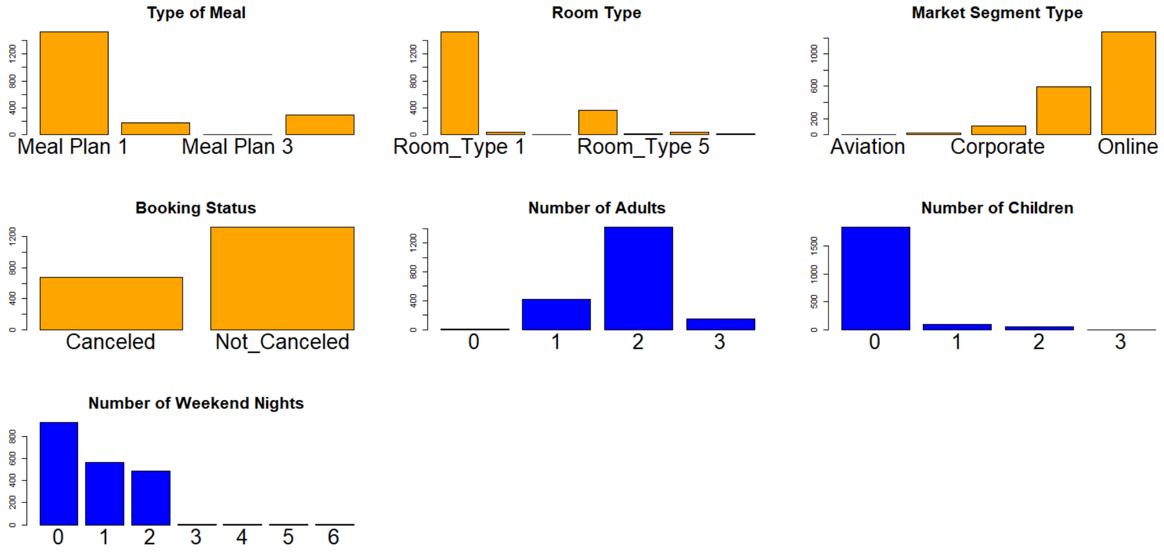


Diagram 1

From the barplot we can see that the majority type of meal is meal plan 1,the preference room type is room type 1.Also a lot of peaple book their reservation online and almost zero via aviation.In the data set most of the people dont cancel their reservation,the ususal number of adults is 2 and very few are three or 1 or even zero adults.In most reservation the number of children are zero and finaly the number of weekend nights are zero at their majority but also many people book for 1 or 2 weekend nights.

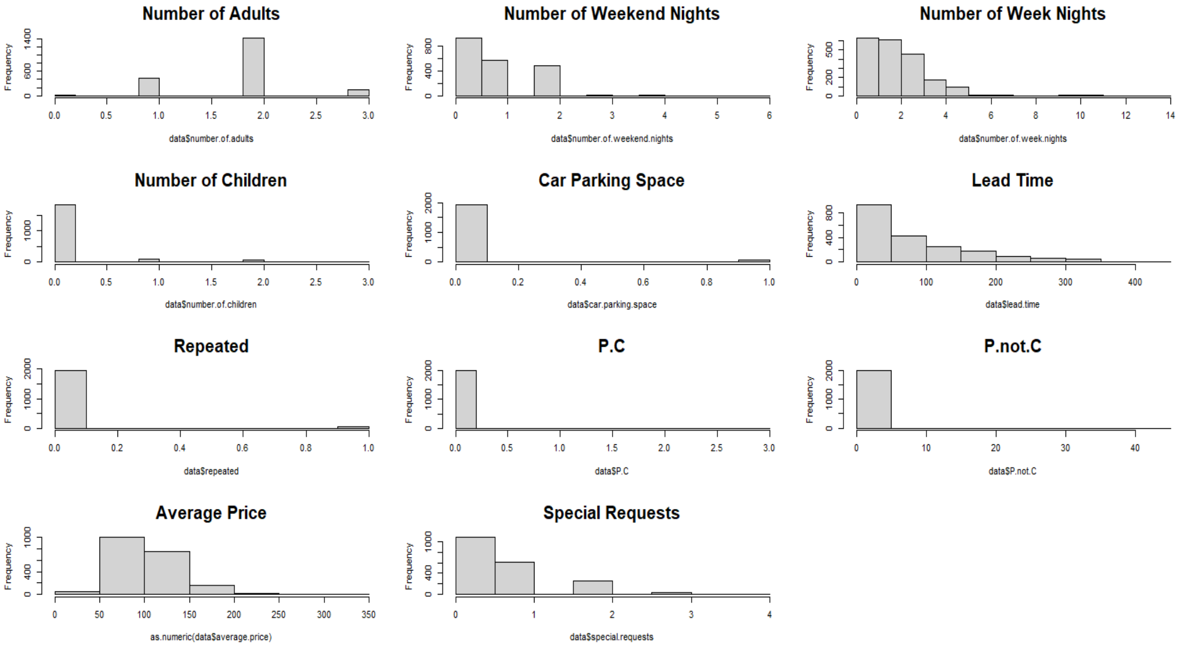


Diagram 2

There are the histograms of the variables that show us the distribution of each variable

Next it presented some graphs with the combination of the booking_status “canceled” or “Not canceled”

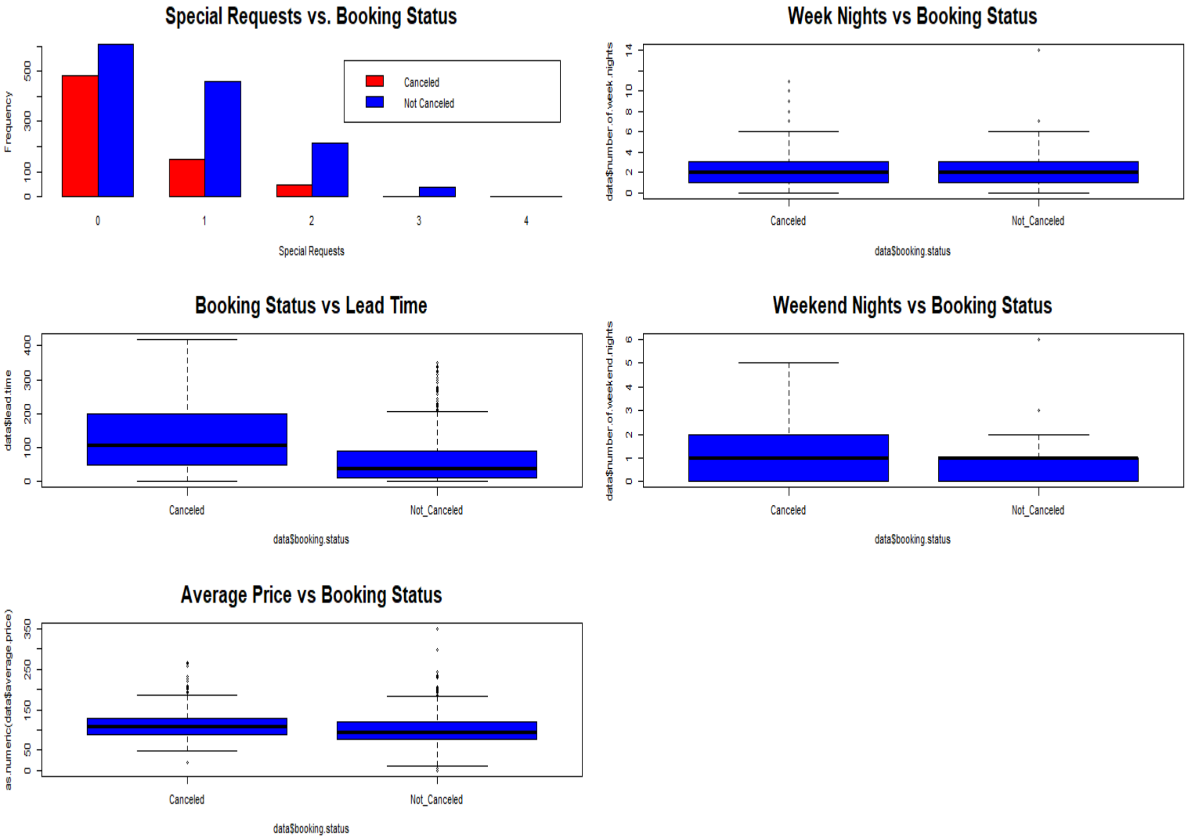


Diagram 3

We see in the barplot that as the number of special request increases the cancel or not cancel reservation decreases.In the other boxplots we see that relation of the variables with the booking status,as we see that the variability of the variables are bigger at the category canceled bookings.

Feature selection

In this section it will be presented the variable selection so to find the variables that contributes to the classify of the canceled or not.We do this procedure because we must to improve model prediction,avoid overfitting to the training data and select only the most important predictor variable.The model that do this job is the lasso logistic regrassion because we have a binary response.Above is the mathematical model that we implement.

$$L1(w,b|(x)|) = -\Sigma\left[y(i)\log\left(\sigma(z^{i})\right)+1-y(i)\log\left(1-\sigma(z^{i})\right)\right]+\lambda\sum|Wj|$$

But wee need to find the optimal λ to do this we do a procedure than is cross validation and the best lambda is the lamda that reduces the binomial deviance.

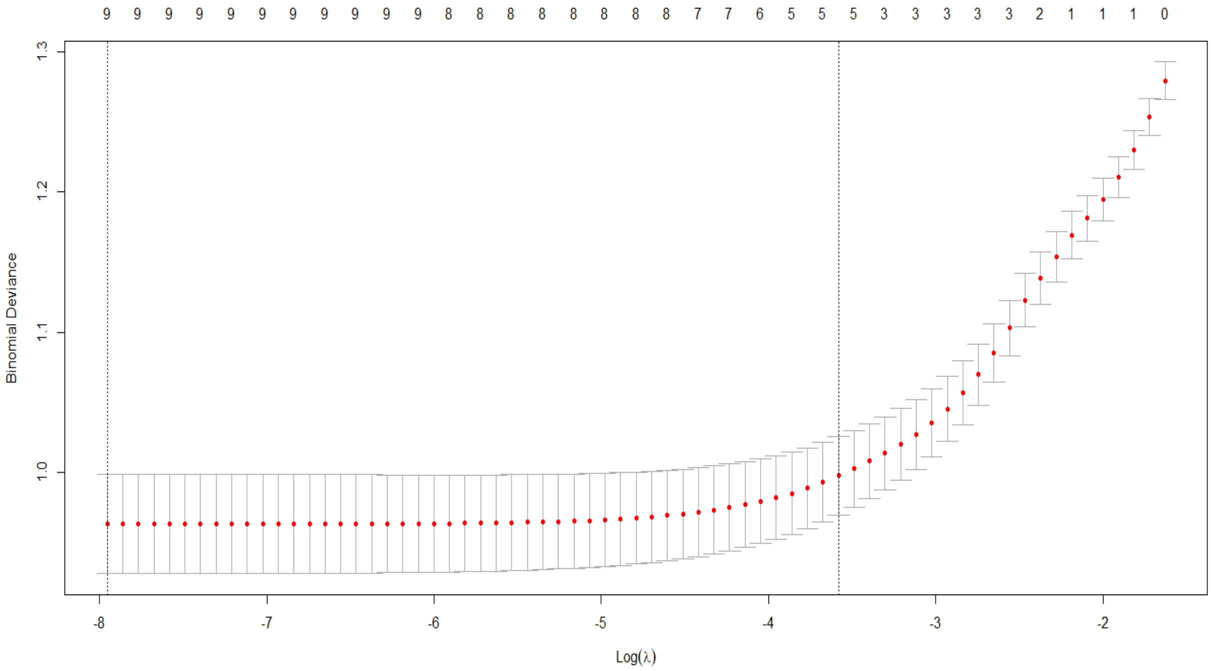


Diagram 4.

The optimum lambda is 0.0003511384 and after we fit the model to the data it reduces the unimportant coeficients to zero.The variable that the keeps it out is the market.segment.type the room type and the type of meal it keeps it out.

Then we see the correlation plot of the remaining variables.

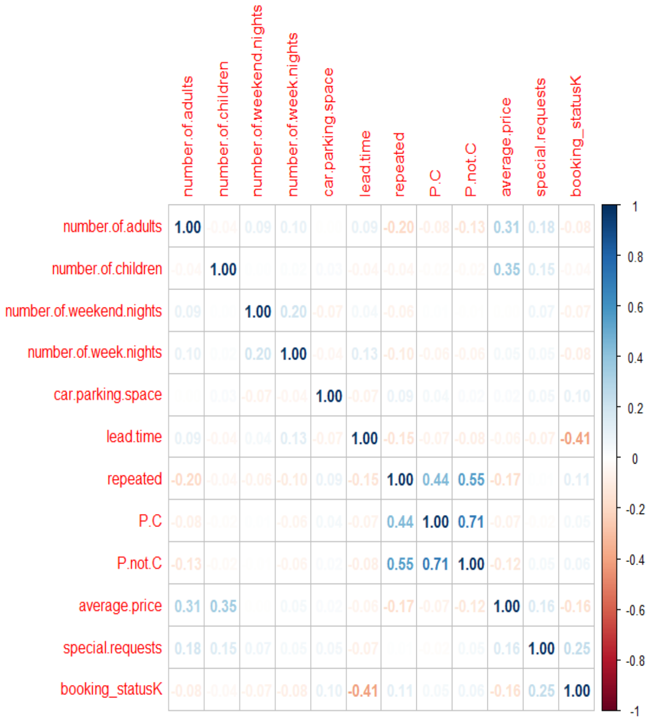


Diagram 5

In the correlation plot we see that we dont have strong correlation in the booking_status with the remaining variables.

Model selection and evaluation

In this section it will be presented the different classification models and their performace to classify whether a new booking will be canceled or not.The method that it will be implement is k- fold cross validation fold

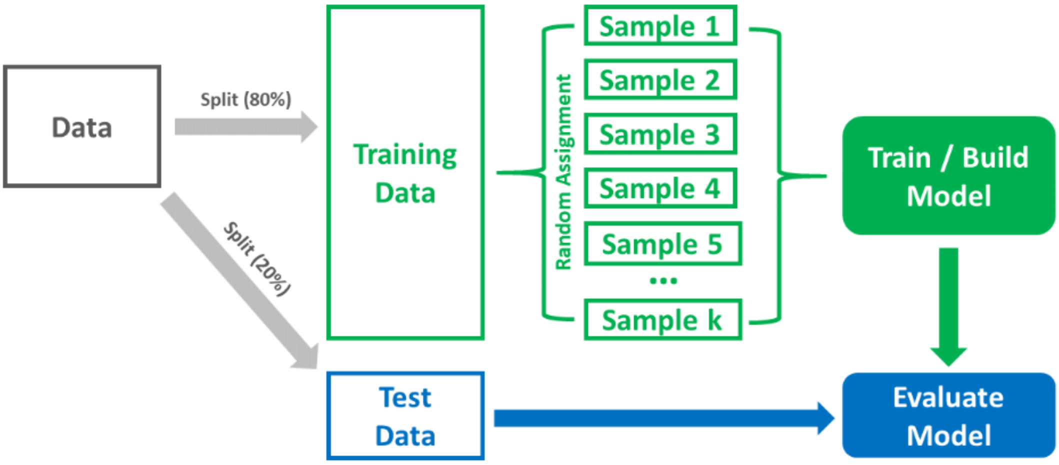


Diagram 5

There is the illustration of the procedure,which is at first to split the dataset 80% to train set,and 20% on test set where we fit the model to see how accurate we are about the prediction base on the training set.Randomly split the dataset into more sub(samples) into K folds.

The the optimal K is defined by grind search,where we take some values of k for example k=5,7,8,10.

For example if we take k=5 fold procedure ,where the model is trained into k-1 parts and 1 test into test part,we repeat the procedure 5 times rotating the test set.Finally determine an expected performance metric the accuracy.

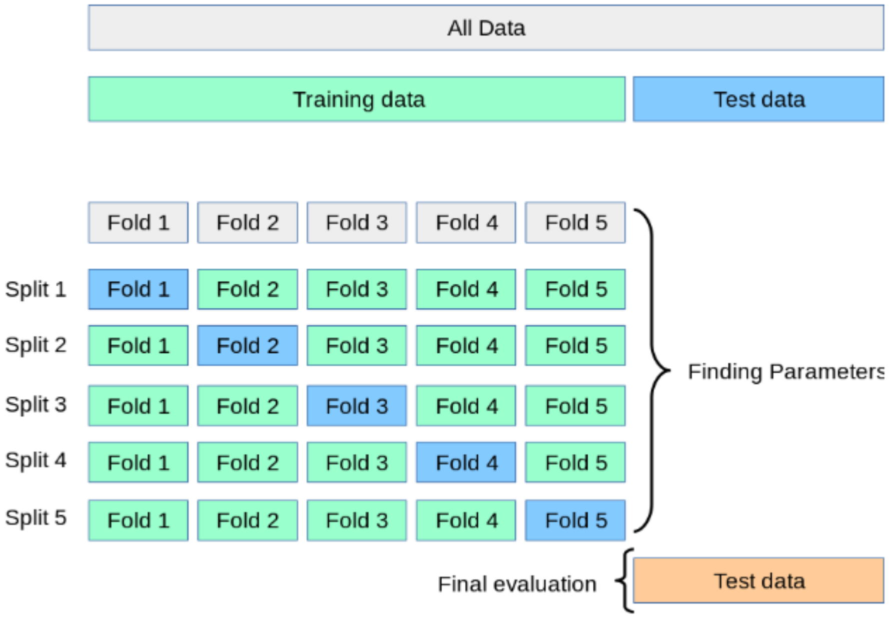


Diagram 6:Example of the procedure into training dataset for k=5

We repeat this procedure to find the optimal K that has the biggest accuracy into the training test and then we fit the model into the test set.

The fist model that it will implement into the above procedure is random forest model for classification.We must to see what is the optimal K into the samples base on the accuracy.

Accuracy	Number
0.8107590	4
0.7944815	6
0.8094838	8
0.8169823	10

We see that the best fold to split the samples is 10 because we have the bigest accuracy now into this K=10,perform the model into the test set to see the performance of the model and the accuracy.

After we implement the model we draw a confusion matrix with the predict values of the model and the actual.

	Actual values	
Prediction	0	1
0	89	21
1	46	243

Confusion matrix

In this matrix the prediction values about the model and the actual values base on the test set.

To indetify that 0 is canceled and 1 is not canceled.We have a big accuracy of the model at 83% and the confidence interval is (79%,86%) about the accuracy,which is big and predict very well.

About the importance of the variables that contributes in the random forest algorith is the above.

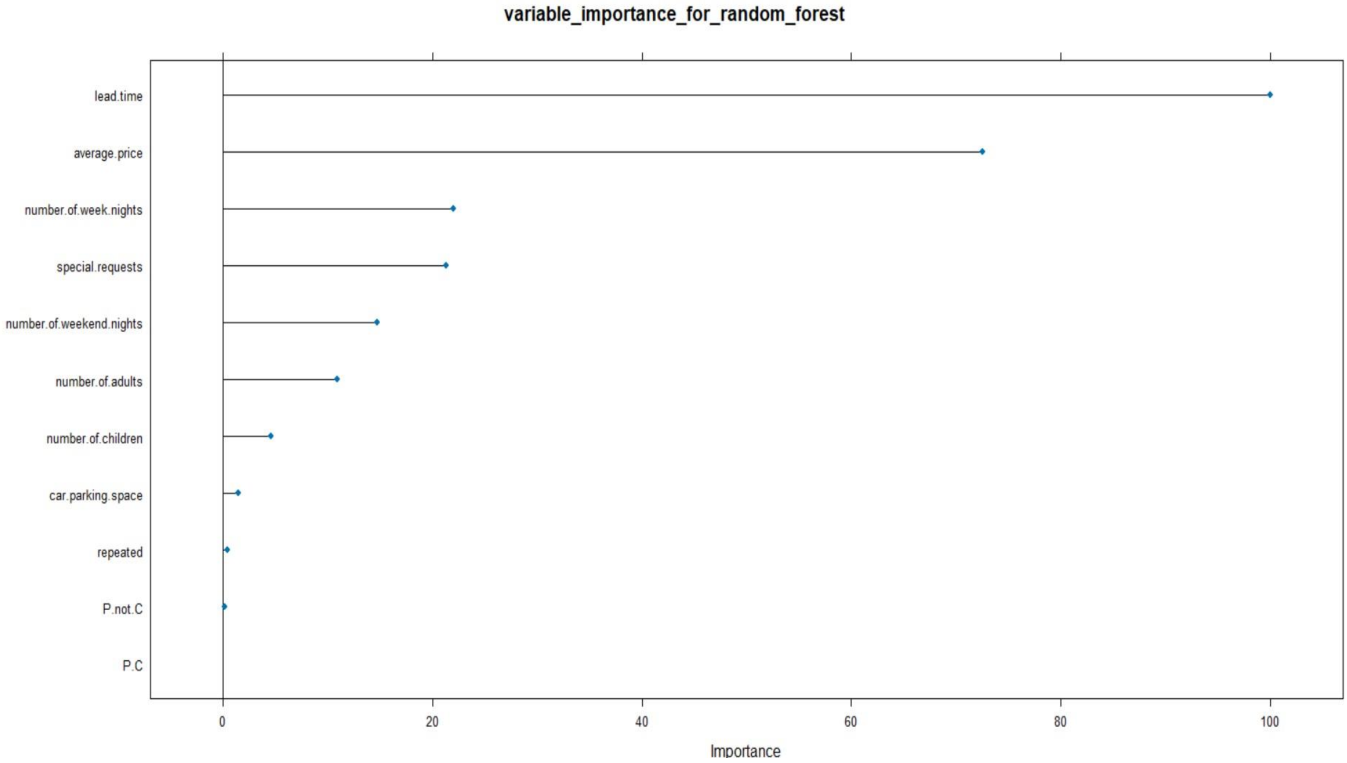


Diagram 7

Base of this diagram we see that the variables that dont play role in the random forest algorithm is PC,P not C and repeated are the variables that have zero importance.

Another performance metric is the ROC curve which we can see if the line is close to 1 we have good predicting ability and if the line is away from 1 is week the predictive ability.

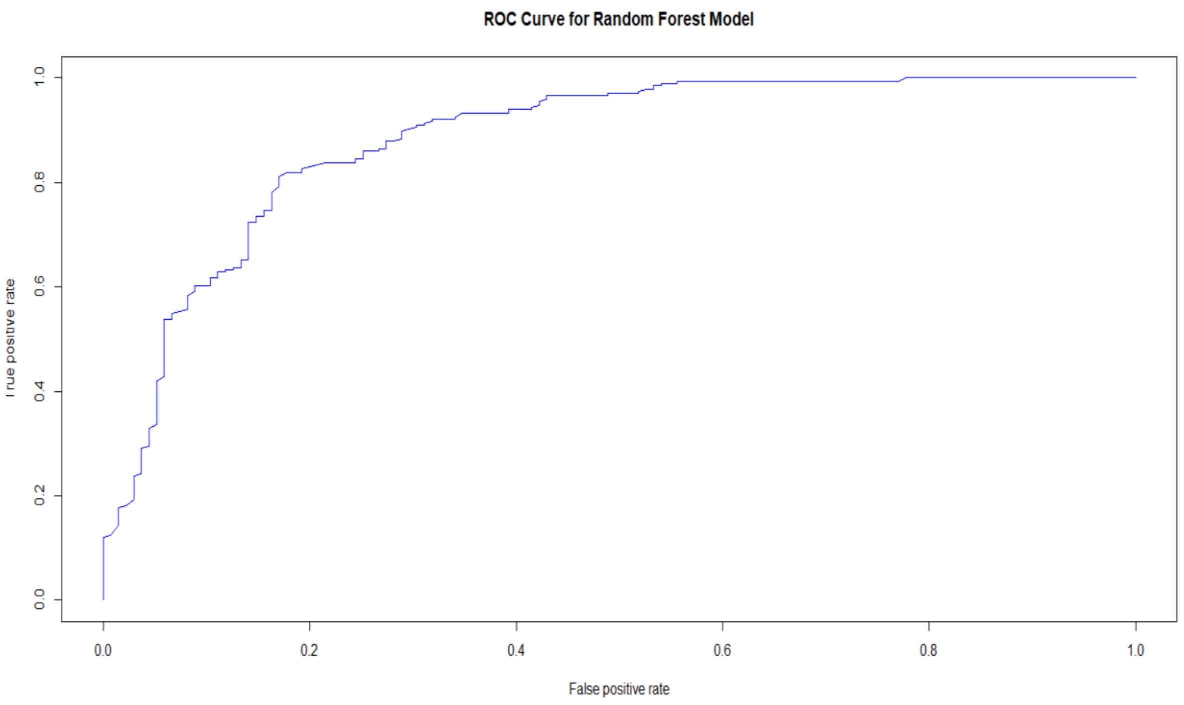


Diagram 8:ROC curve

We see that the line is close to 1 we have strong predicting ability of the random forest

The next we perform is KNN k-nearest neighborhood method to see the predicting ability, the procedure is the same but we scale the data to have the same units of measurements,because the KNN method have the euclidian distance for measure.Now in the K-NN we need to specify the number of neighborhoods and the number of folds as always.

To be more specific,illlustrate what the numver of the neighborhoods means.

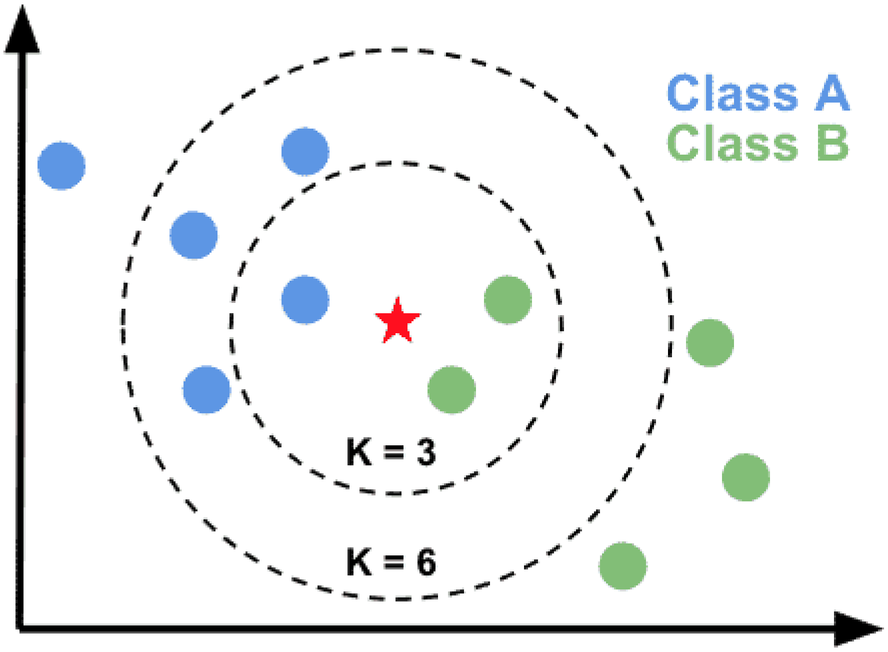


Diagram 9

This is an example of different k neighborhoods and if k=3 the prediction is green and if k=6 the prediction is blue.

So we need to speciffy the optimum k and we will do it base on the accuracy.To summarize,we need to specify the optimal K neighborhood via method that called grind search and the optimal K fold.

First we try for cross validation fold for k=2 and then we do grind search do find the optimal k neighborhood.

K	Accuracy	fold
9	0.7575	2
9	0.766868	4
9	0.7687501	6
7	0.7662455	8
3	0.7662836	10

We see from the table that the optimal fold and neighborhood K is for K=9 and fold 6 that we have the biggest accuracy.

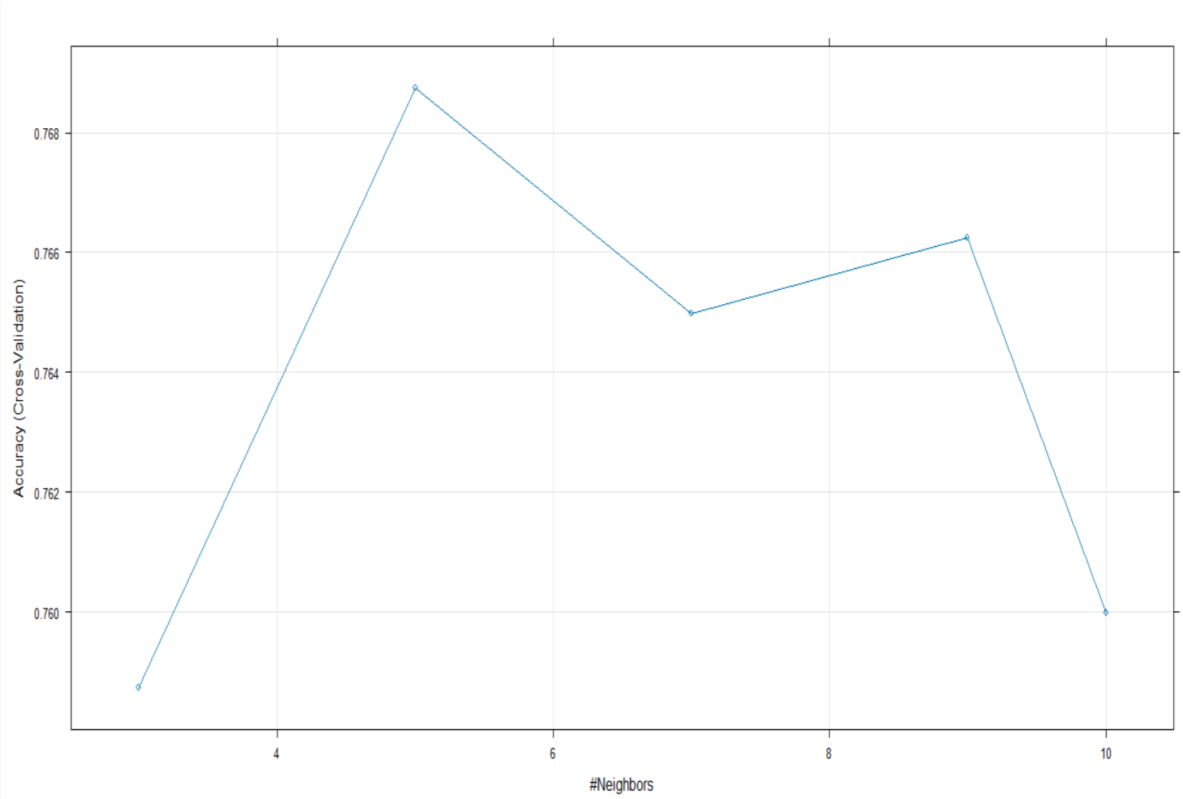


Diagram 10

We see that the biggest accuracy we get it for neighbors=6

When we predict the data in the test data and compare it with test set we get accuracy 0.78 with confidence interval 95% in range (73%,82%)

	Actual values	
prediction	0	1
0	82	25
1	62	231

Confusion matrix

This matrix tell what we have predicted right and what we have predicted wrong and calculate the accuracy.

And base the k-NN we see what select for variable importance

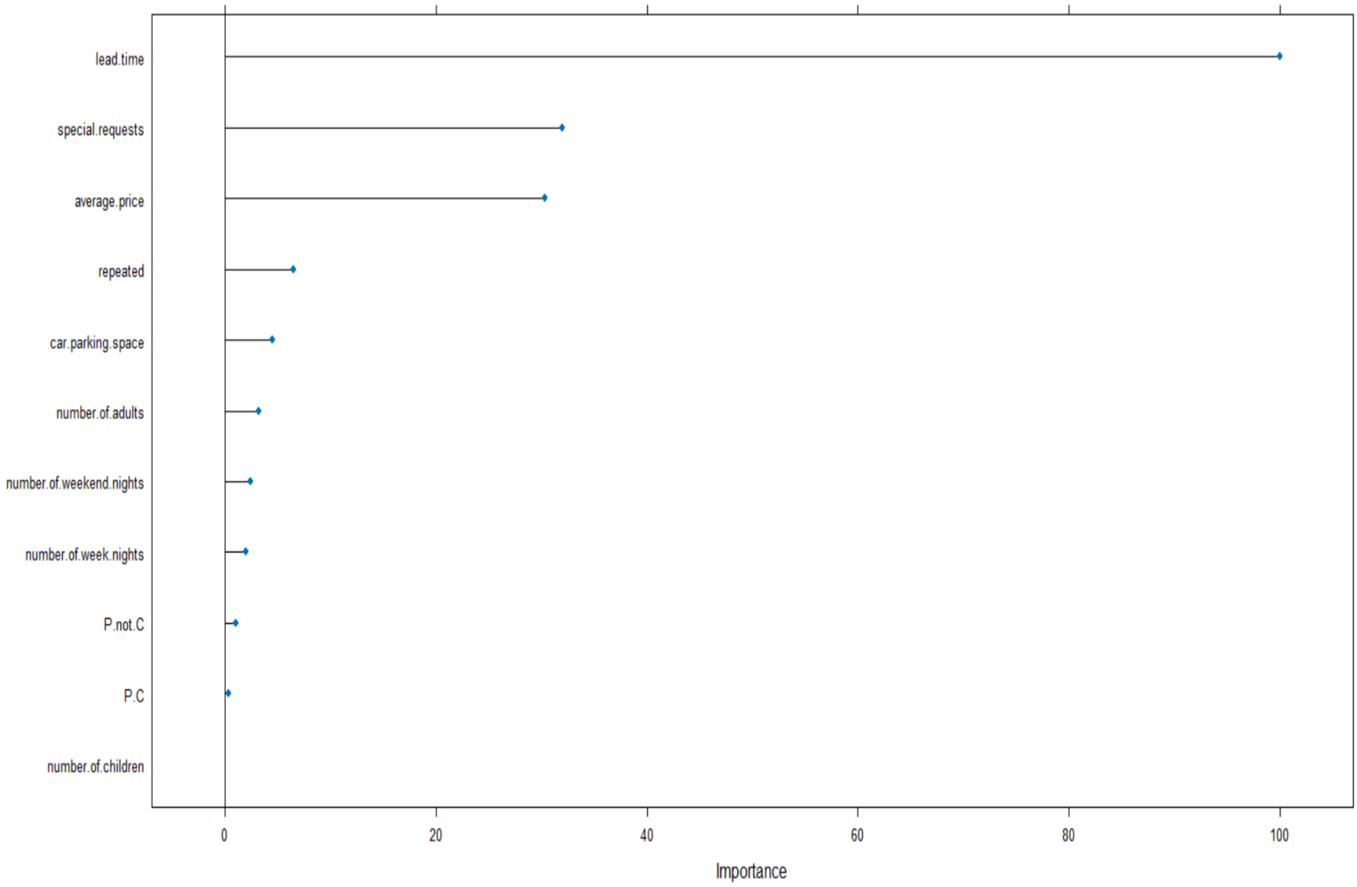


Diagramm 11 the importance of the variables base on the classification K-NN

We can see that zero importance have the number of children the pc,p no c and very little importance have the number of nights the number of weekend nights.

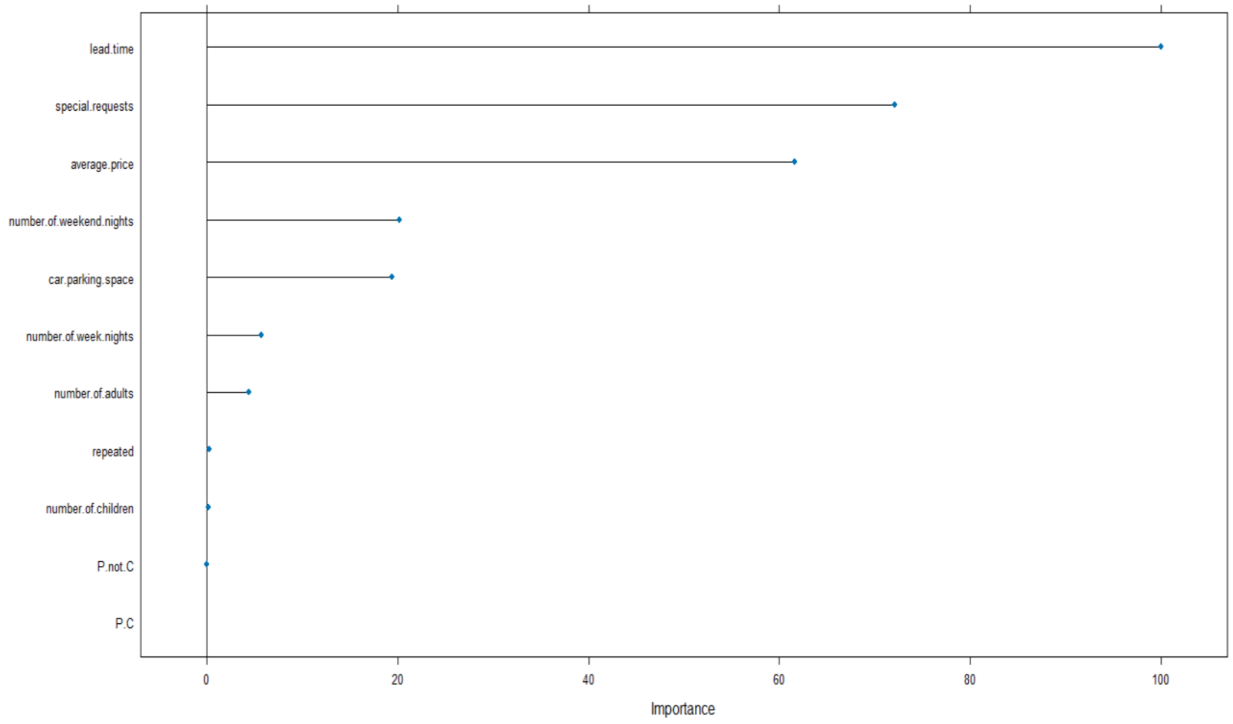
The next classification method is called glm and first we will see what fold has the best accuracy to use.

Accuracy	Number
0.7712	4
07681	6
0.7693	8
0.7713	10

We see that the for fold 10 we have better accuracy,so we use the model glm in fold 10.Following is the confusion matrix between the predicted and actual values.

	Actual values	Actual values
Prediction	0	1
0	75	18
1	69	238

There is the confusion matrix from the predicted and actual values.The accuracy is 78% and the misclasification rate is 22%.There is the variable importance base on the glm model



Conclusion

From the models that we use for the classification the best is the random forest,because have greater accuracy than the others models about 83%.The folds to train the model is 10-fold cross validation .The variables that are more importan to classify whether a booking will cancel or no are the following

Number of adults

Number of children

Number of weekend nights

Number of week nights

Car parking space

Lead time

Repeated

P.C

P not C

Average price

Special request