Airline Passenger Satisfaction Analysis

I. BUSINESS UNDERSTANDING

1.1 Background

The global airline industry is ravaged by COVID-19 as people travel significantly less than previously. Nonetheless, at the end of the pandemic, it is anticipated that demand for air travel will increase dramatically as people rush back from vacations. How can airlines best position themselves for success after the pandemic? Understanding what makes customers happy and satisfied is crucial for solving this business challenge because customers who are unhappy or dissatisfied can lead to fewer rides and lower sales.

However, an airline cannot just improve customer satisfaction without addressing the costs involved, such as providing free onboard services or discount rental cars. Faced with such a great amount of spending, airline companies need to think carefully when making the decision about how to reach optimal profit when putting the budget into improving customer satisfaction.

Moreover, the flight itself is only one part of the overall customer experience. On-time flights, quality in-flight entertainment, better snacks, and so on all contribute to a pleasant experience. Due to the complexity of customers' experiences, it is important to build a machine learning model to better predict customers' satisfaction levels in advance.

1.2 Business Objective

In this project, we want to first know how to keep customers satisfied by predicting when they are prone to be unsatisfied. We aim to build a model to help airline companies make better decisions about how to get the maximum profit by formulating optimal customer service strategies based on customer satisfaction prospects. In summary, the business questions this project aims to answer are:

- What is a model that airline businesses can employ to predict passenger satisfaction?
- How can airlines maximize profits for each customer based on customer satisfaction prospects?

1.3 Data Mining Goals

The data mining goals of this project are to 1) predict how likely a customer will be satisfied with a trip, given their demographic information (age, gender, etc.), flight status, and pre-flight and mid-flight experience. The intended output of this project is to have accurate probability predictions for each customer. On the other hand, airline companies need to balance costs and revenue. So 2), we hope to help them gain an overall picture of how much they should spend on each customer to increase their satisfaction and subsequently reach maximum profit.

II. DATA UNDERSTANDING

2.1 Describe Data

We obtained an airline passenger satisfaction survey from the kaggle website (Airline companies are hidden for confidentiality). The dataset contains 119255 customer records and 23 features, including customers' **demographic** information, their **subjective satisfaction survey results** (scale 1-5, with 5 being the most satisfied) for each aspect during pre-flight and mid-flight, such as their satisfaction with online boarding or in-flight wifi services, and **the flight status**, including flight distance, flight delay, etc. [Appendix 1]. This data's dependent variable is whether or not consumers are satisfied (binary variable), given their demographic information, subjective satisfaction survey results, and flight status. The data has 64569 satisfied customers and 54686 unsatisfied customers, respectively.

2.2 Data Visualization

2.2.1 Customers' demographic with satisfaction

Gender: Each gender is well-represented and balanced, and females are more satisfied with the overall service of the airline company. [figure 1]

Age: People who are 25 to 60 are the most represented in this data set. This difference could be that the young and old groups don't travel much more often compared to the 20–60 age group, and thus there are fewer satisfaction surveys filled out. [figure 2]

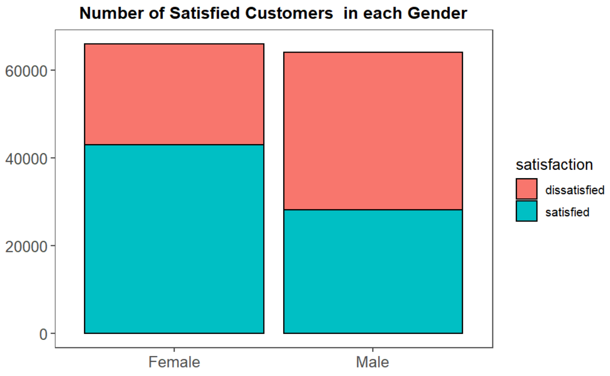


figure 1

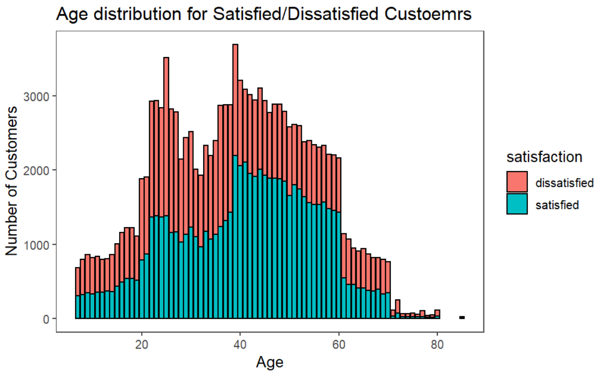


figure 2

2.2.2 Subjective satisfaction survey

From the boxplot, we can see that the most passengers were satisfied with “Online support”, “Online booking”, “Online boarding”, "Inflight.entertainment", "On.board.service", "Leg.room.service", "Baggage.handling", "Cleanliness". [figure 3]

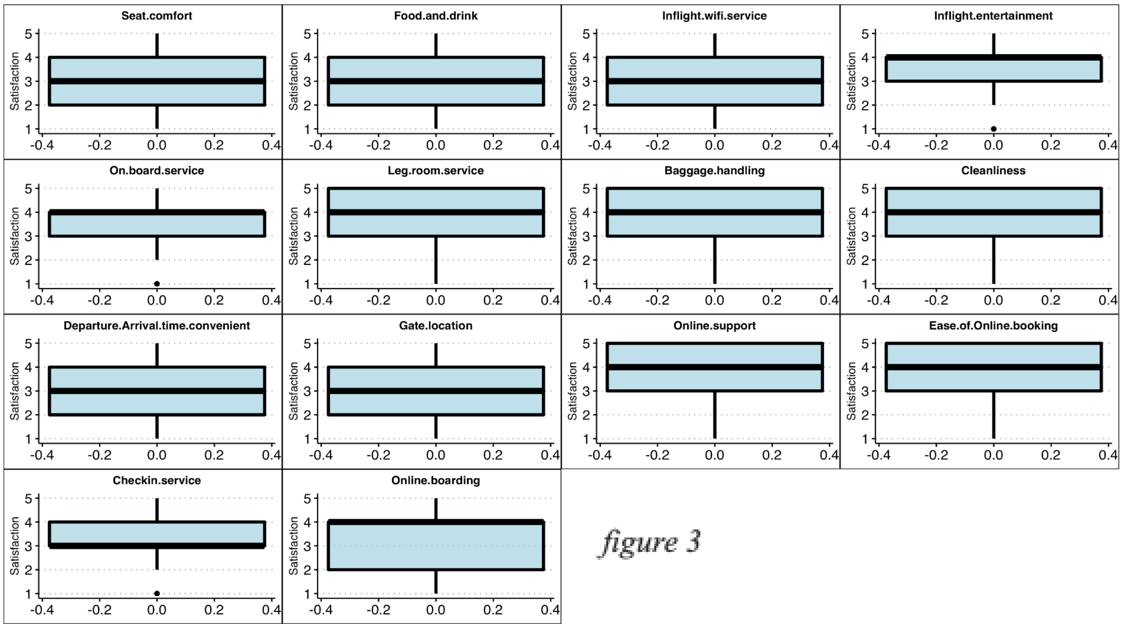


figure 3

2.2.4 Other variables

We also visualized other variables, such as customer type, type of travel and class.

Loyal customers tend to be more satisfied than disloyal customers. The graph [figure 4] indicates that airline companies should convert more customers into loyalty programs to increase the total satisfaction level of customers. When we look at the type of travel and class, we can see that there are more business travelers than personal travelers. Thus, there are more travelers from business class who are more satisfied than Eco and Eco plus. [Figure 5 & 6]

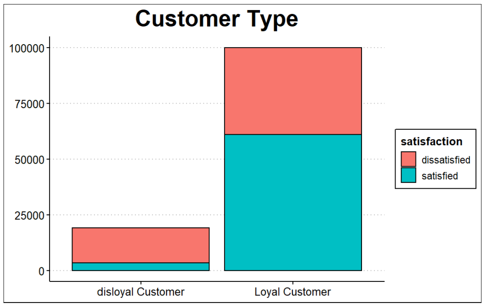


figure 4

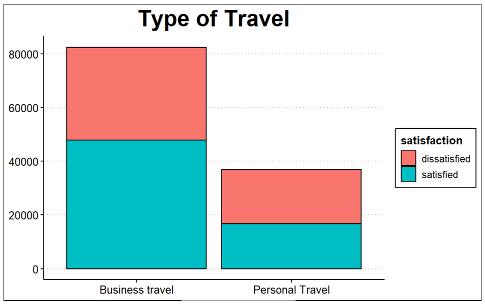


figure 5

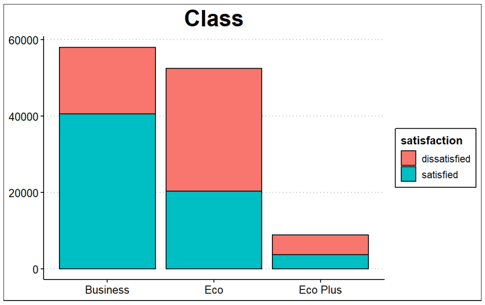
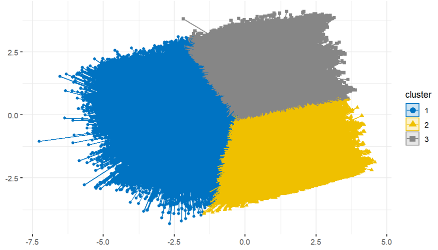


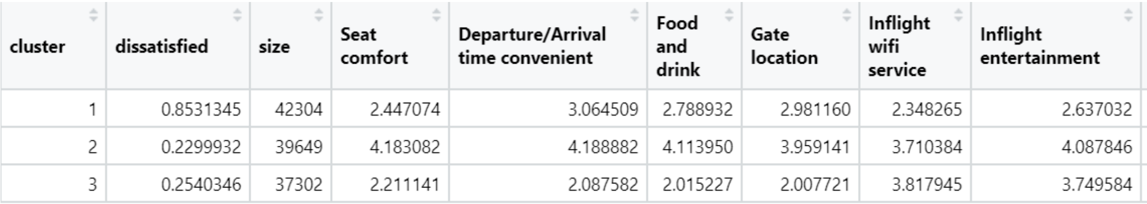
figure 6

2.3 Additional data exploration

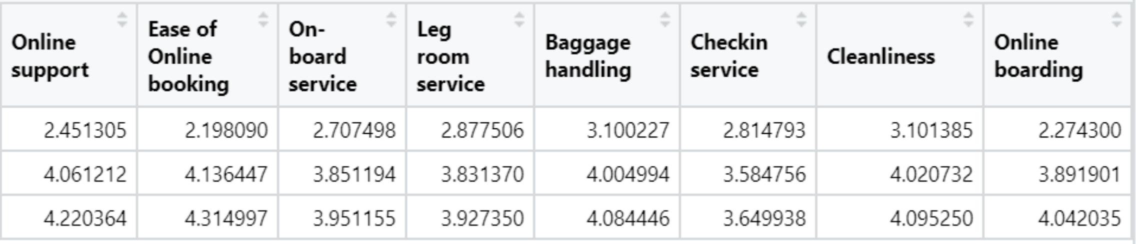
2.3.1. K-means



We want to understand our customers better through segmentation. To see what the features are of each cluster, we divided customers into three clusters. Cluster 1 [blue] has the highest probability of being dissatisfied, and it has the largest size. We can observe which service influences the customers most to their satisfaction, and we can focus on these services: In-flight wifi, in-flight entertainment, online support, online booking convenience, on-board service, leg room service, baggage handling, cleanliness, online boarding



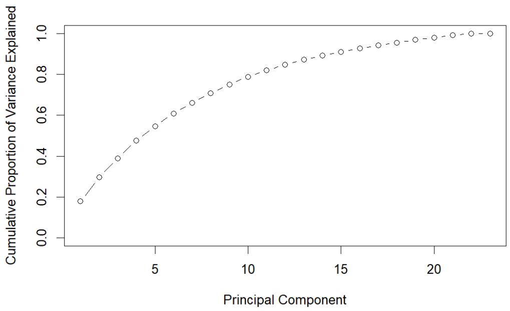
cluster	dissatisfied	size	Seat comfort	Departure/Arrival time convenient	Food and drink	Gate location	Inflight wifi service	Inflight entertainment
1	0.8531345	42304	2.447074	3.064509	2.788932	2.981160	2.348265	2.637032
2	0.2299932	39649	4.183082	4.188882	4.113950	3.959141	3.710384	4.087846
3	0.2540346	37302	2.211141	2.087582	2.015227	2.007721	3.817945	3.749584

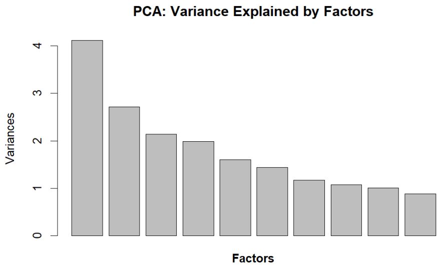


Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling	Checkin service	Cleanliness	Online boarding
2.451305	2.198090	2.707498	2.877506	3.100227	2.814793	3.101385	2.274300
4.061212	4.136447	3.851194	3.831370	4.004994	3.584756	4.020732	3.891901
4.220364	4.314997	3.951155	3.927350	4.084446	3.649938	4.095250	4.042035

2.3.2. PCA

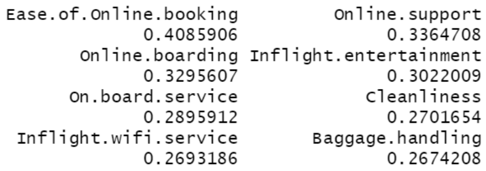
Through the PCA, we can see which variables are significant to explain the model. From the cumulative proportion graph, we can see that we need to choose 8 or 9 components to explain 70% of the variance in the model.



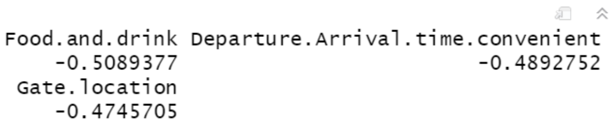


We focus on the first and second principal component and the latent variables.

First principal component: Second principal component:



```
Ease.of.Online.booking         Online.support 
             0.4085906              0.3364708 
       Online.boarding Inflight.entertainment 
             0.3295607              0.3022009 
      On.board.service            Cleanliness 
             0.2895912              0.2701654 
    Inflight.wifi.service     Baggage.handling 
             0.2693186              0.2674208 
```



```
Food.and.drink Departure.Arrival.time.convenient 
    -0.5089377                        -0.4892752 
 Gate.location 
    -0.4745705 
```

In the first principal component, we can see that the variables related to **online service** are important features to explain the satisfaction model. In the second principal component, the **in-person experience** through the flight is important in impacting the model.

III. DATA PREPARATION

3.1 Data cleaning

3.1.1. Missing values

From the summary of the data, we noticed that the “Arrival Delay in Minutes" columns have 393 NAs. It corresponds to 0.3% of the data, which is a small portion. So, we dropped them.

3.1.2. 0 values

There are some survey entries with a score of 0, which were unfilled survey questions. 0 is not meaningful for our study. After removing the 0s, the remaining data has about 119,255 entries.

3.1.3. Data type transformation

The columns “satisfaction”, ”Gender”, “Customer Type”, “Type of Travel”, “Class” are characters. We convert them into factors to prepare for model building.

3.2 Data Construction

Our final goal is to decide how much satisfaction service can be improved individually to get the maximum profit for each customer. We need to assume **ticket price** per customer, estimate a **cost and satisfaction equation** for calculating impact of improving satisfaction on cost and for a simpler calculation, we want to combine all the subjective satisfaction variables into one **overall subjective satisfaction** and **use it as our decision**. So, we made several assumptions as below.

3.2.1. Assumption about airline ticket price

As airline price algorithms are not made public and are closely guarded secrets, we need to build our own airline pricing model in order to calculate airline profit. Although in a realistic situation, airline ticket prices are decided by many factors such as current sales volume, seasonal impact, and so on, because of the limitations of data and resources, we used a simplified model with only two factors determining ticket prices. They are flight distance and class level. We used

RomeAirline [7] as a reference, and considered inflation from 2013 to 2020. Also, ticket prices vary a lot from class to class. We assume that ticket prices need to be multiplied by the class level. We created the following price model and calculated ticket prices for each customer.

$Ticket\ price\ =\ (0.2\ *\ Flight\ distance\ *\ Class\ level)\ +\ 100$

3.2.2. Assumption about overall subjective satisfaction

The research reports by JD Power in 2019 [1] and 2022 [2] indicate that in-flight services—especially food and beverages—are key to passenger satisfaction. Inflight entertainment is also the primary driver of passenger satisfaction. Hence, we consider those two variables to have the highest weight. Moreover, good customer service served as 36% of the contribution to the airline selection. Based on the observations, we divided subjective satisfaction factors into four categories: food and drink, inflight entertainment, customer service (including check-in service, on-board service, inflight wifi service, online support, ease of online booking, baggage handling) and others. We gave these groups specific weights of 0.3, 0.2, 0.3, and 0.2, respectively. We used the following model to summarize overall satisfaction for each customer:

*Overall satisfaction = 0.3 *food and drink + 0.2*inflight entertainment +0.3* 0.2*(checkin service + on-board service + inflight wifi service + online support + Ease of Online booking+Baggage handling) + 0.2* 0.167*(Seat Comfort+ Departure/Arrival time convenient+Gate location + Leg room service + Cleanliness + Online boarding)*

3.2.3. Assumption about cost of improving overall satisfaction

When trying to improve customers' overall satisfaction rate, the airline company will need to spend its budget on many aspects. We hope to get the data about how much it needs to upgrade the customer service when customer overall satisfaction improves by 1 point (1 to 5).

According to research [9], food and drinks, staff costs, and basic amenities are easily measurable for each customer. And these factors contribute to customer satisfaction a lot. Hence, we take those factors into consideration, and build the model at the following (although we are aware that this cost and satisfaction formula is not ideal, it could still be easily modified for actual use):

*Cost = 100 * (△ a change in the value of a overall satisfaction)*

3.2.4. Divide into train data and test data

We divide the dataset into 70% training data and 30% testing data.

IV. MODEL BUILDING

4.1 Model Objectives

Assumption:	If a customer is unsatisfied, he/she will churn
Uncertainty:	Satisfied (S) or not satisfied (NS)
Decision:	Improve overall subjective satisfaction (OSS)
Benefit:	$V\ (X \mid OSS)$
Cost:	$C\ (X \mid OSS)$
Goal:	If we improve each individual customers' satisfaction, how much profit we will expect to gain from the customer.

$$E\,[Profit \mid X,\ OSS]\ =\ P\,(S \mid X, OSS)\ *\ (\,E\,[\,V\,(X,\ S) \mid X,\ OSS,\ S\,]\ -\ E\,[\,C\,(X,\ S) \mid X,\ OSS,\ S\,]\,)$$

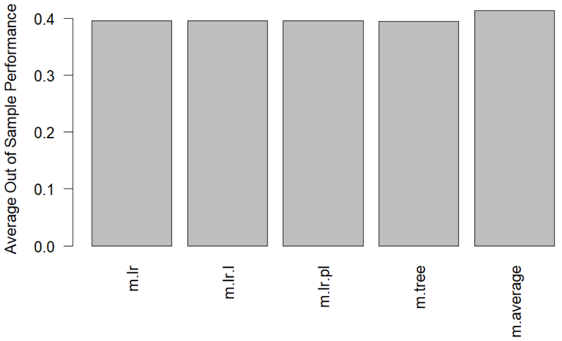
4.2 Build Model

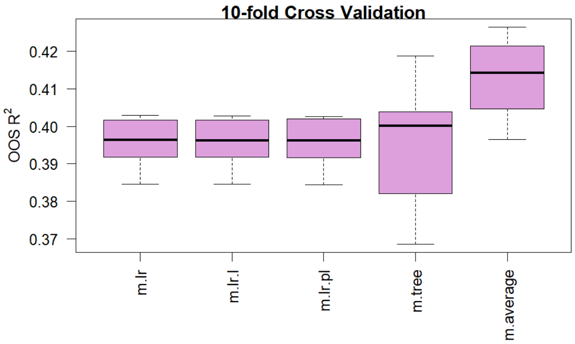
As the dependent variable is binary, we built the following models that airline businesses can employ to predict passenger satisfaction: 1) Logical regression. 2) Lasso logistic regression 3) Post-lasso logistic regression 4) Tree of classification We trained all our models on training data with 10-fold cross-validation. To validate our model, we applied R^2 to data not included in the training sample.

V. EVALUATION

5.1 Evaluate Results

Through the 10 fold cross-validation, the logistic model has the highest OOS R^2.





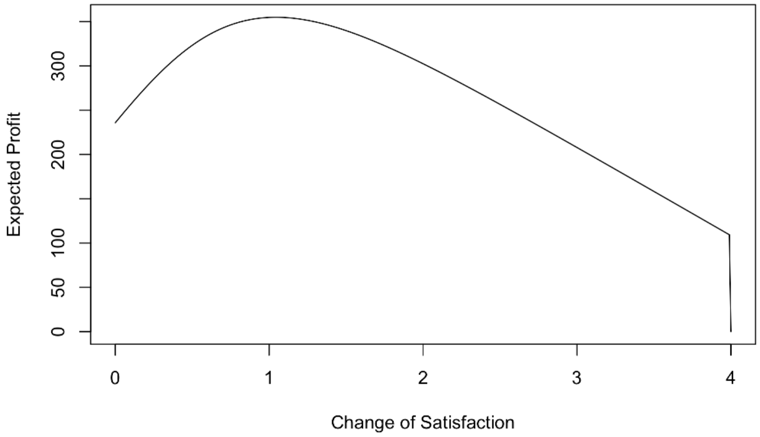
5.2 Model Performance

We used logistic regression for prediction and acquired a corresponding confusion matrix (threshold = 0.5). We multiplied the cost-Benefit Matrix and confusion matrix to predict our expected profit, which is around 1449330$ compared to baseline profit 1265050$ [Appendix 2]. From the profit curve [Appendix 3], we should target 60% of customers to maximize the profit. All other performance curves can be accessed through [Appendix 3].

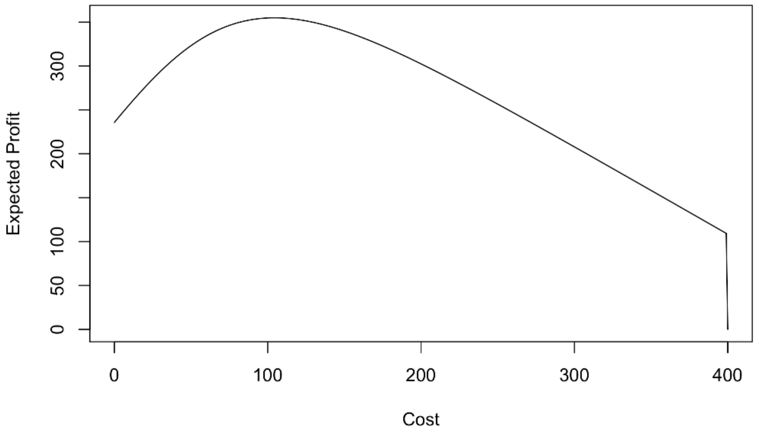
VI. DEPLOYMENT [Answer Business Objective]

Personalization can assist airlines in enhancing customer experiences and customer loyalty, hence increasing revenue. Therefore, we want to help companies make better decisions to increase each customer's satisfaction and achieve their maximum expected profit. We use the test data to predict each customer's probability of satisfaction by increasing overall satisfaction and find out each customer's CLV to calculate the expected profit of each customer.

Change of satisfaction on expected profit



Cost associated with change of satisfaction on expected profit



For example, for this customer [index = 120], if we increase his/her overall satisfaction by 1, we can get the maximum expected profit ($350) from that customer. And we could convert the overall satisfaction improvement to the cost, which the airline should spend on increasing the overall satisfaction [Assumption 3.2.3-cost function]. If we want to increase the overall satisfaction by 1, the airline needs to spend about $100 on this customer to his/her satisfaction. Based on our research, we find that the factors of food and drink and inflight entertainment impact airline customer satisfaction the most. We build a profit-cost curve specifically for each customer. So that airline companies can provide customized service plans for each customer, such as giving discounts, improving food and drink service, providing more inflight entertainment (movies, games), upgrading online booking systems and so on.

Although airline companies can provide customized service for customers to increase each customer's overall satisfaction, there are also many other factors that can impact on customers' overall satisfaction, such as seat comfort and inflight wifi, which airlines need to improve the overall facilities. It is difficult to measure the overall cost of a whole facility upgrade for a specific customer. We need more information and data on the cost of improving each airline's services to get a more precise suggestion on which airline can improve.

TABLE OF CONTENT

VII. REFERENCE

[1] In-Flight Services—Not Ticket Prices—Boost Passenger Satisfaction on International Flights, J.D. Power Finds 2019 Airline International Destination Satisfaction Study | J.D. Power (jdpower.com)

[2] *North American Airline Passenger Satisfaction Declines: Here's Why That's Good News, Says J.D. Power* *2022 North America Airline Satisfaction Study | J.D. Power (jdpower.com)*

[3] *The True Cost of Poor Customer Service to Your Business* The True Cost of Poor Customer Service to Your Business | Midlands Technical College

Based on this research[3]: Today as always, customers remember and communicate bad experiences far more than they communicate good. It is bad enough that approximately 50% of customers will never do business with a company again based on a single bad experience. However, the research shows that roughly 95% of unhappy customers will tell at least one person about their experience.

[4] Calculating the Cost of a Better Airline Cabin for All Calculating the Cost of a Better Airline Cabin for All (skift.com)

[5] A better approach to airline costs A better approach to airline costs | McKinsey

[6] How much fuel do aircraft need? How Much Does It Cost To Fuel A Commercial Airliner? (simpleflying.com)

[7] Airline fare analysis: comparing cost per mile Airline fare analysis: comparing cost per mile - Rome2rio

[8] The Cost Of Flying: What Airlines Have To Pay To Get You In The Air The Cost Of Flying: What Airlines Have To Pay To Get You In The Air (simpleflying.com)

VIII. APPENDIX

Appendix 1. Variable Description:

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 or dissatisfaction)

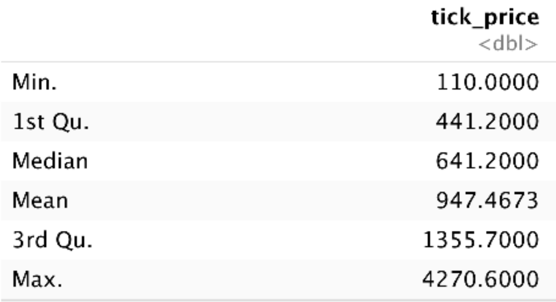
Appendix 2. Cost Benefit and Confusion Matrix

Cost Benefit	Y=1(P)	Y=0(N)	**Confusion (prediction)**	Y = 1(P)	Y= 0(P)	**Confusion (Base Line)**	Y=1(P)	Y=0(P)
Y^ = 1	150	-100	Y^ = 1	16200	4021	Y^ = 1	19371	16406
Y^ = 0	130	-80	Y^ = 0	3171	12835	Y^ = 0	0	0

$$Expected\ profit\ =\ 150\ *\ 16200\ +\ 130\ *\ 3171\ -\ 100\ *\ 4021\ -\ 80\ *\ 12385\ =\ 1449330$$

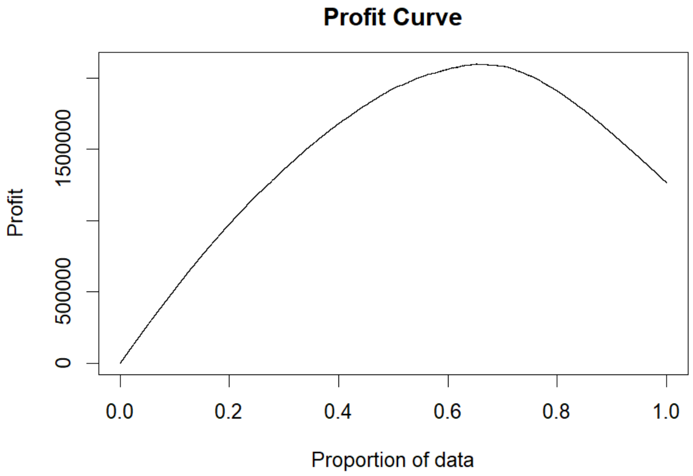
$$Baseline\ Profit\ =\ 150\ *\ 19371\ -\ 100\ *\ 16406\ =\ 1265050$$

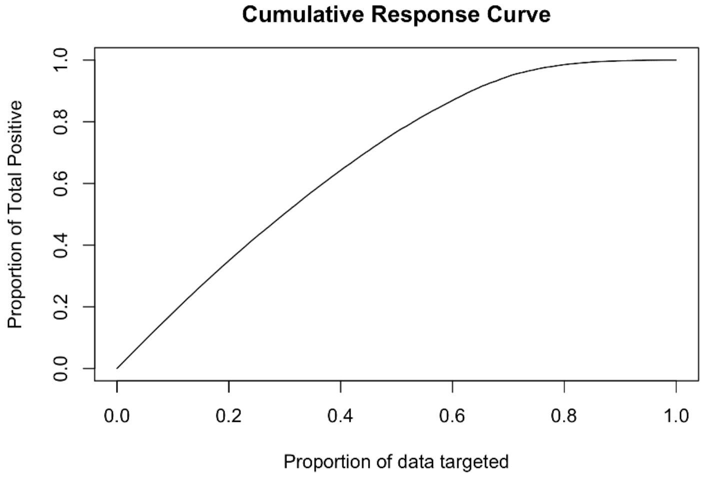
The logic behind the cost benefit matrix is calculated around the 15% of average ticket price.

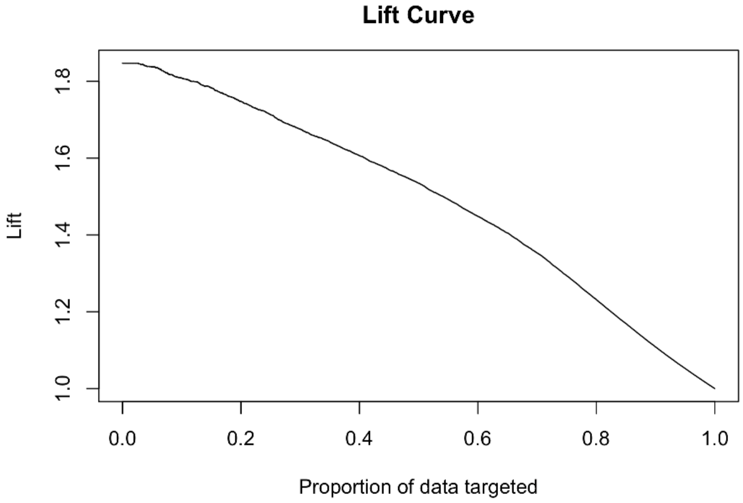


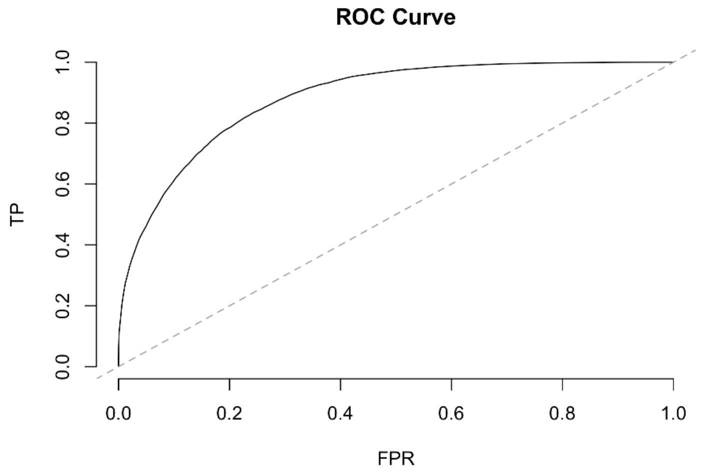
	tick_price <dbl>
Min.	110.0000
1st Qu.	441.2000
Median	641.2000
Mean	947.4673
3rd Qu.	1355.7000
Max.	4270.6000

Appendix 3. Profit, ROC, Cumulative Response, and Lift Curve









Appendix 4. Contribution of each member

Jackson: built model and wrote report

Yun Hsuan: made data visualizations and built deployment model

Fangqing: researched assumption and made powerpoint

Rosalie: organized R code and wrote report

Charuta: made powerpoint and did data transformation