Airport Passenger Survey Analysis

Divyang Jain

16110323

MSc in Data Analytics

X16110323@student.ncirl.ie

1. ABSTRACT

From past few years, customer has become a crucial part of any organization as they help to improve the mistakes, poor qualities by analyzing the customer queries. Customer satisfaction surveys helps to get the feedback which helps companies to work accordingly, it has around 10-15 questions that relates to customer experiences, service delivery and the overall satisfaction. If the customer is not happy with the service then it's the biggest nightmare for a company as customer relationship is a vital part of any organizations. Airport is also one of the largest organization which is owned by the government which is used by the people to travel among different areas and government duty is to make their customers happy by fulfilling their needs, it can be possible by conducting the surveys time to time which leads to my paper study. This paper studies about the results from **Austin-Bergstrom international Airport (ABIM)** customer surveys which can improve the customer satisfaction and to aim the precise areas for the enhancement. The aim of the project is to find out which quality of an airport effects the most to the overall satisfaction from which they can work on their issues by using **multiple regression** and also to find the finest accuracy for the customer response by using different machine learning techniques like **decision trees, random forest, Support vector machine(SVM) and deep learning** with **visualizing** which year received more customer satisfaction.

2. DATASET INTRODUCTION

Dataset which is used in this project is Airport Quarterly Passenger Survey for Austin (Texas). Data is accumulated by Texas government from Austin-Bergstrom International airport website and is available on: https://data.austintexas.gov/browse?q=Airport%20Quarterly%20Passenger%20Survey&sortBy=relevance&utf8 from January 2015 to December 2016. Dataset contains 2450 rows and 37 columns. Extra columns are added for the customer response and some transformation and manipulation has been done with full statistical methods like squaring, using logarithm, inverse cases and some nominal values are assigned to the customer response to find the appropriate result.

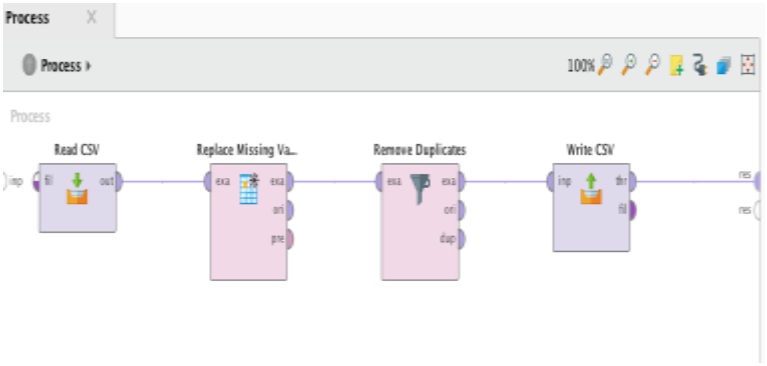
3. FEATURES IN THE DATASET

1.Quarter- Explains the quarter of the year q1, q2, q3, q4.
2. Data recorded- Day on which the data is recorded.
3. Departure Time- Departure time of the customer's flight
4. Ground Transportation to/from airport- Ground facilities of the airport provided.
5. Parking Facilities- Airport parking facilities.
6. Parking facilities (value for money)- Are parking fees worth the facilities.
7. Availability of baggage cars- baggage cars available or not.
8. Efficiency of check-in-staff- Explains the efficiency of check in staff.
9. Check-in wait time- Wait time for check in is too long or short.
10. Courtesy of check-in staff- Nature of the check inn staff with a customer.
11. Wait time at passport inspection- Customer wait time at passport inspection.

12. Courtesy of inspection staff- Nature of inspection staff is good or bad.
13. Courtesy of security staff- Nature of security staff is good or bad.
14. Thoroughness of security inspection- Security thoroughness.
15. Wait time of security inspection- Customer wait time at security check.
16. Feeling of safety and security- Is the customer belongings safe or not.
17. Ease of finding your way through the airport- Does the way navigates throughout the airport.
18. Flight information screens- Flight information is available on screens or not.
19. Walking distance inside terminal- Distance is too long or not.
20. Ease of making connections- Connecting flights are easily catchable or not.
21. Courtesy of airport staff- Nature of airport staff.
22. Restaurants- Are there enough restaurants.
23. Restaurants (value for money)- Are restaurants valuable?
24. Availability of banks/ATM- Are their enough ATMS.
25. Shopping facilities- Are their enough shops for shopping?
26. Shopping facilities (value for money)- Are shops valuable?
27. Internet access- Is there proper internet access (WIFI services).
28. Business Lounges- Are there enough executive lounges?
29. Availability of washrooms- Are there enough washrooms?
30. Cleanliness of washrooms- Are washrooms cleaned?
31. Comfort of waiting/gate areas- Customer is having comfort at waiting gates or not.
32. Cleanliness of airport terminal- Are airport terminal cleaned?
33. Ambience of airport- Do customers satisfy with the infrastructure and ambience of the airport?
34. Arrivals passport and visa inspection- Does it take too long for the visa inspection?
35. Speed of baggage delivery- How is the service for the baggage delivery?
36. Customer Inspection- Are u happy with the customs inspection?
37. **Overall Satisfaction- Does the customer happy with the overall services?**
38. **Customer Response (nominal transformed)- Is the customer satisfied or not (related to the overall satisfaction).**

4 . RESEARCH AND INVESTIGATION

Before performing any analysis, we have to clean the dataset and in real life dataset is not clean. According to Mannila ,1996 we have to perform data cleansing or scrubbing which removes the duplicates, missing values to improve the quality of the data and for that I have used rapid miner to clean the dataset:



After cleaning the data set it's an important part to understand the relationship between attributes and here overall satisfaction is judged by many other survey questions that covers in data exploration (Rahm, 2000).

Analysis techniques are not one-time process it is an iterative process, to avoid this repetitive process a process model is designed which is very crucial. There are different phases of the process model:

1. Business requirement collecting and analysis
2. Data source ident cation
3. Analyzing structure of the data
4. Building the data model

In my project, multiple regression has been done on the Airport survey data set based on the scores ranges from 0 to 10 (0 being the worst and 10 being the best) in the overall satisfaction score using **SPSS**.
Among all the 2450 rows, I have performed the analysis on 158 rows which contains 38 columns.
Objective of the multiple regression is to find how these services affects the most to gain overall satisfaction by correlating with it. Here,

I have **taken four independent variable** (factors) -ground transportation, feeling of safety and security, ambience of airport and custom inspection affecting **one dependent variable (**outcome)- overall satisfaction.

I have used standard regression or forced entry method in which all the values are entered simultaneously (into one go).

Before performing multiple regression, it is important to check some assumptions as if these assumptions are violated then results will be useless.

Preliminary test

1. **Sample size** - N>50+8m which means there should be more then 82 rows.

2. **Multicollinearity and singularity-** data should not be multicollinearity and singular means factor should not be combination of other factors and should not represent too much with the dependent variable.

3. **Outliers-**Data should be constant, it should not have any other external variable which is outside the range and interrupts in the results.

4. **Normality, homoscedasticity, linearity, independence of residuals-**residual is the difference between the outcome and the predicted values. It should be normally distributed with a straight-line relationship. The residual variance of the dependent score should be similar (Pallant, 2013).

Research Question on multiple regression- How well do the 4 measures (ground transportation, feeling of security and safety, ambience of airport, custom inspection) affects the overall satisfaction.

Step 1: <u>Check the assumptions</u>

1.Sample size-this assumption can be seen from descriptive statistics which shows 158 rows. Hence, I haven't violated this assumption.

```
GET
  FILE='/Users/divyang/Desktop/CRM final.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA COLLIN TOL ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Overall_satisfaction
  /METHOD=ENTER Ground_Transportation Feeling_of_safety_and_security Ambience_of_airport
    Custom_inspection
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS DURBIN HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(3).
```

Regression

[DataSet1] /Users/divyang/Desktop/CRM final.sav

Descriptive Statistics

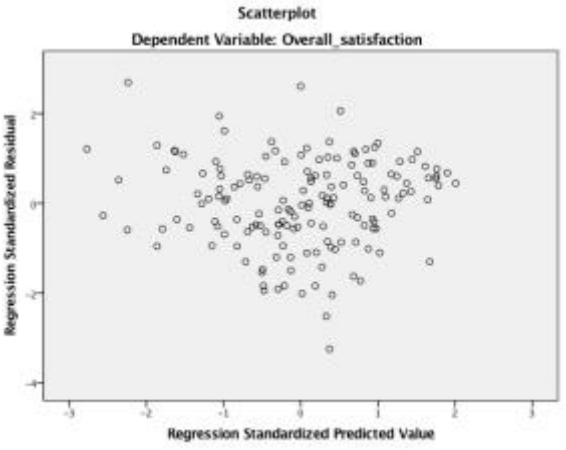
	Mean	Std. Deviation	N
Overall_satisfaction	5.37573	1.145010	158
Ground_Transportation	.99105	.272369	158
Feeling_of_safety_and_security	.25730	.126885	158
Ambience_of_airport	.42861	.150693	158
Custom_inspection	.14342	.120034	158

2.Multicollinearity and singularity- Correlation table shows that the data have met the assumption i.e. four factors have more than 0.3 value means they are correlating substantially and does not correlate too much with each other.

Correlations

		Overall_satisfaction	Ground_Transportation	Feeling_of_safety_and_security	Ambience_of_airport	Custom_inspection
Pearson Correlation	Overall_satisfaction	1.000	.741	.180	.568	.395
	Ground_Transportation	.741	1.000	.088	.442	.206
	Feeling_of_safety_and_security	.180	.088	1.000	.374	.276
	Ambience_of_airport	.568	.442	.374	1.000	.494
	Custom_inspection	.395	.206	.276	.494	1.000
Sig. (1-tailed)	Overall_satisfaction	.	.000	.012	.000	.000
	Ground_Transportation	.000	.	.137	.000	.005
	Feeling_of_safety_and_security	.012	.137	.	.000	.000
	Ambience_of_airport	.000	.000	.000	.	.000
	Custom_inspection	.000	.005	.000	.000	.
N	Overall_satisfaction	158	158	158	158	158
	Ground_Transportation	158	158	158	158	158
	Feeling_of_safety_and_security	158	158	158	158	158
	Ambience_of_airport	158	158	158	158	158
	Custom_inspection	158	158	158	158	158

3.Outliers-it can be seen from the scatter plot in which the values should be between -3.3 and 3.3. Here all the values are between that point and data is scattered like a rectangle.



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Custom_inspection, Ground_Transportation, Feeling_of_safety_and_security, Ambience_of_airport[b]	.	Enter

a. Dependent Variable: Overall_satisfaction
b. All requested variables entered.

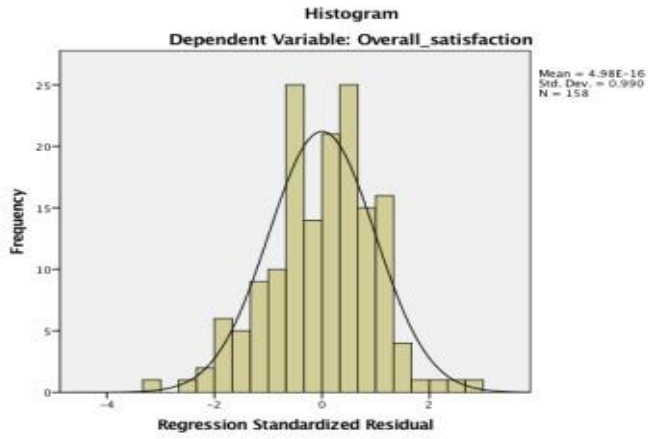
Model Summary[b]

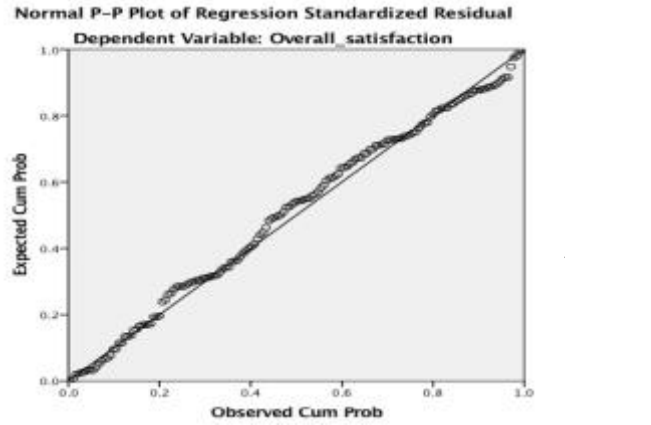
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.800[a]	.641	.631	.695411	1.339

a. Predictors: (Constant), Custom_inspection, Ground_Transportation, Feeling_of_safety_and_security, Ambience_of_airport
b. Dependent Variable: Overall_satisfaction

4.Normality, homoscedasticity, linearity, independence of residuals- It can be seen from Normal pp plot and the histogram. Here, it Is making a straight line with the normally distributed histogram.

Charts





Now, all the assumptions are met its time to evaluate a model.

Step 2: <u>Evaluating a model</u>

First table shows that all the data has been entered through the standard multiple regression method (Enter) or forced entry method.

The model summary table tells how much variance is shown by independent variables (ground transportation, feeling of security and safety, ambience of airport, custom inspection) on dependent variable (overall satisfaction). Here, it is 64.1% which is a respectable result and the value of r is less then .9 i.e. 0.8, hence it is not violating the assumption.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	131.844	3	43.948	91.472	.000[b]
	Residual	73.990	154	.480		
	Total	205.835	157			

a. Dependent Variable: Overall_satisfaction
b. Predictors: (Constant), Internet_access, Ground_Transportation, Shopping_facilities

Here are the hypotheses,

H0: β1= β2= β3= β4

H1: β1=! B2 Not all are equal

We can check the multiple R significance from ANOVA table which shows 0.000 which is less than 0.000, hence it's a significant and **good fit model.**

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Ground_Transportation	Feeling_of_safety_and_security	Ambience_of_airport	Custom_inspection
1	1	4.458	1.000	.00	.00	.01	.00	.01
	2	.298	3.868	.02	.01	.02	.00	.82
	3	.161	5.264	.02	.05	.87	.01	.00
	4	.050	9.422	.26	.03	.02	.90	.16
	5	.033	11.615	.70	.90	.08	.08	.00

a. Dependent Variable: Overall_satisfaction

Casewise Diagnostics[a]

Case Number	Std. Residual	Overall_satisfaction	Predicted Value	Residual
154	-3.241	3.465	5.71860	-2.253602

a. Dependent Variable: Overall_satisfaction

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.83750	7.21610	5.37573	.916391	158
Residual	-2.253603	1.869932	.000000	.686495	158
Std. Predicted Value	-2.770	2.008	.000	1.000	158
Std. Residual	-3.241	2.689	.000	.987	158

a. Dependent Variable: Overall_satisfaction

These three tables show the outliers which doesn't not fit into the model and case wise diagnostics table shows that 154 cases have been removed as it is effecting the model.

Step 3: <u>Evaluating each of the independent variables</u>

Result of the analysis

The aim of the analyses was to know which variable in the model contributes the most to the prediction of overall satisfactions which can checked from the **coefficient table** (Pallant,2013).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.697	.232		8.268	.000	1.438	2.356					
	Ground_Transportation	1.541	.535	.162	2.881	.005	.484	2.599	.395	.227	.140	.347	1.339
	Feeling_of_safety_and_security	1.661	.484	.219	3.434	.001	.705	2.616	.568	.268	.166	.588	1.724
	Ambience_of_airport	2.568	.229	.811	11.218	.000	2.117	3.019	.741	.673	.546	.798	1.253
	Custom_inspection	.089	.477	.001	.019	.992	-.919	.948	.380	.001	.000	.842	1.188

a. Dependent Variable: Overall_satisfaction

To find the contribution of each independent variable then see beta value under standardized coefficients**. Look down the leading (largest) value in this column, here the largest value of bets is ambience of airport followed by feeling of safety, ground transportation and custom inspection** (Pallant, 2013)**.** To find which variable making a unique contribution check the significance value that should be less than 0.05. Here three of the variables are making unique contribution expect the custom inspection which may be due to the overlapping in the independent variable. To see the unique contribution means how much individual variable contributing to the result, we have to square the values of the Part and check the result. Overall from model summary they are contributing= **64.1%.**

Beta values(Contribution)

1.Ground transportation=16.2%
2. Feeling of safety and security=21.9%
3. Ambience of airport= 61.1% (maximum contribution)
4. Customer inspection=0.001% (least contribution)

Part values(uniqueness)

1.Ground transportation=16.2%

2. Feeling of safety and security=2.75%

3. Ambience of airport= 45.29%(maximum unique contribution).

4. Customer inspection=0.001%

5.APPLYING TECHNIQUES

After applying multiple regression through SPSS, I have used rapid miner to find the finest accuracy by using different machine learning techniques. Machine learning is involved in every data analytics thing which helps to acquire new information from the data (Faria et.al.,2010). Using machine learning with rapid miner is a great idea which saves a lot of time and effort. "Rapid Miner Studio combines technology and applicability to serve a user-friendly integration of the latest as well as established data mining techniques" (Chisholm, 2013, p:19). It is used to define the analysis with drag and drop of operators which can set the parameters. In this project, rapid miner is used to check the accuracy with different models.

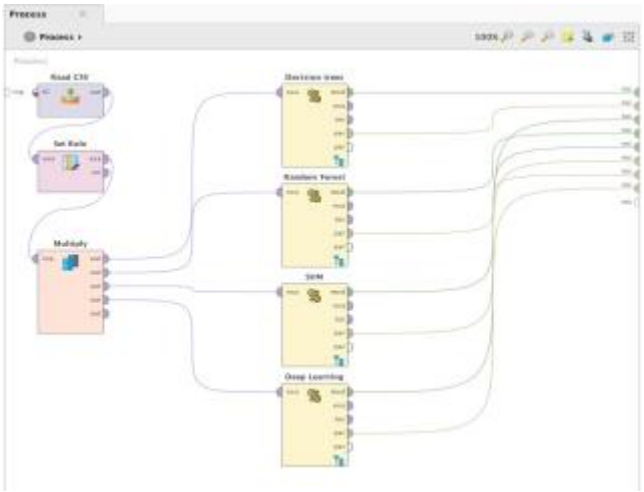
Steps involved:

1.Drag read csv from operators pane and drop it to the process place- read the file cleaned file and mark the attributes.
2.Drag and drop the set role to the process with changing the customer response to the label (nominal)
3.Drag and drop the multiply option to the process.
4.Click on edit on the menu bar and insert 4 blocks and connect it to the multiply.
5.Connect each validation to the result from mod and per output.
6.Double click each validation and put decision trees, random forest, SVM and deep learning in the training pane.

respectively leaving apply model and performance on the test pane.

7. Run the process and calculate each process.

(Chisholm, 2013)



I have used four techniques to see the best performance of the **customer response**.

1. Decision Trees- They are powerful classifiers, which makes a tree structure and check the relationship between the features and the outcomes.

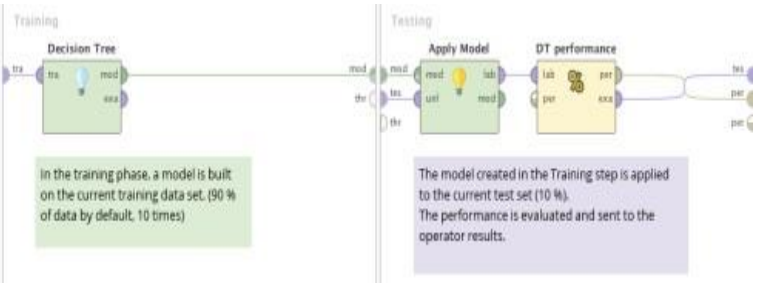
2. Random Forest- They operates after creating the decision trees at the training time and outputting the class prediction (regression) of the individual trees.

3. SVM (Support vector machine)- It creates the boundary between points of the data plotted in a multidimensional which can represent rows as examples and the columns as feature values by creating a hyperplane.

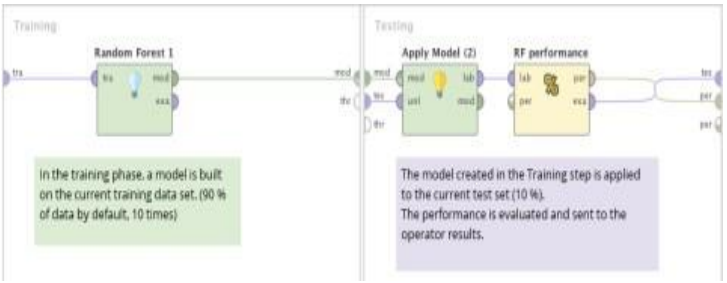
4. Deep learning -It uses Artificial Neural network (ANN) that contains more than 1 hidden layer and performs machine learning to predict the performance on the test set. (Lantz, 2013).

Data is divided into training set (90%) to train the data (learn) and the test data set (10%) which is used to predict the accuracy.

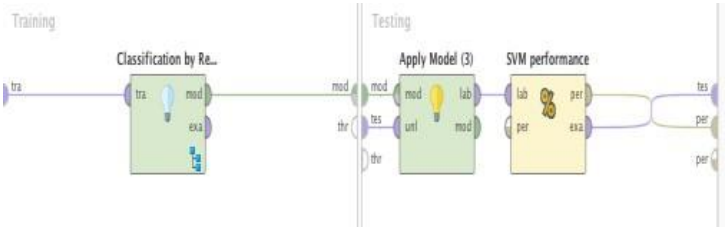
Here are all the snapshots inside the validation process:



DECISION TREES

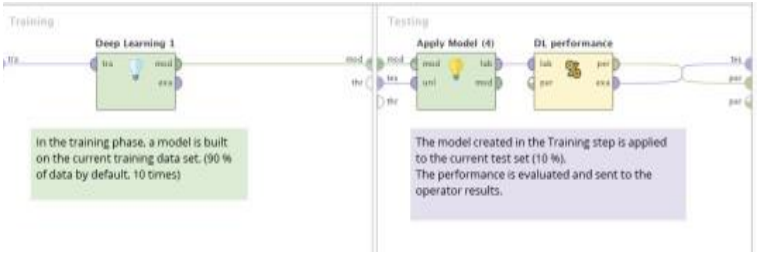


RANDOM FOREST



CLASSIFICATION BY REGRESSION WHICH INCLUDES SVM

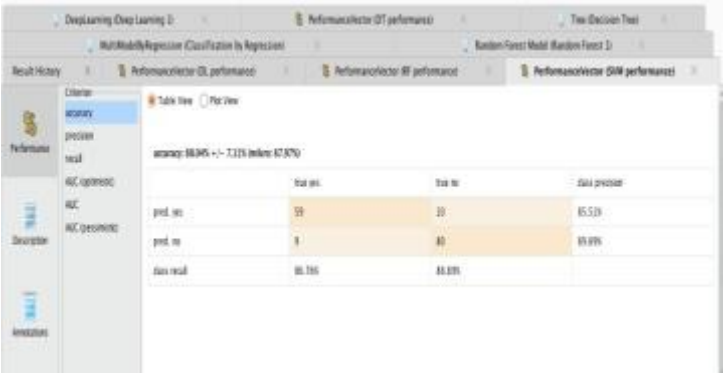
SVM IN CLASSIFICATION BY REGRESSION



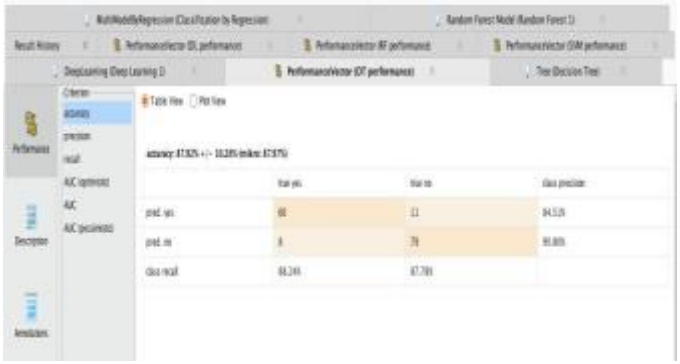
DEEP LEARNING

After running the process, I got this result: -

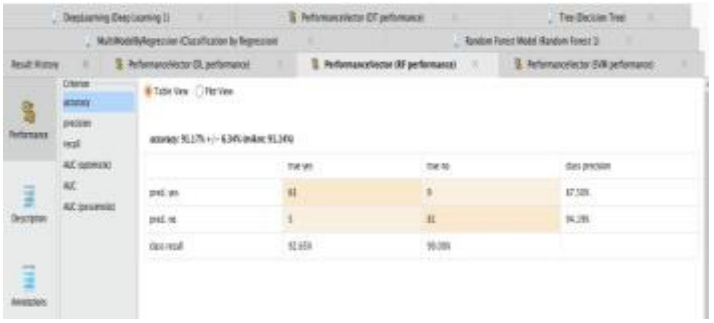
This is the accuracy from deep learning-87.21%



This is the accuracy from Support Vector Machine-88.04%

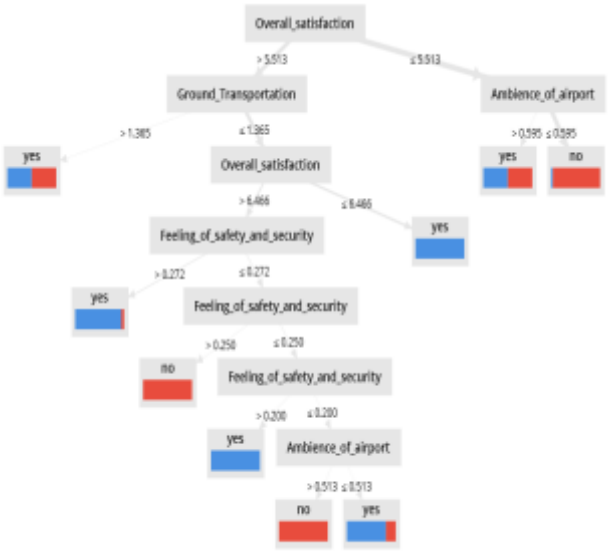


This is accuracy from Decision Trees-87.92%



This is the accuracy from Random Forest-91.17%

As we can see, we got the best result from Random forest method. We will check the random forest description as how the random forest tree was formed:-



This is diagram which is made from decision tree.

Tree

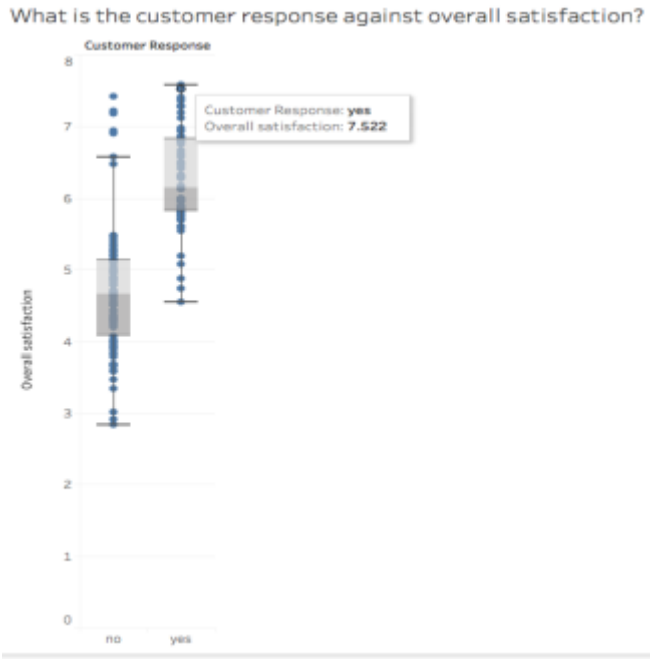
```
Ambience_of_airport > 0.646: yes {yes=5, no=0}
Ambience_of_airport ≤ 0.646
|   Ambience_of_airport > 0.207
|   |   Feeling_of_safety_and_security > 0.073
|   |   |   Ambience_of_airport > 0.246
|   |   |   |   Feeling_of_safety_and_security > 0.186
|   |   |   |   |   Feeling_of_safety_and_security > 0.117
|   |   |   |   |   |   Feeling_of_safety_and_security > 0.133
|   |   |   |   |   |   |   Feeling_of_safety_and_security > 0.137
|   |   |   |   |   |   |   |   Feeling_of_safety_and_security > 0.146
|   |   |   |   |   |   |   |   |   Feeling_of_safety_and_security > 0.165
|   |   |   |   |   |   |   |   |   |   Ambience_of_airport > 0.498
|   |   |   |   |   |   |   |   |   |   |   Ambience_of_airport > 0.641: no {yes=0, no=2}
|   |   |   |   |   |   |   |   |   |   |   Ambience_of_airport ≤ 0.641
|   |   |   |   |   |   |   |   |   |   |   |   Ambience_of_airport > 0.631
|   |   |   |   |   |   |   |   |   |   |   |   |   Ambience_of_airport > 0.637: yes {yes=4, no=0}
|   |   |   |   |   |   |   |   |   |   |   |   |   Ambience_of_airport ≤ 0.637: no {yes=0, no=5}
|   |   |   |   |   |   |   |   |   |   |   |   Ambience_of_airport ≤ 0.631: yes {yes=27, no=6}
|   |   |   |   |   |   |   |   |   |   Ambience_of_airport ≤ 0.498
|   |   |   |   |   |   |   |   |   |   |   Feeling_of_safety_and_security > 0.178: no {yes=7, no=34}
|   |   |   |   |   |   |   |   |   |   |   Feeling_of_safety_and_security ≤ 0.178: yes {yes=4, no=1}
|   |   |   |   |   |   |   |   |   Feeling_of_safety_and_security ≤ 0.165: yes {yes=3, no=0}
|   |   |   |   |   |   |   |   Feeling_of_safety_and_security ≤ 0.146: no {yes=0, no=8}
|   |   |   |   |   |   |   Feeling_of_safety_and_security ≤ 0.137: yes {yes=3, no=0}
|   |   |   |   |   |   Feeling_of_safety_and_security ≤ 0.133: no {yes=0, no=6}
|   |   |   |   |   Feeling_of_safety_and_security ≤ 0.117: yes {yes=6, no=0}
|   |   |   |   Feeling_of_safety_and_security ≤ 0.186: no {yes=0, no=7}
|   |   |   Ambience_of_airport ≤ 0.246: no {yes=0, no=10}
|   |   Feeling_of_safety_and_security ≤ 0.073: yes {yes=3, no=0}
|   Ambience_of_airport ≤ 0.207: no {yes=0, no=17}
```

The tree was divided like this from which we got the best accuracy.

6.VISUALIZATION

Data has been mined now it's time to visualize the data through **tableau.**

Tableau is a wonderful tool for understanding the data and converting data into the information while judging the patterns and by making vital decisions on the dataset (Milligan, 2016).



This whisker pot shows the different quartiles which is minimum, median, maximum (Q1- 25%,50% and 75 %) and shows information on the customer response against overall satisfaction.
Picture says positive response from the customers (Yes) which is 7.522 is more than the negative response (NO).

Second graph shows the different dates of year 2015 and 2016 quarterly as the surveys happened one time in 4 months.
As the graph says that there was less overall satisfaction in the first quarter of 2015 which depleted till second quarter but by the awareness of the customer response it touched the **peak in the third quarter** in the month of July. In 2016, overall satisfaction of customers at airport was hiked in quarter 1 only then keeps on decreasing.

7.CONCLUSION

The project report explains the deep analysis on airport survey of Austin at Texas by comparing different attributes by applying multiple regression and check which attributes affects the most in the analysis. We got to know that ambience of the airport affects the most by having **61.1%** beta value. Hence, airport management team should **target ambience of the airport as the first quality for the customer** and to have **customer relationship management** to enhance the overall satisfaction. After that data was imported into rapid miner to apply the machine learning techniques after applying different techniques we got the best accuracy from Random forest of 91.27%. It was applied on customer response of different attributes and overall satisfaction. Finally, data was imported to tableau to check which quarter received the maximum number of satisfaction the result showed quarter 3 of 2015 and quarter 1 of 2016 showed the highest result. Hence. Customers surveys should be there to enhance the performance of an organization.

8.References

Chisholm, A., 2013. *Exploring Data with RapidMiner*. Packt Publishing Ltd.

Faria, B.M., Reis, L.P., Lau, N. and Castillo, G., 2010, June. Machine Learning algorithms applied to the classification of robotic soccer formations and opponent teams. In *Cybernetics and Intelligent Systems (CIS), 2010 IEEE Conference on* (pp. 344-349). IEEE.

Lantz, B., 2013. *Machine learning with R*. Packt

Publishing Ltd. Milligan, J.N., 2016. *Learning*

Tableau 10. Packt Publishing Ltd.

Mannila, H., 1996, June. Data mining: machine learning, statistics, and databases. In *Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on* (pp. 2-9). IEEE.

Pallant, J., 2013. *SPSS survival manual*. McGraw-Hill Education (UK).

Rahm, E. and Do, H.H., 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), pp.3-13.