2014

Airline Satisfaction Survey

By:-

Group Members: (Group 2)

- Ruchika Narang
- Sophia Kelly
- Xinbei Zhu
- Mikhil Mistry
- Melissa Tran
- Sanchit Singh

Executive Summary

Every day of the year, for every hour of the day, airplanes are transporting over 2.5 million people each day. With such a large volume of people, it is necessary for airline companies to attract new flyers and maintain the loyalties of existing customers. The most popular method is by conducting a survey to see what variables are most likely to affect the overall satisfaction rate a customer experience. The data we obtained is from a data set for an airline company that documents the overall satisfaction rate for over 129,000 passengers with 30 influencing variables. Our goal is to use various data mining techniques to determine which of the 30 variables plays the most crucial factors which will result in either a positive or negative flying experience. To accomplish this, we first used the program R to clean the large volume of data.

After the data has been preprocessed, we will continue to explore the data by using the classification techniques of Logistic Regression, KNN, and Classification Tree using XL Miner. Once we have obtained all the results from XL Miner, we will compare the numbers to determine which model should be recommended for the airline to use. The final model will consist of the top most influential variables that factor towards the passenger's flight experience. The airline can then use our analysis to implement new policies to vastly improve their services to ensure a higher passenger satisfaction rate.

Table of Contents

Introduction

Background

The survey dataset was retrieved from IBM's Data Analytics website which contains responses from an Airline Satisfaction Survey conducted in the first quarter of 2014. The survey was aimed at passengers taking domestic flights in the United States. In this survey, the passengers were asked to rate their overall experience of travel including their experience at the airport. The satisfaction rating scale ranges from 1-5, with 1 indicating lowest level of satisfaction and 5 being the highest. There are a total number of 30 variables used in this dataset. The experience of a traveler is characterized by conducting a satisfaction survey which has 30 factors/variables, out of which 14 are categorical variables, 10 are continuous variables and 6 are character variables. There are few variables which show information about the time spent by the passenger at the airport such as travel duration, flight times, travel class, distance etc. Also, there are variables indicating personal information about an individual such as age, gender including the money spent at the airport on eatables and shopping.

Problem Description

In this project, we will address which factors would affect customers' satisfaction ratings the most. Based on the survey, the factors that affect customer's rating are complicated; some of them may even combine to produce effects on overall satisfaction. Achieving the precise satisfaction rating can be a big challenge for the airline corporation. Our analysis would help management to understand each factor. From the management perspective, knowing the importance of each factor would help them improve their satisfaction ratios and thus, bring them more sales and market occupancy.

Question of Interest and Data Description

The success of the model depends on the output variable selected and in determining the major variables or factors that impact the output variable. The output variable selected for this purpose is called as 'Satisfaction score'. The output variable, 'Satisfaction score', is a categorical variable that holds a value in the range of 0-5.

The main question of interest is to identify those variables that majorly affect the overall satisfaction score. The first task in this identification process is to perform a process called as 'Dichotomization' on the output variable – 'Satisfaction score'. The process of dichotomization classifies a record as either 0 or 1, where 0 (low score) represents that a traveler has given a satisfaction survey score of less than or equal to 3.5 on a scale of 5 and 1 (high score) represents a satisfaction score of above 4 on a scale of 5.

Figure 1 represents the frequency distribution plot graph for the records before dichotomization was performed. As observed in the graph below, value scores of ‘4’ and ‘5’ approximately constitutes most of the data and the remaining scores constitute the rest

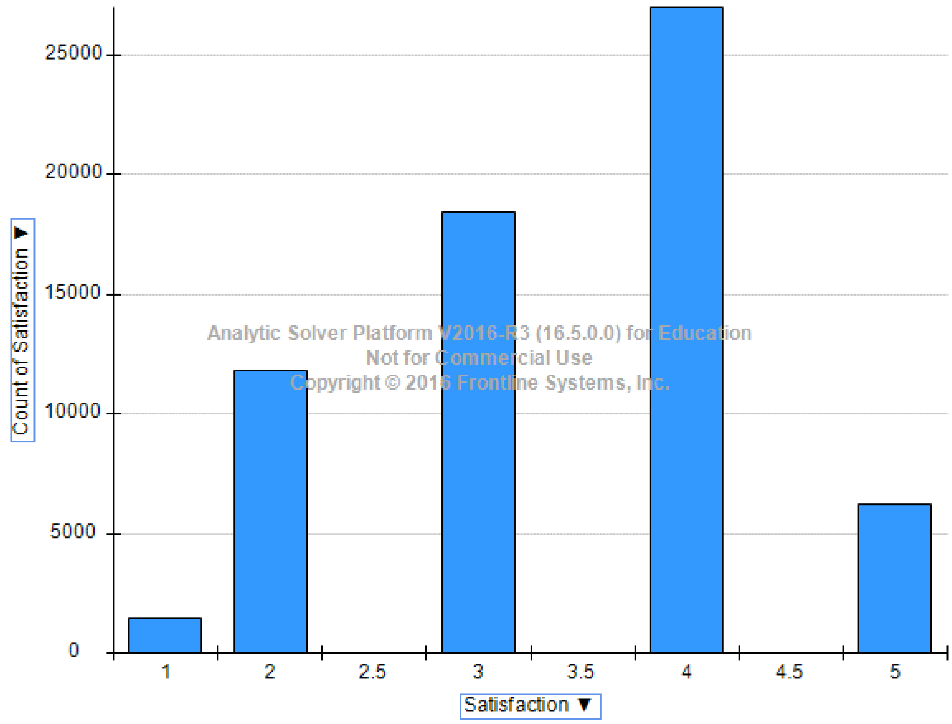


Figure 1: Before Dichotomization

Figure 2 represent the frequency distribution plot graph for the records after dichotomization. For analysis, values between ‘4’ and ‘5’ are dichotomized as ‘1’ and the remaining are dichotomized as ‘0’.

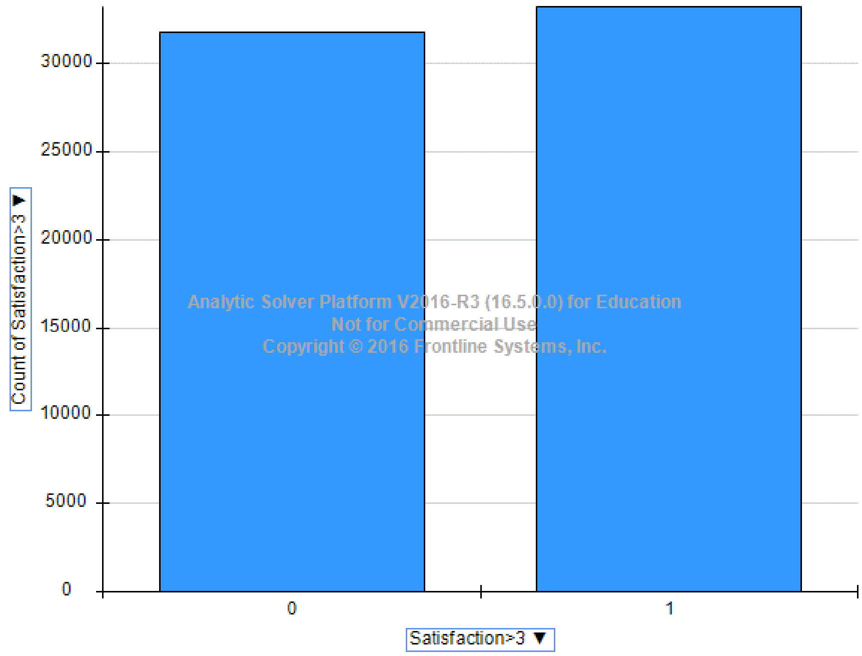


Figure 2: After Dichotomization

This evaluation will be further used to identify the variables that are most likely to affect the level of satisfaction for the passenger and to identify the reasons that account for some airlines performing better than others.

S/No	Variable Name	Variable Description	Type of Variable
1	Satisfaction	Rating scale from 1-5 indicating the level of satisfaction, 1 being the lowest	Categorical
2	Airline Status	Blue, Gold, Platinum, and Silver	Categorical
3	Age	People using airline services with age from 15-85	Continuous
4	Age Range	Eight age groups like 0-19, 30-39, 85+	Categorical
5	Gender	Male and Female	Categorical
6	Price Sensitivity	Degree to which consumers' behaviors are affected by the price of the airline service with 1 being the lowest	Categorical
7	Year of First Flight	In which one has taken their first flight	Categorical
8	No of Flights p.a.	Number of flights travelled per year	Continuous
9	No of Flights p.a. grouped	Frequency of flight travel in range	Categorical
10	% of Flight with other Airlines	The percentage of flights being taken with other airlines	Continuous
11	Type of Travel	Purpose of the travel like business, personal, mileage.	Categorical
12	No. of other Loyalty Cards	Number of loyalty cards a person holds to avail discounts, coupons, or some other rewards	Categorical
13	Shopping Amount at Airport	Total money spent on shopping at the airport	Continuous
14	Eating and Drinking	Total money spent on eating and drinking at the airport	Continuous

	at Airport		
15	Class	Different class of travel such as Business, Eco, and Eco Plus	Categorical
16	Day of Month	The day of the month travelled	Categorical
17	Flight date	Travel date	Categorical
18	Airline Code	Different airline code such as AA, MQ, FL	Character
19	Airline Name	Different airlines like EnjoyFlying Air Services, Oursin Airlines Inc.	Character
20	Origin City	Name of the city from where the flight departed	Character
21	Origin State	Name of the state from where the flight departed	Character
22	Destination City	Name of the city from where the flight arrived	Character
23	Destination State	Name of the state from where the flight arrived	Character
24	Scheduled Departure Hour	Flight departure hour like 15, 5, 9	Continuous
25	Departure Delay in Minutes	The delay in departure of a flight in minutes	Continuous
26	Arrival Delay in Minutes	The delay in arrival of a flight in minutes	Continuous
27	Flight cancelled	Whether flight was cancelled?	Categorical
28	Flight time in minutes	Time (in minutes) spent by the flight to reach the destination	Continuous
29	Flight Distance	Distance travelled by the flight	Continuous
30	Arrival Delay greater 5 Mins	Whether flight got delayed by more than 5 minutes?	Categorical

Data Preprocessing and Exploration

Detection of Missing values and Removal

Almost every time data has to be cleaned and pre-processed before it can move to the next step which is data analysis. Our data was gathered from a public opinion survey and because of this, it was bound to have missing values since people tend to forget answering questions, or they deliberately don't answer. Our main challenge was to have a subset of the dataset that represents the original dataset well. Further, we wanted to use XLMiner for data analysis as it is more intuitive and easy to use.

The data set was huge and had 129K records. Xlminer supports only 65K records for missing value detection and removal. Hence, R was preferred to extract random 65K records. Then, we identify the missing values in the columns and rows using a heatmap, plotted by importing function written in ImageMatrix.R. Below is the R code to take read a file, random sample of 65K records, and build a heat map:

```
dat = read.csv ('Satisfaction Survey.csv', head=T, stringsAsFactors=F)
df = data.frame(dat)
new_csv = df[sample(nrow(dat), 65000),]
class(is.na(dat_sample))
myImagePlot(is.na(dat_sample))
```

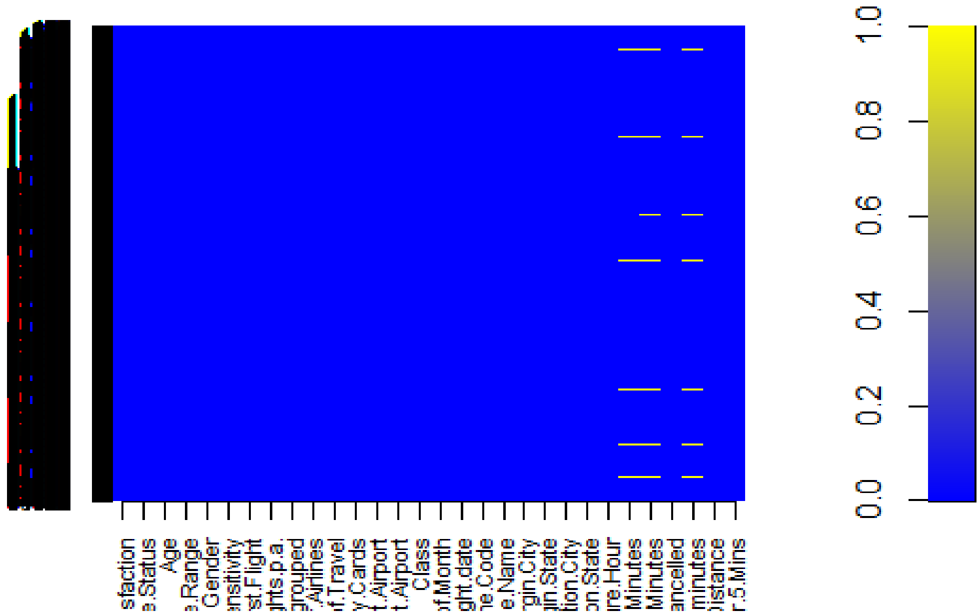


Figure 3: Heat Map

The heatmap showed that there were not major columns or rows that were completely blank or without any values (Refer Fig 3 - heatmap). Considerable amount of missing values were found in three columns – 'Departure.Delay.in.Minutes', 'Arrival.Delay.in.Minutes', and 'Flight.time.in.minutes'. A total of 1355 rows had one or more columns with missing values. All these rows were removed using an R code that stores the row id with missing values and deletes the entire row (Refer Fig 4 – Removing Missing Values). Below is the R code used to check the missing values and take out rows with missing values:

```
## check missing with for loop
for (ii in 1:ncol(dat_sample)) {
  print( colnames(dat_sample)[ii] )
  print( table(is.na(dat_sample[,ii])) )
}
## take out columns with many missings
id.row = apply (dat_sample, 1, function(x) sum(is.na(x)) == 0)
```

Below are the screenshots of the before and after count of the data set:

```
[1] "Departure.Delay.in.Minutes"

FALSE  TRUE
63839  1161
[1] "Arrival.Delay.in.Minutes"

FALSE  TRUE
63645  1355
[1] "Flight.cancelled"

FALSE
65000
[1] "Flight.time.in.minutes"

FALSE  TRUE
63645  1355
```

```
[1] "Departure.Delay.in.Minutes"

FALSE
63645
[1] "Arrival.Delay.in.Minutes"

FALSE
63645
[1] "Flight.cancelled"

FALSE
63645
[1] "Flight.time.in.minutes"

FALSE
63645
```

Figure 4 – Removing Missing Values

Selecting the sample

The original data set was huge with 129K records. The major challenge to work with this data set was the size and count of records in the data set. Also, XLMiner only supports 10K records for model building with a limitation of handling missing values for only 65K records, the obvious choice was to use R programing. We partitioned sampled dataset into training and test set with 13.985% and 86.015% weightage. Below is the R code to partition the data:

```
id.train = sample(1:nrow(dat), nrow(dat)*. 13985) # ncol() gives number of columns
id.test = setdiff(1:nrow(dat), id.train) # setdiff gives the set difference
```

From the original data set of 129K, a random sample of 65K records are treated for missing values and partitioned into training and test data set with 8.9K and 54K respectively. For our data analysis and model building, we will be using training dataset which is easier to analyze both in XLMiner and R. Moreover, the data sampling will be random to avoid any bias-ness or overfitting in the model.

Transformation of Data

The dataset has a total of 30 variables out of which 15 are categorical variables and remaining are numerical variables. Out of the 15 categorical variables, 12 variables are nominal, and the others are ordinal. Not all variables are useful for building a model. Hence some variables will be eliminated during the data mining process. The next step was to perform collinearity check. All the independent variables were checked for multicollinearity with other independent variables. One the basis of the result of the two variables that are highly correlated was removed.

The output variable – Satisfaction, had values ranging from 0 to 1. It signified the satisfaction score given by the customer. We dichotomized the variable with a rule that if satisfaction score is between 0 to 3.5 then the corresponding value in the column is '0' – bad or if the score is between 4 to 5 then the corresponding value in the column is '1' – good. Similar dichotomization techniques were applied other categorical variables in the data set that had one value contributing to half of the count in that variable and others contributing to the other half.

For the categorical variables with equal distribution among values, dummy variables technique was used removing the most prevalent column after the transformation. Few of the character variables that signified quality information were also categorized using combination of aggregation and dummy categorization techniques. Continuous variables were retained without any transformation.

Summarization of data

Overall summary of the cleaned data is that it has 8.9K records, 26 variables out of which 18 are binary variables, 8 are continuous variables and no character variables. Adjoining table compares original data set with the final cleaned data set.

	Original	Final
Total Records	129K	8.9K
Variables count	30	26
Categorical Variables	14	18-binary
Continuous Variables	10	8
Character Variables	6	0

Data Mining Techniques

We would like to use this data to develop a prediction model that will help in the evaluation and improvement of airlines' performance. Also, we know that our output variable being used for the mining task is a categorical variable which leads to use the classification methods to evaluate our model. The overall satisfaction score will be used as the dependent variable and other relevant variables will be evaluated based on their impact on the dependent variable.

A different set of classification techniques that will be applied using XLMiner and R programming to evaluate our model are mentioned below:

- Logistic Regression
  - Forward Selection
  - Backward Selection
  - Stepwise selection
- KNN
- Classification Tree

The above models will be used to identify the important variables that will serve as the core of our model.

Logistic Regression

- **Forward Selection**

For building Logistic Regression model, we use **Excel-XLminer-Classify** function. After setting default cutoff probability for success of 0.5 and confidence level of 95%, the optimistic variable subset was selected by XLminer using forward selection. As shown below, when there are 10 variables, cp is closest to the # of coefficient, so we choose the model with 10 variables.

Choose Subset	9	7464.367	12.632	0.1935	Intercept	Type.of.Tr	Airline.Sta	Arrival.Del	Ag
Choose Subset	10	7458.647	8.7703	0.4682	Intercept	Type.of.Tr	Airline.Sta	Arrival.Del	Ag

Therefore, the model is:

Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	0.519571	0.153679	11.43039625	0.000723	1.681306	1.244042	2.272262
Airline.Stat	-1.23429	0.063764	374.7000704	1.77E-83	0.291041	0.256849	0.329784
Price.Sensi	0.180094	0.061057	8.700198004	0.003182	1.197329	1.062287	1.349538
Gender_M	0.41894	0.057638	52.83088412	3.64E-13	1.520348	1.357944	1.702176
Type.of.Tra	2.119555	0.062427	1152.774675	1.1E-252	8.327433	7.3684	9.411289
Class_Eco	-0.17127	0.07152	5.734507095	0.016635	0.842596	0.732389	0.969387
Arrival.Del	-0.85	0.060048	200.3789731	1.73E-45	0.427413	0.379958	0.480795
Age	-0.01709	0.001769	93.34862504	4.38E-22	0.983056	0.979653	0.98647
No.of.Fligh	-0.01686	0.002146	61.69994365	4E-15	0.983285	0.979159	0.98743
Scheduled.	0.02157	0.006143	12.32853868	0.000446	1.021804	1.009575	1.034182

Plug the number to Logit, we got:

Log(odds)=0.519-1.234Airline.Status_blue+0.18Price.Sensitivity_1+0.419Gender_Male+2.120Type.of.travel_business-0.171Class_Eco-0.85Arrival.Delay.greater.5.Mins-0.017age-0.017No.of.Flight+0.022Scheduled.departure.hour

Here is the evaluation of the model performance:

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix

	Predicted Class	
Actual Clas	1	0
1	3101	776
0	1106	2672

Error Report

Class	# Cases	# Errors	% Error
1	3877	776	20.0154759
0	3778	1106	29.2747485
Overall	7655	1882	24.5852384

Performance

Success Class	1
Precision	0.7371
Recall (Sensitivity)	0.79985
Specificity	0.70725
F1-Score	0.76719

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix

	Predicted Class	
Actual Clas	1	0
1	558	123
0	178	386

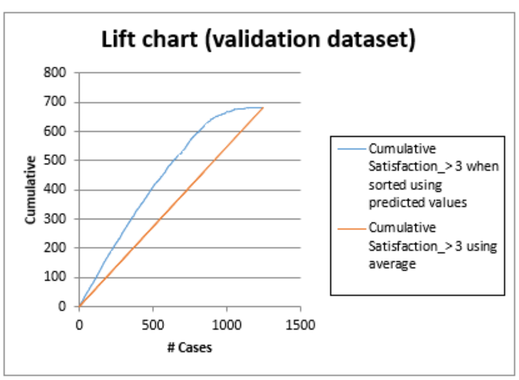
Error Report

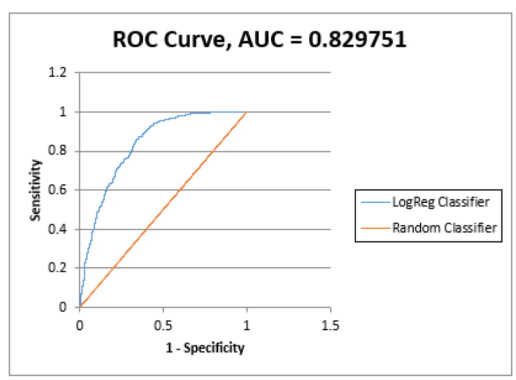
Class	# Cases	# Errors	% Error
1	681	123	18.061674
0	564	178	31.5602837
Overall	1245	301	24.1767068

Performance

Success Class	1
Precision	0.75815
Recall (Sensitivity)	0.81938
Specificity	0.6844
F1-Score	0.78758

The lift chart and ROC visually measured the model performance.





- **Backward Elimination**

Under the Backward Elimination method, 12-variables subset is considered as the optimal model, because its cp is closed to the # of variables.

Choose Subset	13	8694.51	6.5669	0.8848	Intercept	Airline.Sta	Price.Sens	Gender_N	Flight.Date_Jan		
Choose Subset	12	8696.12	6.2096	0.8296	Intercept	Airline.Sta	Price.Sens	Gender_Male			

The system generated the regression model:

Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	0.528306	0.14409	13.44320496	0.000246	1.696058	1.278765	2.249523
Airline.Stat	-1.26525	0.059459	452.8084577	1.8E-100	0.282168	0.251128	0.317044
Price.Sensi	0.162369	0.056678	8.206981959	0.004173	1.176295	1.052621	1.314499
Gender_M	0.456674	0.053721	72.26422406	1.88E-17	1.578814	1.421031	1.754116
Type.of.Tra	2.109512	0.057989	1323.351856	9.5E-290	8.244216	7.358498	9.236545
Class_Eco	-0.16907	0.066149	6.5330222	0.010589	0.844446	0.741765	0.961342
Arrival.Del	-0.79378	0.06391	154.2615912	2.03E-35	0.452133	0.398902	0.512468
Age	-0.01736	0.001644	111.4355615	4.75E-26	0.982793	0.979631	0.985965
No.of.Fligh	-0.01691	0.002003	71.26340231	3.13E-17	0.983234	0.979381	0.987101
Shopping.A	0.001106	0.000483	5.253150113	0.021907	1.001107	1.00016	1.002054
Scheduled.	0.023716	0.005709	17.25492274	3.27E-05	1.024	1.012605	1.035523
Departure.	-0.00238	0.000876	7.409059413	0.00649	0.997619	0.995908	0.999333

Log(odds)=0.519-1.265Airline.Status_blue+0.162Price.Sensitivity_1+0.457Gender_Male+2.11Type.of.travel_business-0.169Class_Eco-0.794Arrival.Delay.greater.5.Mins-0.017age-0.017No.of.Flight+0.001Shopping.Amount.at.Airport+0.024Scheduled.Departure.Hour-0.002Departure.Delay.in.Minutes

Here is the evaluation of the model performance:

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	2180	533
0	773	1854

Error Report			
Class	# Cases	# Errors	% Error
1	2713	533	19.64614818
0	2627	773	29.42519985
Overall	5340	1306	24.45692884

Performance	
Success Class	1
Precision	0.738232
Recall (Sensitivity)	0.803539
Specificity	0.705748
F1-Score	0.769502

Validation Data Scoring - Summary Report

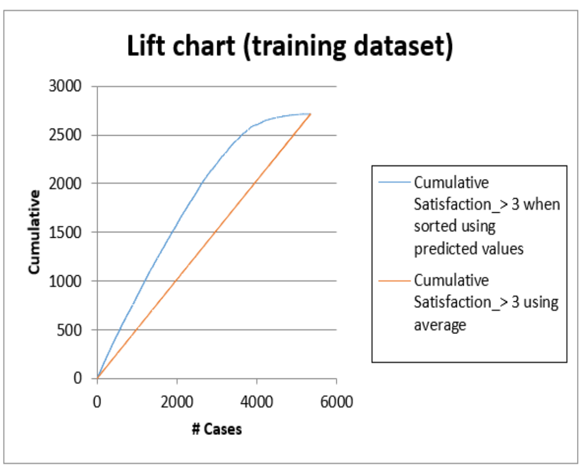
Cutoff probability value for success (UPDATABLE)

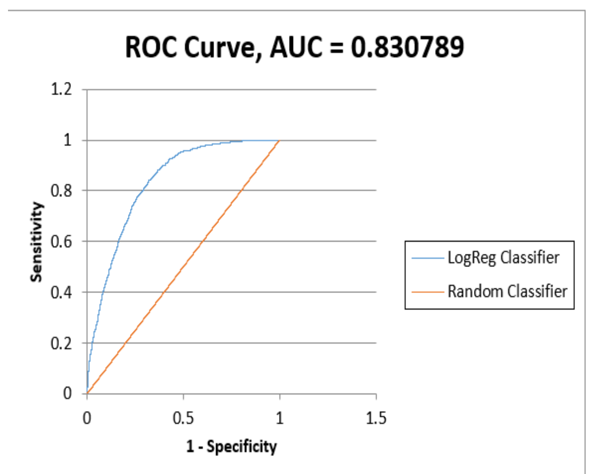
Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	1480	365
0	516	1199

Error Report			
Class	# Cases	# Errors	% Error
1	1845	365	19.78319783
0	1715	516	30.08746356
Overall	3560	881	24.74719101

Performance	
Success Class	1
Precision	0.741483
Recall (Sensitivity)	0.802168
Specificity	0.699125
F1-Score	0.770633

The lift chart and ROC measure the model performance visually.





- **Stepwise selection**

By using Stepwise variable selection, 8-variables subset is selected, because its cp is closed to the # of variables:

Choose Subset	7	5234.271	19.5949	0.0248	Intercept	Airline.Status_Blue	Genc
Choose Subset	8	5225.402	12.5373	0.1655	Intercept	Airline.Status_Blue	Genc

The new regression model is:

Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	0.591735	0.166588	12.6173695	0.000382	1.807121	1.30373	2.504879
Airline.Stat	-1.28269	0.076362	282.1558835	2.55E-63	0.277291	0.238746	0.322058
Gender_M	0.426195	0.06885	38.31882281	6.01E-10	1.53142	1.338102	1.752666
Type.of.Tra	2.112911	0.074466	805.0925196	4.2E-177	8.27229	7.148915	9.572191
Arrival.Del	-0.87237	0.07226	145.7504128	1.47E-33	0.41796	0.362767	0.481551
Age	-0.01768	0.002106	70.43436998	4.76E-17	0.982479	0.978432	0.986543
No.of.Fligh	-0.01721	0.002562	45.10155782	1.87E-11	0.98294	0.978017	0.987889
Scheduled.	0.022118	0.007378	8.987462131	0.002718	1.022365	1.007687	1.037256

Log(odds)=0.592-1.283Airline.Status_blue+0.426Gender_Male+2.113Type.of.travel_business-0.872Arrival.Delay.greater.5.Mins-0.018age-0.017No.of.Flight+0.022Scheduled.Departure.Hour

Here is the evaluation of the model performance:

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix

	Predicted Class	
Actual Class	1	0
1	2176	537
0	768	1859

Error Report

Class	# Cases	# Errors	% Error
1	2713	537	19.79358644
0	2627	768	29.23486867
Overall	5340	1305	24.43820225

Performance

Success Class	1
Precision	0.73913
Recall (Sensitivity)	0.80206
Specificity	0.70765
F1-Score	0.76931

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix

	Predicted Class	
Actual Class	1	0
1	1474	371
0	526	1189

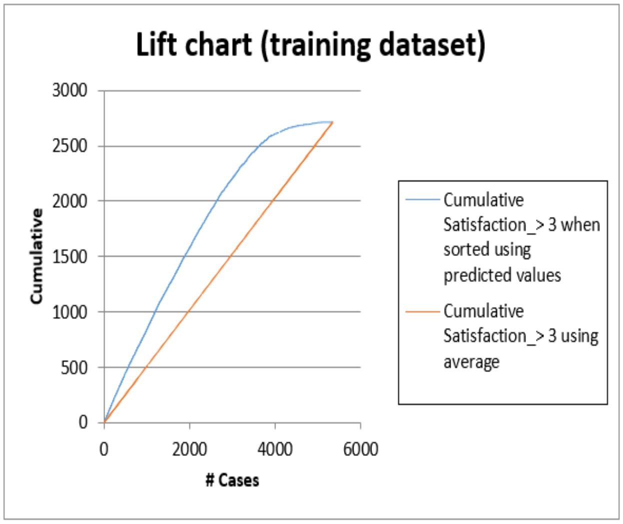
Error Report

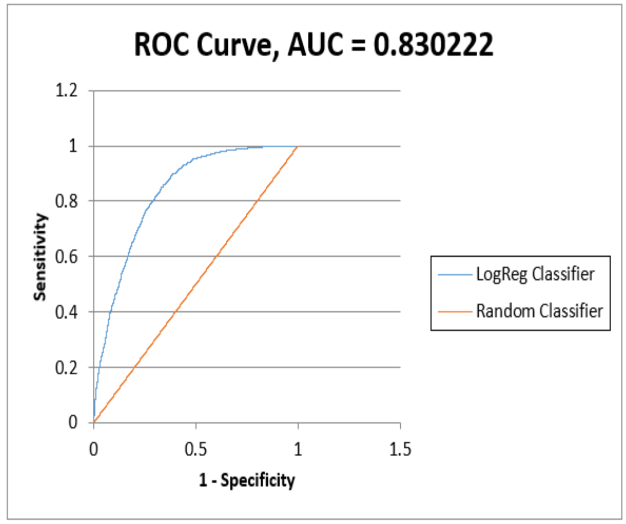
Class	# Cases	# Errors	% Error
1	1845	371	20.1084
0	1715	526	30.6706
Overall	3560	897	25.1966

Performance

Success Class	1
Precision	0.737
Recall (Sensitivity)	0.79892
Specificity	0.69329
F1-Score	0.76671

The lift chart and ROC measure the model performance visually.





KNN (K- Nearest Neighbor)

In this method, XLminer uses “nearest neighbor” to predict the satisfaction, the predicted numbers are determined by records near to them . K=12 (nearest 12 records) was chosen by the system, since its error rate is the lowest at -33.41% (shown below).

Validation error log for different k

Value of k	% Error Training	% Error Validatio	
1	0	39.3574	
2	22.26	38.5542	
3	20.4703	37.5904	
4	25.5258	35.1807	
5	25.4213	35.3414	
6	27.6421	34.5382	
7	27.2371	35.9036	
8	28.6871	34.6988	
9	27.6029	34.2972	
10	28.9223	34.2169	
11	28.7394	35.1807	
12	29.8106	33.4137	<- Best k
13	28.9615	34.3775	
14	30.098	34.1365	
15	29.7975	34.3775	
16	30.1502	33.5743	
17	30.0588	34.1365	
18	30.6074	33.9759	
19	30.3592	34.3775	
20	31.2214	33.8153	

We chose to go with k = 13 instead to have an odd number which will avoid conflicts in classification. Confusion Matrix gives us the first impression about model performance. For the validation dataset, 65% of the total record are correctly predicted. Its Specificity and Sensitivity shows the correctness of prediction on 1s and 0s respectively.

Validation Data Scoring - Summary Report (for k = 13)

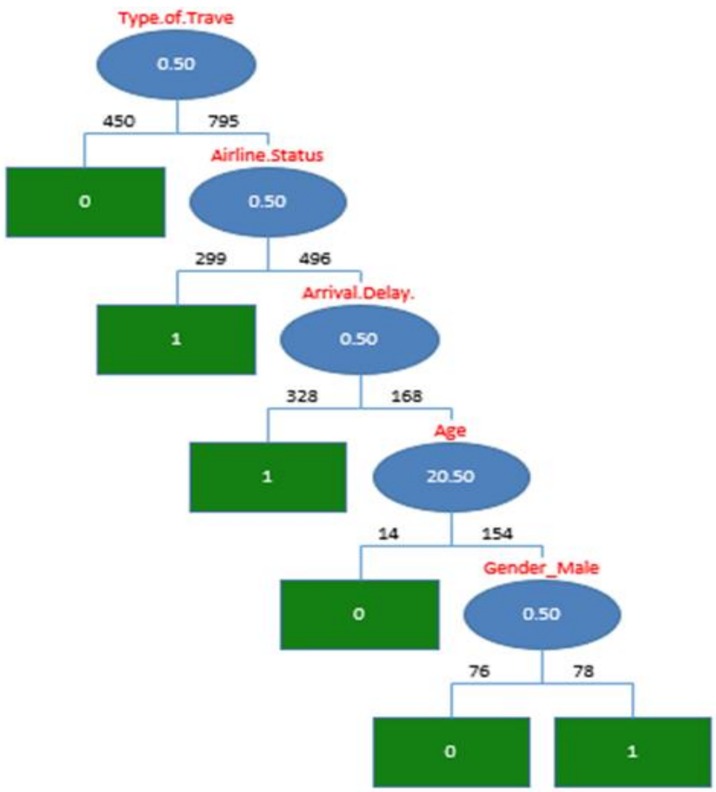
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	520	161
0	265	299

Error Report			
Class	# Cases	# Errors	% Error
1	681	161	23.6417
0	564	265	46.98582
Overall	1245	426	34.21687

Performance	
Success Class	1
Precision	0.66242
Recall (Sensitivity)	0.763583
Specificity	0.530142
F1-Score	0.709413

Classification Tree

In doing the classification tree predictive method, Best Pruned Tree is being used in the following processes. 1 is considered as success class. In this method, dependent variable-airline satisfaction are predicted by its cutoffs (split value) of its every independent variable. As shown below, there are 5 decision nodes, each with different split value. The 5 variables chosen to predict classification are: Type of Travel Business, Airline Status Blue, Arrival Delay > 5 mins, Age, and Gender_male.



Validation Data scoring - Summary Report (Using Best Pruned Tree)

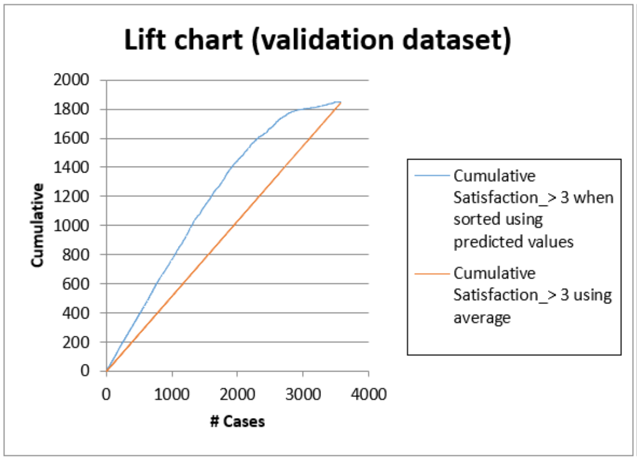
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	540	141
0	165	399

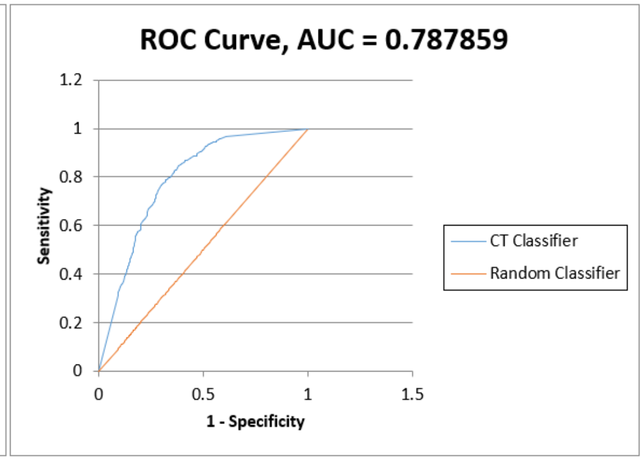
Error Report			
Class	# Cases	# Errors	% Error
1	681	141	20.70485
0	564	165	29.25532
Overall	1245	306	24.57831

Performance	
Success Class	1
Precision	0.765957
Recall (Sensitivity)	0.792952
Specificity	0.707447
F1-Score	0.779221

The confusion matrix above shows us number records that are correctly classified/misclassified. The overall error rate is 24.57%. To be more specific, the Specificity is 0.70, means 70% of 0s are correctly classified, and Sensitivity of 0.79 means 79% of success class are correctly classified.

The lift chart and ROC Curves tell us a bit more about the tree performance:





Compared to the models we have so far, The ROC Curve of classification tree looks good with a large area between two lines and the AUC is higher at 0.79.

Model Interpretation and Comparison

Forward Selection			
Error Report			
Class	# Cases	# Errors	% Error
1	681	123	18.20
0	564	178	31.38
Overall	1245	301	**24.17**
Number of variables			10

Classification Tree			
Error Report			
Class	# Cases	# Errors	% Error
1	861	141	21.24
0	564	165	28.51
Overall	1245	306	**24.57**
Best Prune Nodes			5

KNN			
Error Report			
Class	# Cases	# Errors	% Error
1	681	161	23.6417
0	564	265	46.98582
Overall	1245	426	**34.21687**
Number of K's			13

Conclusion

From the information that we gathered from our models, the logistic regression model with forward selection proves to be the most accurate model in determining the satisfaction rate of passengers for the airline. While most of the models are fairly similar to each other in results, less the KNN model, the forward selection model has a slight advantage in terms of its performance metrics.

Among the five classification methods sampled, the results of the variables produced by forward selection has the most favorable outcomes. This is because of the following reasons:

- The forward selection model has a sensitivity rate of 81.9%, which is a very good percentage in accurately classifying which responses were “1” or a satisfied score.

- The forward selection model has a specificity rate of 68.4%, which is a decent percentage in classifying the response which were “0” or an unsatisfied score.
- Its AUC of ROC Curve may be slightly lower than backward selection; however, at 82.9% it is relatively the same as backward’s 83.1%.
- It has lowest total error rate of 24.17%. The total accuracy rate is 75.8%, highest among three methods.

With these results, we would be confident in presenting to the airline a proficient model that would determine customer satisfaction rate:

Log(odds)=0.519-1.265Airline.Status_blue+0.162Price.Sensitivity_1+0.457Gender_Male+2.11Type.of.travel_business-0.169Class_Eco-0.794Arrival.Delay.greater.5.Mins-0.017age-0.017No.of.Flight+0.001Shopping.Amount.at.Airport+0.024Scheduled.Departure.Hour-0.002Departure.Delay.in.Minutes

With this model at hand, the airline could begin examining the 9 main variables that contribute to the overall satisfaction rate of its passengers. These variables are:

1. Airline Status
2. Price Sensitivity
3. Gender
4. Type of Travel
5. Class
6. Arrival Delay
7. Age
8. Number of Flights
9. Scheduled Departure Hour

Specifically, the airline should try to satisfy people traveling for business because this would greatly affect the satisfaction. According to our equation, the airline should also reduce delays in arrival time because the greater the delay is, the lower the satisfaction rate becomes. By focusing their efforts in developing a strong foundation for the above-mentioned criteria, the airline will see positive results in achieving overall customer satisfaction in its future flights.

References

1. ***Federal Aviation Administration****.* Air Traffic by the Numbers, 2018. https://www.faa.gov/air_traffic/by_the_numbers/
2. **"Measurement of Airline Passenger Satisfaction: A Comparison of Methodology"** by Jason A. Mlady, John P. Young, Purdue University, 12 April 2013
3. "**The Quality – Profitability Link in the US Airline Business: A Study Based on the Airline Quality Rating Index"** by Nicole Kalemba, Fernando Campa-Planas, University Rovira and Virgili, Reus, Spain
4. **"Customer satisfaction in the airline industry: the role of service quality and price"** by Dwi Suhartanto, the department of business administration, Bandung state polytechnic, any Ariani Noor, the department of business administration, Bandung state polytechnic