STAT451 Project Final Report: Predicting Airline Passenger Satisfaction with Machine Learning

Kaiyan Ma kma57@wisc.edu
Wai Tsun Chan wchan34@wisc.edu
Yiyi Zhao yzhao388@wisc.edu

Abstract

In the modern society, flight has become a more common and popular transportation to people. As the airline industry rapidly expand its size over the past years, the question of how to satisfy the passengers has been posted to the airline companies in order to keep their own customers. This project intends to create machine learning models that can accurately predict if an airline passenger feel satisfaction on their flight experience. Using the "Airline Passenger Satisfaction" data set from Kaggle, we explore these 100000 cases and 22 features data by fitting it onto ensemble method machine learning models. For the purpose of this project, we used stacking method and bootstrap methods. Also, we tested the features for feature importance. This test provides a basic idea of what features are important for predicting the passenger's satisfaction. Besides, we created three models with different features in order to explore the effect of objective facts have on passenger's satisfaction, and the effect of subjective values have on passenger's satisfaction. We ranked the models By evaluating the performance of different models and methods. This projects also have interesting findings like gender and age are important features when predicting passenger satisfaction by subjective rating by customers, but less important when predicting passenger satisfaction by objective facts like how many minutes delayed on departure.

1. Introduction

Thanks to transportation, we are living in a smaller world than the people who live in the past, but we also depend on transportation in our daily life, which transportation has become an essential part of our lives. Undeniably, in compare to the past airline industry, flight becomes safer, more convenience, and providing more services and entertainments than it used to be. Nowadays, people no longer only take long distance plane rides as those are almost inevitable, more people preferring flight than drive even in short distance domestic travels. The number of flights worldwide over a year was 23.8 million in 2004, but the number has increased to 40.3 million in 2019 before the COVID-19 pandemic. This expansion makes airline industry become more competitive than years ago, among all the dazzling choices of airline companies nowadays, a factor affecting customer's decision in choosing an airline would be satisfaction. Airline satisfaction provides a significant clue from a customer's point-of-view to the companies on what matters the most.

Factors that affect a passenger to decide a flight experience is satisfied or not can be both subjective and objective. Overall, customer satisfaction reflects the personal experience of flying with an airline company by a subjective opinion. It values every details from the check-in process to baggage claim, the customer service, or even the slightest details of food quality in-flight. And we can further group those factors into two categories, subjective values and objective facts. How would these factors affect passenger's satisfaction all together? and how important are the subjective values and objective facts in affecting passenger's satisfaction separately?

In this project, we use the 100000 cases and 22 features Airline Passenger Satisfaction data set to discover the key factors of how a customer decide their flight experience is satisfied or not. These features are including the satisfaction level of different services that an airline company provides, which rated by the passengers. These services include but are not limited to check-in service, seat comfort, in-flight entertainment, and baggage handling. Other elements such as departure delay, arrival delay, and customer's travel information are also a part of the analysis. Our intention is to build three models to predict if a customer is satisfied with their overall experience or not based on various detailed services. Moreover, we can evaluate the objective factors such as flight class, delayed minutes in departure, delayed minutes in arrival in a model, as these factors are not based on the passengers' own value judgement. Also, we can evaluate the subjective opinions such as rating of food and drink, rating of seat comfort, and rating of in-flight service in another model.

As mentioned above, our goal is to create a machine learning model to predict the overall satisfaction. We want

to identify and understand the most significant elements that affect the overall satisfaction, and we are expecting the predictability of three models would be different.

2. Related Work

There is a need to examine the influence of airline service quality on passenger satisfaction because it holds significant importance in customer loyalty which lies particularly close to the profit.

Customer satisfaction generally means customer reaction to the state of fulfillment and customer judgment of the fulfilled state [7]. It also functions as an antecedent of customer loyalty because it prevents churn and consolidates retention, constituting a solid foundation of customer loyalty [4]. For those customers with high user engagement and loyalty, repetitive same-brand purchasing despite situational influences is more likely to happen. The larger the loyal customer group they have, the steadier revenue will be. Earlier studies also suggest that customer loyalty provides the foundation of a company's sustained competitive edge as well as companies' growth and performance [5]. The level of service to the high-satisfaction customer embodies the whole strategic orientation of airlines; therefore, it is a crucial factor for airline companies to think about.

Aviation is a fast-moving, dynamic, and ever-changing industry requiring airlines to innovate and pre-empt events to remain competitive [9]. Since nowadays passengers have multiple options to choose from, they may judge or evaluate airline service quality by comparing their experiences and expectations over many quality attributes. If the passenger is not satisfied with the quality of service, they will reconsider the buying decision for further flights and will probably switch to another airline [3]. Thus, the delivery of high-quality service becomes a marketing requirement among air carriers due to competitive pressure [8]. That is to say, figuring out what affects customer satisfaction will guide airlines to invest extra funding for some critical aspects to improve their competitiveness.

A Related study, by Juliet Namukasa from 2013 [6], has shown that the quality of pre-flight, in-flight, and post-flight services strongly influence the satisfaction level. The same study also concluded that customer satisfaction varies based on each customer's own will, and different service strategies could take place based on the demographic characteristics of customers.

Moreover, by classifying out those passengers who are likely to have high-level satisfaction, airlines can make exact adjustments to balance the input and outcome to maximize the profit when launching new services. Relating to An and Noh's study, in-flight service quality are important according to the customer seat class based on their satisfaction level [2]. The conclusion also draws attention to the fact that airline companies' in-flight service should have different delivery strategies based on the customer seat class.

All in all, it is significant to the airline companies to understand their service quality. It is also important to understand customers' ideal flight experience in order to adjust their current services or promote new services. Our study could provide proves and suggestions to these companies on which types of service have high correlation with customers' overall satisfaction level. It could also predict customers' overall satisfaction level base on the rating on each type of service. These result could be beneficial to these airline companies to regulate their flight services.

3. Proposed Method

In this project, our group decided to use three ensemble methods, and rank three methods by the test accuracy. Given a zip file of two data sets provided by TJ Klein on Kaggle [10], it contains 103903 cases on the train.csv, and 25975 cases on the test.csv. Our group wants to spilt the data by our own instead of using the pre-spilt data set, so we manually combine two data sets to a full data set, and named it airline.

After we obtain the full data set, we would do data cleaning before using the data set on machine learning models. For the missing values, we decided to drop all the rows that contain missing response since we have a large data set. For the variables that contain character values, as some of the machine learning models would not be able to use character responses, we decided to re-code the responses of those variables to 0,1 or 0,1,2. We will use the label "satisfaction" which come with the data set to evaluate passenger satisfaction, it is a binary variable and we will label "satisfied" as 1 while "neutral or dissatisfied" as 0.

A Logistic regressions will be following the data cleaning, we will regress all the features on the satisfaction variable, to get an estimate and p-value of all the features. As we are estimating a binary label being 0 or 1. The formula is:

$$P(y=0) = \frac{1}{1+\exp(\beta^\top x)},$$

and for each new β, the sample estimate would be calculated by a log-likelihood equation:

$$\begin{aligned} &-\log L(\beta) \\ &= -\log(\prod_{i=1}^{N} P_i^{1-Y_i}(1-P_i)^{Y_i}) \\ &= -\sum_{i=1}^{N}(Y_i\beta^\top x_i - \log(1+\exp(\beta^\top x_i))), \end{aligned}$$

These estimates would tell us the effect of a feature on passenger satisfaction, and the p-value would tell us how significant it is. We would be using all the features that is significant on predicting satisfaction of passengers.

Our interest of how important are the subjective values and objective facts in affecting passenger's satisfaction separately will be answer by splitting the full data set onto two subsets. We would be based on the responses of the variable, to split it into objective subset if the responses of the variable are fact that does not contain subjective value. On the other hand, if the responses of the variable contain any subjective values, the variable would be split into subjective data set. We would also fit the machine learning models with these subsets to see the test accuracy, but the main focus of this project would be the full model.

When all three data sets are ready, we split the data set into training, and testing set by using the train test split method from the sci-kit learn package. By setting the random state equal to the same value, the test set and the train set would be the same through out the model fittings.

The ensemble models we will implement are stacking, XGBoost, and Bootstrap. Each method would fit the full data set one time, and our project would be consider "successful" if our best model among these three can "accurately predict" the binary outcome "satisfaction." Before we fit the the full data set with machine learning methods, we are expecting a test accuracy of over 80 percents, and we consider this model would consider success if the test accuracy could be over 90 percent. We are having such a high expectation as the features on this data set are mostly subjective ratings which are similar to how passengers rate satisfaction.

In addition to our project main focus, we are looking forward to expose the relationships between some of the features and passenger satisfaction when we work through the machine learning process; and also how important are the subjective values and objective facts in affecting passenger's satisfaction. By using correlation test on filtering the features, we should be able to explore some correlations between feature and label. Besides, by fitting the subsets with the bootstrap models, we should be able to compare the test accuracy between different data sets, we are expecting the test accuracy of objective model and subjective model would be different.

For the conclusion, we would rank the methods in predicting the full data set. Also, we would be discussing the relationships between some of the features and passenger satisfaction; and how important are the subjective values and objective facts in affecting passenger's satisfaction if plausible.

Stacking involves making predictions from multiple machine learning models and getting comprehensive results. We first built base estimators based on Random Forest and Gradient Boosting, then inserted the results into our stacking classifier, enabling us to take the most trustful machine learning model. The Scikit-learn classifier is used during this process. We also adopted hyperparameter optimization to improve our staking model, using RaneddomSearchCV. The principle of RaneddomSearchCV is straightforward; the program will try each set of super-parameters and then choose the best group. Sometimes this can be time-consuming and face a dimensional disaster. While with cross-validation, RandomizedSearchCV can be more useful since we have multiple parameters in our airline passenger satisfaction dataset.

For XGBoost, XGBoost stands for "Extreme Gradient Boosting" which is a supervised-learning algorithm that uses the training data to predict a target variable [1]. It is based on gradient boosting which is an ensemble method that takes multiple weak learners with low predictive performance to form a stronger learner with high predictive performance. In this experiment, XGboost is used to create models to find the feature importance score of each feature. The objective function for XGBoost is

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

For bootstrap, it is a method which it resamples the data and replace with a sample set using the same sample size.

4. Experiments

In order to answer the research question of this project, we used logistic regressions on the the data set to test whether a feature is significant in predicting the satisfaction of passengers. And the ensemble methods of XGBoost, bootstrap, and stacking were used in a similar manner to evaluate the capability of each method in predicting an accurate result on airline passenger satisfaction.

For this project, we used the regression model to make two extra data sets along with the full data set, one contains subjective features, another contains objective features. We will test all three data sets with three methods, but we will mainly focus on the full data set.

4.1. Data Set

In this project, we will use the "Airline Passenger Satisfaction" data set. This data set is provided by TJ Klein on Kaggle [10], originally split into two sets by the author, named "test" and "train." The train.csv data set contains 103903 rows with 25 columns; the test.csv data set contains 25975 rows with the same columns. We use the row bind function in r to combine two sets into one set named "airline," it has 129880 passengers' survey response. This data set contains 25 unique features; one is the index, seven are the general information about the flight, sixteen are the subcategories' satisfaction rating (on the scale 0 to 5), and the last column is the overall satisfaction rating (on the scale 0 to 5).

4.2. Cleaning Data

In order to process with different machine learning methods, we cleaned and reformed the data set. First, we removed some columns that should not have any correlation to the satisfaction result, such as id. We re-coded the non-interval data with interval variables accordingly, including Gender, Satisfaction, Customer type, Type of Travel, and Class. For Gender; male is set to 0 and female is set to 1. For Satisfaction, neutral or dissatisfied is set to 0 and satisfied is set to 1. For Customer type, we set Loyal Customer to 0 and disloyal Customer to 1. For Type of Travel, we set Business travel to 0 and Personal Travel to 1. For Class, we set Business to 0, Eco Plus to 1, and Eco to 2.

After the re-coding, we found out that the variable Arrival Delayed in Minutes has 393 missing responses, so we decided to drops all the entire rows that contain missing values as we have a large data set. After dropping the missing values, the final data set contains 129487 unique responses. And this data set is ready to run the linear regression on.

Besides, we standardized all those features by subtracting the mean and scaling to unit variance since we are comparing measurements with different units.

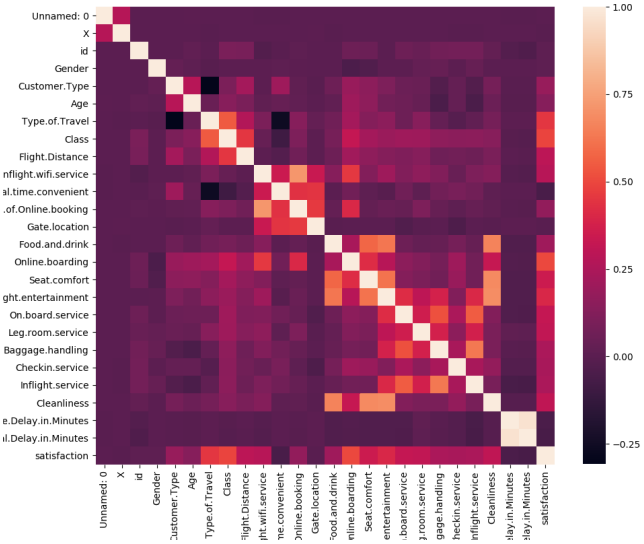

Figure 1. Correlation Plot

4.3. Logistic Regression

By using the glm function in R-studio, we regress all 22 variables on Satisfaction. From the result, we find the the p-value of Flight Distance variable is not significant, so we decided to exclude this variable from the regression and model. After, all the variables are significant on predicting the Satisfaction, part of the result is shown in table 1.

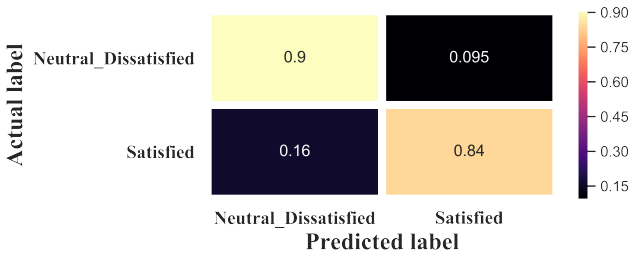

Figure 2. Confusion Matrix

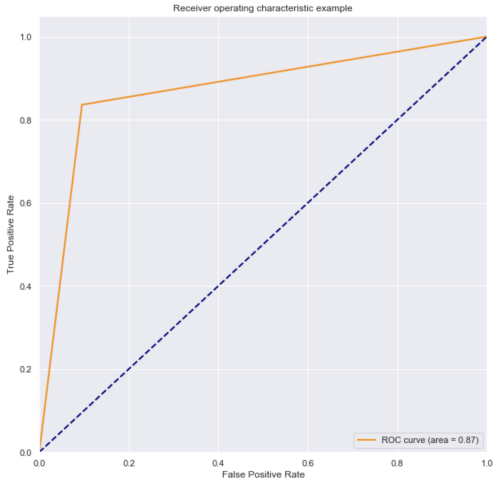

Figure 3. ROC Plot

4.4. Data Subsets and Logistic Regression

To answer the question of what is the effect of objective/ subjective variables on passenger's satisfaction, we further fitted two regression model. The objective model regresses 7 variables (Gender, Customer Type, Age, Type of Travel, Class, Departure Delay in Minutes, Arrival Delay in Minutes) on satisfaction. The subjective model regresses 14 variables (In-flight WiFi service, Departure Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, In-flight entertainment, On board service, Leg room service, Baggage handling, Check-in service, In-flight service, Cleanliness) on satisfaction. We found that gender and age is not significant on the objective model but it is significant on the subjective model, so we decided to create two subsets and place gender and age on the subjective model instead (we will further disscuss. Then, we have a subjective subset with 17 variables, and an objective subset with 6 variables. We will also run the machine learning methods on these sub models to get the test accuracy, but this project still mainly focus on the full model.

Features	P-value
Gender	$0.000244***$
Customer.Type	$< 2e-16***$
Class.Eco	$< 2e-16***$
Leg.room.service	$< 2e-16***$
Arrival.Delay.in.Minutes	$< 2e-16***$

Table 1. Part of the significant features in the full model.

4.5. Software

For this project, the software that we mainly used are Python, Jupyter notebook, Jupyter lab, and R studio; and the packages we mainly used are Numpy, Scikit-Learn, Pandas, and XGBoost.

4.6. Hardware

All the team members mainly used our own laptop(s) as the only hardware(s) for this project.

5. Results and Discussion

5.1. Stacking

We used the default 5-fold cross validation and the accuracy of the best stacking estimator are shown in the following table. After hyperparameter optimization, the best result achieved 96.71% accuracy on the test set. The parameters of staking are separately attached above in Table 2 and Table 3.

Methods	Seed	Parameters
Random Forest	123	-
Gradient Boosting	123	$max.depth$=8 $reg.alpha$=5.77e-05 $use.label.encoder$=False $verbosity$=0
Logistic Regression	123	$use.probas$=True
$drop.proba.col$=last	-	-

Table 2. Stacking Methods and Parameters

Training Accuracy:	0.9929
Test Accuracy:	0.9671

Table 3. Accuracy with the Best Hyperparameters

5.2. XGBoost

We built three XGBoost models with selected 22 columns from the dataset and the objective/subjective data subsets. With the full 22 columns dataset model, this model obtains a test accuracy of 96.14%. We are also able to examined the importance of each feature as shown in Figure 4. feature importance score represents how each feature is in use for constructing decision trees inside the model. The higher the score is, the more important the feature is. From figure..., we can identify the highest feature score in this model is on "Arrival Delays (in minutes)". Followed by "Departure Delays (in minutes)" and "Cleanliness", these three features are significantly important with airline satisfaction level.

Training Accuracy:	0.9754
Test Accuracy:	0.9614

Table 4. Accuracy for Full Dataset Model

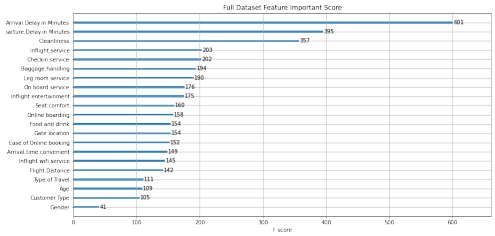

Figure 4. Full Dataset Feature Important Score

Accordingly, the model for objective data subset has a test accuracy of 79.07% and the model for subjective data subset has a test accuracy of 94.86%. As for objective data subset model, "Arrival Delays (in minutes)" and "Departure Delays (in minutes)" still remains the highest feature score, and appears to have more significance as objective features (Figure 5). Meanwhile, the feature score from "In-flight Service" and "Cleanliness" increases in the subjective data subset model comparing with the overall dataset (Figure 6).

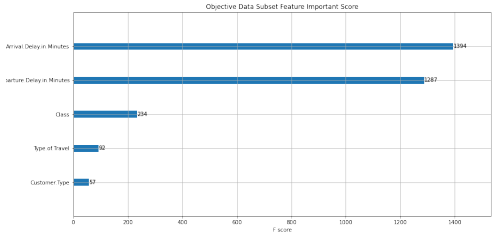

Figure 5. Objective Data Subset Feature Important Score

Even though these four features mentioned above are the top four important features in the overall dataset model, the observation of the increase in subjective data subset model is interesting, especially for feature "In-flight Service". In addition, other features also seem to be rational in the models. We are able to draw a conclusion that "Arrival Delays (in minutes)", "Departure Delays (in minutes)", "Cleanliness", and "In-flight Service" are the most significant fea-

tures relating to airline overall satisfaction rate.

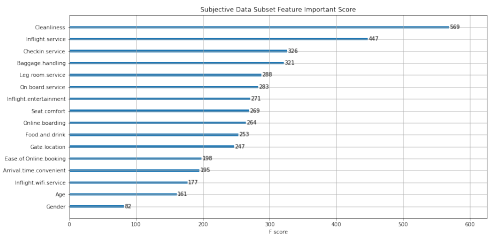

Figure 6. Subjective Data Subset Feature Important Score

5.3. Bootstrap

Data set	Method	Accuracy	Parameters
Full	oob	0.9472	-
Full	.632	0.9538	$max.depth$=15
Objective	oob	0.7879	$criterion$=entropy
Objective	.632	0.7879	$n.split$=50
Subjective	oob	0.9286	$estimator$=tree
Subjective	.632	0.9367	$random.seed$=1

Table 5. Test accuracy of different data set by Bootstrap method

We used the out-of-bag(oob) method and the .632 method of bootstrap on all three data sets. From table 5, we can see that there are differences in test accuracy between methods and data set. The machine learning model performs better with the full data set, than with the subjective data set, the subjective data set ranked the last. Also, within each method, we can see that the .632 method performs better than the oob method. In this bootstrap test, we do not include the .632+ method as the data set is too large.

By this table, we may be able to answer the extra question that this project has. Even though the logistic regression model on the full data set, objective subset, and the subjective subset have all the estimate significant on predicting passenger satisfaction. However, the test accuracy of the models using objective data set are obviously having a lower test accuracy than other data sets.

6. Conclusions

A high level of customer gratification is an important marketing goal for every brand and the key to success. In this report, we highlighted the performance of the airline companies and the customers' satisfaction grades. Specifically, we used an airline satisfaction dataset to identify the relationship between service features and the overall passenger satisfaction while predicting the full dataset with stacking and XGBoost models. We are able to find the features which are most significant to the overall passenger satisfaction level through XGBoost method. They are "Arrival Delays (in minutes)", "Departure Delays (in minutes)", "Cleanliness", and "In-flight Service". In addition, the highest accuracy on the test set in the full dataset is from stacking with 5-fold validation. We also encountered some interesting findings. For example, we identified that the "Flight Distance" feature is not as significant as other features through its p-value from logistic regression analysis. We also found a relationship between objective ans subjective features. Some features, such as "In-flight Service", did not show its importance in the full dataset model but rather its significance increases in the data subset.

Overall, in order to answer our research purpose, the result show that the stacking method has the highest test accuracy 0.9671, XGBoost has the second highest with a test accuracy of 0.9614, .632 Bootstrap has the lowest among these methods, but still having a test accuracy of 0.9538. As all the methods have a high test accuracy and over our expectation of 0.9, we think that we have created a model that can accurate predict the passenger satisfaction. Besides, for our interest of comparing the objective facts and subjective values on effecting the passenger satisfaction; I think there are clues that suggesting we are more difficult to use objective facts to predict the passenger satisfaction. I think it because of the passenger satisfaction is based on people subjective value.

7. Acknowledgements

We would like to acknowledge Professor Raschka for providing support and guidance. We learned various extra knowledge about machine learning and practical applications about classifiers, which will be very helpful for our further studies. And we are also thankful for all the TAs' hard work.

8. Contributions

Each group member from this group contributed similar amount of work on this project. For data analysis, Kaiyan was responsible for the stacking method, Yiyi was responsible for the XGBoost method, and Wai Tsun was responsible for the bootstrap method. For the written portion of the report, each member contributed to their own method. In addition, Wai Tsun focused on the abstract, introduction, and experiment data description. He did much of the main work for the proposed method and raised plenty of new ideas; Yiyi focused on related work, proposed method-XGBoost, software/hardware, result, and conclusion. She also did a lot of graph creation for feature importance and the final conclusions; Kaiyan focused on experiment, result, and conclusion.She was responsible for fitting and tuning the staking models as well as coding up some of the graphs in logistic regression. Both Kaiyan and Yiyi worked on the figures and tables in the report.

Kaiyan was responsible for fitting and tuning the staking models as well as coding up some of the graphs and tables in the data processing and logistic regression. She also helped write for each of the writing sub-sections.

References

[1] Introduction to boosted trees.

[2] M. An and Y. Noh. Airline customer satisfaction and loyalty: impact of in-flight service quality. *Service Business*, 3(3):293–307, 2009.

[3] R. Archana and M. Subha. A study on service quality and passenger satisfaction on indian airlines. *International Journal of Multidisciplinary Research*, 2(2):50–63, 2012.

[4] C. Fornell. A national customer satisfaction barometer: The swedish experience. *Journal of marketing*, 56(1):6–21, 1992.

[5] M. Lee and L. F. Cunningham. A cost/benefit approach to understanding service loyalty. *Journal of services Marketing*, 2001.

[6] J. Namukasa. The influence of airline service quality on passenger satisfaction and loyalty: The case of uganda airline industry. *The TQM Journal*, 2013.

[7] R. L. Oliver, R. T. Rust, and S. Varki. Customer delight: foundations, findings, and managerial insight. *Journal of retailing*, 73(3):311–336, 1997.

[8] P. L. Ostrowski, T. V. O'Brien, and G. L. Gordon. Service quality and customer loyalty in the commercial airline industry. *Journal of travel research*, 32(2):16–24, 1993.

[9] C. Raynes and K. W. H. Tsui. Review of airline-within-airline strategy: Case studies of the singapore airlines group and qantas group. *Case studies on transport policy*, 7(1):150–165, 2019.

[10] K. TJ. Airline passenger satisfaction. *Kaggle*, 2019.