Passengers‘ Satisfaction about Airline Business

Qinren Zhou(qz142), Jieqian Liu(jl2750), Zidong Xu(zx92), Jieyi Sun(js4958), Zixuan Wang(zw306)

Introduction

As the coronavirus pandemic continues, building and maintaining customer satisfaction with flying will be the key to getting back to 2019 levels of commercial aviation industry. From those sentiment of social media, an irreversible trend comes out to be that customers are caring about the flying experience more than ever. In order to study the specific influences on customer satisfaction for airline business, this project focuses on solving 10 major problems using machine learning approaches.

Analysis/Statistical methods

- **Background**

Covid 19's Impact on Airport Traffic in 2020

COVID-19 has severely crippled the global airline industry with air service reductions widespread throughout 2020. This map shows the overall traffic to and from the Airport as a Percentage of the Traffic volume during the baseline period. The baseline period used for computing this metric is from 1st Feb to 15th March 2020. Compared to the first 3 months, the traffic volume in the rest of the year 2020 has decreased a lot. From this map plot, it is obviously that the airports in Boston, College Park, Grapevine and Los Angeles has the minimal traffic. They may suffer a lot from Covid 19 and had to cancelled many domestic and international flights.

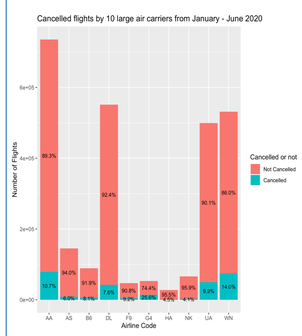

Figure 2.

This plot aid those seeking to visualize the impact that the virus has had on the domestic United States airline industry through cancellation data. It tracks the performance of domestic flights operated by 10 large air carriers from January - June 2020. The meaning of those Airline Carrier Code: AA: American Airlines, AS: Alaska Airlines, B6: JetBlue, DL: Delta Air Lines, F9: Frontier Airlines, G4: Allegiant Air, HA: Hawaiian Airlines, NK: Spirit Airlines, UA: United Airlines, WN: Southwest Airlines. We can see during this period, the AA and DL has the highest traffic volume: they have the most flights but also cancelled a lot. But when it comes to the percent of cancelled, The G4 and WN is the maximum. So the satisfaction of G4 and WN airlines may be affected more than AA and DL although they have more cancelled flights. So we will explore what is the most important factor for passenger satisfaction in the following part.

- **Data Cleaning**

  - The Twitter US Airline Sentiment dataset consists of text tweets with label ‘Positive’, ‘Negative’ and ‘Neural’. In order to do the further classification, tokenization and stemming were utilized to clean the text data, and then useless stop words were removed.
  - The wage dataset consists of Airlines companies’ employee compensation and customer satisfaction combined by year. The label for SVM classifier is generated by customer satisfaction rate.(“high ” for satisfaction rate higher than the average and vice versa)

- **Method**

The following is the machine learning methods used in this project:

  - **Random Forest** : using Gini index to perform classification of labeled text data.
  - **Decision Trees**
  - **LDA/Logistic Regression**: to explore which services have a positive effect on improving customer satisfaction
  - **Best subset selection**: objectively selecting the model that fits the best
  - **Support Vector Machine** with a linear kernel
    - suitable for classify n continuous variables
    - repeated cross-validation method with a resampling iteration of 10 and 3 complete sets of folds to compute

Models and Results

- **Panoramic View of Sentiment for US Airlines**

In order to obtain the most realistic and direct voice as well as the satisfaction from airline customers’ opinion, a labeled set of Twitter data were applied to measure the sentiment of airline customers in this section.

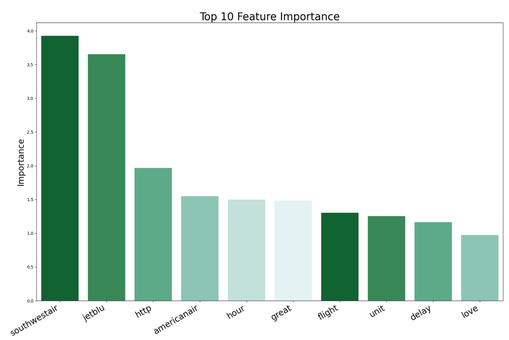

Figure 3.

Random Forest algorithm was applied to classify the Twitter US Airline Sentiment data of label ‘Positive’ and ‘Negative’. The feature importance outcome is shown in the bar chart above.

- The Top 10 most important feature for sentiment classification includes ‘Hour’ and ‘Delay’, which indicates that customers care about the time-related factors when evaluating airlines in the US.
- ‘Southwest’ and ‘Jetblue’ ranked as the first two key words for classifying positive tweets and negative tweets about airlines..

- **Inflight Service and Ticket Prices**

These variables were used to evaluate how inflight service and ticket prices influence satisfaction levels: customer type, the actual age of passengers, purpose of the flight of the passengers, travel class in the plane of the passengers.

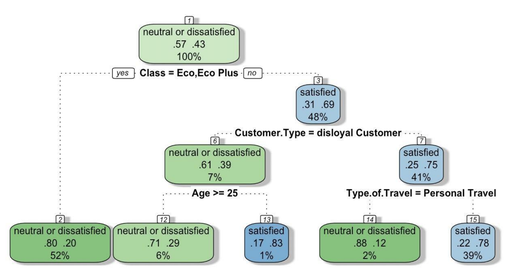

Figure 4. Decision Tree for satisfaction level prediction

Main Strategies to improve satisfaction level according to the Decision Tree:

1. Improve the inflight service for customers in economic and economic plus classes.
2. Encourage more customers in business class to become loyal customers.

- **Customers’ Patience**

These variables were used to build logistical model for evaluating the relationship between customers’ patience and satisfaction level: Flight distance, Departure Delayed, Arrival Delayed

	Accuracy	Sensitivity	Specificity
cut=0.47	0.6626059	0.4757111	0.8055608
cut=0.48	0.6632141	0.4643295	0.8153398
cut=0.49	0.6630307	0.4493841	0.8264477

Table 1. Accuracy, Sensitivity and Specificity of Different Threshold in the Logistical Regression Model

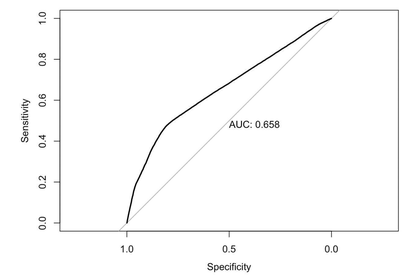

Figure 5. Receiver Operating Characteristic Curve

The area under curve indicates that flight distance, departure delay and arrival delay are insufficient for predicting satisfaction level.

- **Analyze the factors that affect user satisfaction**

The purpose of this section is to explore which services have a positive effect on improving customer satisfaction, we use variables 9-22 of Airline Passenger Satisfaction dataset (each sub-item is scored on a scale of 1-5)

First we use LDA to analyze service items that have a significant impact on consumer satisfaction. The figure below shows the distribution of scores for two types of consumers with different attitudes.

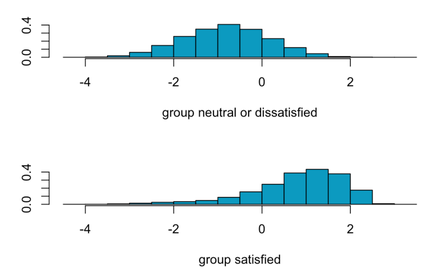

Figure 6. The distribution of consumer questionnaire scores

By looking at the graph below, we can see that the questionnaire scores of customers who are satisfied with the airline service are significantly higher than those who are neutral or dissatisfied. Then we use logistic regression to analyze the influence of important variables in detail, and the results are as follows.

Positive Effect	
Items	CC
Online.boarding	0.83
Leg.room.service	0.34
Inflight.entertainment	0.31
Inflight.wifi.service	0.31
On.board.service	0.27
Checkin.service	0.23
Gate.location	0.12
Seat.comfort	0.15
Baggage.handling	0.062
Cleanliness	0.03

Negative Effect	
Items	CC
Departure.Arrival.time.convenient	-0.30
Departure.Arrival.time.convenien	-0.13
Food.and.drink	-0.08

Table 2. Coefficient of each important factor

- **Airline Company Employees Compensation**

To explore the relationship between airline companies’ employee compensation and customer satisfaction, a dataset containing customer satisfaction rate and employee equivalents and wages for each departments of 4 U.S. airlines from 1995 to 2020 was used.

- **Best Subsets regression** is applied to compares all possible models using airline company relevant information mentioned above to predict customer satisfaction rate.

Model Selection	
	Number of Variables
RSS (min)	6
Adjusted RSq (max)	6
Cp (min)	6

Table 3. Best subsets regression model selection metrics

According to the results of the model selection metrics RSS, Adjusted Rsq, and Cp, the model with 6 predictors (Year, Southwest Airlines, United Airlines, total flight attendant employee equivalents, total pilot employees equivalents, and pilot average wage) is the best.

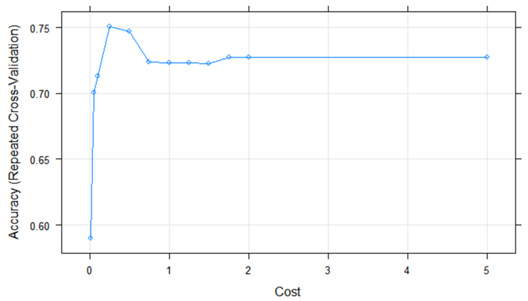

Figure 7. Cost parameter selection for SVM model

- **A SVM model** using is built to classify high or low customer satisfaction rate on the Airline companies’ employment compensation data.

C = 0.25	High Satisfaction	Low Satisfaction
High Satisfaction	12	1
Low Satisfaction	1	16

Accuracy	0.9333
Sensitivity	0.9231
Specificity	0.9412

Table 4. SVM model prediction performance with a cost of 0.25.

The optimal model is selected by GridSearchCV with a cost of 0.25.The prediction perform--ance of the model is superior (Table 4).

Conclusion

1. From the sentiment analysis based on Random Forest of Twitter US Airline Sentiment data, It comes out that airline customers care about the **punctuality of aircrafts**.
1. From the voice of social media, **Southwest** and **Jetblue** are the Top 2 airline with the highest satisfaction in the US
1. The area under curve is only 0.658 indicating that flight distance, departure delay and arrival delay are insufficient for predicting satisfaction level.
1. Through LDA and logistic regression analysis, we found that online boarding is the main service that airlines need to improve. This may be because people's life rhythm is getting faster and faster, so online boarding can not only bring consumers a more convenient flight experience, but also help consumers save time.
1. From the result of the best subset regression, the variables related to employee compensation for management fee and ground staff sector are not selected in the model, while variables related to employee compensation for **flight attendants** sector and **pilot** sector are selected.

References

- Covid 19’s impact on airport traffic: https://www.kaggle.com/datasets/terenceshin/covid19s-impact-on-airport-traffic
- Covid 19 airline flight delays and cancellations: https://www.kaggle.com/datasets/akulbahl/covid19-airline-flight-delays-and-cancellations
- Statista. (2021). American customer satisfaction index scores for airlines in the United States from 1995 to 2021.Retrieved from https://www.statista.com/statistics- -/194941/customer-satisfaction-with-us-airlines-since-1995/
- Massachusetts Institute of Technology. (2021). Airline Data Project.Retrieved from https://web.mit.edu/airlinedata/www/Resources.html
- Twitter US Airline Sentiment https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com