Binary Classification on the Airline Passenger Classification Dataset

Darnell Kikoo[1], Matthew Adrianus Mulyono[2], Winson Allan Wijaya[3]

School of Computer Science, BINUS University
Tangerang, Banten, Indonesia

[1]darnell.kikoo@binus.ac.id

[2]matthew.mulyono001@binus.ac.id

[3]winson.wijaya@binus.ac.id

***Abstract*— This document serves as a report to an experiment we did in determining whether passengers are satisfied with airline flight or not, based on the answers of a survey, compiled into a single data set. The problem we would like to solve could be considered a binary classification problem. We opted to create a python based catboost-assisted binary classification program to solve this problem. The model was then deployed as a web-based program with the assistance of streamlit.**

***Keywords*— Binary Classification, Catboost Classifier, Flight Satisfaction, Machine Learning**

I. Introduction

Airplane has always been one of the most favorite choices to travel abroad, with its speed and safety for the passengers. However, the COVID-19 Pandemic that has been ongoing for the last 2 years has caused a break in ticket sales for almost all airlines, causing this business sector to drop significantly [1].

With the current situation of the pandemic, some business sectors have been in the process of full recovery, including travel. Along with the growth of internationalization, air-based travels have seen a substantial increase in passenger numbers over the years. With improvement in life qualities, more and more companies are seeking to increase their global outreach. The airline business has become increasingly lucrative.

As living standards improve [2], the higher the customer demands become. Now, quality of flight and passenger satisfaction serves as an integral part in maintaining continuous passenger influx. Airline companies have also seen the importance of this factor in recent years, culminating in surveys done to grasp the quality of their service.

Despite their efforts, it might prove to be difficult to ultimately decide whether the customers truly feel satisfied or not, as the numerous questions in the survey result in possibly complex deliberations in making decisions. Due to that, we attempt to find a way to answer the age-old question “Are the passengers satisfied, or not?”.

II. Methodology

In the effort of creating an efficient and performant model, we decided to approach this problem in a systematic manner. In essence, we performed the activities: Data Preparation, Exploratory Data Analysis, Data Preprocessing, Modeling, and Evaluation.

A. Data Preparation

For this experiment, we use the airline passenger satisfaction dataset from Kaggle [3]. We used only the training set for our prediction, which consists of 25 columns and 104,000 rows. The dataset has 2 labels: neutral/dissatisfied (57%) and satisfied (43%). The data preparation phase itself consists of 3 activities. First, the unnecessary columns are removed so that they won’t affect the model in a negative manner. One example would be the “id” column as each row is uniquely identified. Then, we examine each column for missing values. Missing values in categorical columns are filled in with their mode, while the numerical counterparts are filled in with their median value. The last method of preparation is data type conversion. Each column is examined once again to ensure that the data type is correct. Columns that are found to be categorical instead of numerical will be converted to category data type and vice versa.

B. Exploratory Data Analysis

The data exploration process is divided into 2 equally important sections: categorical data analysis and numerical data analysis. Bar graphs were used to map the distribution of the label for each categorical column. We use the matplotlib and seaborn library to create these bar graphs from the Python programming language. There are 2 types of categorical columns in the dataset: rating columns and non-rating columns. The distribution of labels for each column that contains ratings is shown in Fig. 1.

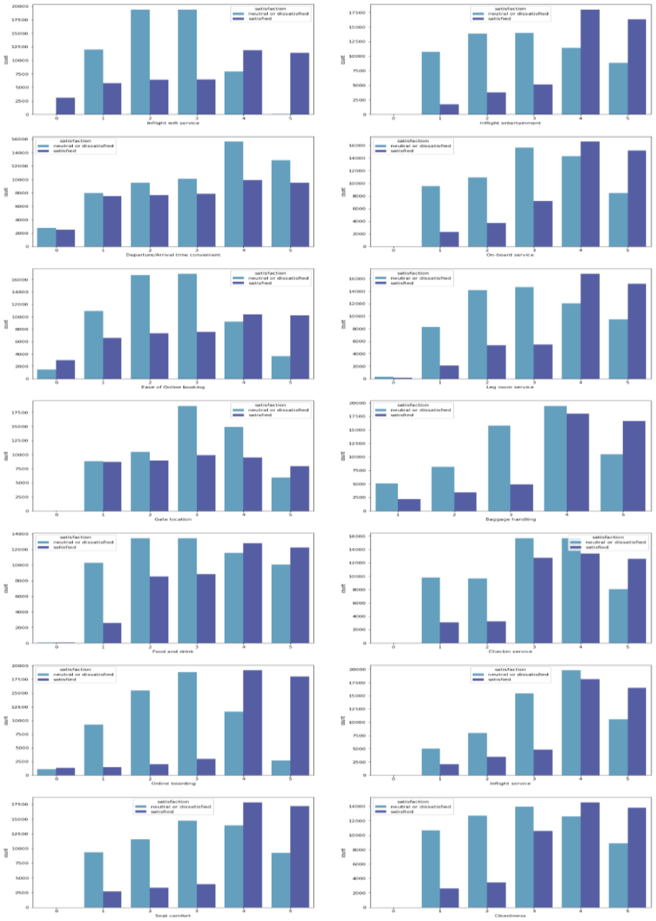


Fig. 1 Bar graphs showing the distribution of labels in each categorical rating column.

Through these bar graphs, we can conclude that those who are satisfied with the inflight Wi-Fi services are more likely to be satisfied with the flight. However, whether they depart or arrive on time doesn't have a significant impact on the passenger's satisfaction. The distribution of labels in non-rating columns is shown in Fig. 2.

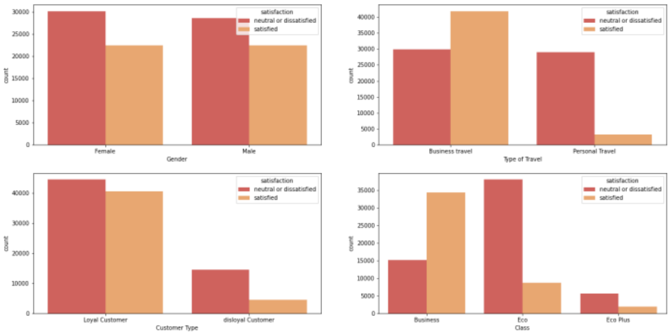


Fig. 2 Bar graphs showing the distribution of labels in each non-rating categorical column.

Based on these charts, we came to an understanding that most passengers whose flight classes aren't Business Class are more likely to be neutral or dissatisfied. We also discovered that most passengers who went for business are usually satisfied, while those who went for personal travel are mostly unsatisfied with their flight. Finally, we also found that loyalty towards an airline doesn't really affect passenger satisfaction.

To understand the numerical columns, we use the kdeplot and boxplot from the seaborn library to visualize their distribution. A kernel density estimate (KDE) plot is used to visualize the distribution of a series in a dataset. Its functions are like a histogram but less cluttered and easier to interpret. The distribution of numerical values in each numerical column in the kdeplot is shown in Fig. 3 and the spread of these columns is shown in the boxplots in Fig. 4.

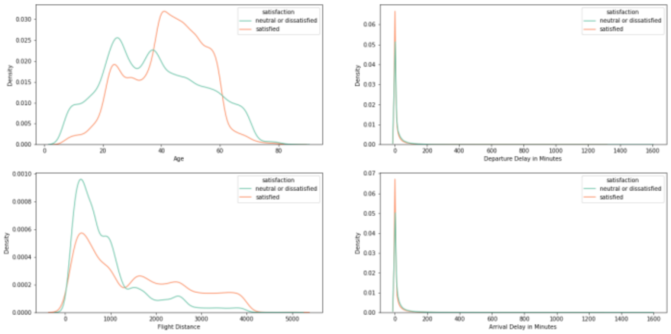


Fig. 3 Distribution of labels in each numerical column using kdeplot

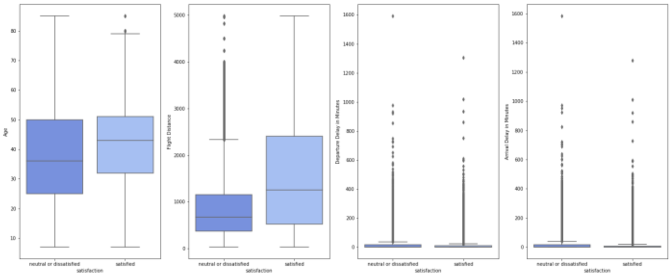


Fig. 4 Distribution of labels in each numerical column in boxplots

From these graphs, we discovered that passengers between the ages of 40 and 60 are more likely to be satisfied, while passengers on flights with distances lower than 1200 km are less likely to be satisfied with their flight. We also discovered that those whose departure/arrival time is late for more than 200 minutes are 60% more likely to be dissatisfied with their flight.

Lastly, we also converted the categorical rating columns back into integer data types to check if they have any correlation with each other using a heatmap. The heatmap is shown in Fig. 5.

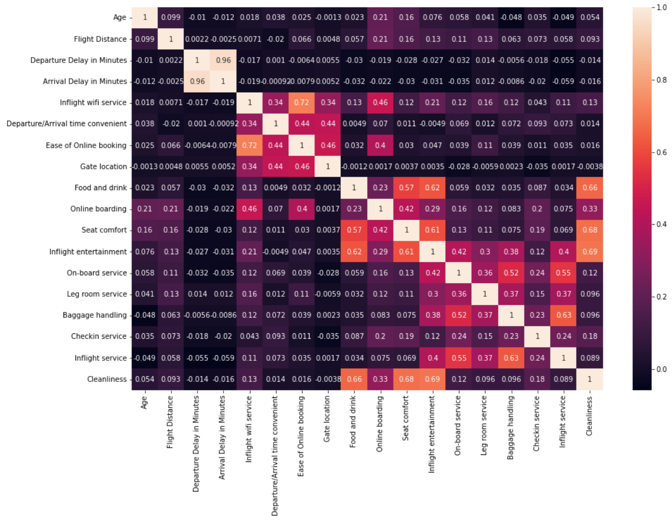


Fig. 5 Heatmap showing the correlation between numerical columns

After the analysis was complete, the processed dataset was saved into a .csv file for the next step.

C. *Data Preprocessing*

Before we start the modeling process, we need to preprocess the data first. The processed dataset will undergo 4 steps of preprocessing. First, the dataset will be split into a training dataset and a testing dataset. We split them using the train_test_split function from the sklearn library. After that, we continue by removing the outlier rows from the training set. To find the outliers, we used the Interquartile rule (IQR). After removing the outliers, we continue by scaling the numerical columns using a MinMaxScaler from sklearn. Scaling is done to make the data points more generalized and easier to calculate. A MinMaxScaler scales the values based on the minimum and maximum values of each column. Lastly, we encode the categorical columns so that they can be fitted into our classifiers. For nominal data, we use the get_dummies function from the pandas library to one-hot encode the columns, while for ordinal data, we used a map function to ordinal encode the columns. After the categorical columns are encoded, we begin the modeling process.

D. *Modeling and Evaluation*

The modeling process involves multiple classification algorithms.

We tested more than 11 algorithms that are mainly used for binary classification. We used a python library called Sklearn, which is commonly used to ease the modeling and evaluation process in machine learning. Those modules include Logistic Regression, Decision Tree, Linear Discriminant Analysis (LDA), Stochastic Gradient Descent (SGD), Gaussian Naive Bayes, Random Forest Classifier, Gradient Boosting Classifier, XGBoost, CatBoost, Light Gradient Boosting Machine (LGBM) and K-Nearest Neighbors (KNN) Classifier with 3 neighbors.

After comparing 11 algorithms with each other, the best algorithm was CatBoost Classifier after getting compared using the predefined metrics, which were then used for further modeling and hyperparameter tuning.

1. *Metrics*

For this classification process, 5 types of classification metrics are used, which are accuracy, precision, recall, F1-Score, and ROC. The main metric will be the recall, where it is better to focus on customers who are dissatisfied with the flight.

2. *Hyperparameter Tuning*

Using the parameters from the Cat Boost algorithm, 3 parameters from this algorithm are tuned, which are `max_depth`, `n_estimators`, `learning_rate`.

Parameter Name	**Parameter Input**
max_depth	[5, 6, 7]
n_estimators	[300, 400, 500]
learning_rate	[0.01, 0.05, 0.1, 0.15]

Fig. 6 Parameters for hyperparameter tuning

From the Fig. 7, GridSearchCV by sklearn.model_selection, are used to do the hyperparameter tuning, with the focus on accuracy. Accuracy is used instead of recall since the training data is quite balanced and is assumed to be able to generalize the model using accuracy metrics.

After using 5 cross-validations, the tuning model earned a result of 0.1 `learning_rate`, 7 `max_depth` and 400 `n_estimators`.

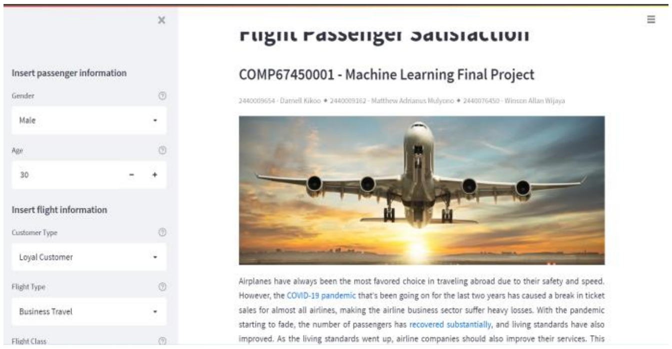
III. Results and Discussion

After modeling and evaluation, the best model from across 11 algorithms is CatBoost, with some specified hyperparameter of 0.1 `learning_rate`, 7 `max_depth`, and 400 `n_estimators`.

Metric	Performance
Accuracy	0.9628
Precision	0.9708
Recall	0.9431
F1 Score	0.9627

Fig. 7 Catboost Model Performance

The performance of this model is represented by Fig. 7, where it achieved 96.28% accuracy, supported by 97% precision, 94% recall, and 96% F1 Score.



Insert passenger information

Gender

Male

Age

30

Insert flight information

Customer Type

Loyal Customer

Flight Type

Business Travel

Flight Class

Flight Passenger Satisfaction

COMP67450001 - Machine Learning Final Project

2440009654 - Darnell Kikoo ♦ 2440009162 - Matthew Adrianus Mulyono ♦ 2440076450 - Winson Allan Wijaya

Airplanes have always been the most favored choice in traveling abroad due to their safety and speed. However, the COVID-19 pandemic that's been going on for the last two years has caused a break in ticket sales for almost all airlines, making the airline business sector suffer heavy losses. With the pandemic starting to fade, the number of passengers has recovered substantially, and living standards have also improved. As the living standards went up, airline companies should also improve their services. This

Fig. 8 Web application preview

So that everyone can try our model, we deployed our model in the form of a web application to Heroku using the streamlit library in Python. Here is the link to the web application: https://flight-satisfaction.herokuapp.com/. To ensure that our model works fine, we also provided a review survey in the application to check if our model works well enough. Every time a user uses the model to predict satisfaction, they can choose 1 of three buttons: “Yes”, “No”, and “No Idea”. If the model prediction checks out, click “Yes”. If it doesn’t, click “No”. Lastly, if the user tried out our model without knowing/remembering if they are satisfied or not, they can click “No Idea”. Fig 8. Shows the preview of the deployed model.

Another way to improve this model is by exploring the data of the airplane flight datasets, where further data analysis is required, and some addition to the hyperparameter tuning.

IV. Conclusions

Being one of the most relied public transports, airplanes have been chosen by most people as the transportation for international travels. Due to that, the ability to determine a customer’s satisfaction level will help airline companies to decide on actions that could increase both the customer’s retention and satisfaction, which are believed to be able to improve sales and the company’s brand.

In this paper, we tried to create a machine learning model using one of the tree algorithms, CatBoost, to predict the satisfactory rate of the airplane customers. Some independent variables from `customer_age` to the rating of the airplane services are used for the training input for the machine learning model.

After doing some actions as seen in the methodology section, the best CatBoost model for the training data was obtained with up to 96% accuracy performance on the testing dataset. Some actions could also be done to further improve the performance, from deep diving more into the data and traversing the hyperparameters of the CatBoost algorithm.

References

[1] S. Maneenop and S. Kotcharin, “The impacts of COVID-19 on the global airline industry: An event study approach,” Journal of Air Transport Management, vol. 89. Elsevier BV, p. 101920, Oct. 2020. doi: 10.1016/j.jairtraman.2020.101920.

[2] M. E. S. El Keshky, S. S. Basyouni, and A. M. Al Sabban, “Getting Through COVID-19: The Pandemic’s Impact on the Psychology of Sustainability, Quality of Life, and the Global Economy – A Systematic Review,” Frontiers in Psychology, vol. 11. Frontiers Media SA, Nov. 12, 2020. doi: 10.3389/fpsyg.2020.585897.

[3] Klein, T.J. (2020, February). Airline Passenger Satisfaction, Version 1. Retrieved March 30, 2022 from https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.