SOUTHEAST AIRLINES

The Spirit of Liberty™

Business Utilization of Net Promoter Score

Zequn Che | Jaishree Palaniswamy | Rashmitha Varma Pandati
Patrick Prioletti | Shreyas Raghavan Sadagopan

THE
iSCHOOL
Syracuse University

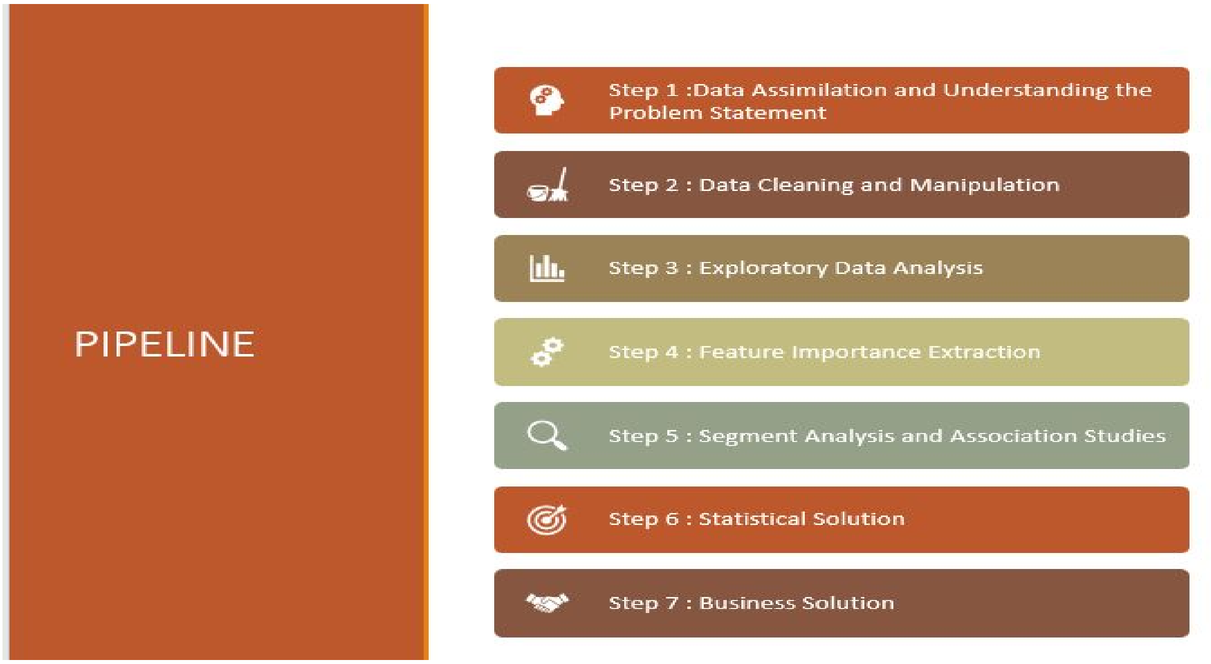
Table of Contents

Problem Statement

For helping the client Southeast Airline to enhance their quality of service, we designed this project for researching the elements which will influence the satisfaction of their clients. In this project, we will create meaningful segments of travellers and to study how their ‘Likelihood to recommend’ varies within that segment with respect to the other variables and then make several recommendations on how to increase the number of NPS based on our data analysis.

Net Promoter Score

Southeast Airlines handed out surveys to their customers to measure how likely they are to recommend the airline to a friend or colleague. The idea behind this is to gather sentiment of what the customers’ experience was like. There are many implications of using Net Promoter Score (NPS) as a measurement. NPS can be categorized by three groups with distinctly different characteristics and impacts on customer brand perception and churn rate. Customers who rated their likelihood to recommend from 1-6 are considered detractors, 7 and 8 are considered passive, and 9 and 10 are considered promoters. Detractors are 1.5 times more likely to stop using a product in comparison to promoters.



PIPELINE

Step 1 :Data Assimilation and Understanding the Problem Statement

Step 2 : Data Cleaning and Manipulation

Step 3 : Exploratory Data Analysis

Step 4 : Feature Importance Extraction

Step 5 : Segment Analysis and Association Studies

Step 6 : Statistical Solution

Step 7 : Business Solution

Data Understanding

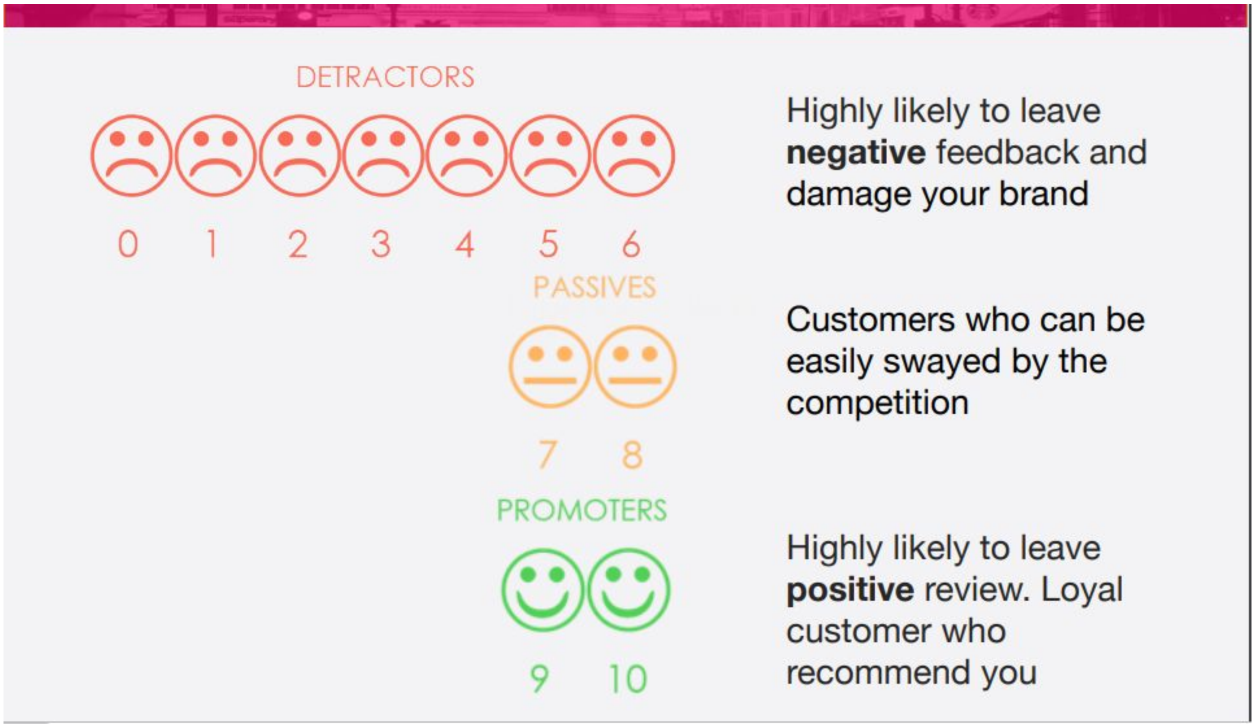
Southeast Airline gathered a lot of raw data by doing survey. The data contains more than 20 variables.

The data set given is at a ‘Flight Journey’ level .

The Explanatory variables can be classified into:

- Traveller Demographic Variables : Age , Gender etc
- Traveller Persona Variables : Price Sensitivity, Loyalty , Amount spent at airports during travel, Frequent Flyer Tags etc.
- Air Carrier Variables : Airline partner, Airline Status etc.
- Travel Variables : Origin City, Flight Cancelled, Duration , Delay etc.
- Other Variables : Latitude, Longitude, Date time etc.

The Likelihood to Recommend is taken as our Y variable for the Analysis.



Highly likely to leave **negative** feedback and damage your brand

Customers who can be easily swayed by the competition

Highly likely to leave **positive** review. Loyal customer who recommend you

Data Manipulation (Binning and NA imputations)

For categorical variables, we converted all of them into factors. For Continuous variables with large ranges, we generated the quantiles to look at the distribution. Then we bucketed the variables into 4 to a maximum of 10 levels ensuring there is a pattern in the population across the buckets. For example, we bucketed the age variable as 24 ,25-35, 36-46, 47-58, 57-71, 71+.

For continuous variables with small range we first generated the quantiles to look at the distribution. Then we created a maximum of 3 levels ensuring logic and the population density is maintained across the buckets. For example, the arrival delay is bucketed as 0, 0-60 Mins, 60+ Minutes.

For cleaning the missing data, we imputed all the NAs with the mode of them. For the arrival and delay minutes, if they are NA, they will be set to “0” (when the flight is cancelled).

VARIABLE SELECTION

1 . Reduction Using Business Logic and Correlation matrix:

- Removed variables which have more than multiple levels Date, Latitude/Longitudes, Origin and Destination city
- Removed Eating amount spent, Shopping at airport,Partner code with respect to the Corr Matrix and business requirements

Total variables 16 at the End of this Step

2. Feature Importance Generation to find the most significant Variables

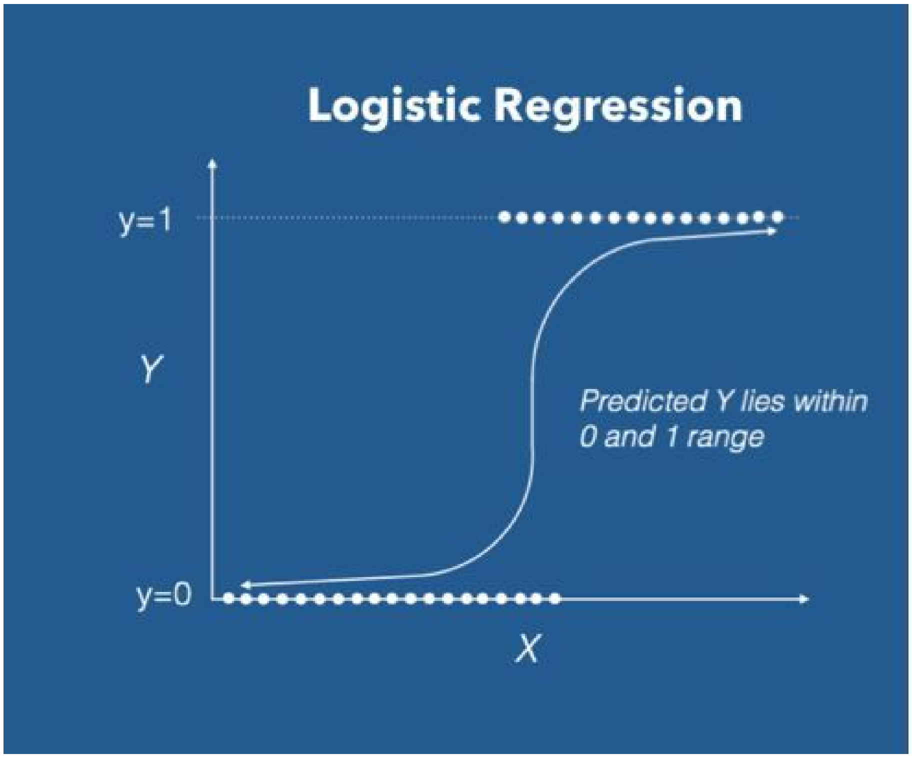
- In order to get optimum Feature Importance, we have built 3 models
  - Logistic Regression
  - Random Forest Classifier
  - Gradient Boost Classifier
- We compare the Relative Importance in each of the models to figure out the best combination of variables

LOGISTIC MODEL

Logistic regression is yet another technique borrowed by machine learning from the field of statistics. It's a powerful statistical way of modeling a binomial outcome with one or more explanatory variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

As the name already indicates, logistic regression is a regression analysis technique. Regression analysis is a set of statistical processes that you can use to estimate the relationships among variables. More specifically, you use this set of techniques to model and analyze the relationship between a dependent variable and one or more independent variables. Regression analysis helps you to understand how the typical value of the dependent variable changes when one of the independent variables is adjusted and others are held fixed.

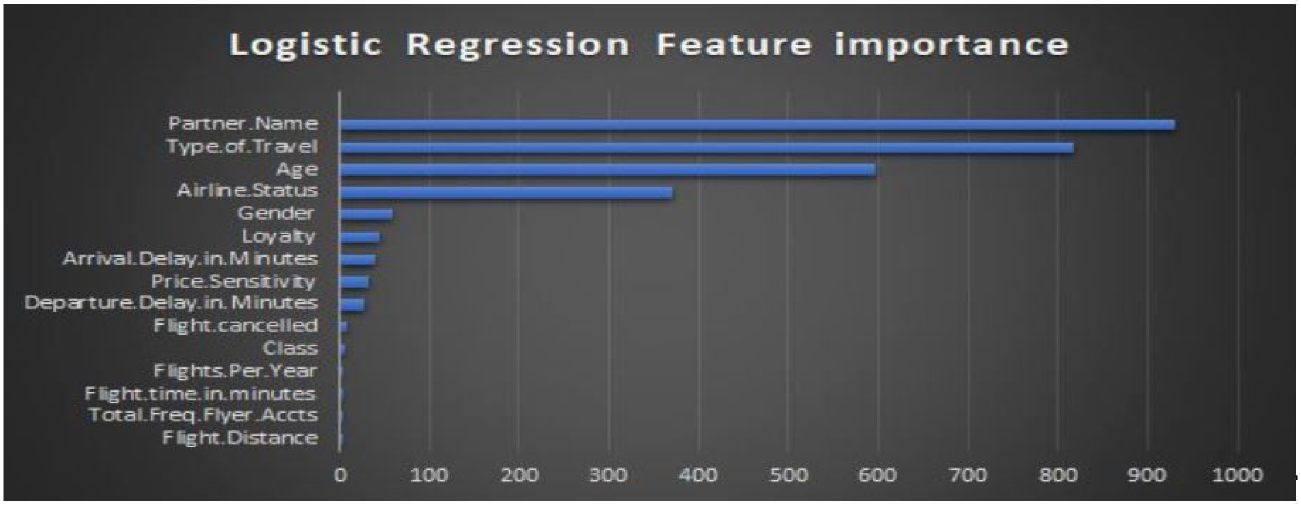
Logistic regression is an instance of classification technique that you can use to predict a qualitative response. The logistic function will always produce an S-shaped curve, so regardless of the value of *X*, we will obtain a sensible prediction.



Models : Logistic Regressor

Model Stats : Gave a **74 %** Accuracy when validated using a test set.

AUC: **0.8376627**



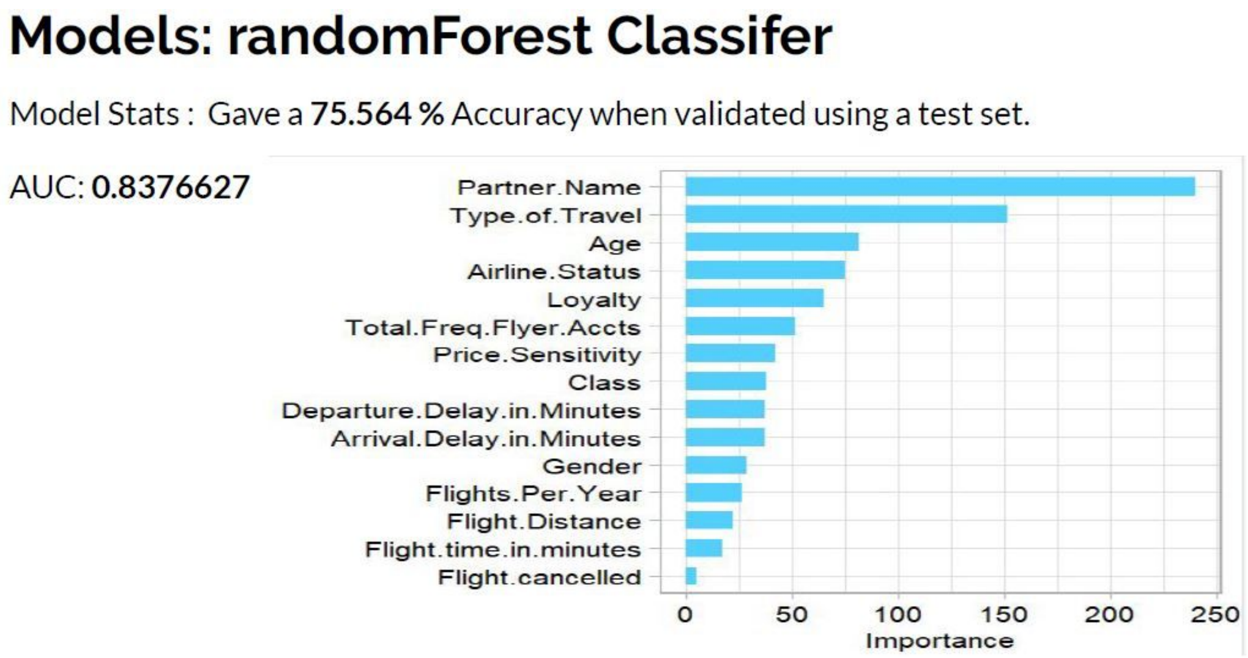
RANDOM FOREST MODEL

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science language, the reason that the random forest model works so well is: ***A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.*** The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. **The reason for this wonderful effect is that the trees protect each other from their individual errors** (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a "forest"), this model uses two key concepts that gives it the name *random*:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes



Models: randomForest Classifer

Model Stats : Gave a **75.564 %** Accuracy when validated using a test set.

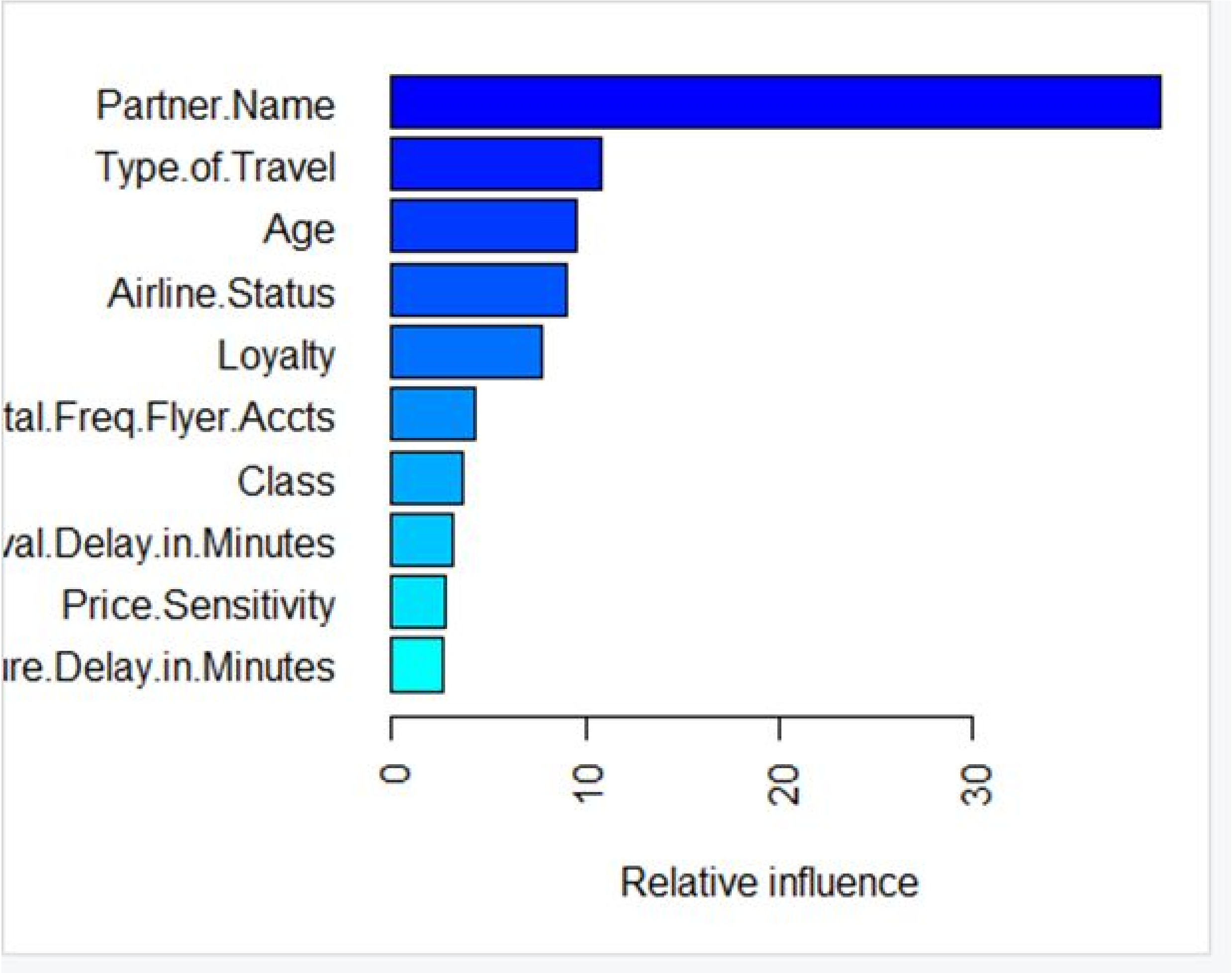
AUC: **0.8376627**

GRADIENT BOOST MODEL:

Gradient Boost Model is which produces a prediction model in the form of an ensemble of weak prediction models. The objective of any supervised learning algorithm is to define a loss function and minimize it.By using gradient descent and updating our predictions based on a learning rate, we can find the values where MSE is minimum.

The gradient is used to minimize a loss function. In each round of training, the weak learner is built and its predictions are compared to the correct outcome that we expect. The distance between prediction and truth represents the error rate of our model. These errors can now be used to calculate the gradient. The gradient is nothing fancy, it is basically the partial derivative of our loss function - so it describes the steepness of our error function. The gradient can be used to find the direction in which to change the model parameters in order to (maximally) reduce the error in the next round of training by “descending the gradient”.

Based on that our model is ran and got the following output.



A gradient boosted model with gaussian loss function. 5000 iterations were performed. The best cross-validation iteration was 80. There were 15 predictors of which 13 had non-zero influence.

ANALYSIS OVERVIEW

In the airlines dataset we have around twenty seven variables, based on which we can analyse the net promoter score. To narrow down and to concentrate on the variables who are highly significant, we have built three models (Logistic Regression, Random forest classifier and Gradient boost classifier). Here, based on the results of the above model we have picked up 8 common variables for further analysis.

The top eight common variables are:

1. PartnerName
2. Age
3. Airline Status
4. Type of travel
5. Loyalty
6. Total freq Flyer
7. Price Sensitivity
8. Class.

With the picked up 8 variables we have done the segment analysis and association studies with the below two models:

i. Apriori Algorithm

ii. Clustering Using KPrototype

MODEL: APRIORI ALGORITHM

What is Apriori Algorithm ?

Apriori algorithm, a classic algorithm, is useful in mining frequent itemsets and relevant association rules. Usually, we can operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a supermarket. It helps customers buy their items with ease, and enhances the sales performance of the departmental store.

There are three significant components that comprises the apriori algorithm. They are as follows.

- Support
- Confidence
- Lift

Let us suppose that we have 2000 customer transactions in a supermarket. You have to find the Support, Confidence, and Lift for two items, say bread and jam. It is because people frequently bundle these two items together.Out of the 2000 transactions, 200 contain jam whereas 300 contain bread. These 300 transactions include a 100 that includes bread as well as jam. Using this data, we shall find out the support, confidence, and lift.

Support

Support is the default popularity of any item. We can calculate the support as a quotient of the division of the number of transactions containing that item by the total number of transactions.

Hence, in our example,

Support (Jam) = (Transactions involving jam) / (Total Transactions)

= 200/2000 = 10%

Confidence

In our example, Confidence is the likelihood that customers bought both bread and jam. Dividing the number of transactions that include both bread and jam by the total number of transactions will give the Confidence figure.

Confidence = (Transactions involving both bread and jam) / (Total Transactions involving jam)

= 100/200 = 50%

It implies that 50% of customers who bought jam bought bread as well.

Lift

According to our example, Lift is the increase in the ratio of the sale of bread when you sell jam. The mathematical formula of Lift is as follows.

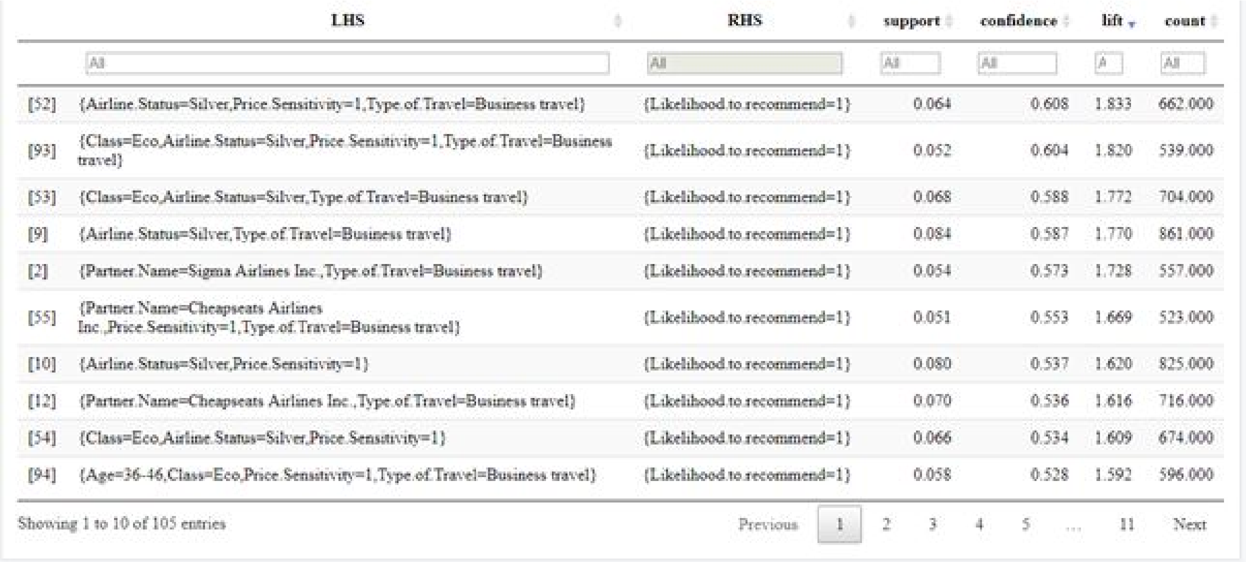
Lift = (Confidence (Jam→– Bread)) / (Support (Jam))

= 50 / 10 = 5

It says that the likelihood of a customer buying both jam and bread together is 5 times more than the chance of purchasing jam alone. If the Lift value is less than 1, it entails that the customers are unlikely to buy both the items together. Greater the value, the better is the combination.

Similarly, in our airlines database, we ran the algorithm both for Likelihood to recommend = Yes and No. We have formulated the values for support, confidence and Lift for each of the combinations of levels of the X variables to the Y variable. We have got the best possible combinations by filtering for the higher Lift values for these associations. We choose the associations which make business sense and can be developed into actionable insights.

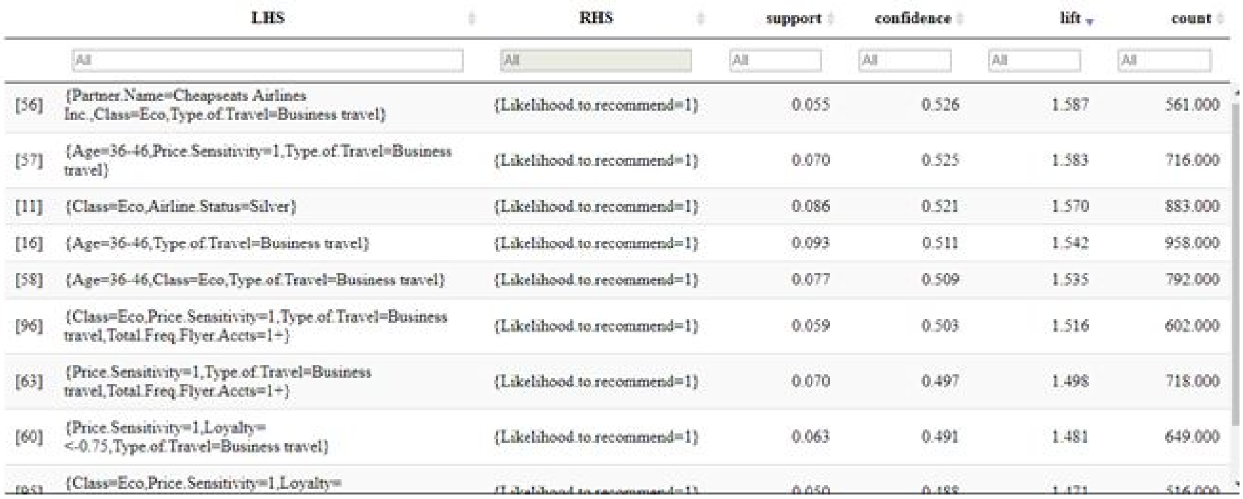
TOP 10 ASSOCIATIONS FOR Y = 1(Promoters)



	LHS	RHS	support	confidence	lift	count
	All	All	All	All	A	All
[52]	{Airline.Status=Silver,Price.Sensitivity=1,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.064	0.608	1.833	662.000
[93]	{Class=Eco,Airline.Status=Silver,Price.Sensitivity=1,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.052	0.604	1.820	539.000
[53]	{Class=Eco,Airline.Status=Silver,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.068	0.588	1.772	704.000
[9]	{Airline.Status=Silver,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.084	0.587	1.770	861.000
[2]	{Partner.Name=Sigma Airlines Inc.,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.054	0.573	1.728	557.000
[55]	{Partner.Name=Cheapseats Airlines Inc.,Price.Sensitivity=1,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.051	0.553	1.669	523.000
[10]	{Airline.Status=Silver,Price.Sensitivity=1}	{Likelihood.to.recommend=1}	0.080	0.537	1.620	825.000
[12]	{Partner.Name=Cheapseats Airlines Inc.,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.070	0.536	1.616	716.000
[54]	{Class=Eco,Airline.Status=Silver,Price.Sensitivity=1}	{Likelihood.to.recommend=1}	0.066	0.534	1.609	674.000
[94]	{Age=36-46,Class=Eco,Price.Sensitivity=1,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.058	0.528	1.592	596.000

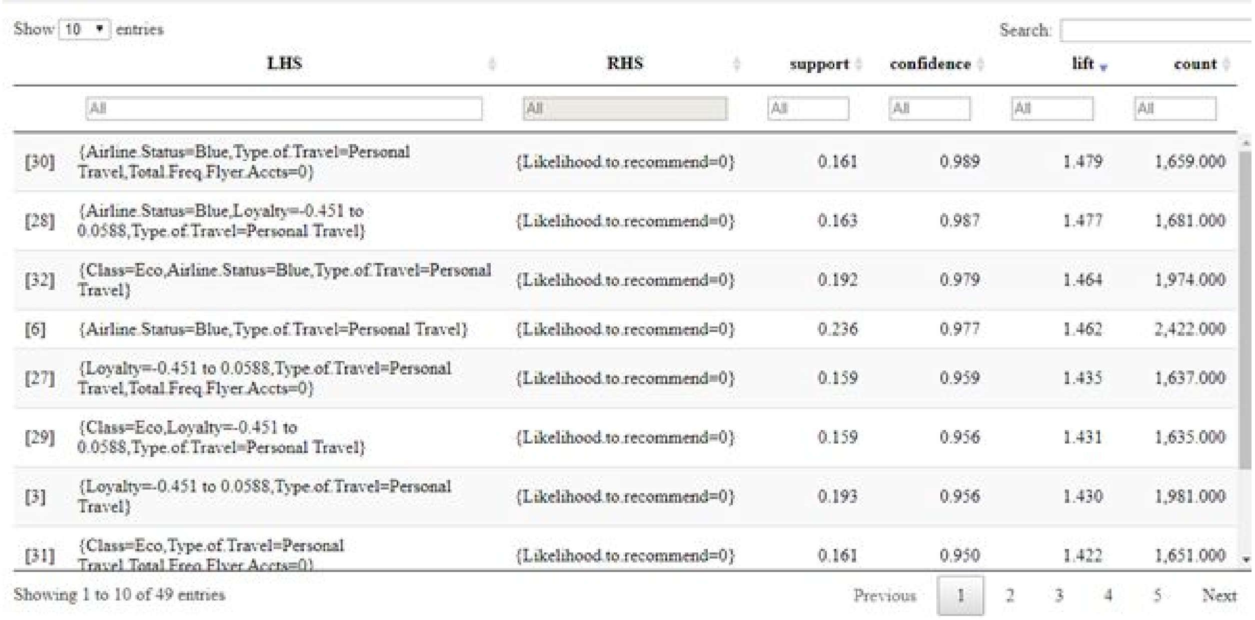
Showing 1 to 10 of 105 entries

For a promoter(Y=1) , we have highest lift value(1.833) associated with the Airline Status-Silver, Price Sensitivity- and type of travel as business travel. So these variables are highly significant to promote the net promoter score



	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[56]	{Partner.Name=Cheapseats Airlines Inc.,Class=Eco,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.055	0.526	1.587	561.000
[57]	{Age=36-46,Price.Sensitivity=1,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.070	0.525	1.583	716.000
[11]	{Class=Eco,Airline.Status=Silver}	{Likelihood.to.recommend=1}	0.086	0.521	1.570	883.000
[16]	{Age=36-46,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.093	0.511	1.542	958.000
[58]	{Age=36-46,Class=Eco,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.077	0.509	1.535	792.000
[96]	{Class=Eco,Price.Sensitivity=1,Type.of.Travel=Business travel,Total.Freq.Flyer.Accts=1+}	{Likelihood.to.recommend=1}	0.059	0.503	1.516	602.000
[63]	{Price.Sensitivity=1,Type.of.Travel=Business travel,Total.Freq.Flyer.Accts=1+}	{Likelihood.to.recommend=1}	0.070	0.497	1.498	718.000
[60]	{Price.Sensitivity=1,Loyalty=<-0.75,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.063	0.491	1.481	649.000
[95]	{Class=Eco,Price.Sensitivity=1,Loyalty=	{Likelihood.to.recommend=1}	0.050	0.488	1.471	516.000

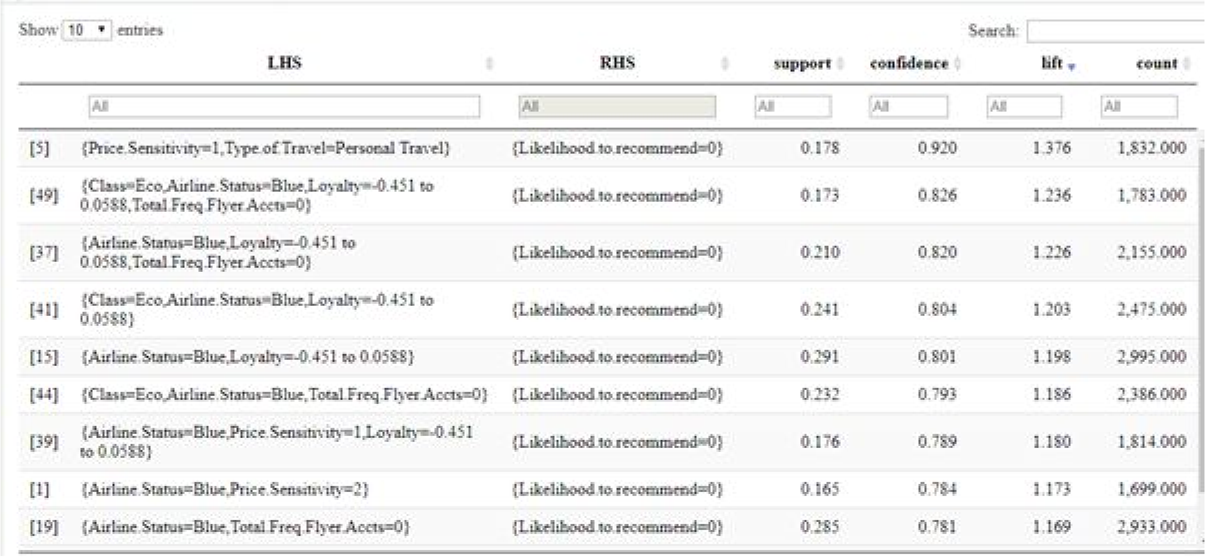
TOP 10 ASSOCIATIONS FOR Y = 0 (Detractors) :



Show 10 entries

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[30]	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.161	0.989	1.479	1,659.000
[28]	{Airline.Status=Blue,Loyalty=-0.451 to 0.0588,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.163	0.987	1.477	1,681.000
[32]	{Class=Eco,Airline.Status=Blue,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.192	0.979	1.464	1,974.000
[6]	{Airline.Status=Blue,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.236	0.977	1.462	2,422.000
[27]	{Loyalty=-0.451 to 0.0588,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.159	0.959	1.435	1,637.000
[29]	{Class=Eco,Loyalty=-0.451 to 0.0588,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.159	0.956	1.431	1,635.000
[3]	{Loyalty=-0.451 to 0.0588,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.193	0.956	1.430	1,981.000
[31]	{Class=Eco,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.161	0.950	1.422	1,651.000

Showing 1 to 10 of 49 entries



	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[5]	{Price.Sensitivity=1,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.178	0.920	1.376	1,832.000
[49]	{Class=Eco,Airline.Status=Blue,Loyalty=-0.451 to 0.0588,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.173	0.826	1.236	1,783.000
[37]	{Airline.Status=Blue,Loyalty=-0.451 to 0.0588,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.210	0.820	1.226	2,155.000
[41]	{Class=Eco,Airline.Status=Blue,Loyalty=-0.451 to 0.0588}	{Likelihood.to.recommend=0}	0.241	0.804	1.203	2,475.000
[15]	{Airline.Status=Blue,Loyalty=-0.451 to 0.0588}	{Likelihood.to.recommend=0}	0.291	0.801	1.198	2,995.000
[44]	{Class=Eco,Airline.Status=Blue,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.232	0.793	1.186	2,386.000
[39]	{Airline.Status=Blue,Price.Sensitivity=1,Loyalty=-0.451 to 0.0588}	{Likelihood.to.recommend=0}	0.176	0.789	1.180	1,814.000
[1]	{Airline.Status=Blue,Price.Sensitivity=2}	{Likelihood.to.recommend=0}	0.165	0.784	1.173	1,699.000
[19]	{Airline.Status=Blue,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.285	0.781	1.169	2,933.000

MODEL 2 : Segmentation (Clustering Algorithm)

What is Segmentation Algorithm?

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Clustering algorithms can automatically recognize the pattern inside the data so as to analyze the collected data without their labels.

Classification VS Segmentation:

In general segmentation means grouping similar things together using unsupervised learning (similar to clustering). Classification on the other hand will have predefined classes and supervised learning is used.

In the context of image processing, the distinction is very clear:

- Segmentation is the process of extracting smaller segments out of one image with the intent of identifying different parts/objects within an image. So for example segmentation of an image can give you back ground and fore ground separately. This is usually done using traditional image processing algorithms like edge detection, PCA etc.
- Classification on the other hand is identifying if an image forms part of a class. For example classifying an image as cartoon or photo. This is usually done using machine learning algorithms with tagged images.

Types of segmentation algorithms:

1 . K Means Clustering : For only continuous data variables

2. K Models CLustering and : For categorical data variable

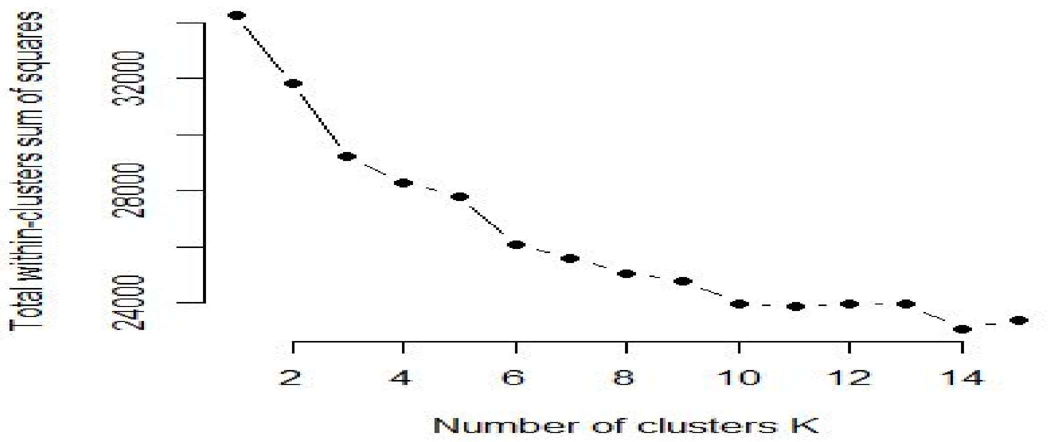
3. K Prototype Clustering : For a combination of continuous and Categorical variables

Leavering Clustering for our Analysis:

Implemented a segmentation algorithm to study the following :

- To club travellers with similar characteristics together. It helps us to study how the Likelihood to recommend varies within that cluster of travellers for different variables.
- We want to profile each of the clusters to see what kind of travellers they are.Profiling helps us notice trends of the likelihood to recommend within the cluster.

To get the optimal number of clusters to explain our data set, we use the elbow curve .



Here, there is a significant bend in the curve around K= 4. This is the optimal number with which we can best segment our data set .

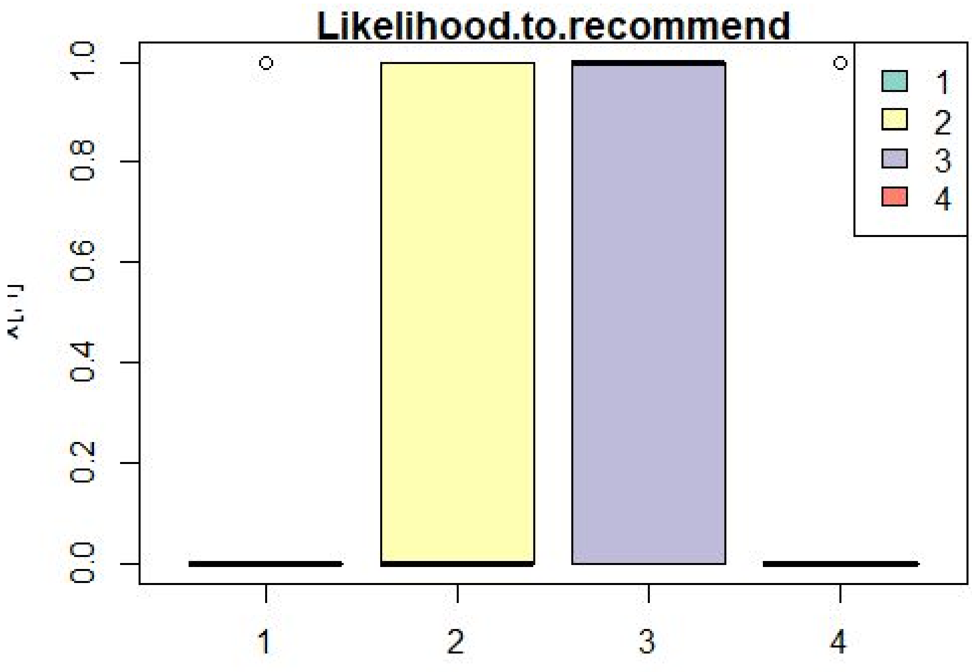
We segmented each of the data points into clusters , and profiled each of the clusters.

What is Profiling?

Having decided (for now) how many clusters to use, we would like to get a better understanding of who the customers in those clusters are and interpret the segments.After validating the convergence of cluster analysis, we need to identify behavior of each cluster.

ClusterPlots and Profiles Across the clusters and Variables

1. **The Likelihood to recommend**

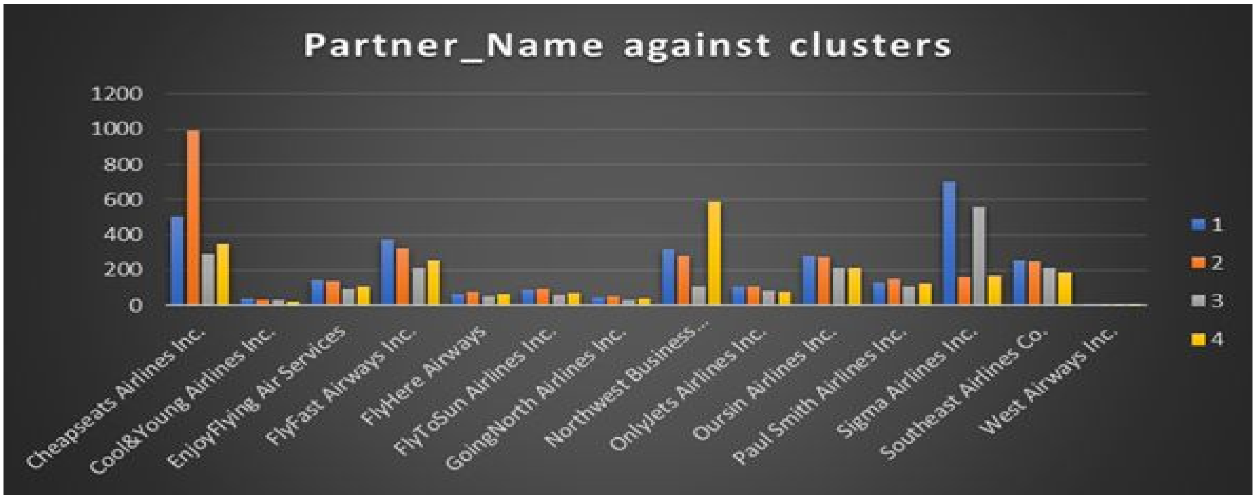


The 4 colours represent the 4 clusters and the bars represent the value of Likelihood to Recommend in each of the 4 clusters.

This plot tells us that Cluster 1 and 4 are Detractor clusters ,ie: Likelihood to recommend = 0
and cluster 2 and 3 are promoter clusters ,ie : Likelihood to recommend = 1.

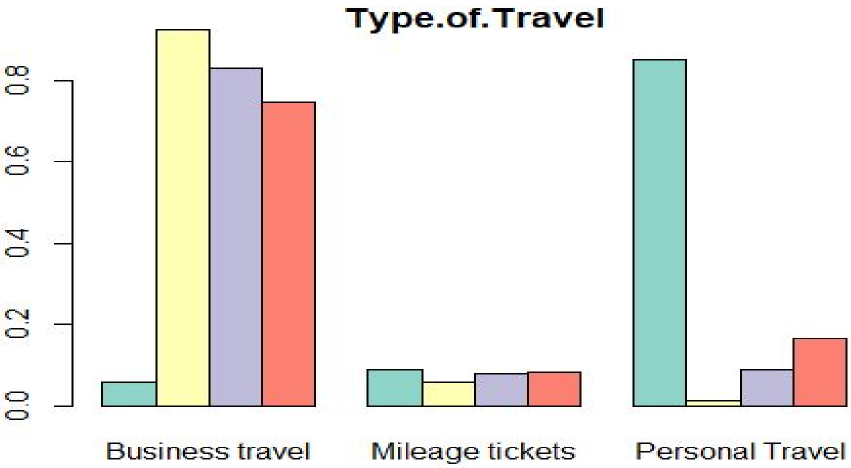
We look at the other X variable's distribution across clusters. If they are significantly populated in clusters 1 and 4 , they are detractors, else promoters

2. **Partner Name against Clusters**



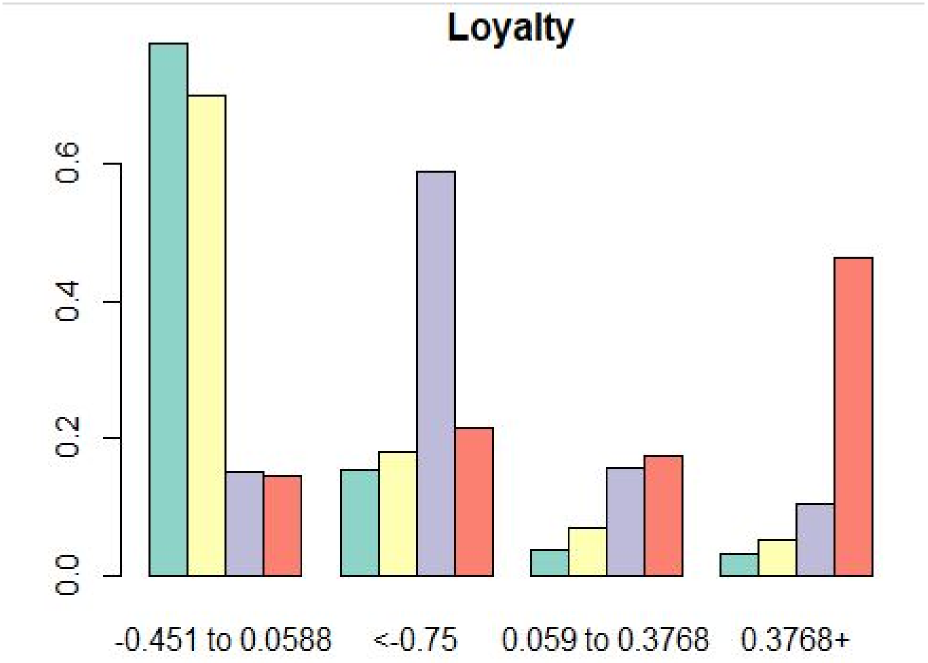
We have conclusive evidence here that Cheapseats Airline Travellers are Promoters. Northwest Business Airline travellers are Detractors.

3. **Type of Travel**



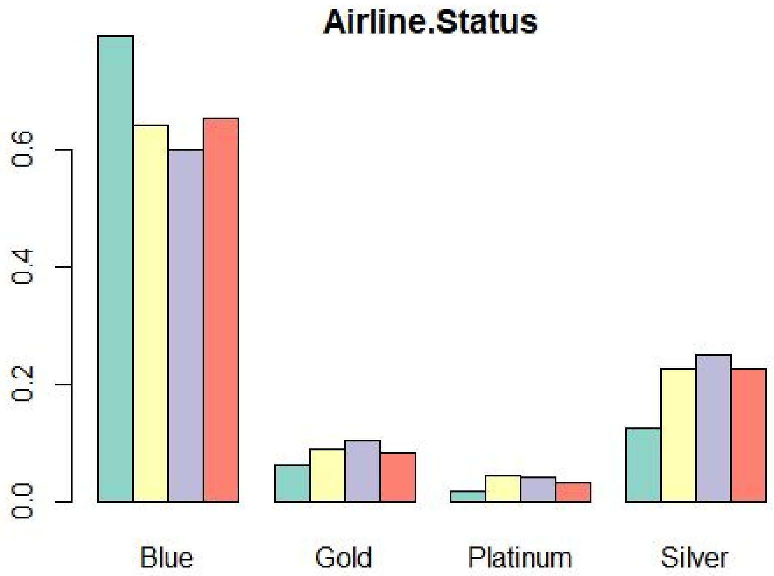
Personal Travellers are detractors and Business Travellers are predominantly promoters

4 . **Loyalty**



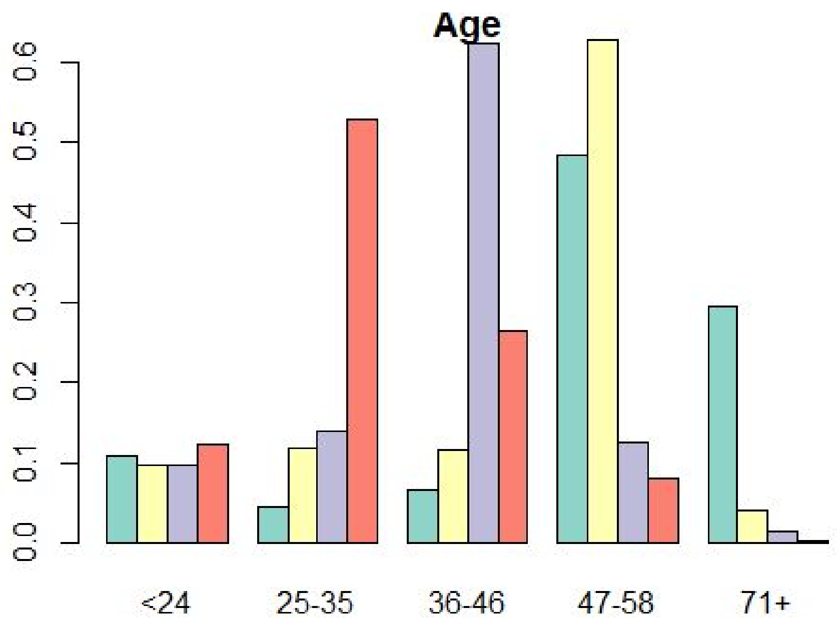
Travellers with high Loyalty are surprisingly detractors and less loyal customers are promoters

5. **Airline Status**



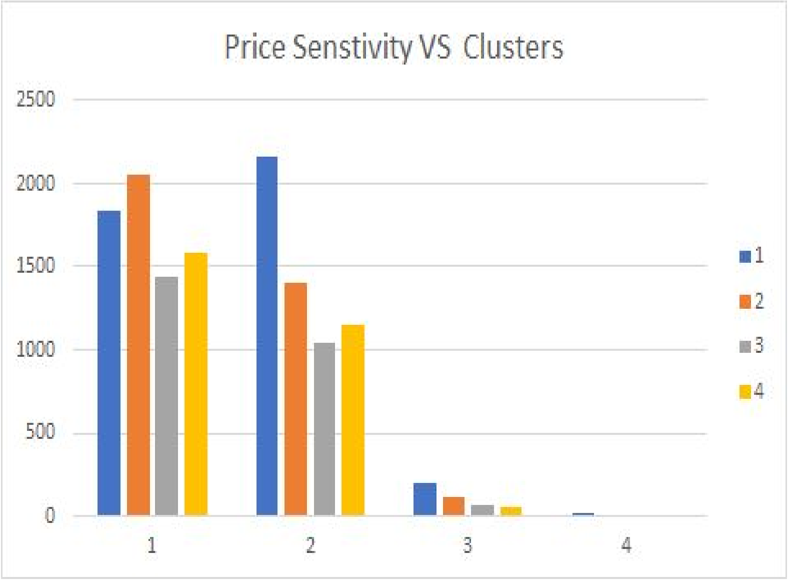
Platinum ,Silver and Gold Airline Status and are promoters. Blue Travellers are detractors

6. **AGE**

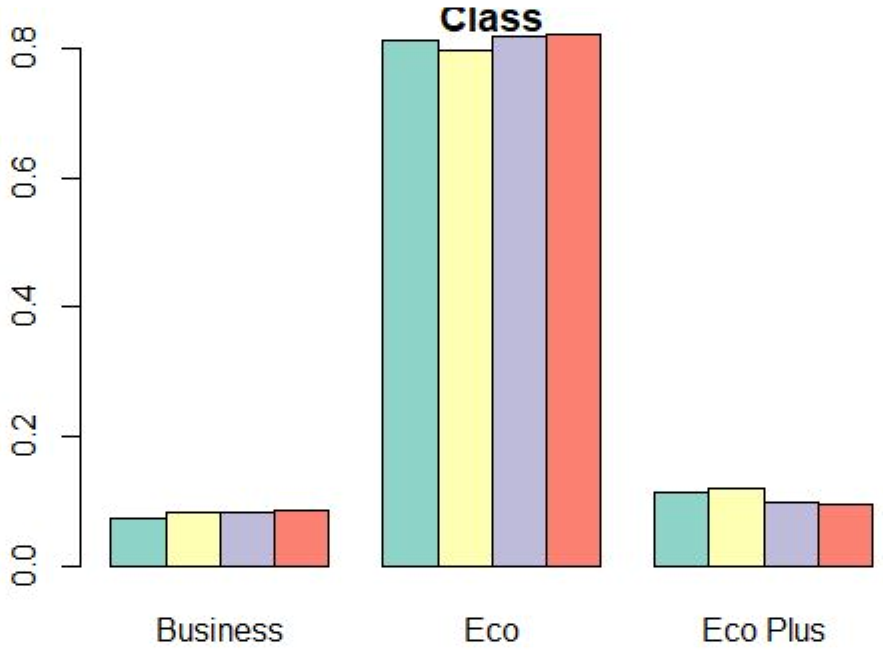


AgeAge group 25- 35 and 71 + detractors.
36-58 age group are Promoters

7. **Price Sensitivity : Promoters Have high Price Sentivity(1)**



8) **Class : Eco Plus Users are slight promoters**



Conclusions and Business Implications

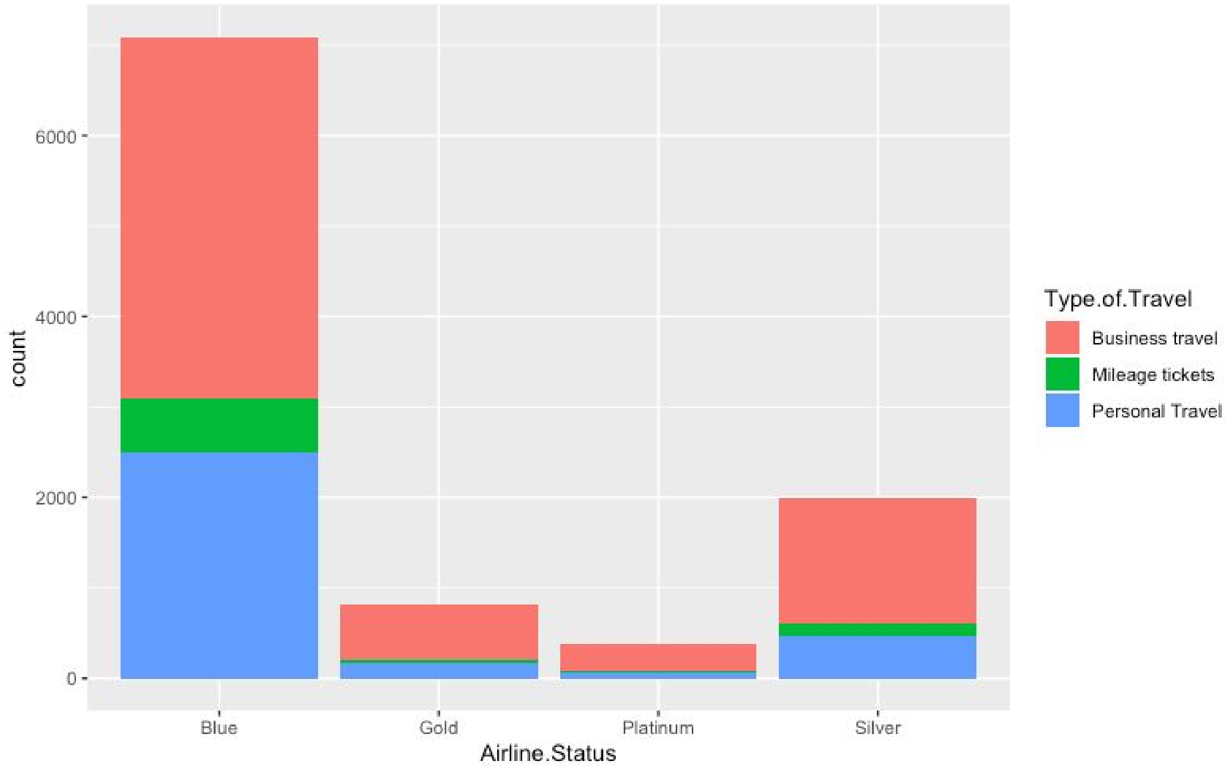
The airline industry as a whole has a wide variety of customers, and when developing a strategy to increase profits and market share, it's important to focus on the consumer. As this is where the money is generated, customer retention rates and generating repeat customers from first-time experiences with an airline is important. Considering contextual information about the industry is important in what potential decisions may stem from the data analyses. Specifically, Southeast airlines oriented themselves in the market similar to others in the industry, following industry standard, though some of the analyses conducted show that these common business practices cannot be effective at best, and have a negative impact business performance at worst.

As airlines and their partners find themselves with ever-increasing debt in the form of frequent flier mileage points, the approach previously taken of rewarding return customers seem to be more financially burdensome all while the valuation of these points are depreciating. In fact, within the data analysis of customer loyalty, the higher the loyalty rating the airline gave the customer, the lower the likelihood to recommend. Continuing to offer mileage points not only appears to be a non-starter, but the risk associated with continuing to use a loyalty rewards program may actually increase churn rate on two fronts.

1. Decreasing value to the customer of mileage points have increased debt while increasing customer attrition.
2. There is a clear pattern that shows the airlines assigned loyalty value to a customer decreases their likelihood to recommend.

When considering other avenues to use, likelihood to recommend, the rating that the customer gives from 1-10 that they'd tell a friend or colleague to use the same service, seems like a better indicator of future business success. Rather than using a loyalty rating, the likelihood to recommend score that a customer gives can be used to evaluate NPS. NPS has been used as a great predictor of customer churn rates. Bucketing customers into three categories: promoters, passives and detractors, has turned into a common analytical tool to assess and predict business performance. Being 1.5 times more likely to stop using a service, detractors are a great signifier of what can be done better in the business or what can be inferred from looking at who detractors are.

When considering new customer acquisition, Southeast places customers into status groups. For first-time travellers, they're automatically placed into Blue status. Regardless of the airline partner, Southeast airline passengers that are blue status are predominantly detractors, being (x times more likely to be a detractor). This doesn't bode well for the airline performance in the future, as it signifies a low repeat customer rate is more likely in the future, and the churn rate will increase. It's imperative to offer better service to first-time or infrequent fliers to decrease the attrition rate going forward. In addition to offering better services, discounts for first-time fliers, or a discount for second-time fliers may help decrease the attrition rate. Because NPS is also a good indicator of brand image and organic marketing, there's the potential to acquire more customers by making a good impression on blue status fliers. Blue status fliers also make up the majority of customers, so to increase market share, blue status fliers must be a focus going forward.



While business travellers are predominantly promoters, personal travellers are detractors. When assuming that low cost is the business model being pursued by Southeast airlines, it doesn’t seem to have a net positive impact for personal travel passengers. Price sensitivity seems to be a great indicator of a passengers likelihood to recommend, though how this score was determined is open to questioning. While business travellers seem to be promoters, the partner, Northwest Business Airline seems to have accrued primarily detractors. Whether they’re not following the business model correctly is unclear, but their poor performance has increased the likelihood of a tarnished brand image, leaving a tough decision to be made with Northwest Business Airline. Terms of the partnership should be brought up at the least, and the NPS as an indicator of overall churn should be a factor in continuing whether the partner’s performance is up to standard or not.

Much like the incentives for first and second-time fliers that could be offered to blue class fliers, the situation is similar for age groups. There seems to be no correlation with middle aged passengers and likelihood to recommend, however, the ages 24 and less and 71 and greater seem to be detractors. A discount campaign focusing on customers within these age groups may help positively impact the attrition rate for these demographics. Moving passengers up to Eco Plus may also be an effective strategy, as Eco Plus passengers tend to be promoters, while the other two flight classes are not. Re-assigning these seats to these specific age groups would both help management and demand response in addition to increasing NPS.