**(Yousef Mutari & Nassar Alanazi)**

**Title:** Are Southeast Airline Customers Satisfied Compared to other Airlines? What Drives Low Satisfaction? Can IBM Watson Answer these Questions?

**Research Objective:**

In this research we represent a specific airline, Southeast. We have data from a survey available from IBM Watson Platform. This data represents answers to questions collected from 63,640 customers flying within United States and using major airlines. The objective of this research is to discover if our customers are equally satisfied compared to other airlines and finally predict the reason of low satisfaction if any.

**Background**

customer satisfaction in general and how if our customer are satisfied compared to other competition is important factor to survive the struggle in this competitive market.

We assume that this survey was most likely requested by the recent news article citing new air fare increases in the industry. After we received the survey results with information on satisfaction, the loyalty status they have with Southeast and the competitor airlines, age, gender, how price sensitive they are, last time they flew with your airline, how many flights they do, how they spend their time at the airports, what type of travel they did, what kind of class they travel in, trip origin and destination, if they were delayed or not and so on.

We loaded the results of the survey into RapidMiner and did some preparation such as dealing with Missing Data and Outliers. Finally, we used IBM Watson Analytics to Perform three tasks:

1. Explore what the respondents are saying in their responses. How satisfied are flyers with our airline, which customers are not happy and what other things stand out in their responses?
2. Identify what is influencing satisfaction so you can develop projects and take actions to alter future results.
3. Develop a dashboard to showcase your findings and use it to tell a story of what you discovered and any actions you might take from this analysis.

**Background of IBM Watson**

IBM Watson is considered one of the smartest computer system that builds today. It uses artificial intelligent and machine learning. To remember in the past for how the computer created, IBM was the first company that was able to create the machine can win the chess game in 1990s. For more idea of creating machine, IBM Watson can defy the human minds. IBM continued working on the same purpose to have designed a computer that is capable to read a lot of data. Also, IBM Watson able to analyze data and understand the human language, which is called the natural language processing. So, it can understand when you ask a question, and Watson will be able to answer the question based on the knowledge that he learned over those years.

**Data Overview**

This Airlines dataset has:

* Number of Cases (customers): 63640
* Number of Attributes: 31

**Attributes Name:**

* **Satisfaction** – it is rated from 1 to 5, that how satisfied is the customer?

5 means higher satisfied, and 1 is lowest level of satisfaction.

* **Satisfaction Top2** – only filtered out this customer’s satisfaction was 4 or 5, and also the answer will be (Yes) If the answer is 3 or lower that will be (N0)
* **Airline Status** – each customer has a different type of airline status or package, which are platinum, gold, silver, and blue.
* **Age** – the specific customer’s age. That is starting from 15 to 85 years old.
* **Age Range** – it shows the people between two particular ages, considered as a group.
* **Gender** – male or female.
* **Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.
* **Year of First Flight** – this attributes shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.
* **No of Flights p. a.** – this could be the number of flights that each customer has taken. The range starting from 0 to 100.
* **No of Flights p.a. grouped** – those flights are then grouped into groups (0, 1 to 10, 21 to 30 and so on.) For the reason they're grouped is to make analysis simpler. For example, if customers with 1 to 10 flights were more satisfied than customers with 41 to 50 flights, it would be easier to look to instead of looking at individual numbers of flights.
* **Percent of Flight with other Airlines** – if we were Southeast Airline, we would like to know how many time that customer fly with other Airlines.
* **Type of Travel** – is provide three traveling purpose for each consumer, which are business travel, mileage tickets that based on loyalty card, and personal travel like to see the family or in vacation
* **No. Of other Loyalty Cards** – it is kind of membership card of each customer, that for retail establishment to gain a benefits such as, discounts.
* **Shopping Amount at Airport** – showing the costumer’s result of how many products have been purchased. The range of shopping amount is from 0 to 875.
* **Eating and Drinking at Airport** – it is the quantity eating and drinking per each consumer at the airport. The masseur of how often for eating and drinking, which is 0 to 895.
* **Class** – it consisted of three different kinds of service level such as, business, and economy plus, economy. Moreover, customers have optional to choose their seat.
* **Day of Month** – it means the traveling day of each costumer. In this attribute, shows total of 31 days of the month.
* **Flight date** – all of these data are abbreviate the passenger’s flight date travel, which were since 2014 and only in January, February, and March.
* **Airline Code** – basically, it is unique two or three digits that mean what is the specific type of airline. There are several codes that consumers have been going with. For example, AA, AS, B6, and DL.
* **Airline Name** – There are several airlines company names such as, West Airways, Southeast Airlines Co, and FlyToSun Airlines Inc. This attribute provide what airline name that passenger have been used.
* **Origin City** – refers to actual city that customers have departed from. For example, Yuma AZ, Waco TX, and Toledo HO.
* **Origin State** – same thing as origin city such as, what state that customers have departed from? A good example, Texas, Ohio, Alaska, and Utah.
* **Destination City** – the place to which passenger travels to. For example, Akron HO, Alpena MI, Austin TX, and Boston MA.
* **Destination State** – also, it is the same thing as origin city, such as, to what state passenger travel to? Some example of destination states, Alaska, Kentucky, Iowa, and Florida.
* **Scheduled Departure Hour** – the specific time at which passengers are scheduled to depart. In this data in scheduled departure hour is starting at 1 am until 23 pm.
* **Departure Delay in Minutes** – which are minutes of departure delayed for each passenger, when compared to schedule. In this data the rage are starting from 0 until 1128 minutes.
* **Arrival Delay in Minutes** – how many minutes of arrival delayed of each passenger. Rang of delayed minutes in this data are starting from 0 until 1115 minutes.
* **Flight Cancelled** – occurs when the airline dose not operates the flight at all, and that is for a certain reason.
* **Flight time in minutes** – indicate to period time to the destination.
* **Flight Distance** – the extent of space between two places. Also, that means how many minutes are passenger traveling between two different places. Rang in this data starting from 31 until 4983 minutes.
* **Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.

**Research Methods and Model Implementation**

First, cleaning the missing data in Rapidminer, and then upload the Excel Airlines file in the IBM Watson website. After that, explore the dataset by ask a question for how do the values of satisfaction compare by Airline name? Then, in the diagram we can see all the Airlines names and average on of satisfaction customers. Specifically, we are interesting in Southeast Airlines.

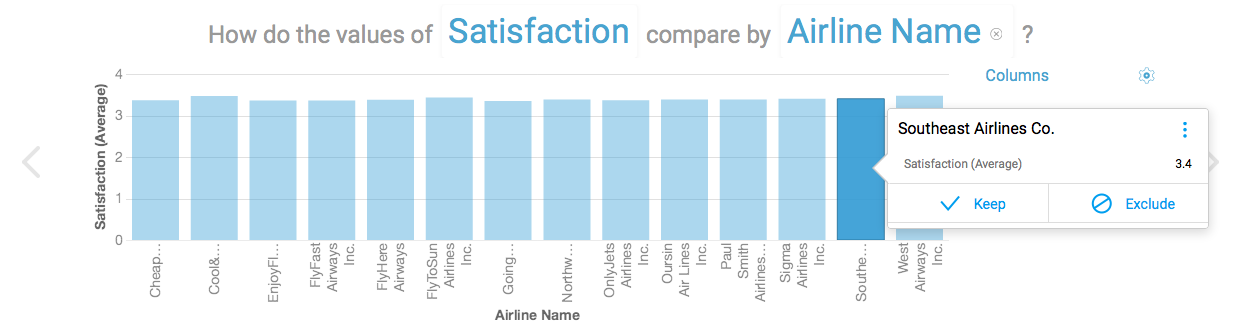
The result was no different between Southeast Airlines and other Airlines. It might be the issue is not only the Airline name, but also maybe the Airline status. There are specific group such as, platinum, gold, silver, and blue customers, and then find out which status who is not satisfied.

In IBM Watson, we were able to predicting the customer’s satisfaction based on inputs. All inputs built in decision tree. Also, we can review profiles with the strongest low predictions for satisfaction. However, we can also predict for what satisfaction associated with?

**Research Questions:**

**Research Question #1:**

**How Do the Values of Satisfaction Compare by Airline name?**



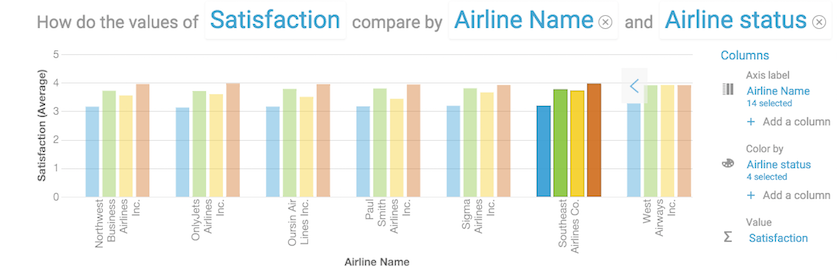
**Figure 1**

**Conclusion #1**

From Figure 1 we conclude that there is no difference between the average of Southeast Customers’ Satisfaction and other Airlines

**Research Question #2:**

**How Do the Values of Satisfaction compare by Airline name and Airline status?**

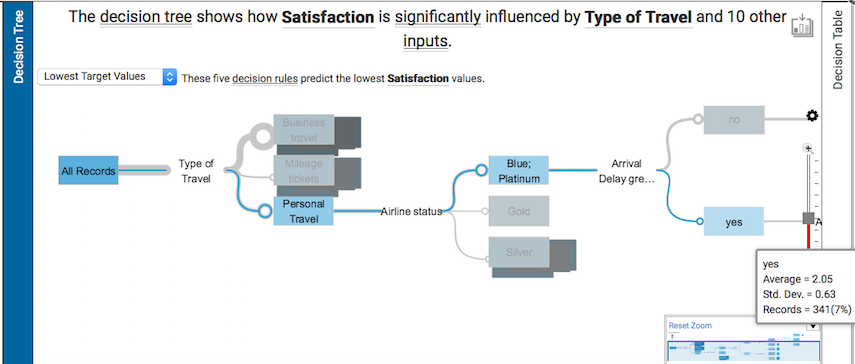


**Figure 2**

**Conclusion #2**

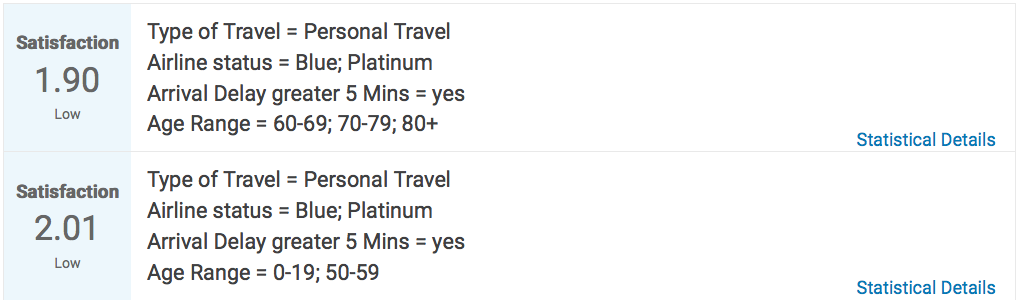
The lowest satisfied group is the blue status customers, with average of 3.18.

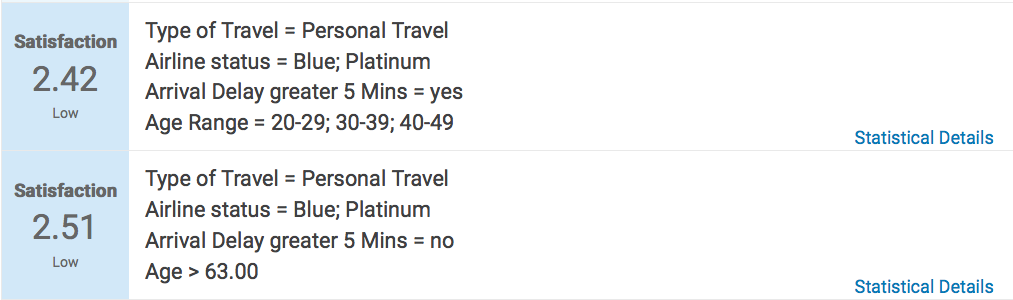
**Research Question #3: After refining our model to include only Southeast we built a Decision Tree to show how Satisfaction is influenced by the Type of Travel and other variables?**

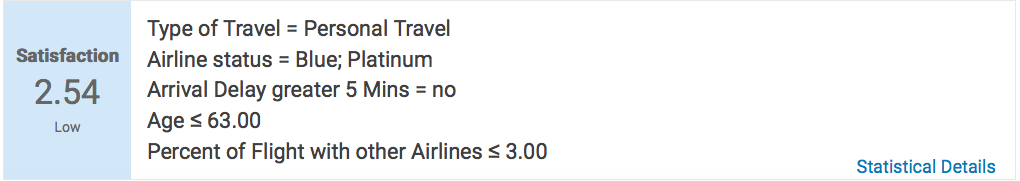


**Conclusion #3**

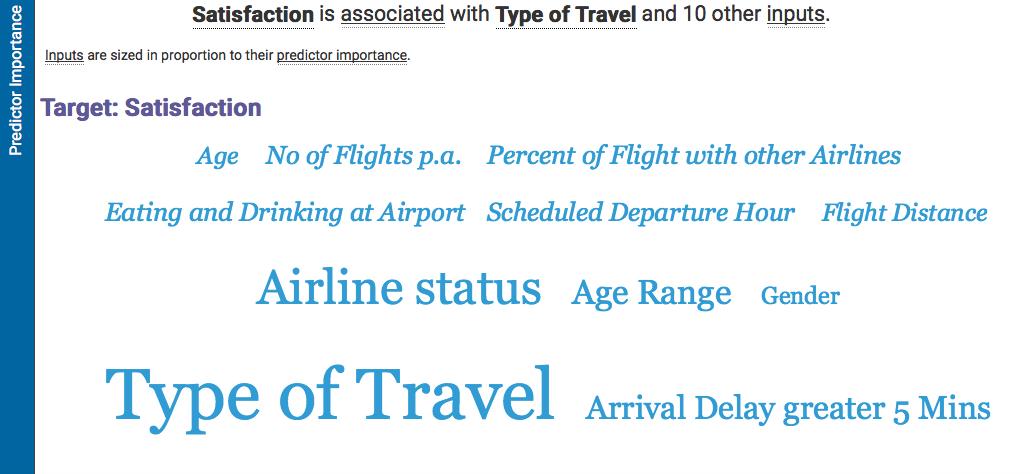
**Top design Rules**







**Predictor Importance**

****

**Final Conclusion**

As we targeted the lowest value of explaining satisfaction of customers in Southeast Airlines, we found through Decision Trees model how this satisfaction significantly influenced by **type of travel** and other inputs. The customers who have blue and platinum loyalty cards and travel for personal, and also they have **arrival delay greater than 5 minutes average** of (2.05) satisfied customers.

Also, this is the lowest satisfaction that effects Southwest Airlines. To review the lowest five decision rules predicting satisfaction. First, (1.90) Type of Travel = Personal Travel, Airline status = Blue; Platinum, Arrival Delay greater 5 Mins = yes, Age Range = 60-69; 70-79; 80+. Second, (2.01) Type of Travel = Personal Travel, Airline status = Blue; Platinum, Arrival Delay greater 5 Mins = yes, Age Range = 0-19; 50-59. Third, (2.42) Type of Travel = Personal Travel, Airline status = Blue; Platinum, Arrival Delay greater 5 Mins = yes, Age Range = 20-29; 30-39; 40-49. Fourth, (2.52) Type of Travel = Personal Travel, Airline status = Blue; Platinum, Arrival Delay greater 5 Mins = no, Age > 63.00. Fifth, (2.54) Type of Travel = Personal Travel, Airline status = Blue; Platinum, Arrival Delay greater 5 Mins = no, Age ≤ 63.00, Percent of Flight with other Airlines ≤ 3.00. Finally, the predictor importance satisfaction is associated with Type of Travel and other inputs.

**We will discuss our solution to this issues in our Poster Session on Dec the 2nd!**