CPS: Data Analytics & AI

Northeastern University

Topic: The Analysis of US Airline Passengers' Satisfaction

Name: Nam Ho Phan

Course: Integrated Experiential Learn

CRN: 22974

Year: 10 April 2021

Introduction:

The standard of airline service now is the prerequisite for any companies if they improve the popularity of company to more customers (Ostrowski et al., 1993). Hence, the passengers nowadays have become more aware of the experience with any airline brands, this makes airline company must put more efforts into improving the quality of every flights; otherwise, the company will not longer be the first choice of passengers. Gronross (2000) supposed that the difference of quality in airline service created the customer's expectation and their actual observation throughout each flight. As a result, the elements having an impact on the satisfaction of customers and their loyalty should be particularly investigated to improve the quality effectively. Also, the solutions will be suggested to make the airline business focus on the necessary elements during their development.

This paper will analyze the customers' attitude towards US airline by using analysis technique of using big data, then provide the meaningful information of patterns of target customers. After evaluating the issues which need to be improved for better quality, I will apply machine learning method which aims to suggest the good predictive model for the airline. The predictive model will be useful for forecasting the result of customer's reviews in the future due to the adequate understanding of past records. Lastly, I will evaluate the effectiveness of model for company's suggestions and recommend the factors which enhance analysis performance.

Literature Review:

The importance of reviews by customers

To improve the quality of service, the appropriate strategy must be concerned as it will take more time to investigate into the detail of issues. Obviously, spending money on the change with personal feelings will cost company a lot without any profitable benefits. According to Guler & Bayrak (n.d.), the inability in understanding customer will lead to the failure of adapting to customer's demand, maximize the profit and build the long-term relationship between customers and company. As Guler & Bayrak (n.d.) said that the improvement might be done effectively if companies would easily recognize the issue of customers due to the dissatisfaction feedback, the customer survey is completely necessary, and it should encourage customers to provide the valuable feedback for service enhancement. Meanwhile, Guler & Bayrak have also shared that categorizing the customer's feature should be assigned to the goals of business. Continuous change of service quality for adapting customer's satisfaction and coping with the feedbacks proactively are the productive way to remain the loyal customer number and increase new ones.

The effects of on-time performance on the airline industry

On-time performance has been an issue to most of airline companies as it plays the most important role in keeping the loyal customers. Apart from the prices, the passengers now have more concerns about one-time arrival/departure, which is part of the business's success in competition (Yimga, 2017). According to Ball et al. (2010), not only airline industry lost billions of dollars because of the increase of flight delays, but the passengers and society also have been affected hugely. This proved that the improvement of on-time flights should be the priority of

strategy as it may have the long-term negative impact on both of economic and social aspect. More obviously, the failure of on-time flights will make the passengers become more frustrated and disappointed easily.

Methodology:

For this project, I decided to analyze the data to diagnose issues and build predictive model, then recommend better solution for business to enhance their service quality. The data should come from the reliable resource, specially I took it from Kaggle. For the accuracy of analysis, I must ensure the size of data to be at least 120,000 records with different reviews. The data was from 2015 survey so I must remind it to evaluate correctly according to the profile of business. For columns, it has totally 18 columns including the information of participants, their evaluation of airline services and the delay time of flights. To clean data set, I check its missing values and outliers, then clear it to make the data set become less skewed or bias.

Description of dataset:

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Flight distance: The flight distance of this journey

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of inflight service

Cleanliness: Satisfaction level of Cleanliness

Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

To analyze data, I will use python, the most common tools for analysis. Critical thinking will be necessary to find out the interesting questions, which lead the business to understand the pattern of their passengers and solve potential issues in future; while the visualization will be always applied to make the result become more comprehensive for managers or stakeholders. For model building, I will use machine learning packages which are already existing in python. The packages will help me to run model quickly based on the algorithms such as regression or classification. After running model, I will find the optimized result to boost the performance.

Business questions:

1/ What is the background of loyal & disloyal passengers?

2/How the delay flight impacts on the different types of passengers?

3/Which segments we should improve to make disloyal customers become loyal?

4/Whether we should improve the quality of far distance flights?

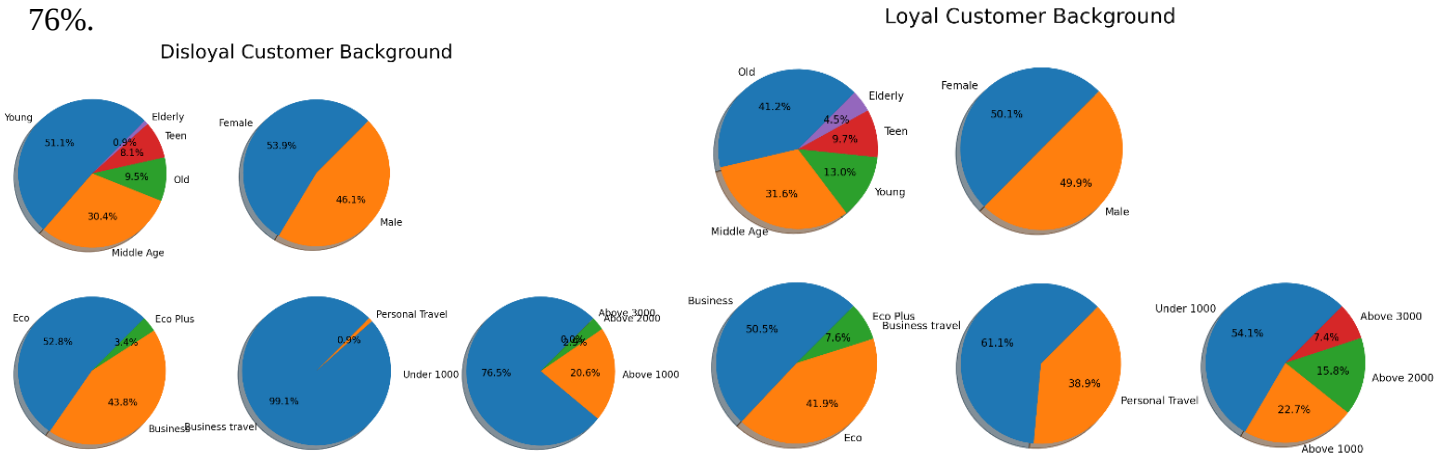
Findings:

EDA:

The Background of Loyal & Disloyal Passengers

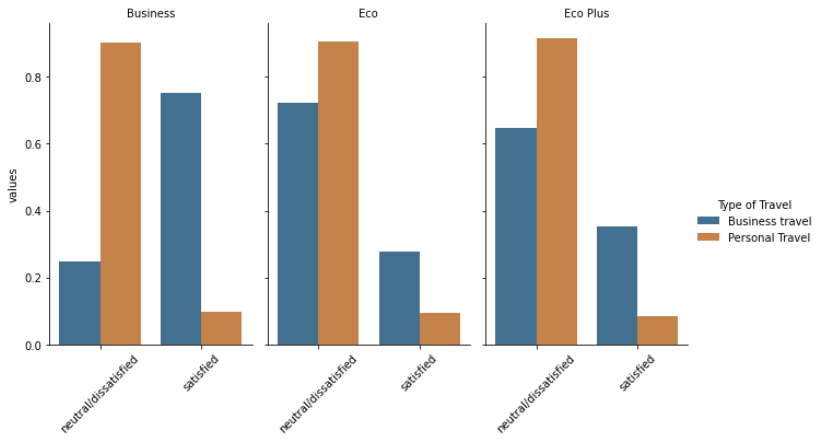
Guler & Bayrak have also shared that the customer's feature should be assigned to the goals of business as it helps in keeping loyal passengers and bringing back new passengers. Therefore, I must understand the structure of our loyal and disloyal customer patterns. Due to that, it will give business more appropriate strategies to keep loyal customers or encourage more new customers within our knowledge of them.

The pia chart will give us the general background of passengers including their age, gender, type of travel, seat class and flight distance. For loyal passengers, they are mostly 35-65 years old with around 72% of passengers and the passenger's gender is quite equivalent. The common class of loyal passengers is Business and Economy, while most of passengers will use airline for their business travel purpose with 61%. For distance of flights, loyal passengers will often choose flights with distance below 2000 miles (76%). For disloyal passengers, the difference between them and loyal ones is that they have more people with age from 35-65, they just only use airline for business travel and their distance of flights mainly is below 1000 with 76%.



The Impacts the Delay Flight Have on The Different Types of Passengers

According to Ball et al. (2010), the management of flight delays will save business billion dollars and have a positive impact on macroeconomy. Thus, I would like to find out the different feelings of passengers based on their travel purpose and different class as it is relevant to their purpose of using airline and ticket price. For the business class, we can see that although passengers using airline for personal travel has more neutral/dissatisfied feelings with more than 90% of personal travel passenger, the passengers traveling for their business purpose feel more satisfied with nearly 75%. This is opposite to the survey on passengers using Economy and Economy plus class, most of them who have flights because of business and personal purpose feel neutral/dissatisfied, with more than 60% and 90% separately.



The Proper Changes to Satisfy Loyal and Disloyal Passengers

Disloyal customers are essential part of business profit because they will potentially become loyal customers or invite more friends to use service, which adapt to the general goal of business. Understanding the reasons that make customer feel dissatisfied is needed because it will help business improve on the right way. Moreover, the feed back from loyal passengers will also give us the advice to enhance the service quality better because their reviews may be quite

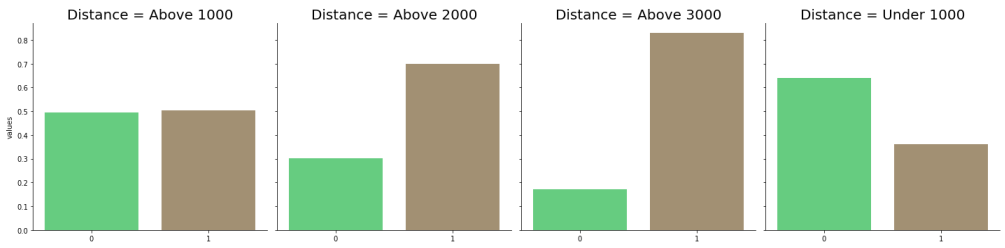
genuine due to their frequency of using. Lack of improper ways for servicing customers will cause the negative result for company in competitive environment (Guler & Bayrak, n.d.). Therefore, the discovery of passenger's reviews will help me to figure out the possible problems that make them consider changing to other airline companies.

Customer Type	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness
Loyal Customer	2.75	3.25	2.77	2.96	3.28	3.44	3.61	3.48	3.50	3.43	3.69	3.65	3.71	3.41
disloyal Customer	2.75	2.41	2.74	2.96	3.04	2.75	2.99	3.05	3.33	3.25	3.79	3.60	3.80	3.06

According to the mean scores of each segment, Inflight Wi-Fi service, Ease of Online booking and Gate Location are the most concern for both loyal and disloyal passengers with lower 3 point. For disloyal customers, they are also having bad impression about the quality of plane seating, Departure/Arrival time convenient and Online Boarding (2.99, 2.41 & 2.75). Meanwhile, this is quite comfortable to the loyal customers.

The Quality of Different Distance Flights

The reason I think of this because the flight distance will have different effects on customers' experience. The flight distance is related to the type or interior design of planes which also help technical managers understand the issues. Also, the quality of service would be different depending on each type of distance flight. The bar chart shows that passengers traveled above 2000 and above 3000 have more satisfaction with nearly 70% and 90%, while passengers using flight below 1000 is likely to feel neutral or dissatisfied with 60%. This has proved that the farer the flight distance, the better the quality of flight.



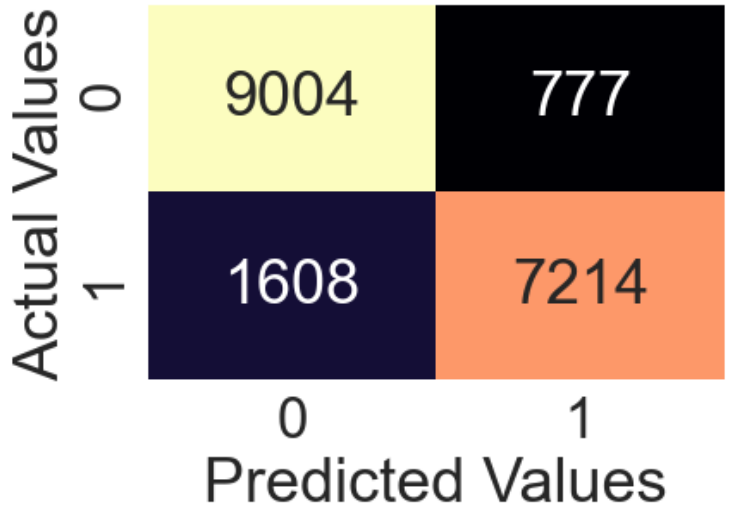
Building predictive model:

First, this model is suitable for regression and classification, so I will use logistic regression, random forest, and gradient boosting to make good prediction.

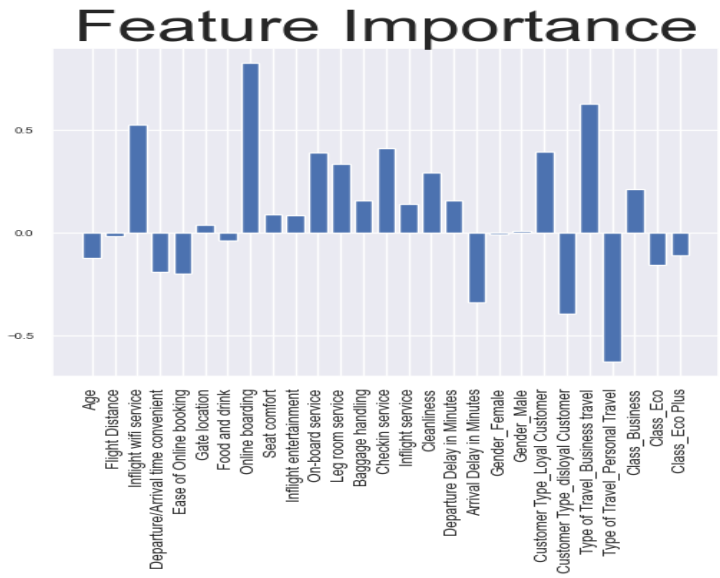
Logistic regression:

The accuracy of logistic model is 87%. The variables of feature importance are Inflight Wi-fi service, Online boarding, Business travel or Loyal customer, which contributed to the satisfaction of customers. Arrival Delay, Disloyal customer or Personal Travel are the weakest variables.

```
Logistic regression
Train Accuracy Score: 87.51%
Test Accuracy Score: 87.18%
```



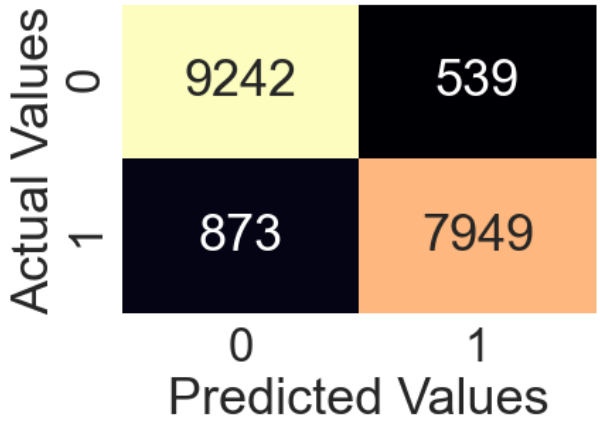
	precision	recall	f1-score	support
0	0.85	0.92	0.88	9781
1	0.90	0.82	0.86	8822
accuracy			0.87	18603
macro avg	0.88	0.87	0.87	18603
weighted avg	0.87	0.87	0.87	18603



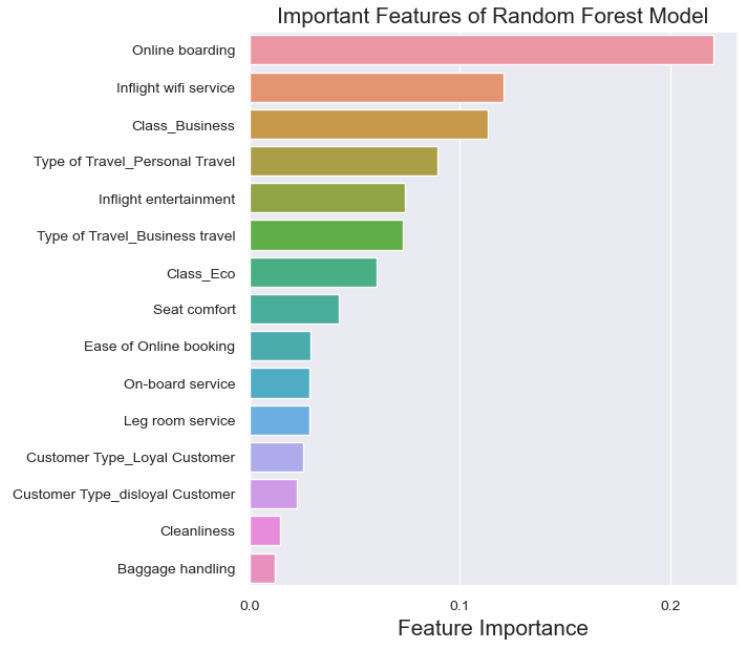
Random Forest Classifier

The accuracy score: 92.41 as model parameters {'max_depth': 5, 'n_estimators': 300}

The Random Forest Model gives me 92% of accuracy, with its optimized parameters such as max_depth as 5 and number of estimators as 300. The variables of feature importance include Online boarding, Inflight Wi-Fi service and Class Business. The weakest variables are removed as it had zero impact on this model.



	precision	recall	f1-score	support
0	0.91	0.94	0.93	9781
1	0.94	0.90	0.92	8822
accuracy			0.92	18603
macro avg	0.93	0.92	0.92	18603
weighted avg	0.92	0.92	0.92	18603

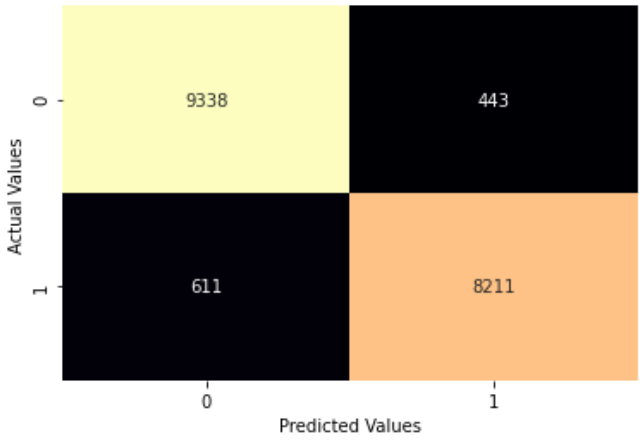


Gradient Boosting:

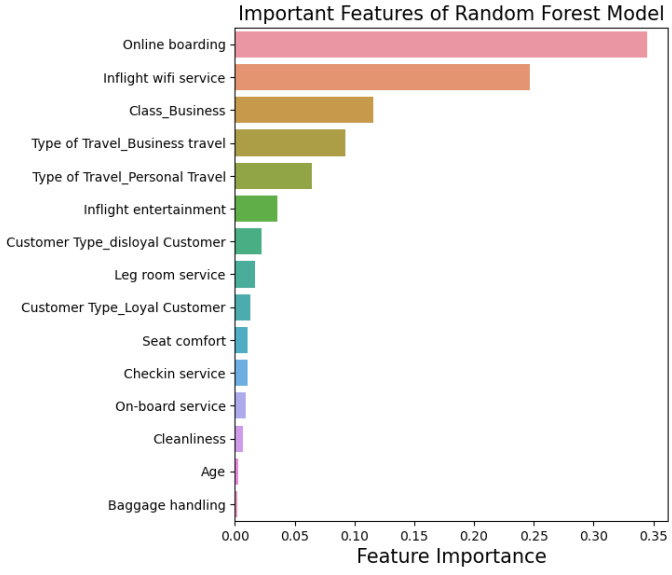
`The accuracy score: 94.33 as model parameters {'learning_rate': 0.1}`

The Gradient Boosting gives me the highest accuracy with 94%. As my computer is not strong enough to run throughout all parameters, I just only set the fine-tuning parameter for learning rate. So, I have learning rate as 0.1 is the ideal score to get that high accuracy. For

feature importance, Online boarding, Inflight Wi Fi service and Class Business are also the most important variables.



	precision	recall	f1-score	support
0	0.94	0.95	0.95	9781
1	0.95	0.93	0.94	8822
accuracy			0.94	18603
macro avg	0.94	0.94	0.94	18603
weighted avg	0.94	0.94	0.94	18603



Discussion:

After finding out the interesting result of analysis, the common features that US airline have dissatisfied their customers are the quality of ticket booking service and the Wi Fi service. Hence, the company should focus on improving the quality of these services. Meanwhile, the seating of planes, on-time performance and online boarding should be considered to invest more to keep the disloyal customers. The Eco plus class should be taken care more as the delay problems occurs because they seem to be loyal customers, so we will not want to lose these customers. For the flights with distance below 1000, I would like to recommend that companies should investigate more about the reasons that most of passengers did not feel satisfied. To increase the number of passengers, companies should create the marketing campaigns to widen their business in market by encouraging more young people to use US airlines.

For predictive model, Gradient Boosting has shown the highest accuracy with 94% compared to 92% of Random Forest Classifier and 87% of Logistic Regression Model. Hence, I will recommend using Gradient Boosting for this data set as it will be more effective. However, the decision would be more accurate if I can find the time lapse of each model. The variables of feature importance help me to use as less variables as possible and remains the high accuracy, this will save company more time and avoid overfitting when building model. The predictive model will help company forecast the behavior of passengers effectively. Additionally, the success of predictive model will help them save time a lot because it can give the result quickly instead of waiting for the review collection.

Limitations:

Throughout the project, I also faced many challenges. Firstly, the size of data set is not ideal for me to build the qualified model because it cannot reflect on the general opinion. The second challenge is that my computer is not strong enough for me to build model effectively. With the better computer, I will be able to optimize the result of model. Moreover, the lack of domain knowledge is one weakness of this project so I cannot make the further recommendation for company thought I already have the result of analysis. In future, if these challenges can be handled, I believe I can give better results for the company.

Reference:

Yimga. J (2017). Airline on-time performance and its effects on consumer choice behavior. Research in Transportation Economics. 12-25. Retrieved from **https://doi.org/10.1016/j.retrec.2017.06.001**

Guler. A & Bayrak. T (n.d.). MANAGING CUSTOMER COMPLAINTS THROUGH CUSTOMER SEGMENTATION. Retrieved from chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://decisionsciences.org/wp-content/uploads/2017/11/p669806.pdf

Ball. M, Barnhart. C, Dresner. M, Hansen. M, Neels. K, Odoni. A, Peterson. E, Sherry. L, Trani. A & Zou. Bo (2010). A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States. Total Delay Impact Study. Retrieved from https://rosap.ntl.bts.gov/view/dot/6234

John D (2018). US Airline Passenger Satisfaction, Version 2. Retrieved from https://www.kaggle.com/johnddddd/customer-satisfaction