**Capstone Project**

**“Falcone Airlines Customer Satisfaction”**

**By: khaled Majzoub**

**Table of Content:**

**Executive Summary…………..………………………...Pages 1,2,3**

**Model Selection.………………………………………....**

**EDA…………………………………………………Pages 4,5,6,7,8,9**

**CART……………………………………………….Pages 10,11,12**

**Random Forest…………………………………...Pages 13,14,15**

**Logistic Regression……………………………..Pages 16,17**

**Models Comparisons……………………………Pages 18,19,20**

**Model Interpretation………….………………....Page 21**

**Conclusions and Recommendations……………….Pages 22,23**

**Executive Summary:**

* **Problem Statement:**

“Falcon Airlines” a US airline carrier aims to determine the relative importance of each parameter with regards to their contribution to passenger satisfaction.

They made a survey to around 90k of their customers and asked them 23 questions, including a question if they are (in overall flight experience) “satisfied” or “neutral or dissatisfied”.

The Answers for 20 questions are based on likert scale (Excellent, good, etc)

The remaining 3 questions were based on numeric answers.

However, the dilemma is that they need to know the most significant variables that affect customers’ satisfaction and the relationship between the variable and satisfaction i.e, if the airlines manages to increase the satisfaction of one of the variables by 1 point the overall satisfaction result will differ.

* **Brief Description of Methods:**

First, I explore data and analyze it, fix outliers and structure and missing data, understand each variable and correlations between them.

Then I start with variable transformation by transforming the variables to only “positive” and “negative.

I transform “excellent”, “good”, “acceptable” to “positive’’.

“extremely poor”, “poor’’, “needs improvement” and “missing values” to “negative”.

After that I make another analysis for the new data and the correlations between them to get new insights.

After studying all variables and focusing on the most important, I excluded 10 variables based on my intuition and kept only the most significant ones. This led to reduced complication and led to the excluded variables not affecting the overall customer satisfaction.

Starting with models and after splitting the data to train and test, I make a CART, Random Forest and logistic regression models then evaluate them.

I make comparisons based on KPIs and the most important variables that fit my logic.

Finally, I will interpret the models and make final recommendations.

* **Final Insights:**

I focus on the most important 4 variables to minimize the effort and time and maximize the benefit and return on investment.

All three models have Inflight entertainment (all number 1).

**Variables Importance:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CART** | **FOREST** | **Logistic Regression** |
| 1 | Inflight Entertainment | Inflight Entertainment | Inflight Entertainment |
| 2 | Ease of Online booking | Food n Drink | Ease of online booking |
| 3 | Check in service | Check in service | Online boarding |
| 4 | Online boarding | Ease of online booking | Check in service |

Forest model had the variable “food n drink” which is more reasonable to satisfy customers than “online boarding”.

**Model comparisons:**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric/Model | **CART Train** | **Forest Train** | **Logistic Regression Train** |
| Error rate | 0.2563 | 0.2011081 | 0.2544623 |
| Accuracy | 0.7437 | 0.7988919 | 0.7455377 |
| Sensitivity | 0.7154 | 0.7984987 | 0.6989455 |
| Specificity | 0.7652 | 0.7991585 | 0.7872234 |
| AUC | 0.7692007 | 0.8408412 | 0.808 |
| KS | 0.4842302 | 0.5907522 |  |
| Gini | 0.2817875 | 0.486449 |  |

**From Coefficient variance increase percentage of variables**

|  |  |
| --- | --- |
| inflight\_entertainmentpositive | 5.66501491 |
| ease\_of\_onlinebookingpositive | 4.39055963 |
| checkin\_servicepositive | 2.64212727 |
| online\_boardingpositive | 1.70330809 |

* **Recommendations:**

I recommend “Falcon Airline” should concentrate on enhancing the following variables:

1- Inflight Entertainment

2- Food and Drink

3- Ease of online booking

4- Online boarding

5- Check in Service.

The most important 2 variables are “inflight entertainment” and “food and drink”, because “inflight Entertainment” is easy to be enhanced and does not require a lot of time and money and enhancing it is essential to all customers.

As for “Food and Drink” also this is an easy variable to be upgraded with little effort and money, they can satisfy the 50% of the passengers that are dissatisfied from this service.

80% of the passengers are satisfied from the last 3 variables, which means enhancing them will benefit the overall outcome.

Enhancing online booking and online boarding is easy and doesn't require a lot of money and time.

Overall customer satisfaction is good, “Falcon Airflight” is doing good and little effort can increase the overall satisfaction which will positively affect the reputation of “Falcon Airline” and the return on investment**.**

**Model Selection:**

* **Checking data and preparing it:**

I checked both datasets ( flight data ) and (survey data) changed variables to correct format ( factors ) and left “delays in mins” to integers.

I merged both datasets into 1 new dataset named it (fli\_sur) using the common “ID” variable between. And deleted the ID variable.

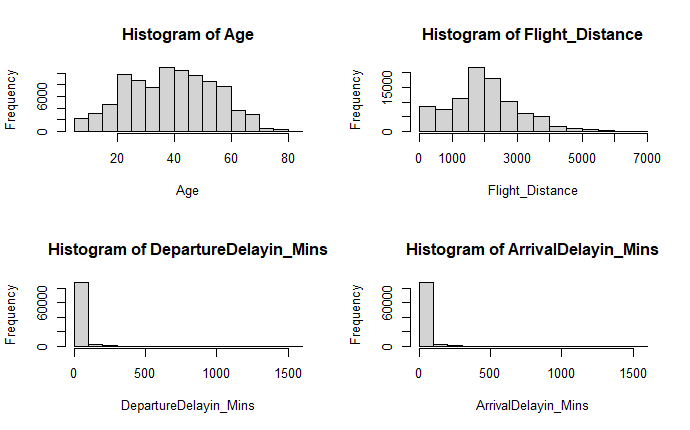
We have missing values, less than 0.004 from whole data, so i deleted all those observations.

There are dashes in some of the variables, maybe customers skipped or forgot to fill it. I will replace the dashed into “neutral”

Now the data is ready to EDA

* **EDA**

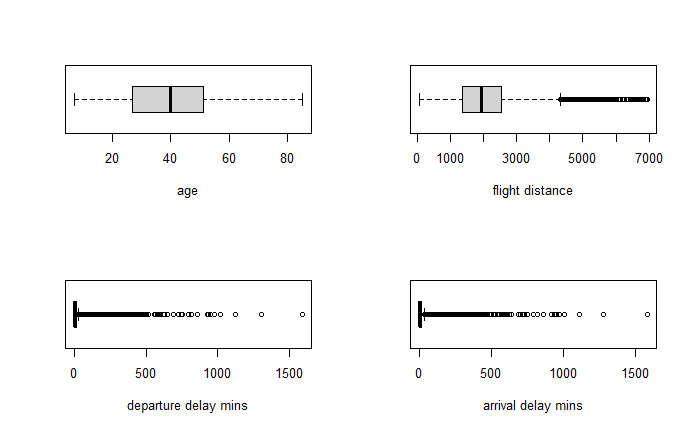
**Univariate:**



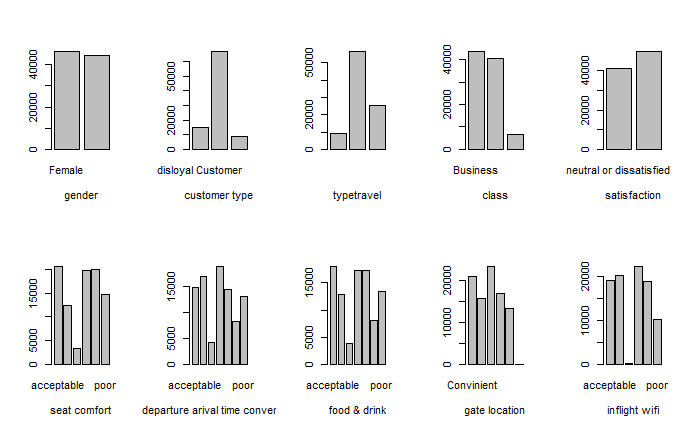
**\* Age is normal distribution**

**\* Flight distance have outliers**

**\* delays are skewed a lot and it makes sense, because of unexpected delay**



**\* Outliers in the delayed must be treated**



**\* genders are almost equal**

**\* loyal customer are more than not loyal**

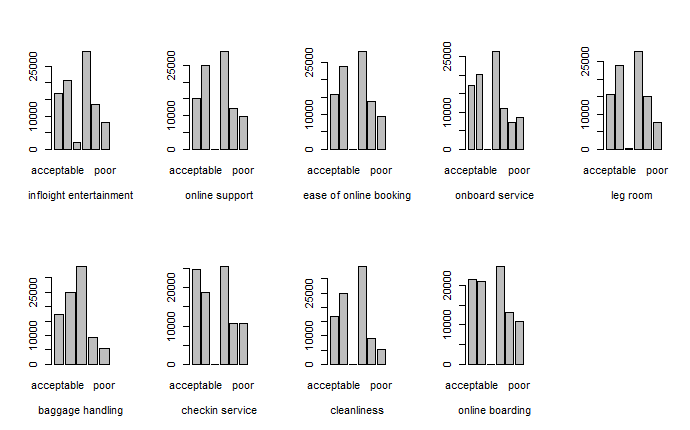
**\* business travel is more**

**\* more satisfied than not satisfied**

**\* more seat comfort positive**

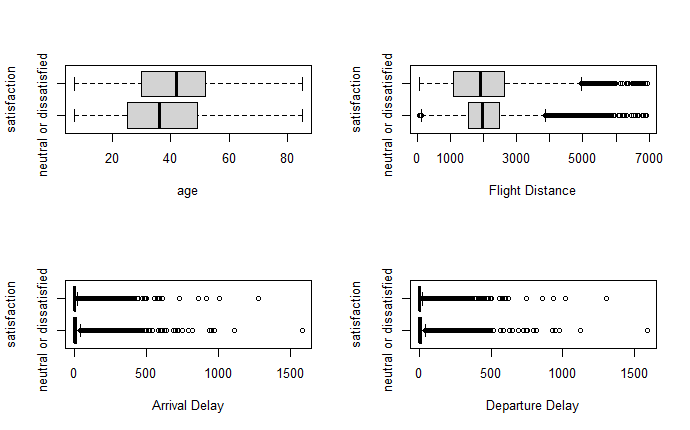
**\* more positive inflight**

**\* inflight wifi**

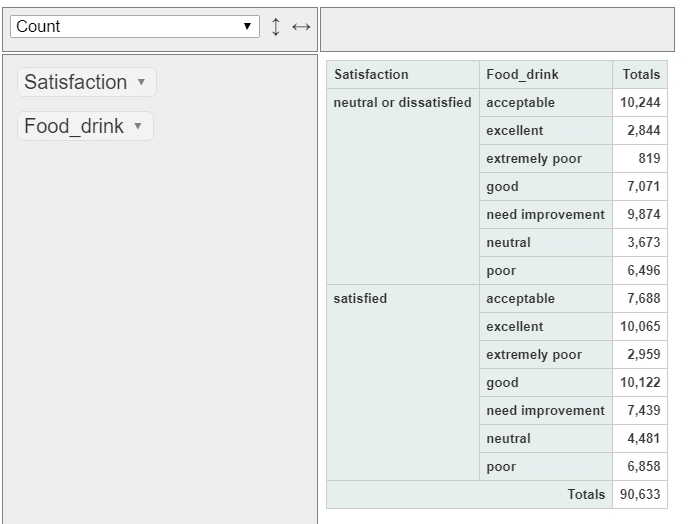


**More positive**

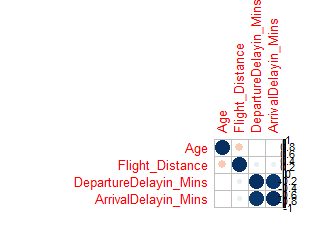
**Bivariate:**

**Correlations between Satisfaction and** 

**numeric variables**

****

**\* food n drink are almost 50% \* around 10 dissatisfied says it needs improvement , so easily to satisfy them**

****

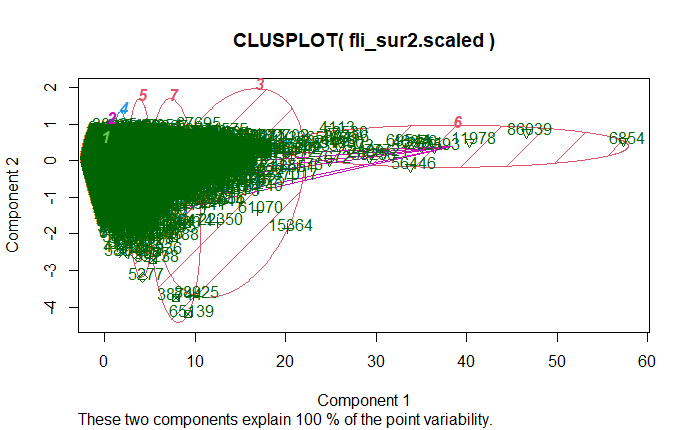
**\* negative and small correlations between them.**

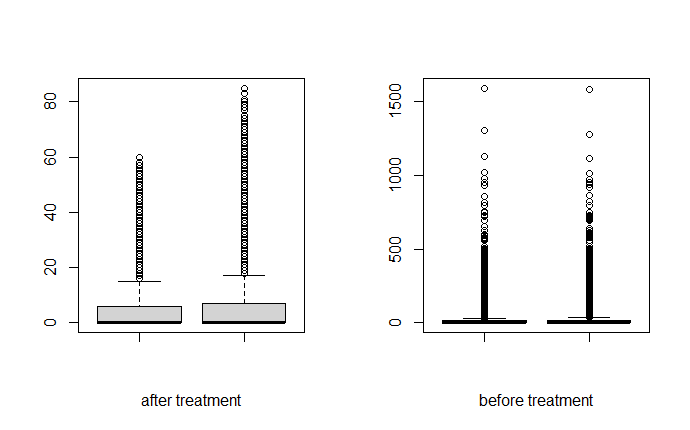
**\* not show a lot of details**

* **Removal of unwanted variables**

I removed based on my intuition and my flight experience variables that will not affect customer satisfaction and kept only: 1- departure delays 2- arrival delays 3- satisfaction 4- time conventinet 5- food n drink 6- wifi - inflight entertainment 7- online boarding 8- ease of booking 9- baggages 10 - check in 11- cleanliness 12- online booking 13- check in service

* **Outlier Treatment**

****

****

**After treating the outliers , this is the best I could do and deleted around 9000 observations!**

* **Variable Transformation**

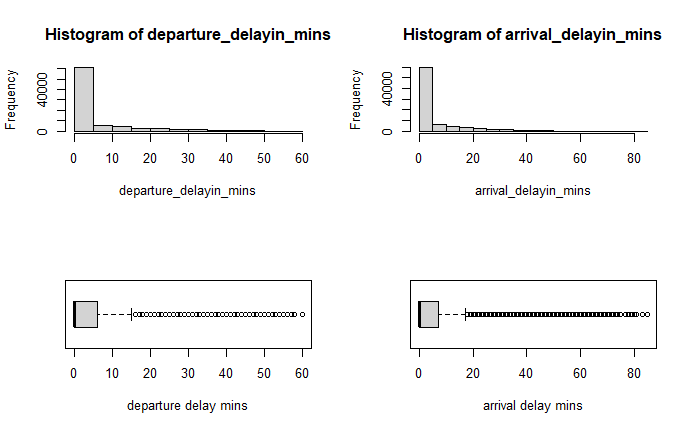
I transformed all factors into 2:

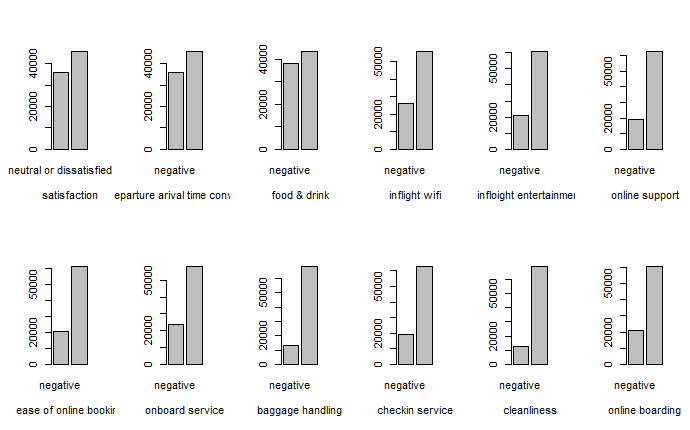
“Positive” = “exellent” , “good” and “acceptable”

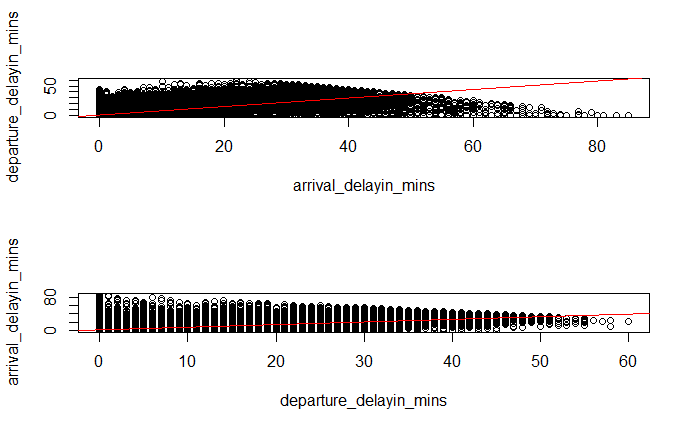
“Negative” = “extremelly poor” , “poor” , “need importvment” and “neutral”

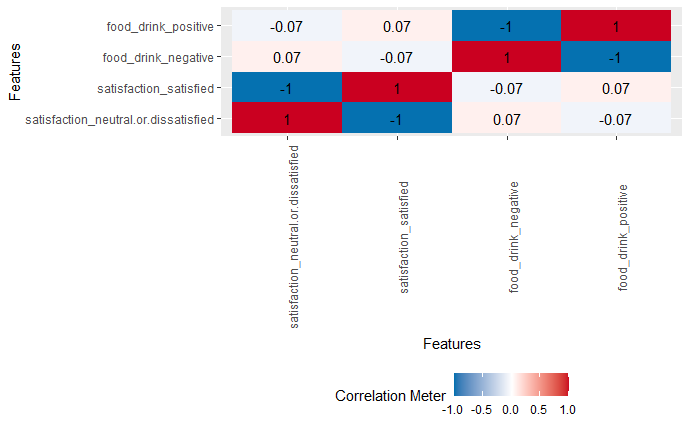
So now I have only positive and negative.

* **EDA**

****

****

****

****

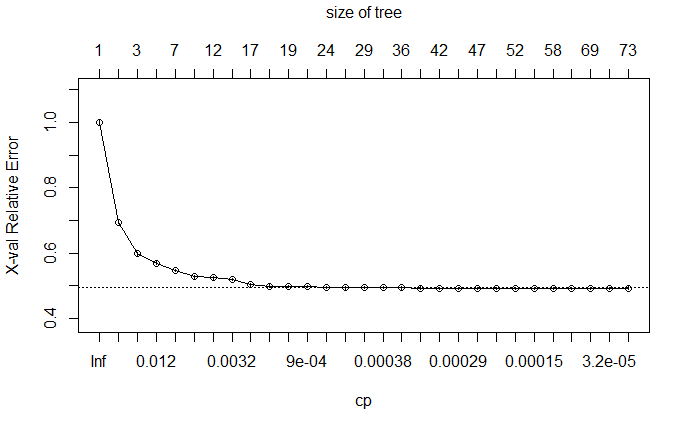
* **Modeling Process**

**First I split the data in train and test datasets so I make the model on train and check the validation of the random observations on the test dataset, I will make 3 models - CART - RandomForest - Logistic Regression and we will compare results for all models on train and test using the most important metrics.**

**\* All results shown in this report are final results after choosing the correct parameters,**

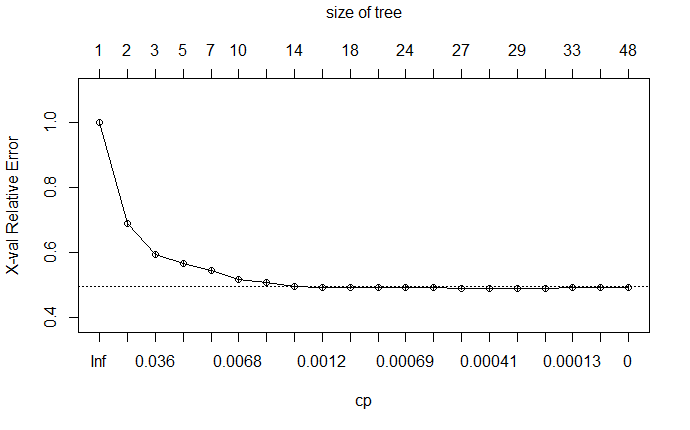
**\* Check R code for coding process**

* **CART: Comparisons between Train and Test**



**Train data set CP plot**

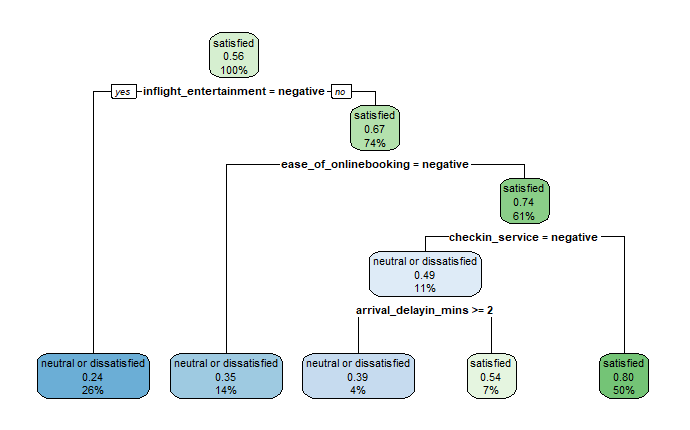
**CP = 1.7837e-04**



**Test data set CP plot**

**CP = 6.4743e-04**

After Pruning and increasing the minbuckets to 2000 the result is:

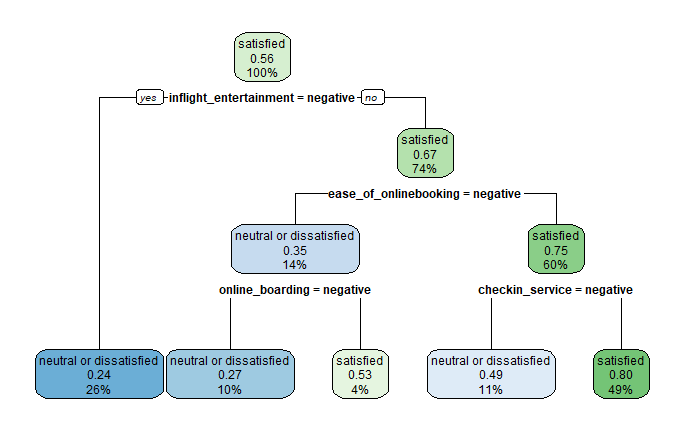


**Train Cart**

1- Entertainment is at first split 0.56 satisfied - 2- ease of online booking -

3- checkin service - 4- arrival delays

I increased the buckets to show the first 4 important variables to understand which variables we need to know on how can we enhance customer satisfaction from the right prediction



**Test Cart**

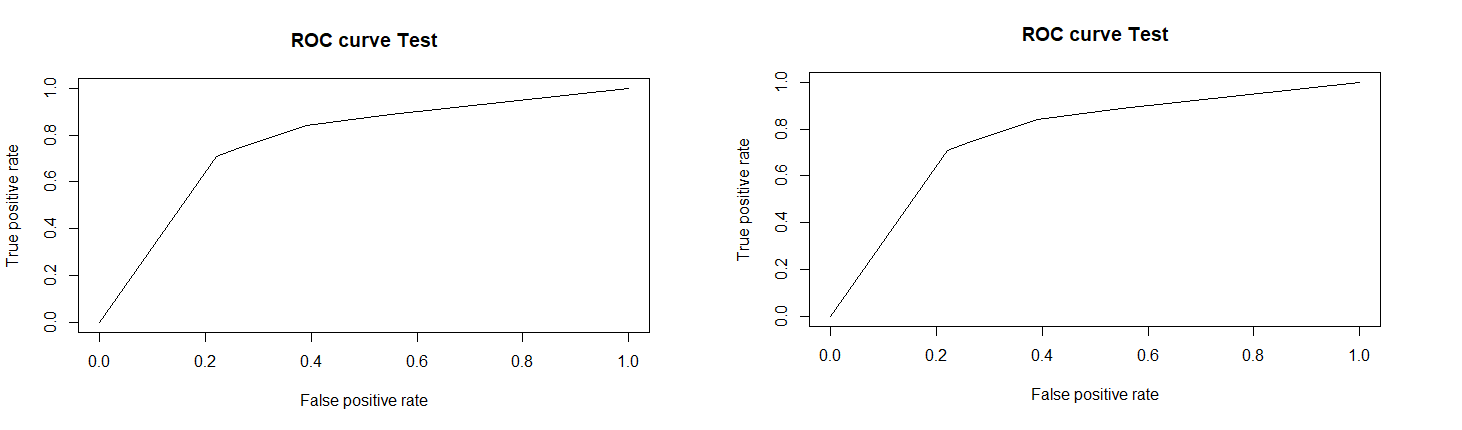
1- Entertainment 0.56 satisfied - 2- ease of online booking - 3- checkin service - 4- online boarding

|  |  |  |
| --- | --- | --- |
| **Prediction (Train)** | neutral or dissatisfied | satisfied |
| neutral or dissatisfied | 17620 | 7609 |
| satisfied | 7010 | 24795 |

:

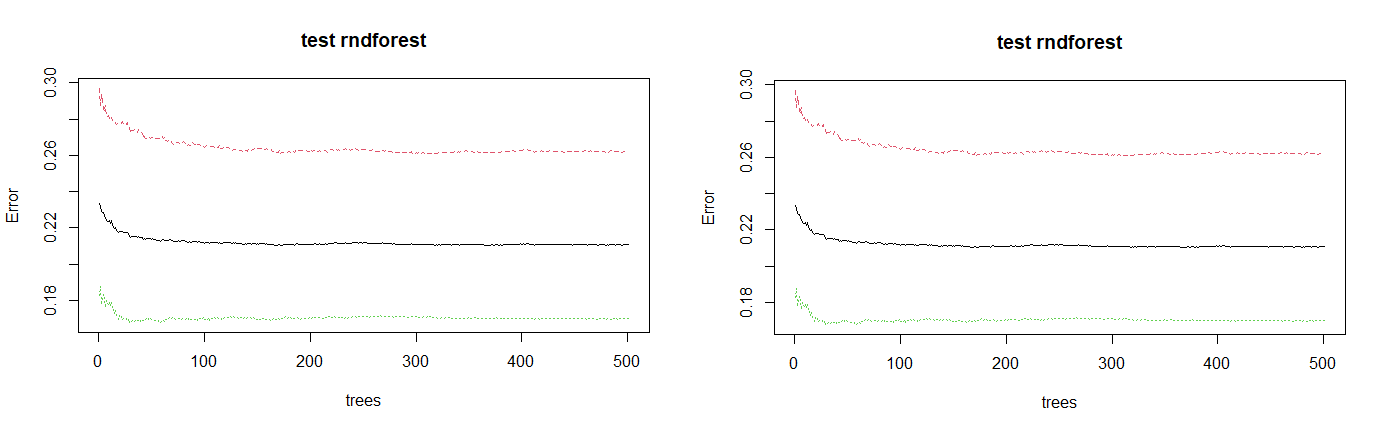
|  |  |  |
| --- | --- | --- |
| **Prediction (Test)** | neutral or dissatisfied | satisfied |
| neutral or dissatisfied | 7930 | 2882 |
| satisfied | 3443 | 10188 |

|  |  |  |
| --- | --- | --- |
| Metric/Model | **CART Train** | **CART Test** |
| Error rate | 0.2563 | 0.2587 |
| Accuracy | 0.7437 | 0.7413 |
| Sensitivity | 0.7154 | 0.7174 |
| Specificity | 0.7652 | 0.7591 |
| AUC | 0.7692007 | 0.7727487 |
| KS | 0.4842302 | 0.4893509 |
| Gini | 0.2817875 | 0.2819473 |



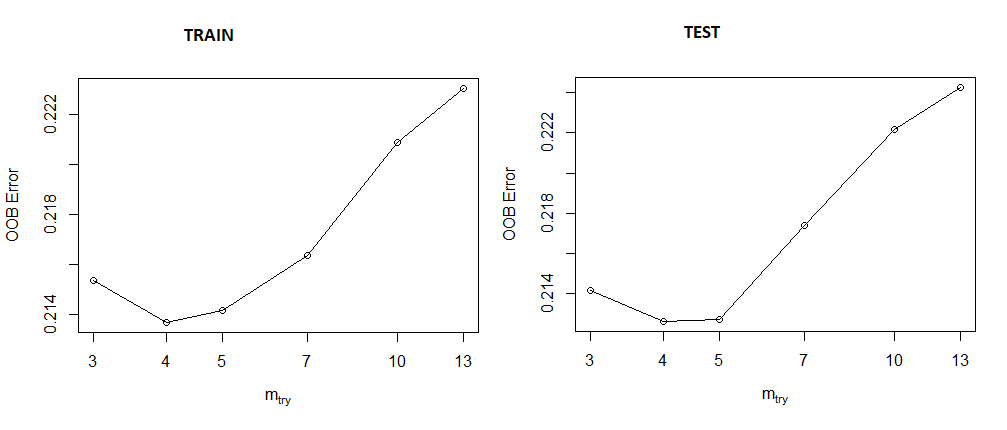
Both models match , not the best error rate.

* **RandomForest: comparison between train and test datasets**

I will use 51 trees.

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction (Train)**  **OBB 21.57%** | neutral or dissatisfied | satisfied | class.error |
| neutral or dissatisfied | 17956 | 7273 | 0.2882794 |
| satisfied | 4893 | 26912 | 0.1538437 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction (Test)**  **OBB 21.09%** | neutral or dissatisfied | satisfied | class.error |
| neutral or dissatisfied | 7974 | 2838 | 0.2624861 |
| satisfied | 2318 | 11313 | 0.1700536 |



**I will build a refined tree**

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction (Train)**  **21.33%** | neutral or dissatisfied | satisfied | class.error |
| neutral or dissatisfied | 18051 | 7178 | 0.2845139 |
| satisfied | 4951 | 26854 | 0.1556673 |

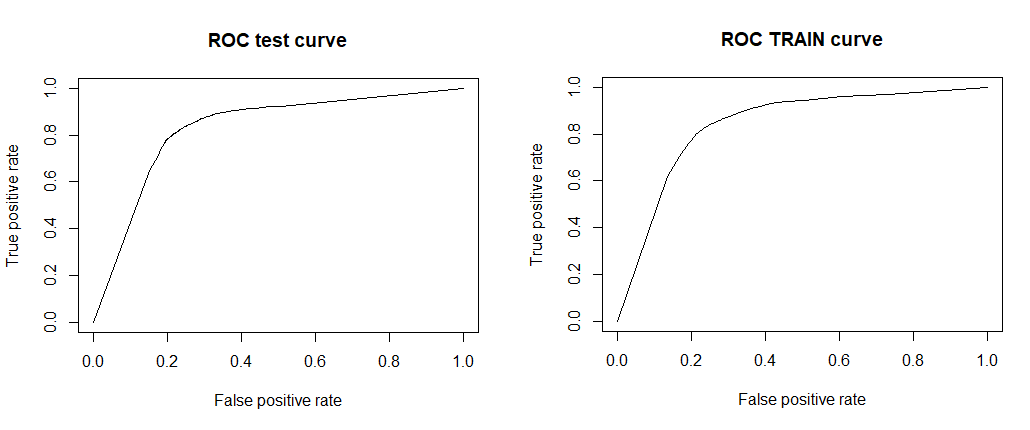
|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction (Test)**  **21.19%** | neutral or dissatisfied | satisfied | class.error |
| neutral or dissatisfied | 7966 | 2846 | 0.2632260 |
| satisfied | 2334 | 11297 | 0.1712273 |

Comparing variables importance between train and test

|  |  |  |
| --- | --- | --- |
|  | **importance(train.rndforest)** | **importance(test.rndforest)** |
| 1 | inflight entertainment | inflight entertainmanet |
| 2 | food n drink | ease of online booking |
| 3 | checkin servioce | food n drink |
| 4 | easeofonlinebooking | check in service |

\* same variables shown in both datasets , but in different sort except for entertainment

|  |  |  |
| --- | --- | --- |
| Metric/Model | **Forest Train** | **Forest Test** |
| Error rate | 0.2011081 | 0.1910567 |
| Accuracy | 0.7988919 | 0.8089433 |
| Sensitivity | 0.7984987 | 0.7983871 |
| Specificity | 0.7991585 | 0.8166207 |
| AUC | 0.8408412 | 0.8287476 |
| KS | 0.5907522 | 0.5894085 |
| Gini | 0.486449 | 0.4924352 |



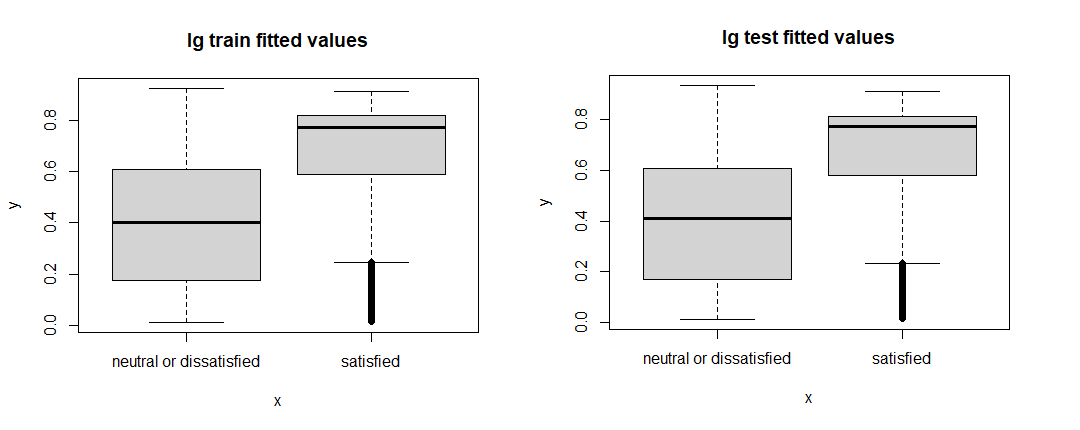
Both models match , and still not the best error rate and AUC but still the importance of variables makes sense and prediction is fair

* ***Logistic Regression: comparison between train and test***

Comparing variable importance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **varImp(fit.train)** | **varImp(fit.test)** | **Lg.model.train**  **Coefficients** | **Lg.model.test Coefficients** |
| 1 | inflight entertainment | inflight entertainment | Entertainment | Entertainment |
| 2 | easeofonlinebooking | ease of online booking | easeofonlinebooking | easeofonlinebooking |
| 3 | online\_boarding | online\_boarding | Check in service | Check in service |
| 4 | checkin\_service | checkin\_service | Onboard service | Onboard service |

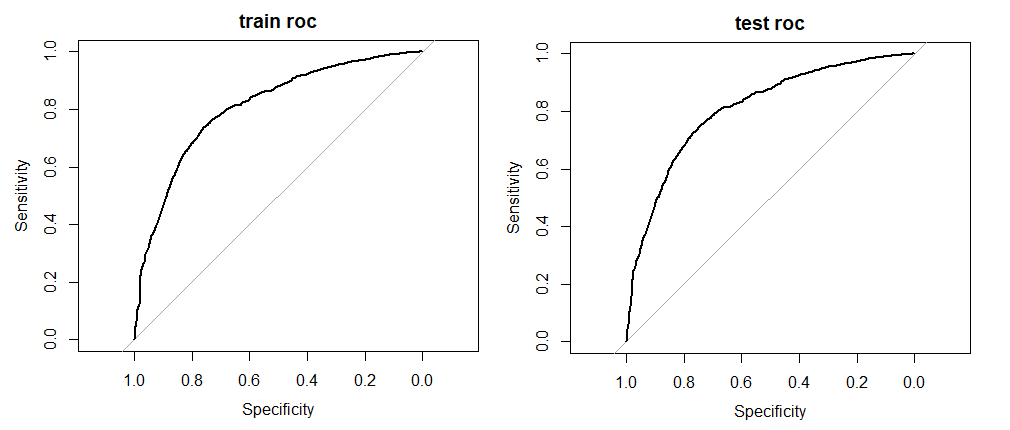
Both datasets variable inflations are below 5



|  |  |  |
| --- | --- | --- |
| **Prediction (Train)**  **> 0.6** | neutral or dissatisfied | satisfied |
| neutral or dissatisfied | 18824 | 6405 |
| satisfied | 8108 | 23697 |

|  |  |  |
| --- | --- | --- |
| **Prediction (Test)**  **>0.6** | neutral or dissatisfied | satisfied |
| neutral or dissatisfied | 8080 | 2732 |
| satisfied | 3479 | 10152 |

|  |  |  |
| --- | --- | --- |
| Metric/Model | **Logistic Regression Train** | **Logistic Regression Test** |
| Error rate | 0.2544623 | 0.2541014 |
| Accuracy | 0.7455377 | 0.7458986 |
| Sensitivity | 0.6989455 | 0.6990224 |
| Specificity | 0.7872234 | 0.7879541 |
| AUC | 0.808 | 0.8093 |



|  |  |  |
| --- | --- | --- |
| Variable/ Model | **Logistic Regression Train** | **Logistic Regression Test** |
| inflight\_entertainmentpositive | 5.66501491 | 5.39446106 |
| ease\_of\_onlinebookingpositive | 4.39055963 | 4.74197011 |
| checkin\_servicepositive | 2.64212727 | 2.55791667 |
| online\_boardingpositive | 1.70330809 | 1.78865377 |

This shows the increase of satisfaction when increasing those variables

**Model comparisons:**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric/Model | **CART Train** | **Forest Train** | **Logistic Regression Train** |
| Error rate | 0.2563 | 0.2011081 | 0.2544623 |
| Accuracy | 0.7437 | 0.7988919 | 0.7455377 |
| Sensitivity | 0.7154 | 0.7984987 | 0.6989455 |
| Specificity | 0.7652 | 0.7991585 | 0.7872234 |
| AUC | 0.7692007 | 0.8408412 | 0.808 |
| KS | 0.4842302 | 0.5907522 |  |
| Gini | 0.2817875 | 0.486449 |  |

**Variables Importance:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **CART** | **FOREST** | **Logistic Regression** |
| 1 | Inflight Entertainment | Inflight Entertainment | Inflight Entertainment |
| 2 | Ease of Online booking | Food n Drink | Ease of online booking |
| 3 | Check in service | Check in service | Online boarding |
| 4 | Online boarding | Ease of online booking | Check in service |

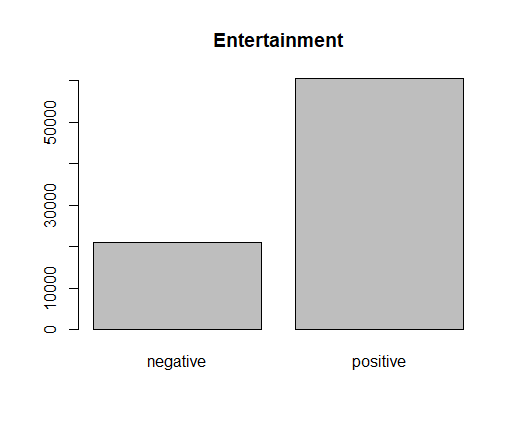
* When reading numbers , randomforest had the best numbers and the most stable model , in regard of train and test
* When comparing the importance of variables and the weight it affects customer satisfaction forest had the most reasonable variables
* this table will give us a good indication on how we can improve the satisfaction of the customers , just by knowing the most 4 important variables will give us the opportunity to enhance them, or maybe the least costly ones and easy to enhance that it would need little time and little money to upgrade it, or maybe the 4 variables that mostly affects dissatisfaction. I,e Whichever approach we chose will definitely enhance the overall

satisfaction just by investing in them..

* The 3 models intersect in:
  + Inflight entertainment (all number 1)
  + Ease of online booking
  + Check in service .
* CART and LR models intersect in:
  + Online boarding

This variable and due to experience , online boarding, is not important that much, we will have to investigate more

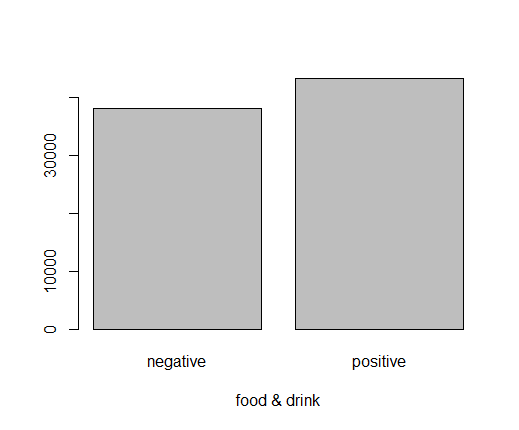
* Forest model had a unique variable, food and drink, which is more reasonable and it fit my experience on inflights satisfaction.
* Having in mind that when i investigated in the correlations between satisfaction and dissatisfaction and entertainment.I went back to submission 2 and found a high correlations between the 2 variables as the below :



**Positive is 3 times negative**

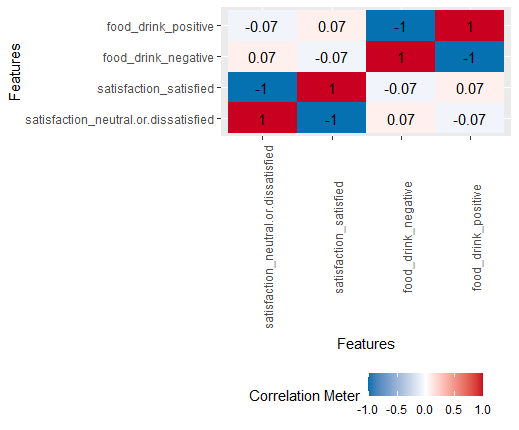
|  |  |  |  |
| --- | --- | --- | --- |
|  | negative | positive | positive/total satisfied or dissatisfied |
| neutral or dissatisfied | 16012 | 20029 | 0.55 |
| satisfied | 4952 | 40484 | 0.89 |

* Investigating in food and drink and found the correlation is very low as below:



**It is almost 50% 50%.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | negative | positive | positive/total satisfied or dissatisfied |
| neutral or dissatisfied | 18282 | 17759 | 0.49 |
| satisfied | 19874 | 25562 | 0.56 |



We need to investigate, because if food and drink is significant, upgrading it might be simple

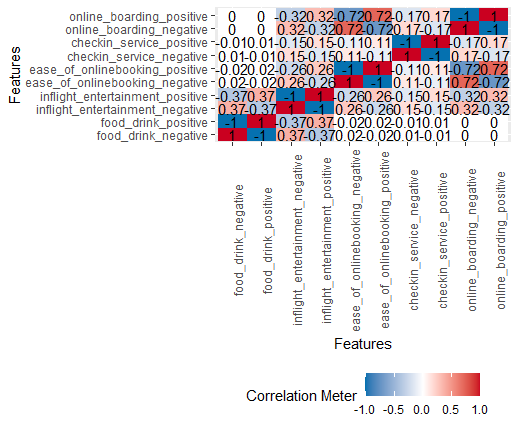
* I will choose the forest model because it has better insights and more reasonable ones.

**Model Interpretation:**

**Random Forest**

|  |  |  |
| --- | --- | --- |
|  | **importance(train.rndforest)** | **importance(test.rndforest)** |
| 1 | inflight entertainment | inflight entertainment |
| 2 | food n drink | ease of online booking |
| 3 | checkin service | food n drink |
| 4 | easeofonlinebooking | check in service |

* I was concerned here because of the difference of the importance between train and test ,

****

* \* all important variables that need to be investigated .

**Conclusions and Recommendations:**

Falcon Airlines now can focus on the most important variables in regards to affecting customer satisfaction and in regards to the increase of satisfaction percentage if those variables were enhanced.

They can start with **Inflight Entertainment**:

* it can be upgraded with least time and money if the system they have is upgradable, but if they need new hardware, a new analysis regarding the ROI must be implemented.
* I suggest focusing on targeting people ages between 7-30 (30% of total passengers) They follow trends and stay up to date with new technologies (ex. Adding new games, music and movies or adding network gaming between the passengers will add up to the overall passenger satisfaction.
* Older passengers might care about news, documentaries and movies, which means; other channels of entertainment should be of reach.
* As predicted from the logistic regression model, if we increase the satisfaction for the ‘’Inflight entertainment’’ by 5 points, the customer satisfaction will increase by 66% which proves that it has a good impact. Achieving these results will not be easy.
* Overall satisfaction of the inflight entertainment is good, because 70% of the passengers have positive feedback on this variable.
* Inflight Entertainment took the first position in all models, so this variable is the most important.

**Food And Drinks:**

* It can be upgraded with minimum time and effort, by adding some treats or making some changes on the served menu.
* It is an important variable, from my experience, even though it was only mentioned in RF model but I gave it second importance after Inflight entertainment
* Around 50% of the customers have positive feedback, which is a good base to start from towards decreasing the number of unsatisfied customers. By making small changes in the served menu or adding extra treats the airlines will be able to to satisfy a larger percentage.

**Ease Of online Booking and online boarding**

* It is important to update and upgrade those variables, as the online services must always be up to date and must always be easy, effective and user friendly.
* 80% of the customers have positive feedback on those variables, which means that the company has a good online service, but enhancing those services will affect positively to the company.
* As in the logistic regression coefficients, if we increase online booking by 4 points the odds of satisfaction will increase by 39% and online boarding will increase by 1 point the odds will increase by 70%.
* I recommend they hire a reputable consulting online and user experience company and make some A/B testings to finally reach the best and easiest and more convenient online booking and boarding system.
* Both variables are easy and cheap to upgrade and the effect will be positive.

**Check in Service**

* All models predicted about the importance of this variable even though it is in the grey area, so they need to make another deeper analysis to understand the significance of this variable.

Overall customer satisfaction is good, “Falcon Airflight” is doing good and little effort can increase the overall satisfaction which will positively affect the reputation of “Falcon Airline” and the return on investment**.**

.

**The End**