# Table of Content

# 1. Introduction

The project focuses on analyzing airline passenger satisfaction through multiple machine- learning methods, including to identify what factors lead to customer contentment for an airline, to select which model can generate the best performance, and to deliver business strategies that can help an airline maintain or even improve customer retention rate. The whole paper can be roughly divided into five parts: (1) Statement of problem and literature review; (2) Data description and data pre-processing; (3) Modeling and model comparison; (4) Customer satisfaction strategies; (5) Conclusion and future studies. Detailed discussions on algorithms and deep insights will be found in the third and fourth parts.

# 2. Problem Statement

Cross all the industries, customer satisfaction plays an essential role for the sustainable growth and development of a company. Good customer relationship can bring direct and indirect benefits for an organization. The impact of high customer satisfaction mainly reflects on price premium, word-of-mouth, low operation costs, high turnover rate, and strong customer loyalty. According the "Pareto's Principle[1]" in economics, the main profit of a company is only in the hands of some consumers, and customer loyalty is the main source of corporate profits. Through an academic research, if the annual customer relationship retention rate is increased by 5%, the total profits of the organization may increase up to 85%, and 60% of the company's new customers are related to existing customers (Li, 2010). Therefore, companies usually utilize the customer contentment surveys/questionnaires to dig out how current customers think about their products and services to enhance brand image and, at the same time, gain competitive advantages (García, 2019).

As for the airline industry, referring to the Porter's Five Forces[2] analysis, although the bargaining power of buyers is moderate, the overall rivalry is quite high. The switching costs for common customers are relatively low and it is easy for a customer to purchase tickets from another dominant player in the industry, particularly in this "Internet+" era. Hence, airline companies scramble to establish membership scheme and improve traveling experience to strengthen customer loyalty. Meanwhile, prices also become one of the important points of acquiring customers and market share, which is always the "magic weapon" of low-cost airlines (i.e., Jetstar, HK Express). As the number of airline brands increases, the competition has been intensified, which means new effective strategies should be taken for winning.

Under these circumstances, we think it is valuable and interesting to analyze factors that can influence customer contentment, do a relatively accurate prediction via various machine learning models and provide useful recommendations.

# 3. Literature Review

In this project, a total of four published literature is reviewed about the essential factors and applied machine learning methods in the customer satisfaction analysis in the airline industry.

Saadat et al. focused on doing a quantitative research about the impact of service strategy on customer

[1] Pareto's Principle points out that only about 20% of variables control 80% of the situation.
[2] Porter's Five Forces is a framework for analyzing a company's competitive environment and the industry situation.

satisfaction in AirAsia Malaysia in their paper published in 2018. They first collected responses from 111 current customers and identified five independent features (including services provided by flight attendants, tangible features, food service, online service and ground staff) that may lead to different levels of customer satisfaction. Then, using SPSS software, they did a multiple linear regression to test hypothesis and found that food service and ground staff are two most significant factors. They finally concluded that AirAsia Malaysia should focus more on these two areas in their future operations, and in the future research, they should consider using a larger dataset to generate a more accurate finding.

In a 2019 published paper, Baswardono et al. did a comparative analysis of decision tree algorithms between random forests and C4.5. The researchers chose a US Airline Passenger Satisfaction dataset (129,881 records and 23 attributes) from Kaggle.com site, and under the condition of ten-fold cross-validation, they applied the two models on the dataset. Furthermore, the researchers also tried different splitting ratio (70:30, 80:20, 90:10). Multiple performance measurements are adopted including Accuracy, Class Error, Precision, Recall and AUC (Area Under Curve). As a result, the two models perform similarly under most criteria, but overall speaking, random forests model has a better performance.

Garcia predicted airline customer satisfaction in the published literature in 2019 using BAGGING (as a representative of ensemble models) and KNN algorithm (as the base learner). In the dataset, the target y variable, “Satisfaction” is a five-point score measurement (from 1 to 5). Therefore, the researcher identified this as a regression problem instead of a classification one, and used average RMSE and average MAE (five-fold cross-validation) as performance measurements. It shows that the ensemble regression model generates the best result.

In 2020, Hong et al. established machine learning models to analyze customer evaluation data for airlines. The applied models include Support Vector Machine (SVM), Random Forests (RF) and Deep Neural Network (DNN), with RF generates the highest test accuracy of 95.7%. The researchers then base on RF model to get top eight and top four important features. In the future studies, they plan to involve domestic airlines data and at the same time contain more management and marketing strategies.

## 4. Data Description & Data Pre-processing

To explore deep into the above-mentioned problem, our group used a dataset downloaded from Kaggle. The dataset captures the records of an airline passenger satisfaction survey. In total, it contains 22 features, excluding “ID” and “satisfaction”, which is the response variable. These available features offer relevant information required for the prediction of customer satisfaction comprehensively by recording customer ratings on areas such as sea comfortable and inflight WIFI service. On the basis of these 22 predictors, we would like to predict whether a customer is satisfied with the airline’s services or not, which is reflected from the “satisfaction” column with binary results. On top of that, the original dataset is split into a training set and a test set; and each of them has 103903 rows and 25976 rows respectively, meaning that the train-test split is around 4:1.

To obtain a more in-depth understanding towards the training set, we generated a data analysis report with pandas profiling. Before putting data into the pandas profiling module, we first dropped out the first two columns, namely 'Unnamed: 0' and 'id', because they contain almost no useful information.

| Variable types | |
|---|---|
| Categorical | 6 |
| Numeric | 17 |

*Figure 1: Table summarizes variable types*

The report shows that, among the 23 variables including the response variable, only 6 are categorical, while the remaining 17 variables are numeric. The 6 categorical variables are 'Gender', 'Customer Type', 'Type of Travel', 'Class', 'Baggage handling', and 'satisfaction'. Since some machine learning models do not accept categorical variables as input, we used label encoder to convert each value in a categorical column from either train set and test set to a number. However, label encoding of a column may introduce a new relationship between the number: there is no hierarchy between the gender types, but when one examines the numeric labels, one might think "Male" take precedence over "Female". Nevertheless, label encoding technique allows us to better interpret the data through renaming "satisfied" customers into "1" and "neutral or dissatisfied" passengers into "0".

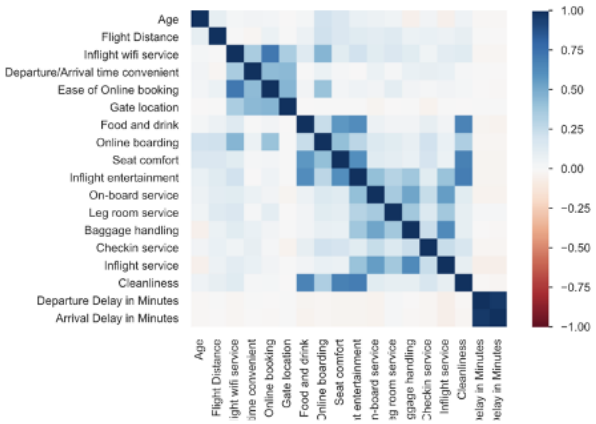



*Figure 2: correlation matrix*

From the correlation matrix in the figure 2, we can observe that the colour of the boxes linking "Departure Delay in Minutes" and "Arrival Delay in Minutes" is dark blue, indicating the presence of a strong positive correlation between these two variables. Highly correlated predictors introduce data multicollinearity issue which undermines the statistical power of the prediction models. Therefore, one considers removing one of the two variables before diving into the modelling part.

```
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Inflight wifi service                 0
Departure/Arrival time convenient     0
Ease of Online booking                0
Gate location                         0
Food and drink                        0
Online boarding                       0
Seat comfort                          0
Inflight entertainment                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Inflight service                      0
Cleanliness                           0
Departure Delay in Minutes            0
Arrival Delay in Minutes            310
satisfaction                          0
dtype: int64
```

```
Gender                               0
Customer Type                        0
Age                                  0
Type of Travel                       0
Class                                0
Flight Distance                      0
Inflight wifi service                0
Departure/Arrival time convenient    0
Ease of Online booking               0
Gate location                        0
Food and drink                       0
Online boarding                      0
Seat comfort                         0
Inflight entertainment               0
On-board service                     0
Leg room service                     0
Baggage handling                     0
Checkin service                      0
Inflight service                     0
Cleanliness                          0
Departure Delay in Minutes           0
Arrival Delay in Minutes            83
satisfaction                         0
dtype: int64
```

*Figure 3: Number of missing values in train set (left) and test set(right)*

Similar to the highly correlated predictors, missing data also can weaken the statistical power of a model and produce biased results. The figure 3 above display the number of missing records in both train set and test set; and we can notice that all the missing values are under the "Arrival Delay in Minutes" column. By combining this observation and the findings shown in the correlation matrix, we decided to drop the "Arrival Delay in Minutes" column to facilitate our data analytics process.

Though the presence of outliers would increase the variability of data, we decided not to drop outliers for this dataset due to the following two reasons: firstly, the high ratings and low ratings given by customers contains useful information, dropping out these extreme values will produce biased results; secondly, all records are of reasonable ranges, so we have reasons to believe that the dataset contains no typos.

After separating the response variable from the predictors in both train and test set, we used "pre-processing" model from sklearn to do the feature scaling for constructing logistic regression. We first used fit method to estimate the mean and standard deviation of each feature from the train set. By using this set of scaling parameters, we used to transform method to standardize both train set and test set. Feature scaling enables the optimal performance of a logistic regression model and prevents the number of iterations from reaching maximum limit.

# 5. Modeling

## 5.1 Model Comparison

- **Logistics Regression Model**

Logistic regression model the probability that Y belongs to a particular category, that is

$$Pr(Y|X)$$

For a given **X**, a prediction can be made.
The logistics model applies the logistics function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Which has several properties:

1. The output is between 0 and 1.
2. Doe low balances now we predict the probability of default as close to, but never below, zero
3. For high balance, we predict a default probability close to, but never above, one
4. Use "maximum likelihood" method to fit the function

We can reorder the equation to:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

The LSH of the equation is the odds, which can take on any value between 0 and infinity. Value of the odds close to 0 indicates very low probability.
We can take log on both sides of the equation:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

The left-hand-side of the equation is called Log-odds/Logit:
a. logistic regression has a logit that is linear in X.
Increasing X by one unit changes log odds by $\beta_1$/the odds is multiplies by $e^{\beta_1}$
b. The amount that p(X) changes due to one-unit change in X will depend on the current value of X.
c. If $\beta_1$ is positive then increasing X will be associated with increasing $p(X)$, if $\beta_1$ is negative then increasing X will be associated with decreasing p(X)
To estimate the regression coefficients, we can apply the maximum likelihood method:

1. Estimating $\beta_0, \beta_1$ such that plugging these estimates in the model for p(X) yields a number close to one for all individuals who defaulted, and a number close to zero for all individuals who did not.
2. Likelihood Function:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

3. $\beta_0, \beta_1$ are chosen to maximize the likelihood function

Result:

| Model | Training Accuracy |
|---|---|
| Logistic Regression | 0.8753945950107792 |

- **Random Forest Classifier**

1. At each split, a random sample of m predictors is chosen as split candidates from the full set of p predictors.
2. The split is allowed to use only one of those m predictors.
   * The number of predictors considered at each split is approximately equal to the square root of the total number of predictors→The algorithm is not even allowed to consider a majority of the available predictors
   *Decorrelating the tree: Making the average of the resulting trees less variable and hence more reliable
3. The main difference between bagging and random forests is the choice of predictor subset size m.
4. Using a small value of m in building a random forest will typically be helpful when we have a large number of predictors.

Result:

| Model | Training Accuracy |
|---|---|
| Random Forest Classifier | 0.9525234832152757 |

- **Boosting**

Works similar to bagging, except that trees are grown sequentially:

1. Each tree is grown using information from previously grown trees.
2. Boosting does not involve bootstrap sampling; instead, each tree is fit on a modified version of the original data set.

Idea behind boosting:
1. The boosting approach learns slowly→ We fit a decision tree to the residuals from the model, rather than the outcome Y, as the response.

a. Each tree can be quite small, with just a few terminal nodes, determined by the parameter d (splits)
b. The shrinkage parameter $\lambda$ slows the process down even further, allowing more and different shaped trees to attack the residuals
c. In boosting, the construction of each tree depends strongly on the trees that have already been grown.

Boosting Parameters:

1. Number of trees B Can result in overfitting if B is too large, although this overfitting tends to occur slowly→Use cross-validation to select B
2. Shrinkage parameter, $\lambda$(between 0.01 or 0.001)→Controls the rate at which boosting learns, very small $\lambda$ can require using a very large B in order to achieve good performance
3. The number of splits, d, in each tree: Controls the complexity of the boosted ensemble.
    a. D=1, each tree is a stump, consisting of a single split→The boosted ensemble is fitting an additive model, since each term involves only a single variable
    b. D is the interaction depth, and controls the interaction order of the boosted model, since d splits can involve at most d variables
4. Boosting vs. Random Forests:
    a. In boosting, because the growth of a particular tree takes into account the other trees that have already been grown, smaller trees are typically sufficient→ Aid the interpretability

Result

| Model | Training Accuracy |
|---|---|
| XGBoost | 0.9525234832152757 |

## 5.2 Performance Measurement

Generally, accuracy is not the best performance measure for classifiers. For different cases, people may consider that true positive or true negative rate is more important. The definition of some measurement we will use later are below:

- Precision: The fraction of relevant instances among the retrieved instances

$$Precision = \frac{True\ Positives}{True\ Positives + \ False\ Positives}$$

- Recall/Sensitivity/TPR: The fraction of the total amount of relevant instances that were retrieved.

$$Recall = \frac{True\ Positives}{True\ Positives + \ False\ Negatives}$$

- F1 Score: conveys the balance between the precision and the recall

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

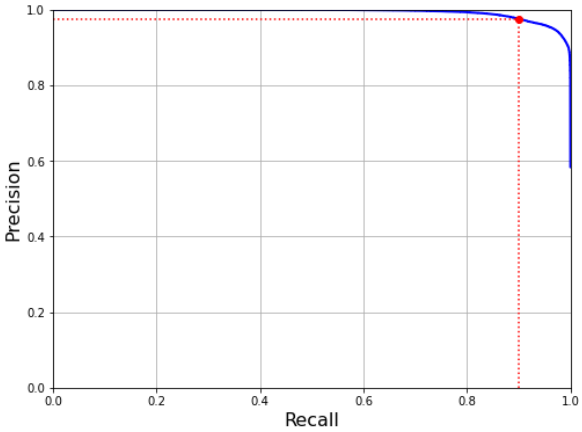
- Receiving Operating Characteristics (ROC)and Area Under the Curve (AUC):
    - ROC plots the TPR(Recall) against FPR where FPR is on the y-axis and FPR is on the x-axis.
    - AUC measures the degree or measure of separability; it tells how much model can distinguish between classes. The higher the AUC, the better the model is at predicting 0s and 1s.

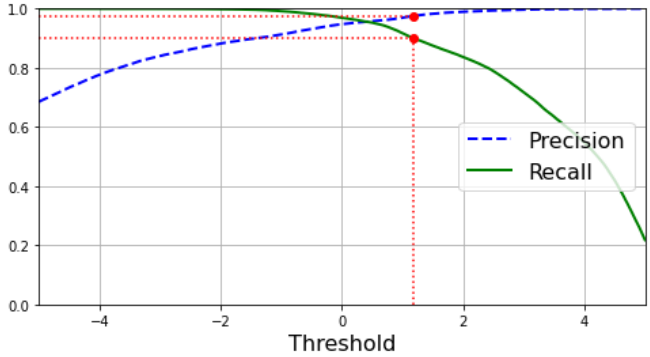
      In our case study, we will be utilizing all the metrics in choosing the best model for our binary classification problem.

In our case, there is no doubt that we need a good precision. We could then find out what the airlines can improve to lower the dissatisfaction rate. But at the same time, the recall should not be too low, or the airlines may be distracted by the irrelevant factors, spending a lot of money on these things but get little effect.

Our group will use cross-validation to get the precision and recall scores for different thresholds. As the sklearn does not support this plot, our group build our own function to plot the relationship of them.

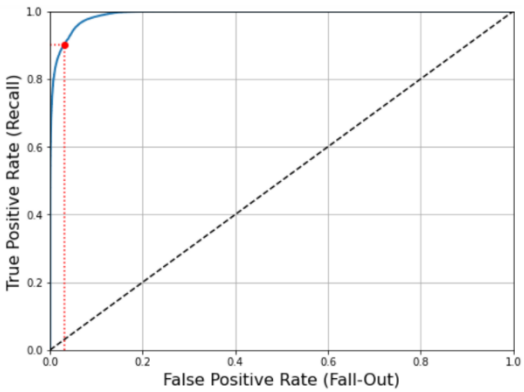
The relationship between precision and recall under different thresholds is shown below:




Our group will also use ROC curve to find the optimal threshold.

The ROC curve plot is below:



As XGBoost has very good performance on our dataset. The precision and recall can be both relatively high at the same time. The AUC score is also as high as 0.9910.

Since our goal is to find out the reasons that affect dissatisfaction the most, we could tolerate a relatively lower recall rate. So, our group set the recall rate to be 0.9. The red dot is the situation when recall rate is 0.9. From its location in the three plots above we could say that it is a reasonable rate. The optimal threshold is 1.1823.

The performance measurement of our model on train set is shown below:

| Precision Score | Recall Score | F1 score | Accuracy |
|---|---|---|---|
| 0.9752 | 0.9000 | 0.9361 | 0.9304 |

| Actual / Predicted | Satisfied | Neutral / Dissatisfied |
|---|---|---|
| Satisfied | 43679 | 1346 |
| Neutral / Dissatisfied | 5888 | 52991 |

Apply our model to test set, the performance is shown below:

| Precision Score | Recall Score | F1 score | Accuracy |
|---|---|---|---|
| 0.9760 | 0.9055 | 0.9394 | 0.9345 |

| Actual / Predicted | Satisfied | Neutral / Dissatisfied |
|---|---|---|
| Satisfied | 11078 | 325 |
| Neutral / Dissatisfied | 1377 | 13196 |

The performance on test set is quite similar with it on train set. The model does not have underfitting or overfitting problem and could be applied to real-world business problem.

## 6. Business Solutions

Based on the feature importance table, we can select the five most important factors which are highly correlated to airline customer satisfaction. They are online boarding, class, inflight wifi service, type of travel and inflight entertainment.

More specifically, passengers with lower classes are more likely to be dissatisfied. For type of travel, dissatisfied passengers mainly come from personal travel. To reduce dissatisfaction, airlines should pay more attention to these two groups. Although these two groups do not bring much profit to airlines, for almost any airline, these two groups are the main groups. Effectively reducing their dissatisfaction is more beneficial to the company.

In addition, the remaining three factors are all related to services quality. Although Ng et al. (2011) stated that in-flight services offered by flight attendants affected customer satisfaction, in this research, we found the most significant service factors have nothing to do with flight attendants. On the contrary, it seems like passengers are more concerned about facilities or system with flight. It's not hard to understand, after all, modern people like efficiency. Before takeoff or during the flight, people can actually do a lot of

fragmented things. If there is no internet or entertainment facilities at this time, people may feel waste of time. In addition, if the system operation is unfriendly or the process is cumbersome, it will further affect the passenger's flight experience and thus cause dissatisfaction.

Therefore, if airlines can only improve limited services, we strongly recommend that these three aspects of services be thoroughly evaluated and optimized. After all, most of the dissatisfaction comes from these areas, indicating that the company still has a lot of room for improvement in these areas.

Given that we have identified what factors lead to customer contentment for an airline, we would like to further deliver business strategies that can help an airline maintain or even improve customer retention rate.

Firstly, airlines should focus on ease of online boarding. As the primary factor that leads to passenger's dissatisfaction, online boarding is still very imperfect. Although online boarding can reduce labor costs and may save everyone queuing time, if the online process is complicated and there are no other facilities such as Wi-Fi support, then the existence of online boarding is meaningless. We recommend that airlines conduct a detailed survey for online boarding to understand the reasons for the dissatisfaction, and then make specific optimizations for online boarding.

In addition, improving the inflight Wi-Fi service and entertainment experience is also significant to the airlines. On the one hand, airlines could develop better software to allow easier access to inflight Wi-Fi, or lower the cost to access inflight Wi-Fi such that more economy class customers can enjoy the service. On the other hand, the inflight entertainment should also keep up with the times. Actually, what people want on the ground is what people want in the air. Inflight entertainment should not be limited to movies, games and shopping, airlines should develop new entertainment methods. For instance, the in-flight diploma is being explored as there's a lot of potential, especially on an 18-hour flight where passengers could finish most of a course.

Last but not least, for the passengers in economy class or personal travel, their dissatisfaction reasons are unknown and vary from person to person. On the one hand, airlines should listen to the voice of these two groups and understand their needs. More importantly, companies need to balance the cost of meeting their needs with the potential benefits. If the new strategy is riskier and has little effect, it is recommended not to take risks.

## 7. Conclusion & Future Studies

This project makes use of machine learning algorithms (Logistic Regression, Random Forest, Xgboost) on inflight passenger dissatisfaction prediction. According to the feature importance given by Xgboost, online boarding, class, inflight wifi service, type of travel and inflight entertainment are crucial features for dissatisfaction of passengers.

Among all models, XGBoost delivered the best result overall, with highest precision score 0.9752. At the same time, the model does not have underfitting or overfitting problem and could be applied to real-world business.

Finally, to make our project more practical, we also provided some recommendation aimed at the vulnerable passengers.

Actually, there are several limitations in our project.

First of all, our dataset comes from Kaggle, and the authenticity of its source is unknown. At the same

time, this dataset does not indicate which type of airline it is applicable to. As airlines are of different types and sizes, our models and recommendations are based on the entire industry and may not suit to specific companies. Therefore, in the future, we would like to focus on a type of airline or conduct a separate study on these significant variables to understand the specific dissatisfaction factors.

Since this report is related to machine learning, we focused on technical analysis and did not assess the real business environment. However, in the real world, resolving customer dissatisfaction is a complicated matter, and many factors need to be considered before making a final decision. For example, research into the cost and benefit of improving Wi-Fi service would be highly beneficial for an airline looking to improve sales through Wi-Fi service.

At last, we'd like to employ more advanced machine learning techniques to help predict flight customer dissatisfaction and optimize the models by applying them into reality.

## 8. References


Baswardono, W., Kurniadi, D., Mulyani, A., & Arifin, D. M. (2019, December). Comparative analysis of decision tree algorithms: Random forest and C4. 5 for airlines customer satisfaction classification. In Journal of Physics: Conference Series (Vol. 1402, No. 6, p. 066055). IOP Publishing.

Bing Li. (2010). Thinking about improving customer loyalty. Market research, 2.

García, V. (2019). Predicting airline customer satisfaction using k-nn ensemble regression models. Instituto de Ingeniería y Tecnología.

James, Gareth, Daniela Witten, Trevor Hastie, & Robert Tibshirani.(2017). An Introduction to Statistical Learning: with Applications in R. New York: Springer。

Hong, S. H., Kim, B., & Jung, Y. G. (2020). Correlation Analysis of Airline Customer Satisfaction using Random Forest with Deep Neural Network and Support Vector Machine Model. International Journal of Internet, Broadcasting and Communication, 12(4), 26-32.

Ng, S.I., Sambasivan, M. and Zubaidah, S. (2011), “Antecedents and outcomes of flight attendants’ job satisfaction”, Journal of Air Transport Management, Vol. 17, pp. 309-313.

Saadat, M., Tahbet, T. R., & Mannan, M. A. (2018). Factors That Influence Customer Satisfaction in Airline Industry in Malaysia. IOSR JBM, 20(8), 1-6.