Airline company customer satisfaction

Applied Predictive Analytics

Group 12

सिद्धिमूलं प्रबन्धनम्

भा. प्र. सं. इन्दौर

IIM INDORE

2016IPM034	Davareesh
2016IPM063	Namrata Velivelli
2016IPM060	Mohammed Rizwan
2016IPM118	Wellington Daniel

Airline company customer satisfaction

Executive Summary:

Key Results

1. The factors affecting customers deciding that they is satisfied are identified using **Factor Analysis**. The service expectations are weighted higher in the customer's mind while making decision regarding satisfaction. Service expectation is the most important.
2. There **are four customer segments**. While the young entrepreneurs and Executives are relatively satisfied the regulars are the unsatisfied customers. Blue Delta Airlines should tailor Marketing efforts towards them in order to make sure they are satisfied.
3. The best model to predict customer satisfaction is found to be Random Forest ensemble model with **94.5%** accuracy.
4. Increasing **Convenience** for the customers can increase satisfaction by up to 60%. Refer to slide 8 for more information.

Recommendation:

- Focus on satisfying the "Regulars" segments. They make up half of the customers and Blue Delta is a budget airline company.
- Improve convenience factors such as online boarding and Wi-Fi to increase customer satisfaction above 70%.
- Design advertisements communicating high service levels.

Techniques used

Dimension Reduction: Factor Analysis, Principal Components Analysis

Clustering: K-means clustering, Hierarchical Clustering

Machine Learning: Logistic Regression, Decision Trees, Random Forest, Gradient Boosting

Business Problem Objective

- To find the Parameters which influences Customer satisfaction.
- To develop insights and intelligence on customer behavior and expectations.
- To develop a model to predict satisfaction of customers.

Data

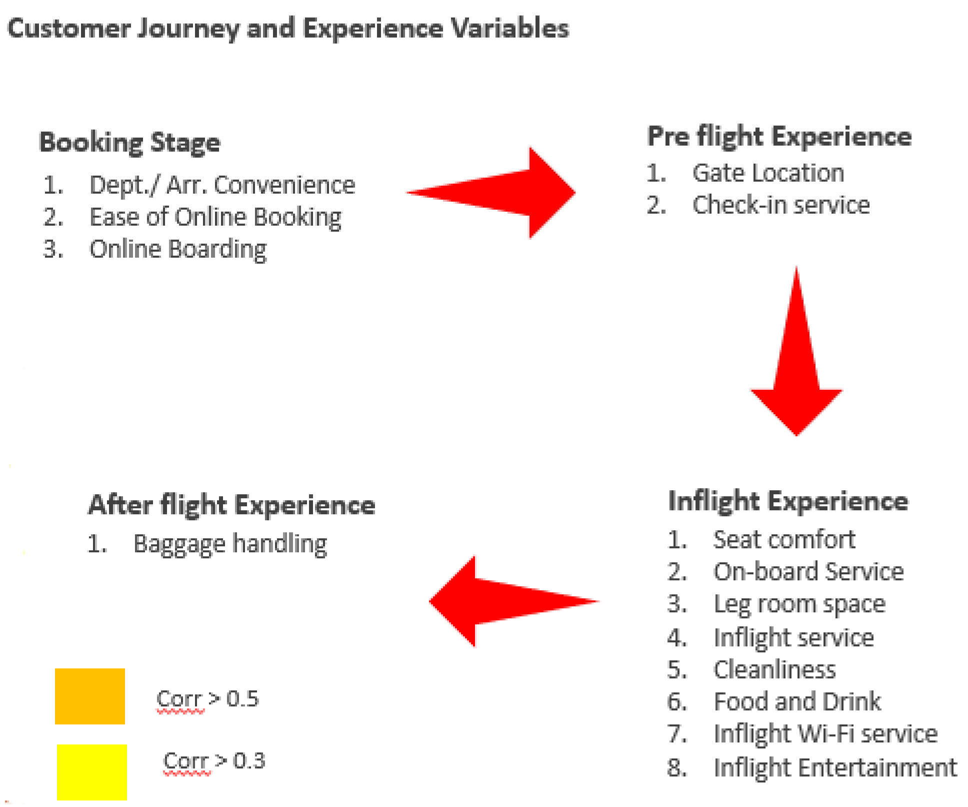
https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

Methodology:

Understanding Causality using PCA and Factor Analysis

- Factor Analysis is a **Dimension Reduction technique** which is widely used in various fields to study causality for variations in responses. It is very effective especially when used for data recorded in a Likert or a sematic difference scale.

- We have used **Principal Component Analysis (PCA)** to decide the number of factors. It can be seen that the variance decreases below 1 with 7 components but for simplicity we have taken 6 factors which explain 50% of the variance.



Customer Journey and Experience Variables

Booking Stage
1. Dept./ Arr. Convenience
2. Ease of Online Booking
3. Online Boarding

Pre flight Experience
1. Gate Location
2. Check-in service

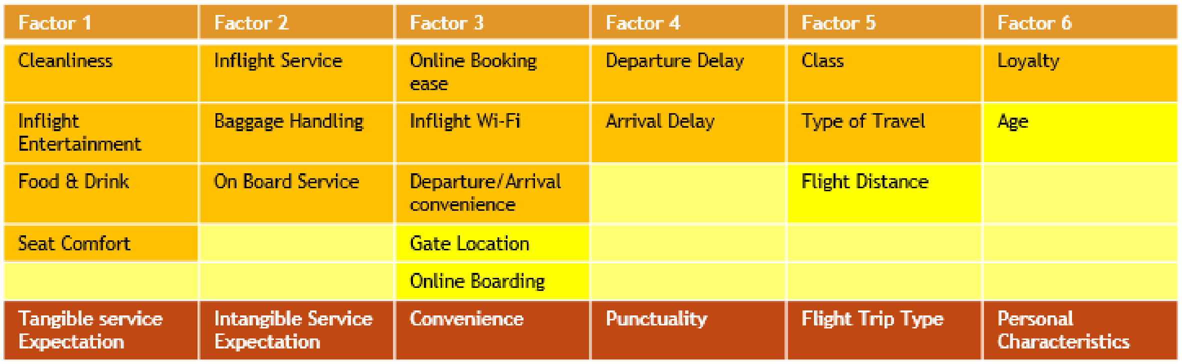
Inflight Experience
1. Seat comfort
2. On-board Service
3. Leg room space
4. Inflight service
5. Cleanliness
6. Food and Drink
7. Inflight Wi-Fi service
8. Inflight Entertainment

After flight Experience
1. Baggage handling

🟧 Corr > 0.5

🟨 Corr > 0.3

Factor Analysis Result:



Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Cleanliness	Inflight Service	Online Booking ease	Departure Delay	Class	Loyalty
Inflight Entertainment	Baggage Handling	Inflight Wi-Fi	Arrival Delay	Type of Travel	Age
Food & Drink	On Board Service	Departure/Arrival convenience		Flight Distance	
Seat Comfort		Gate Location			
		Online Boarding			
Tangible service Expectation	Intangible Service Expectation	Convenience	Punctuality	Flight Trip Type	Personal Characteristics

Dealing with Missing values

- A linear regression model is fitted for Arrival delay using Departure time: (R sq :~ 93%)
- The result is taken as the arrival delay for the entry for completeness. *(Note: Multicollinearity)*
- The other entries are simply removed as they represent less than 0.1% of the sample.

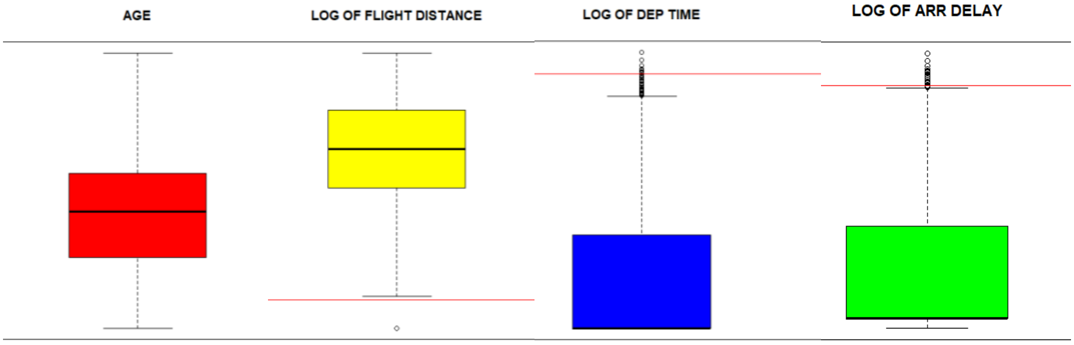
Variable	Missing Values
Inflight wi-fi service	4
Departure/Arrival time convenient	4
Ease of Online booking	3
Gate location	8
Food and drink	5
Online boarding	6
Seat comfort	1
Inflight entertainment	4
On-board service	3
Leg room service	2
Baggage handling	2
Check-in service	3
Inflight service	1
Cleanliness	2
Arrival Delay in Minutes	393 (0.3%)

Normality Assumptions and Transformations

- All the ratings are done in a **semantic differential scale**. We assume that the data is continuous and is normally distributed. This is a safe assumption to make because the data set is large and subjectivity is minimized.
- Arrival/Departure delay are **Exponentially distributed** and their natural logarithms are Normally distributed. It is necessary to transform them before analysis.

Dealing with Outliers

- Age is assumed to be normally distributed and the 4th standard deviation is taken for outliers.
- Flight distance is taken to be log normal for simplicity even though a Gamma distribution would have been a better fit. This is done for simplicity and it doesn't make much difference.
- Departure/Arrival delay are continuous and exponentially distributed. The log value above 3 standard deviations over mean is considered as an outlier.



Variable	No of outliers
Age	0
Flight Distance	11
Departure delay time	20
Arrival delay time	20 (same flights as departure)

Data Transformation and Dummy Variables

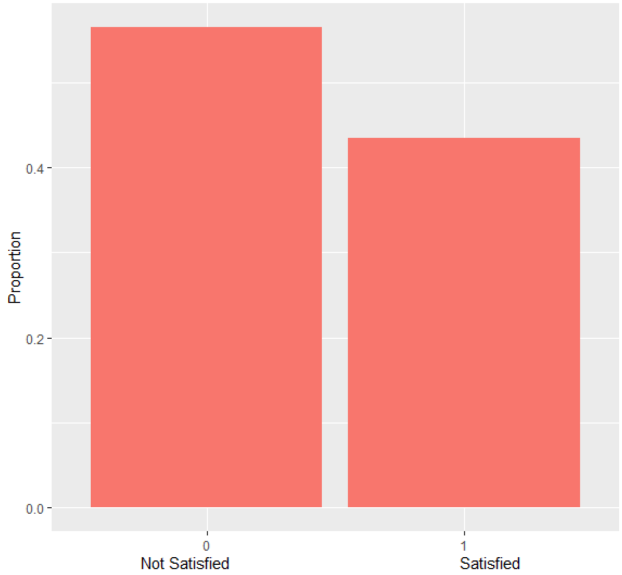
- Gender, Customer Loyalty, **Type of Travel, and Class have to be coded into dummy variables for our analysis. % dummy variables should be created for this**

Exploratory Data Analysis

The Problem we face

- The graph below shows that the company now has more dissatisfied customers than satisfied. In this Exploratory Analysis we will mainly try to understand the extrinsic attributes like age/income of the customers and the details of the flight trip that affect the satisfaction level.
- If these characteristics are properly understood, the company can design Marketing strategies/discounts targeted towards customers who are similar to each other and belong to a certain category.
- This in combination with the results of the Factor Analysis we have previously done will help us understand our customers better.

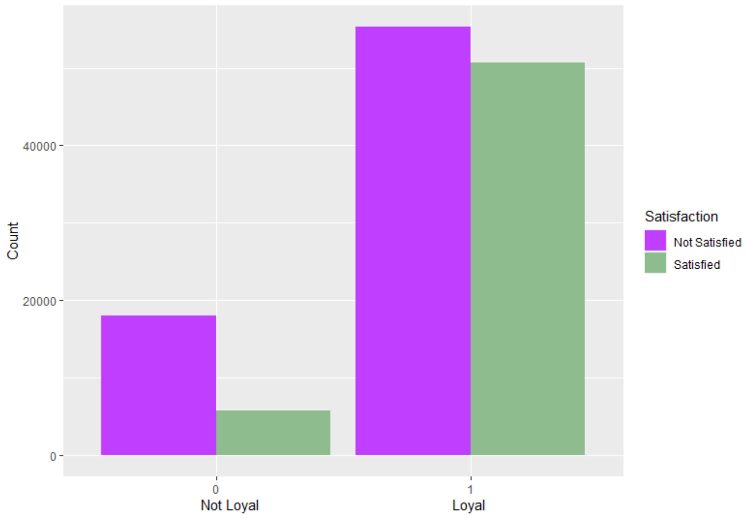
Proportion of satisfied customers



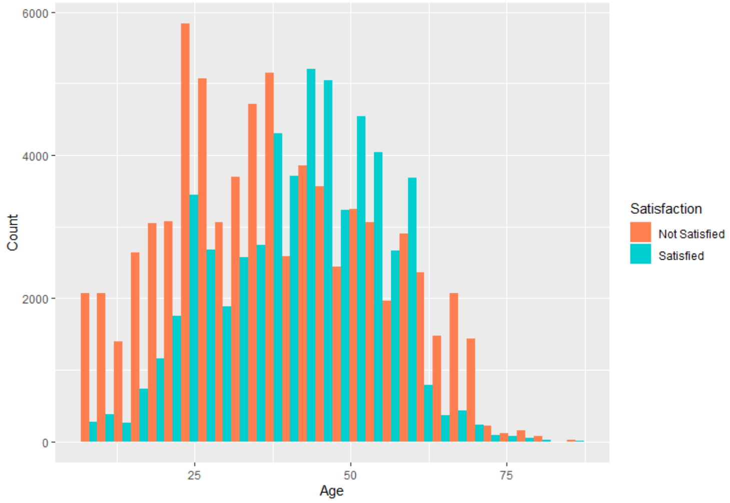
Demographics

- Gender has very low effect in Loyalty of customer and satisfaction.
- From the graphs we can see that older customers are more loyal and satisfied.
- We can also see that the proportion of not satisfied customers who are loyal is comparatively less when compared to the Not Loyal category.
- **Inference**: Older customers are more loyal and more satisfied and loyalty has a slight effect on satisfaction.

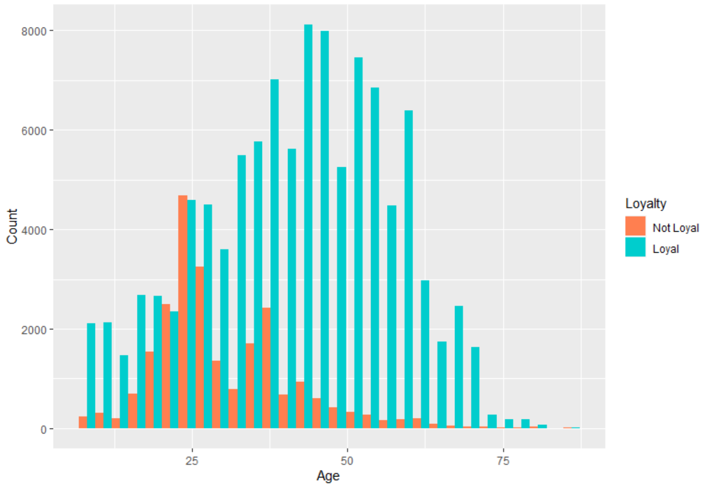
Loyalty and Satisfaction



Age effect on Satisfaction



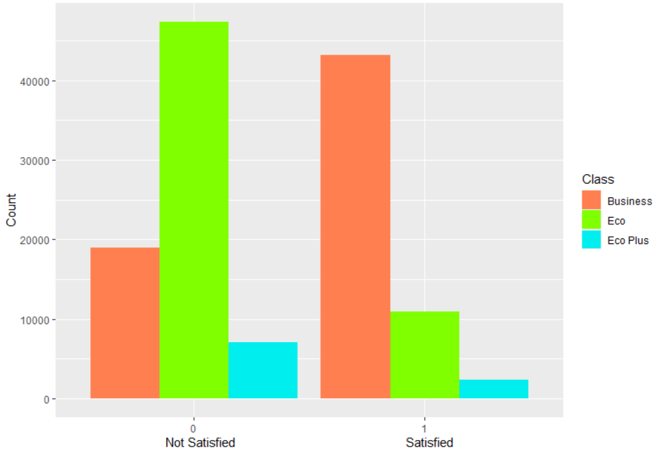
Age effect on loyalty



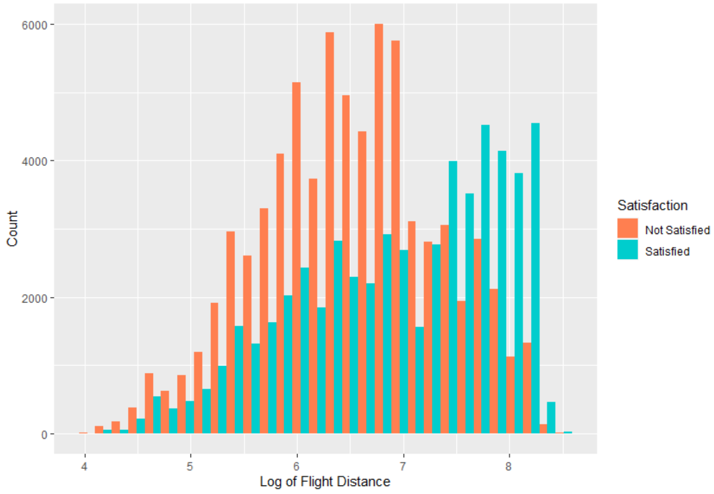
More Visualizations

- The plot between flight distance and Satisfaction shoes that customers who travel more tend to be more satisfied.
- Business travelers and customers travelling for Business purposes are more likely to be satisfied than economy class travelers and customers traveling for personal reasons.

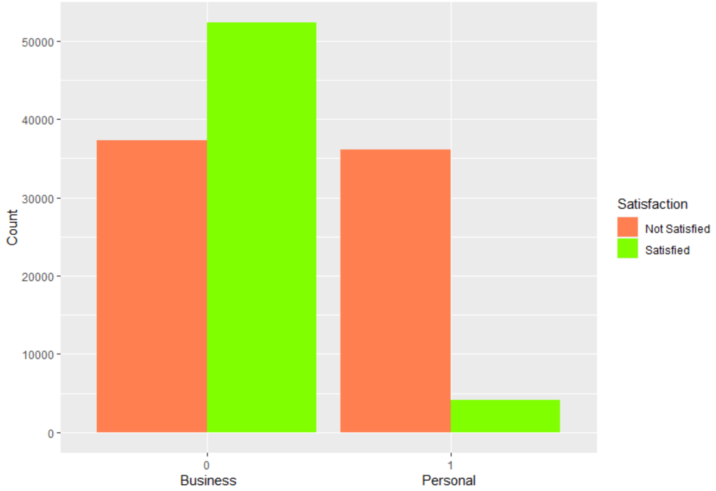
Effect of class on satisfaction



Effect of Flight Distance on satisfaction

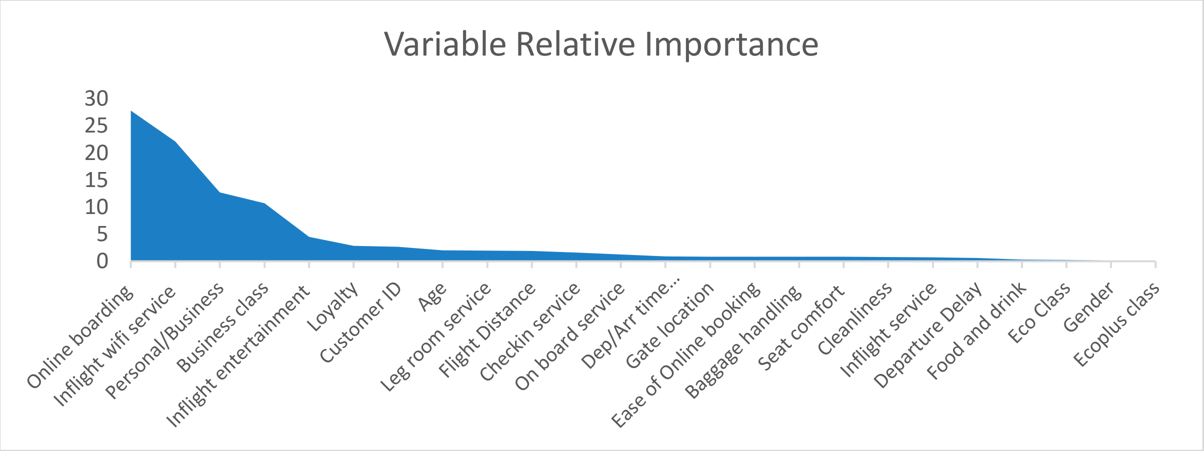


Type of travelers and satisfaction



Boruta:

This relative importance is obtained by running a Boruta, which is a Random Forest based model to study features. All features are confirmed to be important. The physical characteristics and the flight details seem to be of highest importance in predicting satisfaction. Elements determining convenience (Factor 3, Page 2) also seem to be important for satisfaction.



Are we seeing a trend here?

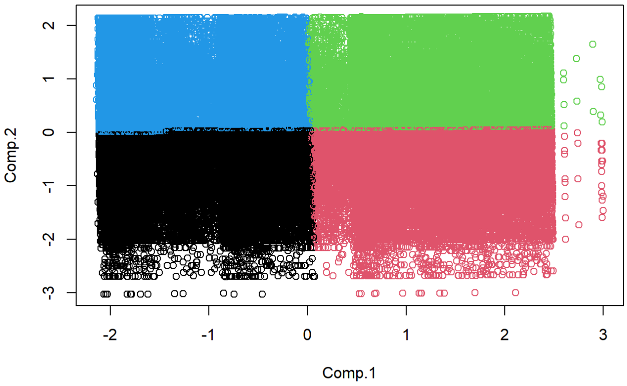
Older customers, customers travelling long distances, Business travelers tend to be more satisfied. Does this mean the company concentrates more on the so called “ High Class travelers”, ignoring the regular economy class travelers who make up almost half of their customer base. We will use clustering techniques to study this.

Clustering of Customer Types

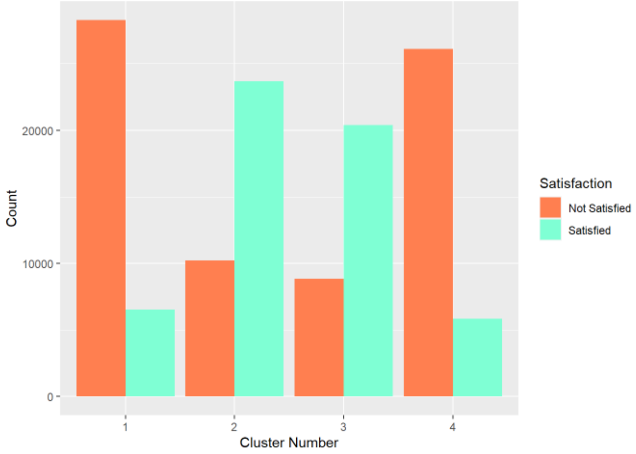
Segmentation using Hierarchical clustering on Principal Components

From our EDA we found that there are multiple type of customers we are dealing with. Therefore we clustered all the customers based on their personal details and flight preferences. Since most of the variables are categorical, we used **Principal Component Analysis** to find the variations, used Hierarchical clustering to determine the number of after which K means clustering was used to cluster the customers using the **Principal components**. This method is referred to as HCPC (Francois Husson, Julie Josse, Jerome Pages 2010). The K-means explained 75% of variance (between/total).

K means clustering on prin. Components 1,2



Satisfaction by cluster



Cluster characteristics

Cluster No (centroid)	Status (Prin. comp 1)	Age (Prin. comp 2)
1 Young Regulars	Low	High
2 Old Regulars	High	High
3 MNC Executives	High	Low
4 Businessmen	Low	Low

Principal Components

Principal Component 1:

Flight Distance

Class Number

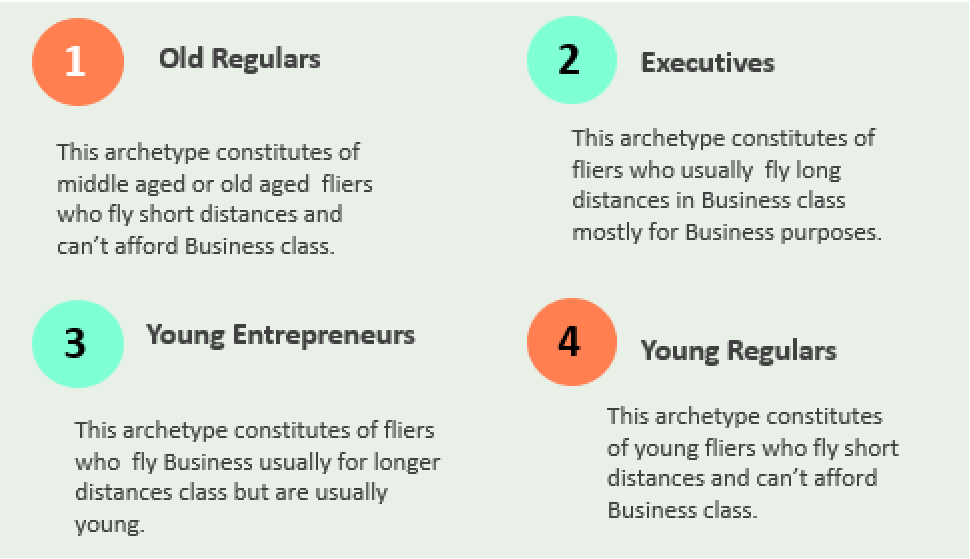
Travel Type

Principal Component 2:

Age

Customer Segments / Archetypes

There are four types of customers/archetypes which are identified. The characteristics are given below. All the four types are comparable in size but the regulars (clusters 1 and 4) are the most dissatisfied.



1 Old Regulars

This archetype constitutes of middle aged or old aged fliers who fly short distances and can't afford Business class.

2 Executives

This archetype constitutes of fliers who usually fly long distances in Business class mostly for Business purposes.

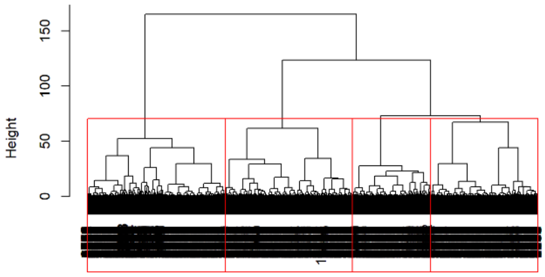
3 Young Entrepreneurs

This archetype constitutes of fliers who fly Business usually for longer distances class but are usually young.

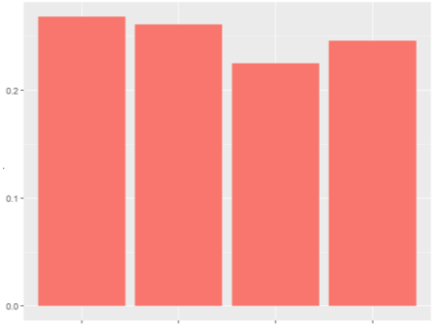
4 Young Regulars

This archetype constitutes of young fliers who fly short distances and can't afford Business class.

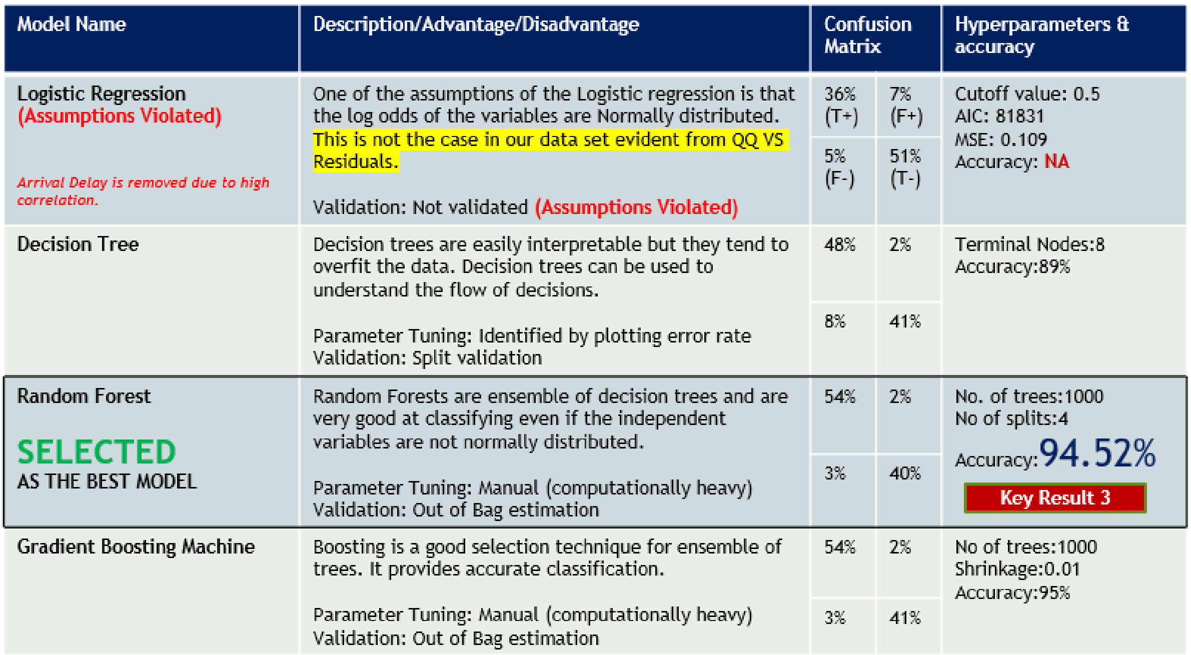
Hierarchical clustering (4 centers)



Total fliers by cluster



Model Selection:



Model Name	Description/Advantage/Disadvantage	Confusion Matrix		Hyperparameters & accuracy
Logistic Regression (Assumptions Violated) *Arrival Delay is removed due to high correlation.*	One of the assumptions of the Logistic regression is that the log odds of the variables are Normally distributed. This is not the case in our data set evident from QQ VS Residuals. Validation: Not validated (Assumptions Violated)	36% (T+)	7% (F+)	Cutoff value: 0.5 AIC: 81831 MSE: 0.109 Accuracy: NA
		5% (F-)	51% (T-)	
Decision Tree	Decision trees are easily interpretable but they tend to overfit the data. Decision trees can be used to understand the flow of decisions. Parameter Tuning: Identified by plotting error rate Validation: Split validation	48%	2%	Terminal Nodes:8 Accuracy:89%
		8%	41%	
Random Forest SELECTED AS THE BEST MODEL	Random Forests are ensemble of decision trees and are very good at classifying even if the independent variables are not normally distributed. Parameter Tuning: Manual (computationally heavy) Validation: Out of Bag estimation	54%	2%	No. of trees:1000 No of splits:4 Accuracy:94.52% Key Result 3
		3%	40%	
Gradient Boosting Machine	Boosting is a good selection technique for ensemble of trees. It provides accurate classification. Parameter Tuning: Manual (computationally heavy) Validation: Out of Bag estimation	54%	2%	No of trees:1000 Shrinkage:0.01 Accuracy:95%
		3%	41%	

Inference

- The above table shows various models fitted to the data to predict the satisfaction of customers. The Logistic Regression model was not used because the plots revealed heteroscedasticity, non-normal distribution and this model cannot be fitted.
- The decision tree has good accuracy and it is also easy to interpret. The Decision tree in the right can help understand customer decisions.
- The Gradient Boosting Method provided a close prediction error rate to the Random Forest but the random forest is chosen to be the best as it provided a better fit in lesser number of trees and less processing time which is important for being computationally efficient.

Most Important Variables in the models

- The most important features in almost all our models were Inflight Wi-Fi service, Online Boarding and Inflight Entertainment.
- These features are part of the Convenience Factor we have identified earlier (Refer slide 2, Factor Analysis) and are disproportionate among high/low class customers.

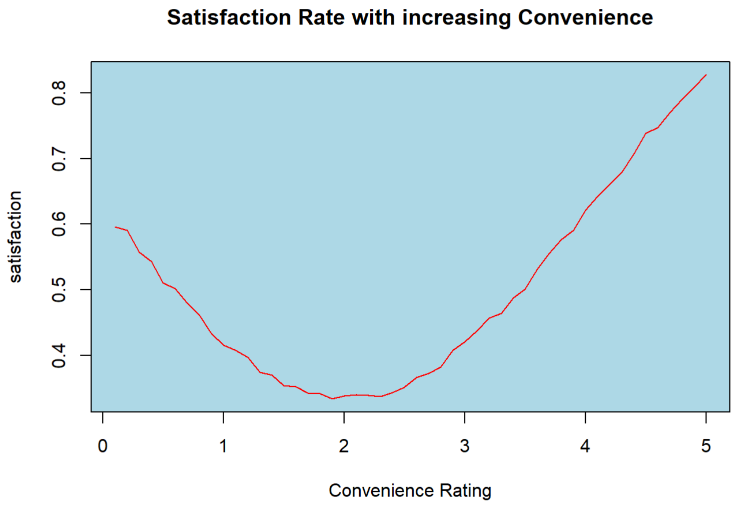
Cluster	Inflight Wi-Fi	Online Boarding	Inflight Entertainment
All clusters	Mean: 2.7 Std. Dev: 1.3	Mean: 3.2 Std. Dev: 1.4	Mean: 3.4 Std. Dev: 1.3
Cluster 1	Mean: 2.7 Std. Dev: 1.2	Mean: 2.8 Std. Dev: 1.3	Mean: 3.1 Std. Dev: 1.4
Cluster 2	Mean: 2.8 Std. Dev:1.4	Mean: 3.6 Std. Dev:1.2	Mean: 3.6 Std. Dev: 1.2
Cluster 3	Mean: 2.8 Std. Dev: 1.4	Mean: 3.6 Std. Dev: 1.2	Mean: 3.6 Std. Dev: 1.2
Cluster 4	Mean: 2.7 Std. Dev: 1.2	Mean: 2.8 Std. Dev: 1.2	Mean: 3.1 Std. Dev: 1.4

Inference

- The Regulars are the most dissatisfied . They have low Convenience score as well.
- The Decision trees of these two clusters are given, which shows that Inflight Wi-Fi is the most important attribute that lower status customers (Cluster 1,4) use to decide, while the higher status (Cluster 2,3) customers decide on the basis of Online Boarding.
- A graph is plotted with different convenience rating (0-5) against the proportion of satisfied customers. The population variance and Normal assumptions were used for calculations.

KEY RESULTS:

- Convenience factors should be improved for improving customer satisfaction.
- Convenience includes Inflight Wi-Fi, Online Boarding and Inflight Entertainment.
- A convenience rating above 4.5 satisfies almost 70% of the customers.
- 60% increase in satisfaction will be achieved if convenience factors are improved.



Reference:

1. Md. Shamim Reza, Sabba Ruhi. Study of Multivariate Data Clustering Based on K-Means and Independent Component Analysis. American Journal of Theoretical and Applied Statistics. Vol. 4, No. 5, 2015, pp. 317-321. doi: 10.11648/j.ajtas.20150405.11.

2. Husson, F., Josse, J., & Pages, J. (2010). Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data. Applied Mathematics Department, 1-17.

3. Jafarzadegan, Mohammad & Safi, Faramarz & Beheshti, Zahra. (2019). Combining Hierarchical Clustering approaches using the PCA Method. Expert Systems with Applications. 137. 10.1016/j.eswa.2019.06.064.

4. Plötz, Patrick & Jakobsson, Niklas & Sprei, Frances. (2017). On the distribution of individual daily driving distances. Transportation Research Part B: Methodological. 101. 213-227. 10.1016/j.trb.2017.04.008.

5. Husson, Francois & Josse, Julie & Lê, Sébastien. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 25. 10.18637/jss.v025.i01.

Contribution:

Devareesh – Data Cleaning, Business Objectives

Mohammed Rizwan – EDA, Clustering

Wellington Daniel – Model Fitting and Selection, Summary

Namrata Velivelli – Compilation and Report