THE iSCHOOL
Syracuse University

SOUTHEAST AIRLINES
The Spirit of Liberty™

Data Analysis for Southeast Airlines

IST 687 M002 Group 3

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

Proposed Plan

In order to Increase the Net Promoter Score of an Airline , We Propose to :

a. Analyse the Airline Promoter Data to figure out patterns and patterns and trends
b. Arrive at the variables which have optimum impact on the Likelihood to Recommend of a traveller.
c. Figure out associations to study Niche groups of travellers and their Likelihood to recommend.
d. Create meaningful segments of travellers and to study how their Likelihood to recommend varies within that segment with respect to the other variables.
e. Put forth a solid recommendation on how to Increase the number of Promoters which in turn will affect the NPS Positively

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 Assimilation and Understanding the Problem Statement

The Data Structure

1. The data set given is at a ‘Flight Journey’ level
2. The Explanatory variables can be classified into:
    a. Traveller Demographic Variables : Age , Gender etc
    b. Traveller Persona Variables : Price Sensitivity, Loyalty , Amount spent at airports during travel, Frequent Flyer Tags etc.
    c. Air Carrier Variables : Airline partner, Airline Status etc.
    d. Travel Variables : Origin City, Flight Cancelled, Duration , Delay etc.
    e. Other Variables : Latitude, Longitude, Date time etc.

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

Total X Vars ; 31
Y Var : 1

Assumption : Price Sensitivity ranges from (1-5) : Hight to low

The Y Variable

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

Business Decisions

- Decreasing value of mileage points means we have to look elsewhere for to provide value to the customers.
- Looking to NPS for important decisions concerning business relationships, potential investments with staff and airport assets.
- Focus on the negatives: Detractors are **1.5** times more likely to **stop** using a service.

Data Cleaning and Manipulation

Data Prep : Binning

Categorical Variables :
1. All the categorical variables have be converted into factors

Continuous Variables :
1. For Continuous variables with large range,for ex: (AGE), we have followed the below mentioned steps :
   a. Generated the quantiles to look at the distribution
   b. Bucket the variables into 4 to a maximum of 10 levels ensuring there is a pattern in the population across the buckets (Fatter in the middle)

Example : The Age Variable is bucketed as <24 ,25-35, 36-46, 47-58, 57-71, 71+

2 . For Continuous variables with small range we have followed the below mentioned steps :
   a. Generated the quantiles to look at the distribution
   b. Create a maximum of 3 levels ensuring logic and the population density is maintained across the buckets

Example : The Arrival Delay is bucketed as 0 , 0-60 Mins ,60+Minutes

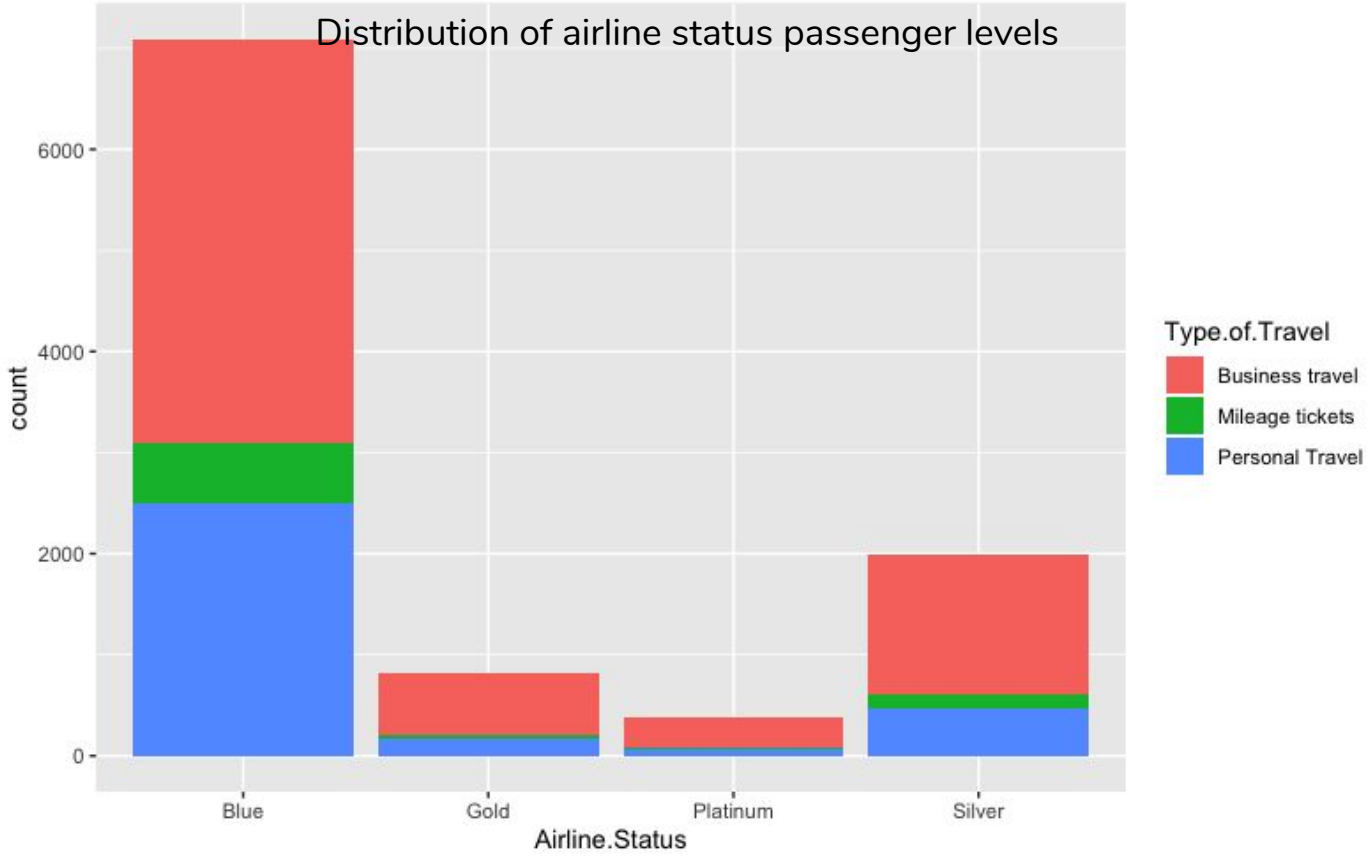
Missing Value Imputation

After the previous step, All our continuous variables have been transformed to character variables.

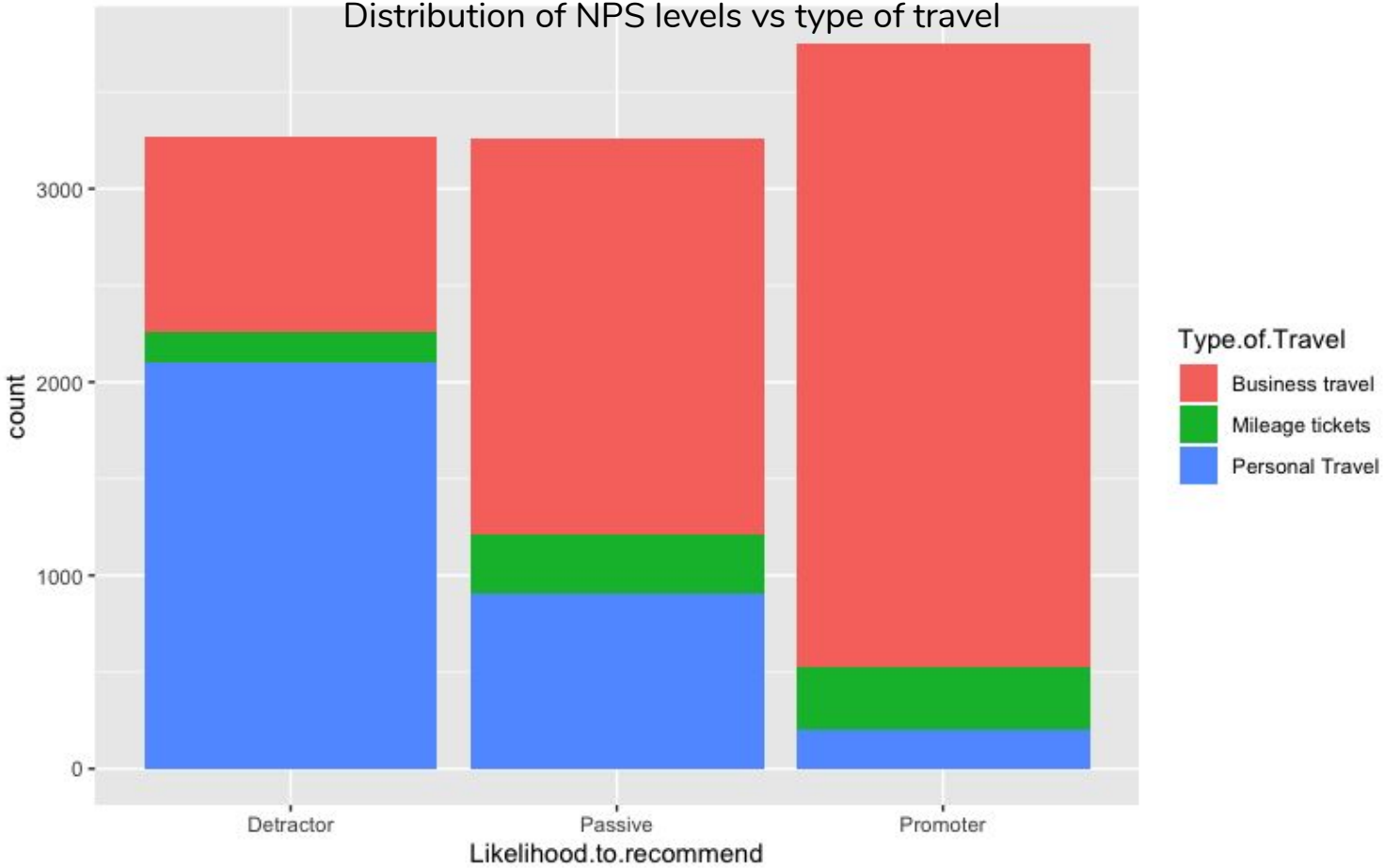
1. Converted All the variables as factors.
2. Imputed NA with the Mode for all the Variables except,
3. The Arrival and Delay Minutes ,if NA, are set to the level “0” if the corresponding flight is cancelled

Exploratory Data Analysis

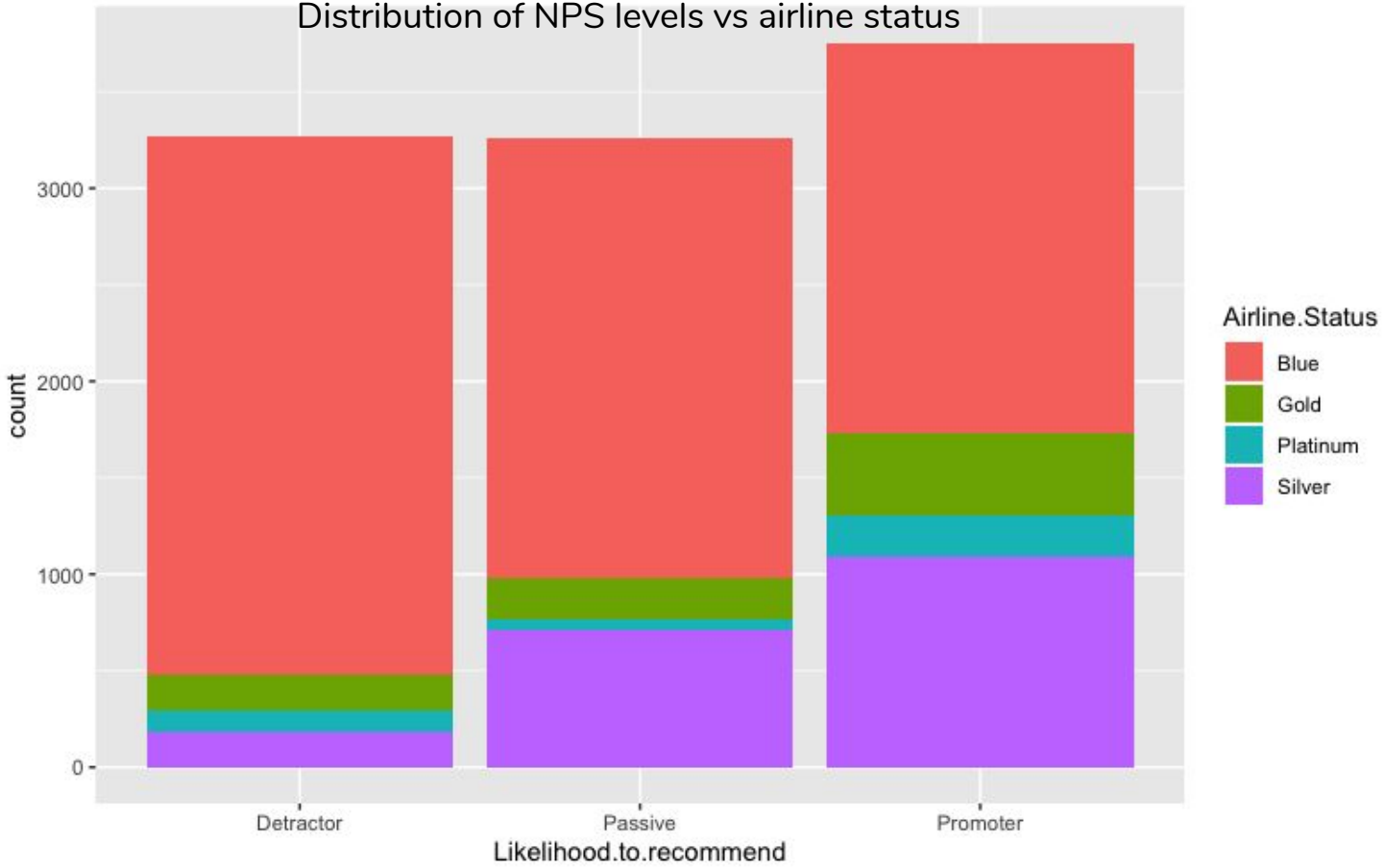
Distribution of airline status passenger levels



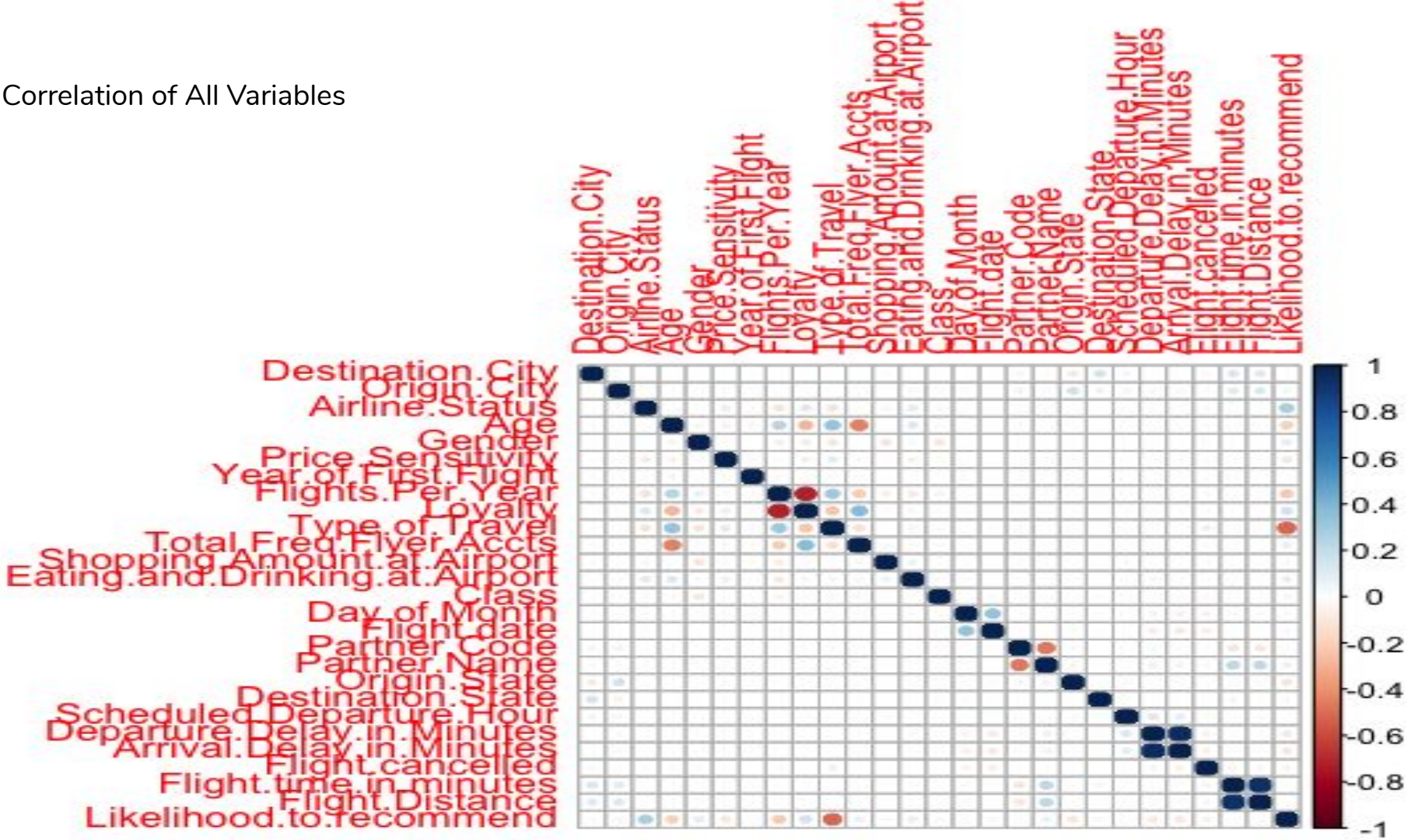
Distribution of NPS levels vs type of travel



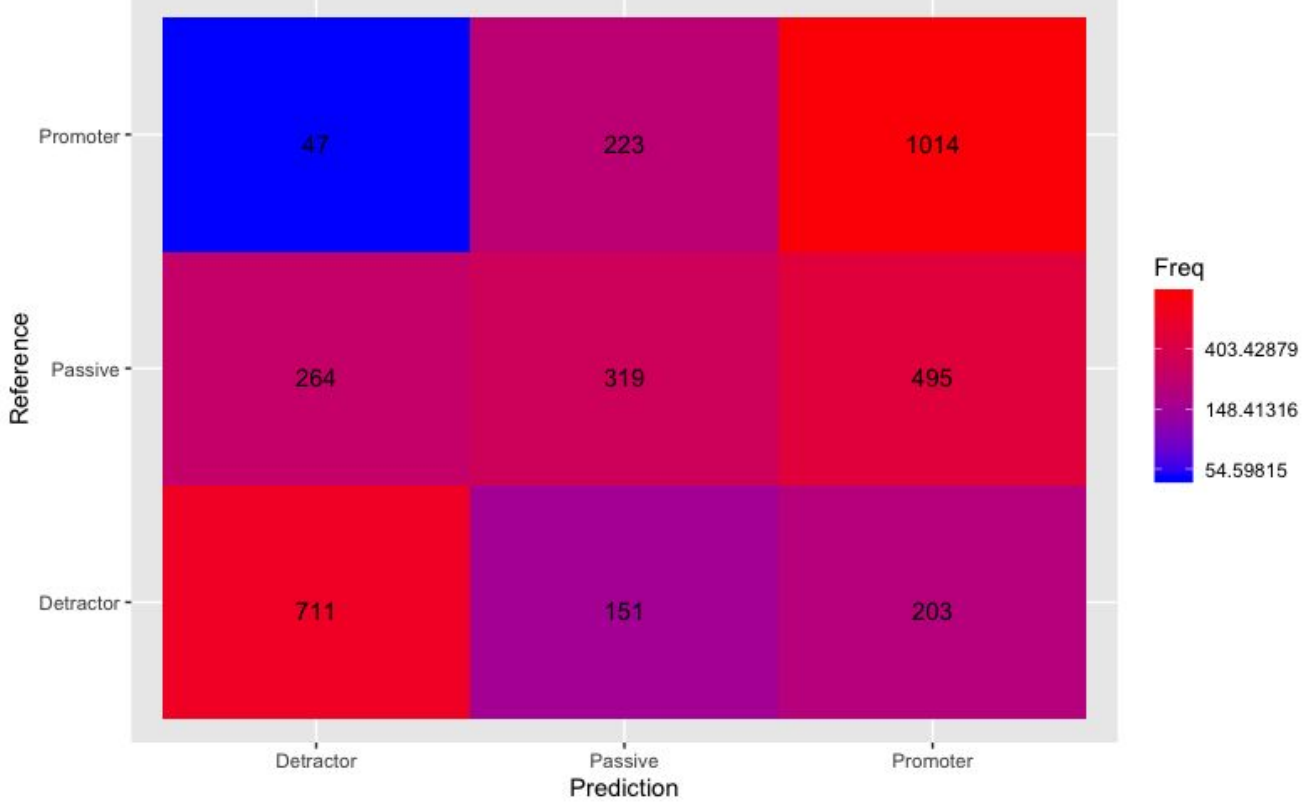
Distribution of NPS levels vs airline status



Correlation of All Variables



Model: SVM



Accuracy: 59.64%

95% Confidence: 58%-61%

No-info Rate: 37.5%

	Detractor	Passive	Promoter
Sensitivity	0.6676	0.29592	0.7897
Specificity	0.8683	0.84078	0.6743

Feature Importance Extraction

Variable Selection Process

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

Variable Selection Process

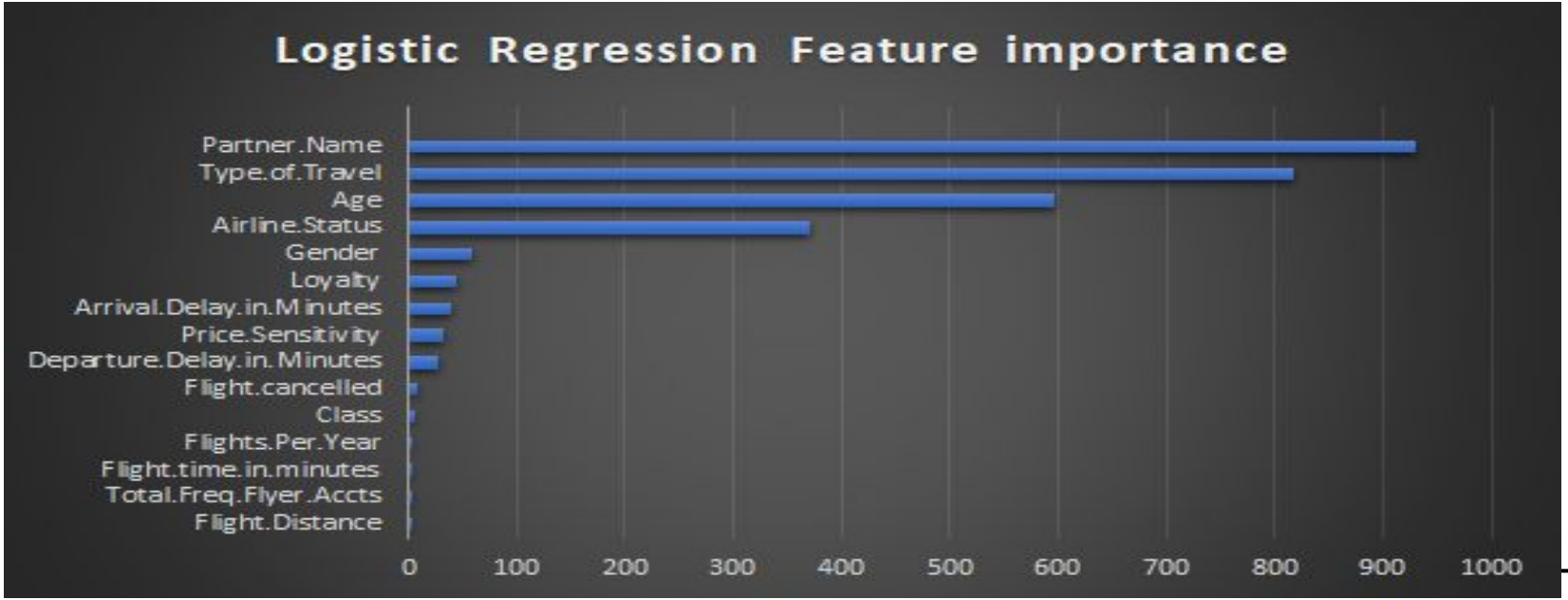
2. Feature Importance Generation to find the most significant Variables

- 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

Models : Logistic Regressor

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

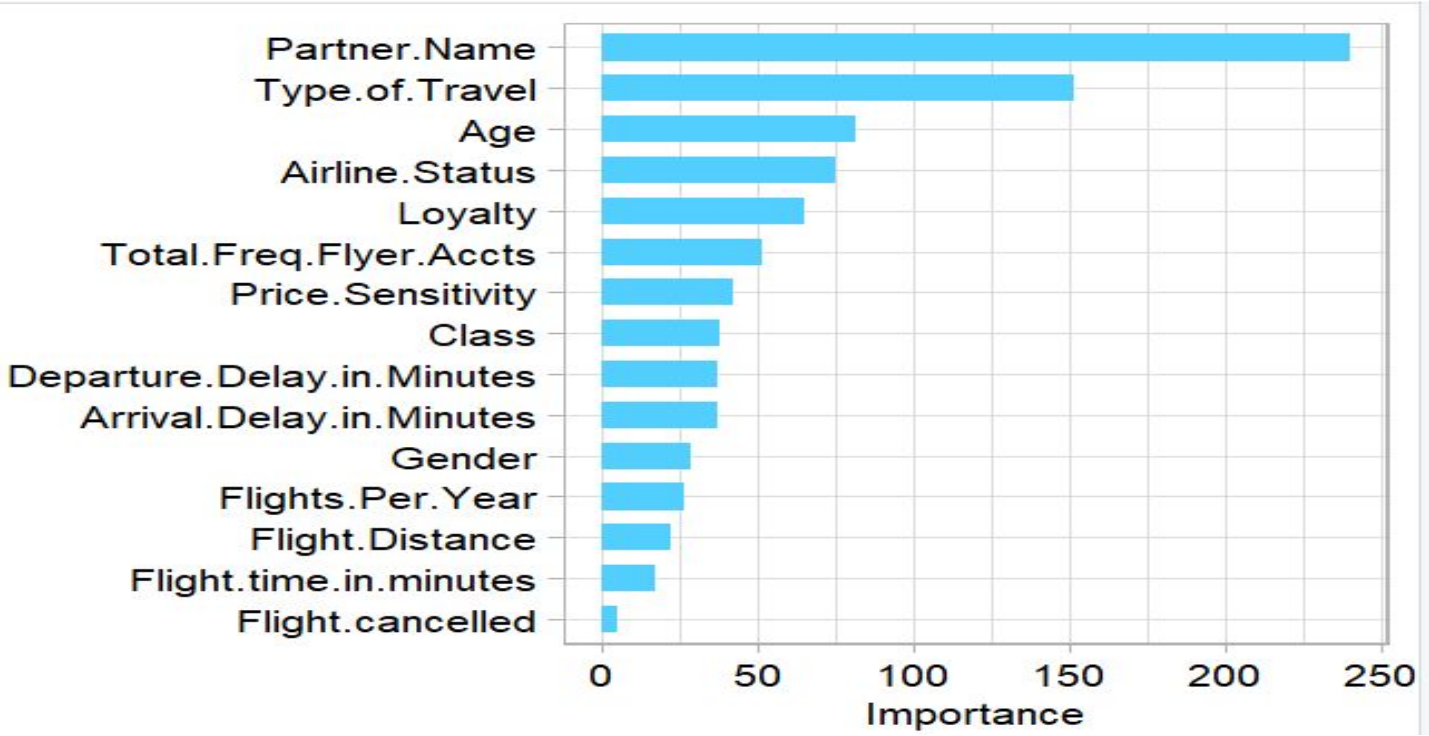
AUC: **0.8376627**



Models: randomForest Classifer

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

AUC: **0.8376627**



Model: Gradient Boost Classifier

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

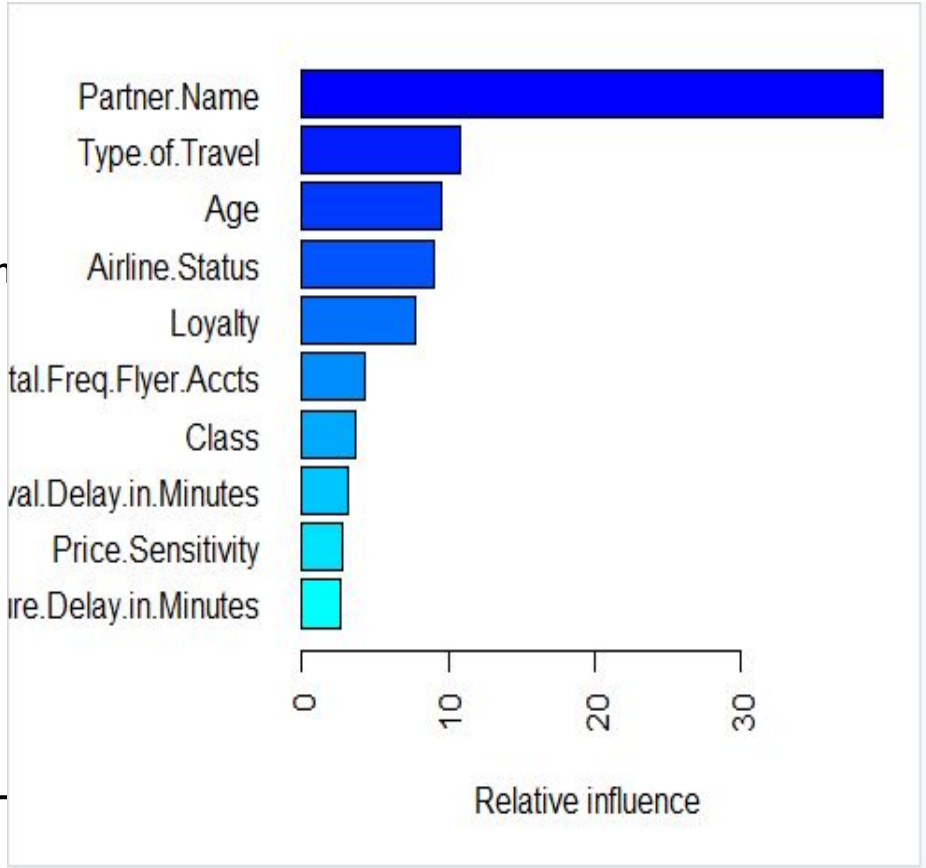
AUC: **0.8376627**

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.



Inference From the Feature Extraction Exercise:

We see that the following variables have high feature importance across all the 3 classifiers. We pick the top 8 common Variables to base our analysis on for the further steps.

a. PartnerName
a. Age
b. Airline Status
c. Type of travel
d. Loyalty
e. Total freq Flyer
f. Price Sensitivity and
g. Class

Segment Analysis and Association Studies

USING KPROTOTYPE CLUSTERING
AND
APRIORI ALGORITHMS

Analysis

1. APRIORI ALGORITHM

Generate Association Rules using APRIORI Algorithm for the reduced data set and the Y variable.

- We formulate values for support,confidence and Lift for each of the combinations of levels of the X variables to the Y variable.
- We get 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.
- We ran the Algorithm both for Likelihood to recommend = Yes and No.

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

Next 10 Associations for Y =1 (Promoters)

	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

Search:

	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

Previous 1 2 3 4 5 Next

Next 10 Associations for Y =0 (Detractors)

Show 10 entries

Search:

	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

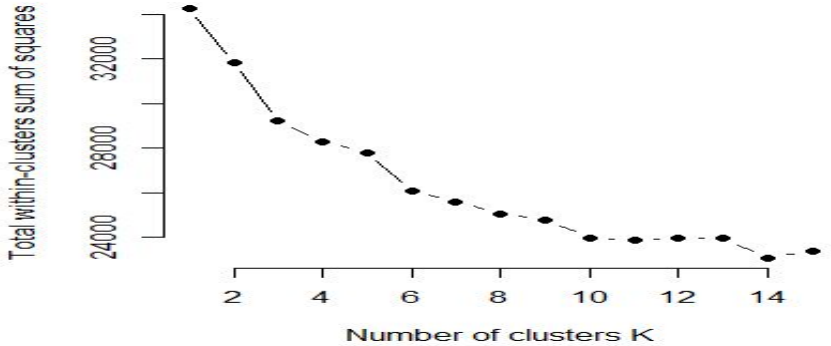
Analysis

2. Clustering Using KPrototype

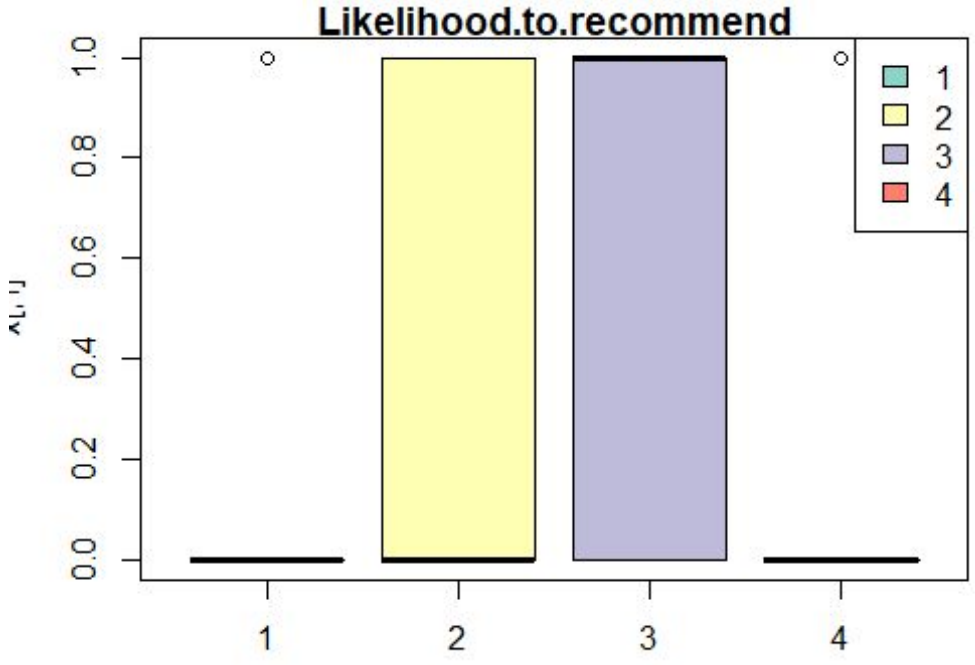
Implemented a segmentation algorithm to study the following :

- **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.**

Optimum Number of clusters : 4



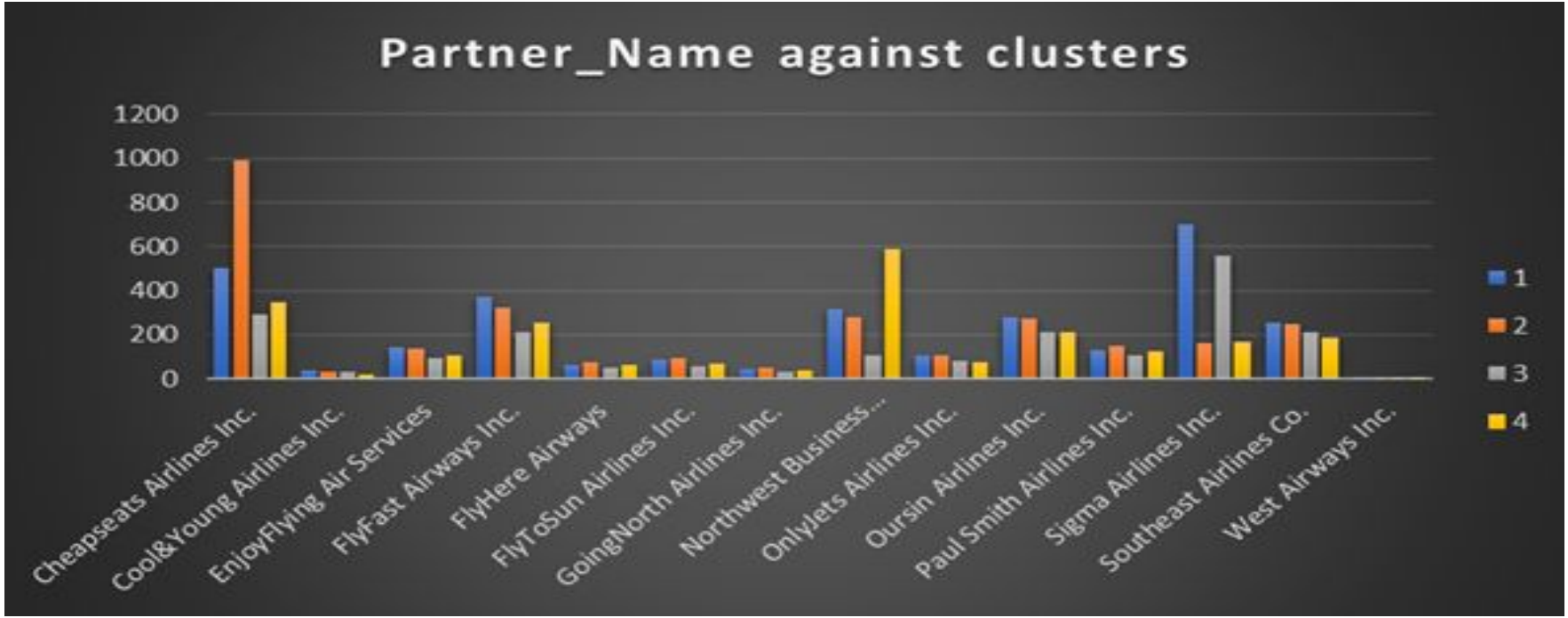
Cluster Profile Plots



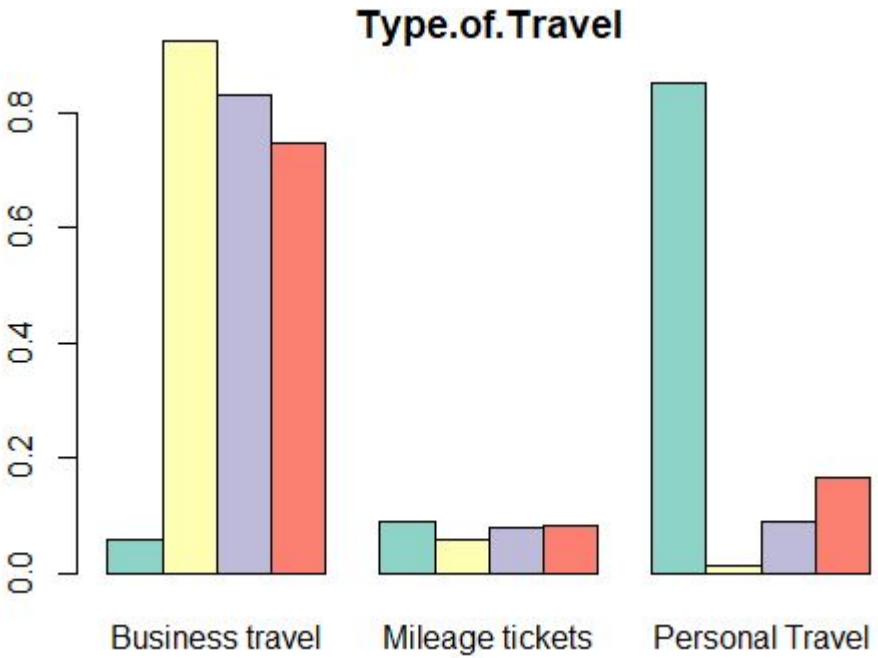
The 4 colours represent the 4 clusters and the bars represent the value of Likelihood to Recommend in each of the 4 clusters.

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.

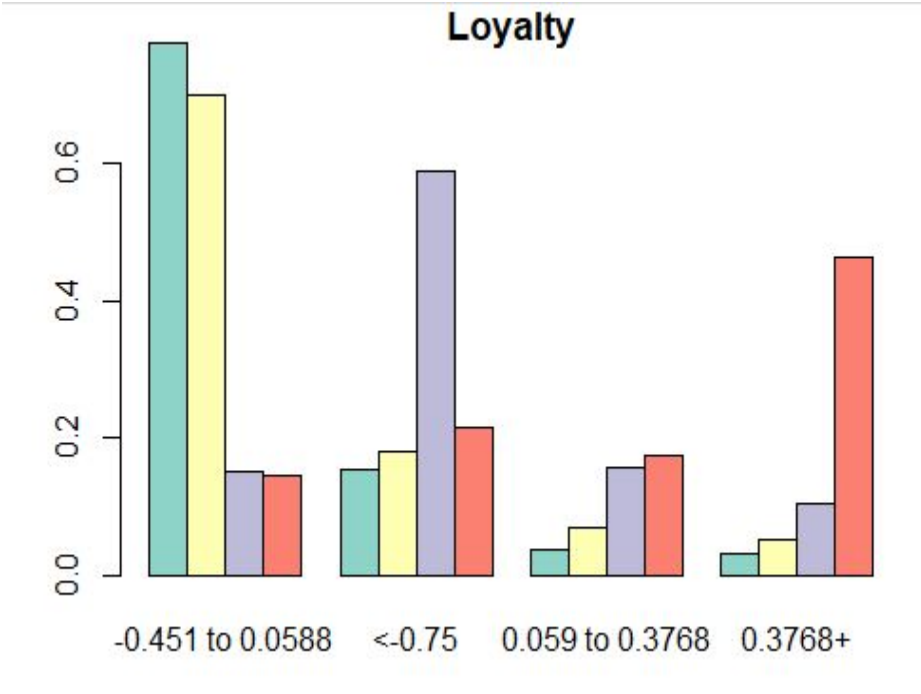
We look at the other X variable's distribution across clusters. If they are significaantly populous in clusters 1 an 4 , they are detractors, else promoters



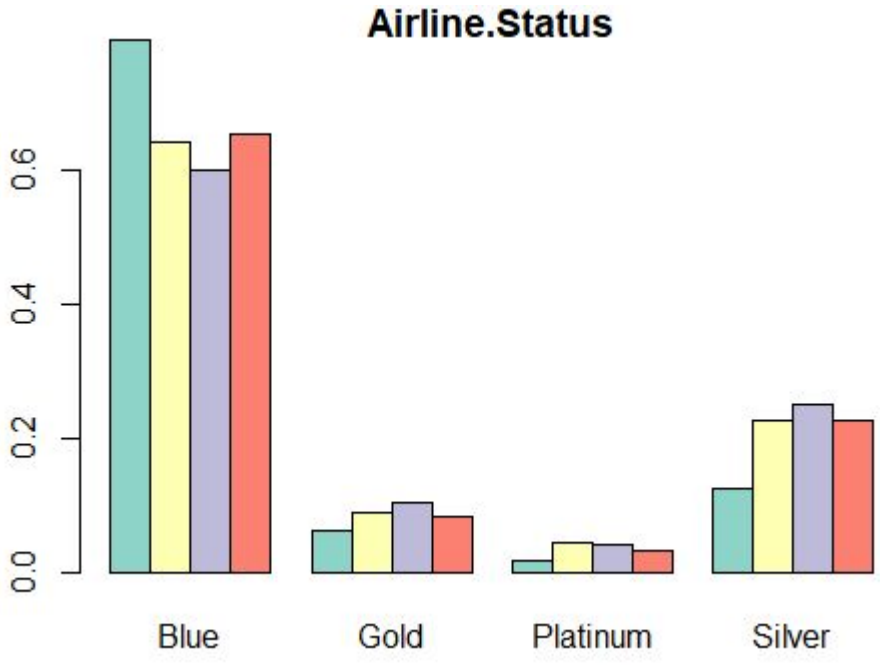
We have conclusive evideence that Cheapseats Airline Travellers are Promoters.
Northwest Business Airline travellers are Detractors.



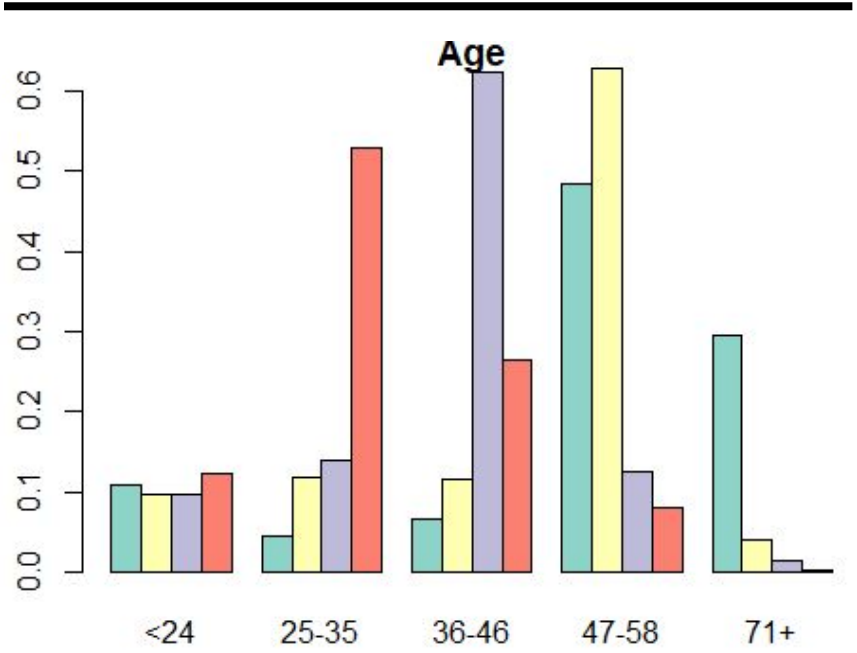
Personal Travellers are detractors and Business Travellers are predominantly promoters



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

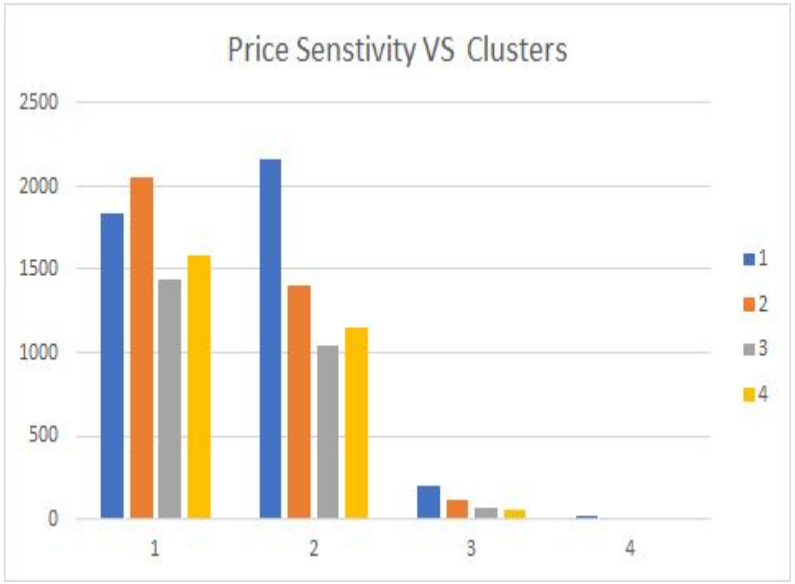


Platinum ,Silver and Gold Airline Status and are promoters
Blue Travellers are detractors

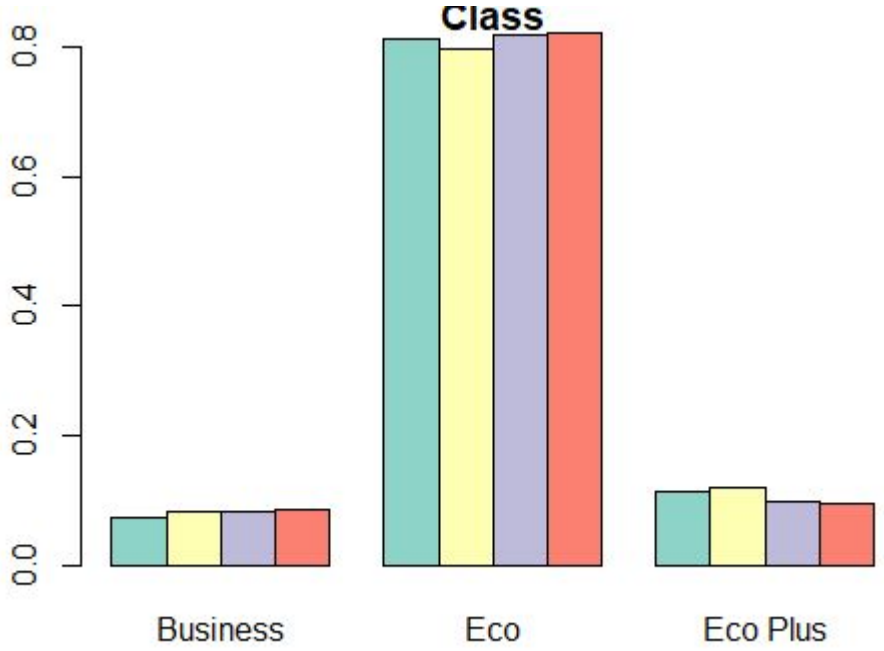


Age group 25- 35 and 71 + detractors.
36-58 age group are Promoters

Promoters Have high Price Sentivity(1)



Promoters Have high Price Sentivity(1)



Eco Plus Users are slight promoters

Statistical Inference from our Models

Conclusively:

- Business travellers and price sensitive customers are more likely to recommend Southeast Airlines.
- Northwest Business Airline Travellers has been failing in comparison to other partners.
- Age based price discrimination may help decrease the number of detractors for 18-24 and 71+ years age groups.
- A ‘first impression’ price discount may help decrease detractors who are Blue class flyers.
- Eco Plus seems to be an effective offering to increase the likelihood to recommend.