Customer Satisfaction in airline Industry

IST687 M005 Group3
Jing Sun, Wesley Stevens, Tian Wang, Jiyu Feng, Yixiong Zheng

INTRODUCTION

Airlines serve hundreds of customers every day. Quality customer experience raises returning-customer rates, which increases revenue. We use flight information from surveyed airline customers to predict how likely the customer was to recommend the airline, on a scale from zero to ten. In this memo we will describe the data, necessary preprocessing, modeling techniques, and reveal the actionable insights we found. We hope to increase revenue by identifying which factors contribute the most to a quality customer experience.

THE DATA

The data has 32 attributes. This includes information about the flight: destination city, origin city, flight date and time, delays, cancellations, price sensitivity, and airline name. The data also includes information about the customer and their activities while at the airport: age, gender, flights per year, airport shopping activity, and likelihood to recommend.

As expected, the data came with missing values. We found missing values in the departure delay, arrival delay, flight time, and likelihood to recommend attributes. We need to fill in each of these missing values to perform further analysis, but we also want to do so in such a way as to keep the data true to itself.

We started by filling the missing values in the “departure delay” attribute with the average delay. The average is a generally accepted method for filling in missing values. There was only one missing value in the “likelihood to recommend” attribute. Since this is the variable we are trying to predict, and there was only one, we dropped the entire datapoint. For the missing values in the “flight time” attribute, we were able to calculate an approximate value based on that flight’s flight distance and the average speed of an airplane.

We found that the “arrival delay” attribute was missing a value whenever the flight was cancelled. No value is the proper input for this attribute since the flight never arrived, but we must replace it with something to do further analysis, so we replaced these values with a zero. We made this choice because if the flight was cancelled, then it could not have been late, since being late implies that it was anticipated to arrive. The rest of the data was in a consumable form and ready for analysis.

Descriptive Statistics & Visualization

Factor Variables

For this part, our team use table() to generate a roughly comprehensive summary of observations and find insights behind the data.

By summarizing starts and destinations, Atlanta and Chicago are two busiest cities in these three months. Atlanta is chosen as the original city and the destination more than 790 times. Frequencies of Chicago as the starting city and the destination are more than 750 times. For states, California is the most favorite state among customers. One possible reason is that metropolises are usually centers of economy and culture. Thus, most customers prefer to choose their flights between these states and cities for their vacations or business meetings.

After analyzing patterns of flight paths, we turn our sights into customer aspects. During these three months, the number of female passengers traveled with flights is higher than that of male passengers and most customers belong to the “Blue” level of travel status. Additionally, most customers’ flights are business travels, which are far more than the sum of mileage tickets and personal travels. When passengers purchase their flights, they prefer to choose the economic class for lower prices and Cheapseats Airlines Inc., whose partner code is WN.

For the airline status, the number of flights every day fluctuates around 100 and only few flights are cancelled.

Numeric Variables

For presenting precise analysis of numeric variables, our team applies histogram as a visual element to display results by using ggplot functions.

For the first part, we pay attention to the customer aspect. Histograms of customer age and traveling days are multimodal graphs, which means that distributions of data we received are spread equally in some sense. When we analyze the first year of flight situation for each customer, we attain a flat distribution which data are inflated around 1000, except the outlier “2003”. But for flights customers taken in the recent 12 months, the distribution is right-tail skewed. Additionally, when customers stay at airports to wait for boarding, most customers prefer to take food or drinks rather than shopping although both histograms display a right-tailed skewness.

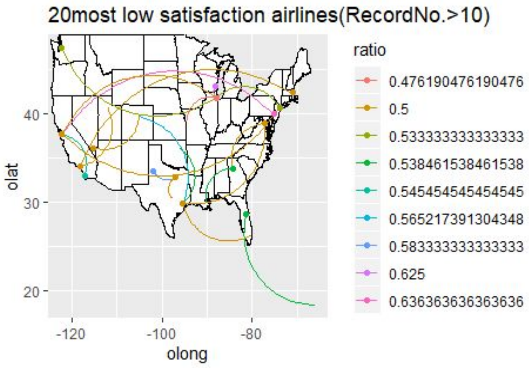
Furthermore, we choose price sensitivity, loyalty of customers and flyer accounts to find insights related to airlines. The histogram of price sensitivity shows a right-tail skewed distribution to some degree, which means that prices of flights are stable in these three months. The loyalty histogram states the low loyalty of customers with a right-tailed skewness pattern. However, 475 customers still show their loyalty to this airline as an outlier. Additionally, the distribution of flyer account is a right-tail skewed histogram.

For flights, most histograms of attributes are positively skewed. The visualization analysis states most flights are punctual. The latitudes of origins and destination are positively skewed and the longitudes are negatively skewed. The graph of scheduled departure hour is multimodal.

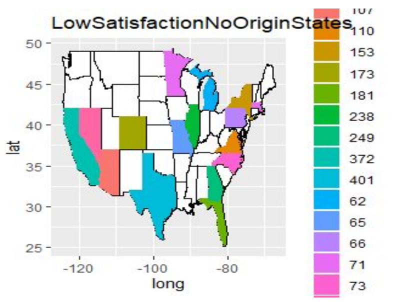
Map Low Satisfaction

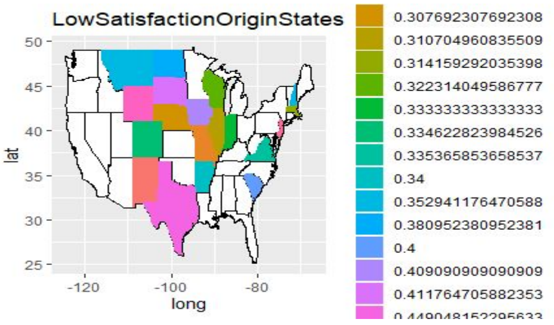
Firstly, we decided to use the attribute ”Likelihood.to.recommend” as the factor to show the customers’ satisfaction of the airlines. And then we found about 50% customers’ “Likelihood.to.recommend” is higher than 7 and half not, so we use the 7 as a flag to determine the low satisfaction. We use two methods to show the condition of low satisfaction - the airlines and the states of origin and destination state of low satisfaction airlines.In addition, in second methods we also use 2 ways to show determine the low satisfaction -low satisfaction number and low satisfaction ratio.

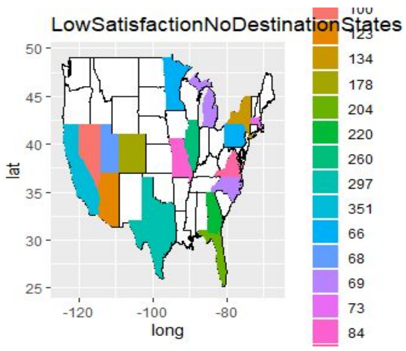
First, we use the ratio of low satisfaction and only show airlines whose records appears more than 10 times to show low satisfaction airlines. The points in the plot show the origin place of the airlines. We can find the origin place of lines appears mainly in the south east and west states and in the north and middle of US there are almost no low satisfaction lines origin.

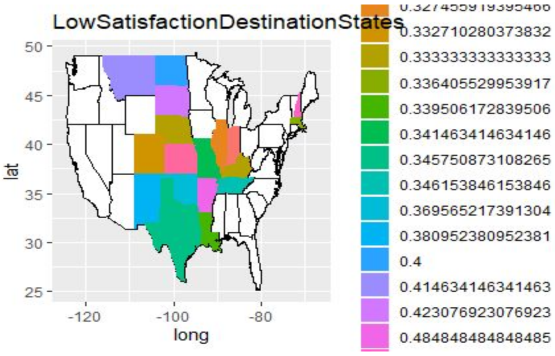


Then we use the number and ratio of low satisfaction lines of origin and destination states of the lines to show low satisfaction condition.









Looking at these four plots vertically we can find plot2 and plot4, plot3 and plot5 have almost the same states. Then we look at the plots horizontally we find in two left plots the west and east coast states have many low satisfaction lines no matter as the origin or the destination states but when we look at the right we find in terms of ratio of low satisfaction it sort of converse with left plots that the high ratio states mainly located in the middle of the US.

So we can get that there are more low satisfaction lines in east and west coast states may because the number of lines is bigger than lines in middle states. And the middle states have less lines but high low satisfaction ratio. Also, Texas appears in all 4 plots and all have a high number.

Modelling techniques

Linear regression

After mitigating the missing data, we could make use of them to predict satisfaction. There are many different types of data which cannot be used in linear regression. Hence we have to convert them into numeric. First we will convert column Airline.Status which contains string data. Such as “Silver” and“Gold” that correspond to different status. Obviously, “Platinum” symbolizes the highest status. So we switch them into ordered number. The bigger value represents higher status. The same method applies to column Class due to the same type of data. 0 represents Eco class, 1 represents Eco Plus class and Business class equals to 2.

Secondly, the column Gender and Flight.cancelled contain mutually exclusive data. Gender consists of female and male, and Flight.cancelled is yes or no. So we decide to convert into dummy variables, using 0 for female and 1 for male. Similarly, 1 in Flight.cancelled represents the cancellation. Considering the same situation in column Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes, we choose to transfer into dummy variables(1 means the flight was delayed). Notice that overall data were collected in January, February and March of year 2014. In order to distinguish flight month with satisfaction, we extracted month of each flight.

We shall also take partner companies into consideration. Every company owns corresponding code. Then we tried to create multiple dummy variables for each company. And leave one dummy out by keeping it as reference category, using the rest for regression. We also create two dummy variables for column Type of Travel, thus distinguishing the satisfaction among different travel types. After examining the data in Flight.Distance, we found that it is reasonable to classify raw data into 3 categories(short, medium, long) by value. Data in category “short” are smaller than second quartile. Finally we convert the categories into corresponding number. Flight.time.in.minutes is processed by this method..

Based on the previous steps, we are capable of integrating majority of columns into regression. We set column Likelihood.to.recommend as dependent variable and the rest numeric columns as independent predictors. First we examine the F-test on R-squared, the value is 2.2e-16. Hence the regression is significant and could be proceeded. The adjusted R-squared is 0.3775 which means predictors account for 37.75% in predicting satisfaction.

For more accurate prediction, we need to remove useless predictors. When including many independent variables in regression, we should ensure that they are correlated to avoid multi-collinearity. Considering that we assume each airplane goes about 470 mph, flight time and distance might be correlated. We remove flight distance from regression and R-squared goes up. Notice that p-value of some predictors are so high which are meaningless. Then we remove them by backward elimination method. Remove the predictor which holds the highest p-value and examine what happened to adjusted R-squared. If R-squared went up, we should drop this predictor until rest predictors contribute to a higher adjusted R-squared. Finally, we obtain a regression includes significant predictors. The F-statistic on p-value is 2.2e-16 and adjusted R-squared is 0.3779. This value 0.3779 is higher than initial regression and indicates the left factors account for 37.79% of client's satisfaction. The coefficients in regression show the strength of prediction. For instance, the higher flight class will result in 8.912e-02 unit increases in client's satisfaction. However, more times of travel per year will lead to 6.055e-03 unit decreases. Additionally, different partner companies generate different influences in satisfaction. In this case, Oursin Airlines Inc, Flyfast Airway Inc and Cheapseats Airlines Inc have a negative impact on client's satisfaction.

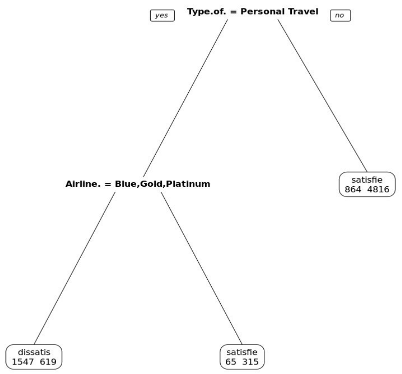
Support Vector Machine

The result of linear regression indicates 16 factors that would influence customers satisfaction. Before we use support vector machine(svm) model to test its accuracy by dividing the data set into a training set and a test set, we exclude [Departure.Delay.in.Minutes] variable because we notice its result is almost the same as [Arrival.Delay.in.Minutes] and include both is not helpful to the explaination. After we put the remaining 15 variables into svm model to predict [Likelihood.to.recommend] variable, we found the accuracy rate is 81.7% while the no information rate is 69.93%. Although this figure is quite good, we try adding and deleting some suspected variables, such as [Loyalty] and [Flight.time.in.minutes]. Finally, we get a svm model that achieves largest accuracy rate 81.9%, which means that 81.9% customers' satisfaction results can be predicted by the 16 factors.

Further Analysis & Actionable Insights

Association Rules & Regression Tree

Based on those 16 factors, we try association rules to find if there are some significant factors that correlate with NPS. By comparing satisfied customers and dissatisfied customers, it seems that dissatisfied customers are associated closely with personal travel, flight delay, blue airline status, no frequent flyer account, female and more flights per year. Among the characters above, travel type is the most significant one. Personal travel customers are more likely to have lower NPS while business travel customers' NPS are much higher.



The regression tree model only sets two branches (travel type and airline status) down and it achieves accuracy rate of

81.56%, which verify what we found from association rules and drive us to explore these two factors.

Subset Data & Key Factors

Recall the association rule we did before, dissatisfied customers are correlated with blue airline status. Combined with regression tree above, we think personal travelers with blue airline status are important labels of detractors and we have to further analyze the inner reason so that we could solve it to improve NPS.

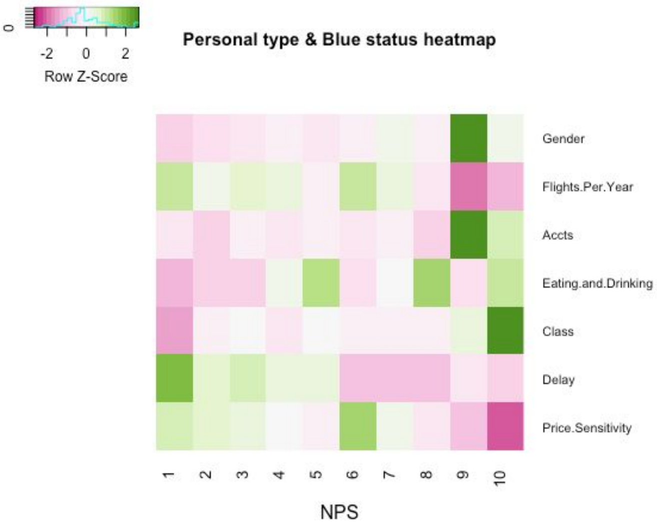
We extract the records of personal travelers with blue airline status and compare dissatisfied customers(NPS<7) with satisfied customers in association rules. By doing so, we found that dissatisfied customers are associated with flight delay, no frequent flyer account, female, economy class, and more flights per year. In comparison, satisfied customers are associated with expensive meal at airport, on-time flight and low price sensitivity.

Now, we narrow the key drives from 16 variables to 7 variables:

[**Arrival delay] [Flight per year] [Gender] [Frequent flyer account]**

[Class] [Eating and drinking at airport] [Price sensitivity]

Key Factor Visualization -- Heatmap



To better understanding the effect mechanism, we draw a heatmap to see how those variables change as the NPS increases. As is shown in heatmap, we can conclude that:

1. Personal travelers with blue airline status should be paid more attention. They tend to give low NPS.
2. In such group, female, less frequent flyer accounts holders, frequent travelers and high price sensitivity travelers are more likely to be dissatisfied.
3. Frequent flyer is in a majority in dissatisfied group.
4. Delayed Flights disrupt customers’ experience. The longer the flight delays, the lower NPS will be.
5. Those take lower level class tend to have lower satisfaction.

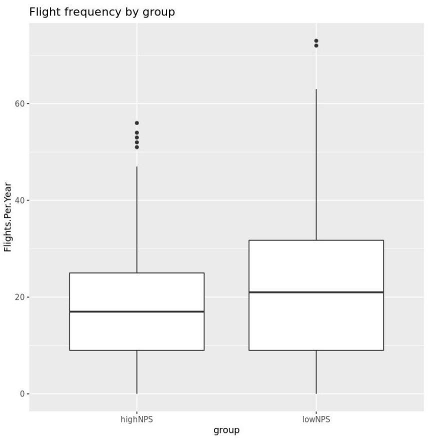
It worth mention again that flight frequency is associated with lower NPS, which should be on the alert. Because those customers should have been targeted to have registered flyer account and be high level airline status. Otherwise, they would be more likely to be dissatisfied when associated with no frequent flyer account and blue status.

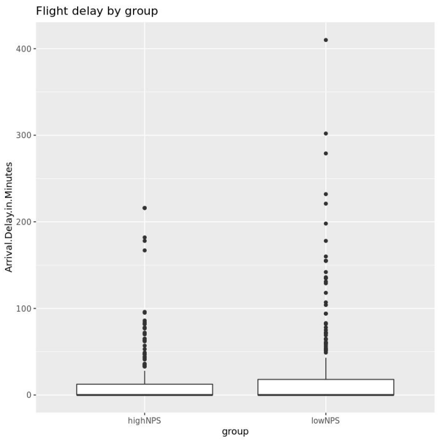
Partner Code Analysis

In the linear regression step, we include some partner code, which means there may exist some partner airline companies that offer good or bad service which affects NPS. We want to look into it by using heatmap in the 7 key factors. It is also a way to test if we can use those factors to explain certain low satisfaction trips. The heatmap implies that, delay, price sensitivity and frequent flyer accounts are the three most distinctive factors that differentiate the customers' satisfaction in different partner company.

Low satisfaction Air Route Analysis

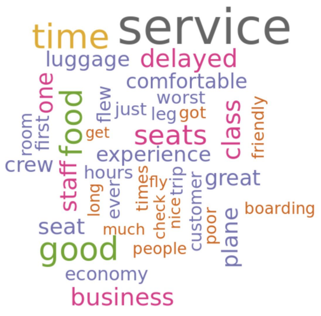
Based on the 20 lowest rate of NPS airline we mapped before, we compare with those 20 air routes with those who have the highest ratio of NPS. We found that punctuality may be the primary cause for low satisfaction air route. What's more, passengers fly more frequently in low NPS air route than those in high NPS air route, which matches the feature we found.





Textmining

Customers' reviews are very informative in improving airline service. In this dataset, 282 customers give review about their airline experience. We use textmining to analyze those reviews. Other than ordinary stopwords in English, we also delete some frequent but useless words, such as "flight", "flying", "southeast", "airline", "plane". Then, by wordcloud graph, we can see high-frequency words in customers' reviews. Other than class, delay, food which we've covered before, customers may also care about luggage, seat room, boarding process and staff.



Recommendations

Incentives for upgrading airline status

Based on the conclusion that lower level airline status and less frequent flyer account disrupt customer experience, we should incentivize customers to register frequent flyer account and upgrade airline status. Besides, it would be better to also take other low satisfaction group's characteristics into account, like gender, price sensitivity, food & beverage consumption, and class.

In this way, I suggest offering coupons for those who register frequent flyer account. These coupons can redeem discount in restaurants or shops at airport, or discount for higher class next trip. When choosing coupon partner, I would suggest considering female's preference because female is more likely to be dissatisfied than male. Once we successfully incentivize passengers to register accounts, it is important to put forward attractive upgrade programs and make those programs more transparent and customer-friendly.

Semi-premium Class

Depending on the analysis above, we know that higher class would improve customers' experience. However, it would cost too much to upgrade customers to higher class. So we could balance it with "semi-premium" class which offers customers a few add-on services to economy ticket. For example, offer customized meal and skip the queue while boarding.

Improve Punctuality Rate

it is no surprising that delayed flights ruin customer's experience. So it is of great significance to decrease delayed flights. But when it comes to a delay flight, we should handle it wisely. Keep passengers informed and give some compensations, such as coupons, upgrading class for their next trip.

Further Analysis on Unexplored Factors

As what we see from text mining, there are still some complaints that we haven't analyzed further, like luggage, staff service. That's where we should collect more data and analyze from.