****

**Wee Kim Wee School of Communication and Information**

K6312 Information Mining & Analysis

The Academic Year 2020-21, Semester 1

Group Project

**Predictive Analysis on Airline Passenger Satisfaction with Classification Methods**

**Link:** [**https://github.com/kzui123/UIC-Team**](%20https:/github.com/kzui123/UIC-Team)

|  |  |
| --- | --- |
| Submitted By Group UIC Team | |
| Members’ Names | Matriculation Number |
| CHEN XIAOFENG | G2001009A |
| TAN JING | G2001007G |
| WEI DONGLIN | G2001033B |
| YANG ZIJING | G2001240B |

**Date: [22th Nov]**

**Predictive Analysis on Airline Passenger Satisfaction with Classification Methods**

UIC team

Chen Xiaofeng, Tan Jing, Wei Donglin, Yang Zijing

**Abstract**

This paper analyzes what airlines services are highly correlated to passenger satisfaction. We evaluate five different classifier model and find that random forest is the best classifier to analyze our data. In data processing, we find that online booking service, seat class, and type of travel positively correlated with customers’ satisfaction attitude. Additionally, gate location, gender, departure or arrival time convenience, and arrival and departure delay negatively correlate with customers’ satisfaction attitude.

**1. Background**

With the global economic integration, the world is more and more connected. People's living standards have continued to improve. The demand for air travel and transportation and the quality requirements of air travel and transportation have also increased. The focus of airlines is not just on flying passengers to their destinations, but also on improving the quality of services and making passengers comfortable during the journey. Airlines provide lots of services such as online booking service, food and drink service as well as check-in service and inflight service. In the aviation market, good services can help airlines win customer’s heart and stand out in the market. What’s more, customer satisfaction plays a crucial role in consuming behavior and the brand image of airlines. In the fierce competition, maintaining customer satisfaction at a good level can help the airline occupy the aviation market and get plenty of repeat customers. Additionally, nowadays, the aviation industry was hit hard by the Covid-19. Many flights have been reduced because of restrictions in some countries. People cannot travel to other countries due to restrictions of some countries. In order to profit and reduce the loss, some airlines introduce “flights to nowhere” which means air travel that take place purely for the purpose of the journey, not the destination. People can enjoy the beautiful views and good services during the flight. In this case, providing good services is very important to make customers satisfied. The question of what services are highly correlated to customer satisfaction is worth to further study.

**2. Purpose of research**

The purpose of this research is to find out what factors lead to customer satisfaction for an Airline. Therefore, airlines can win the loyalty of customers by improving these factors. If customers are very satisfied with the services provided by an Airline, customers will choose this Airline on the next flight, which makes the Airline profit a lot.

**3.** **Data pre-processing**

**3.1 Data Collection**

The dataset involving training and testing data was downloaded from Kaggle Website: <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>. It comes from a detailed online survey regarding the airline passengers’ satisfaction and its possible affecting factors.

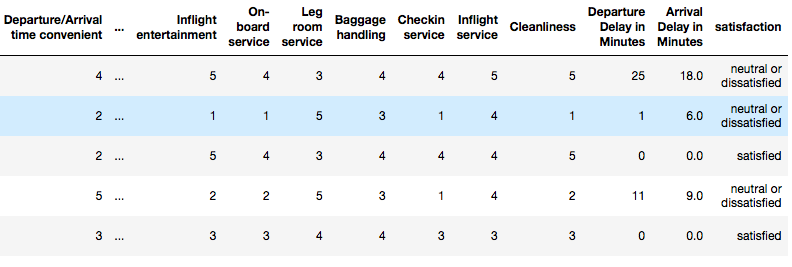
Training data contains 103,904 records and 25,976 for testing data, including 25 labels, namely unnamed 0, ID, Gender, Customer type, Age, Type of travel, Class, Flight distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check in service, Inflight service, Cleanliness, Departure delay in minutes, Arrival delay in minutes, and Satisfaction attitude. Among them, labels from inflight Wi-Fi service to cleanliness are valued by the level of satisfaction ranging from 1-5. Some of the columns are shown in Figure 1:

Fig. 1. Head of training data

**3.2 Data Cleaning**

Firstly, outliers are detected by using box-plot method in the selected database, and they are dropped for both train and test data. After that, missing data are checked in the database. There are initially 309 missing values in train data frame and 83 missing values in test data frame, and they are all removed from the dataset. Next, duplicate data are checked, and no duplicate data is found in the train and test data frame. Finally, the categorical values are all transferred into numerical value for further data mining. Some important transfer labels are listed below:

1. Gender Type

1: Female

0: Male

1. Customer Type

1: Loyal customer

0: Disloyal customer

1. Travel Type

1: Business Travel

0: Personal Travel

1. Seat Class Type

2: Business Class

1: Economic Plus Class

0: Economic Class

1. Satisfaction Attitude Type

1: Satisfied Attitude

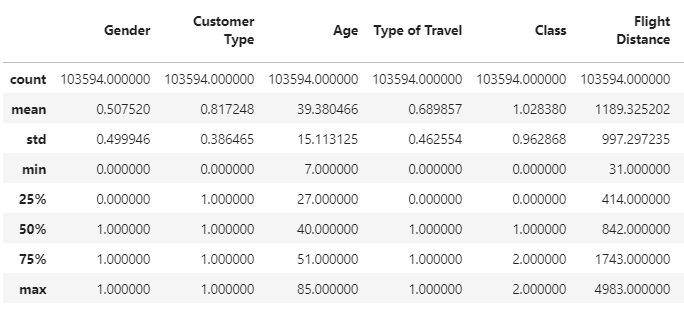
0: Neutral or Dissatisfied Attitude

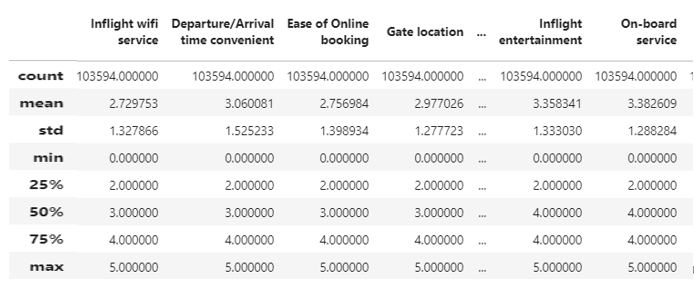
**4. Data Description**

After data cleaning, there are 103,904 samples in the modified dataset. The report will depict data from numerical and categorical features, target feature, general trend, count plot description, and variable correlation.

**4.1 Numerical and Categorical Features**

For numerical data, the average age of surveyed flight customers is 39 years old. The oldest age is 85 years old while the youngest age is 7 years old. In addition, the mean value of flight distance is 1189 km. The longest flight distance in the survey is 4983 km whilst the shortest flight distance is 31 km. In terms of departure and arrival delay in minutes, the average of departure delay is about 14 minutes and the mean value of arrival delay is nearly 15 minutes.





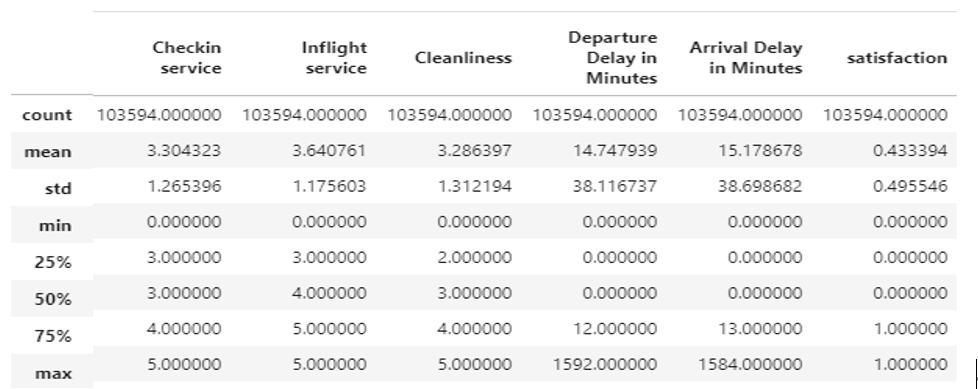
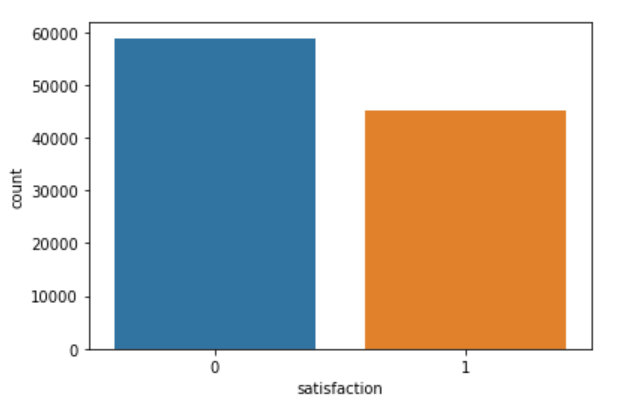


Fig. 2. General Data Description Graph

With regard to categorical data, the mean value of gender type is 0.5, showing that the gender ratio is equal. Moreover, the average of customer type is 0.8 and 75% value is 1, indicating that most surveyed customers are loyal customers. Besides, the mean value of travel type is about 0.6 which illustrates that the proportion of business travel is little higher than that of personal travel. For flight seat class, the average of seat class is 1.02, expressing that most of travelers choose economic plus and economic seat class to take flight.

**4.2 Description of Target Feature**

Significantly, the report describes satisfaction attitude, which is the target feature, in details. At first, the number of satisfied attitude and neutral or dissatisfied attitude are counted separately. In general, there are 45,025 satisfied attitude samples and 58,879 dissatisfied attitude samples. Then the general satisfaction sample distribution is visualized in the following bar chart. We can find that the number of people having neutral or dissatisfied attitude are almost 30% more than that of people having satisfied attitude.

Fig. 3. Satisfaction Level Description

**4.3 General Trend of Dataset**

After describing general data description, the report would like to explore more information about general trend of the dataset.

Figure 4 describes the frequency of customers’ flight distance. Most of the flight distances fall in the range of 0 to 1000 km.

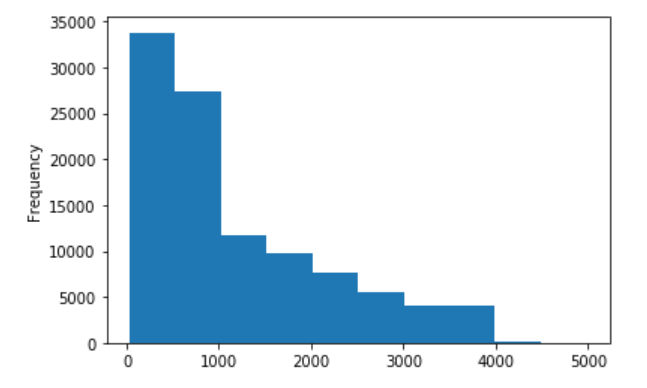


Fig. 4. Flight Distance Frequency Distribution

Moreover, figure 5 illustrates the tendency of relationship between age and satisfaction attitude. We find that young people (7 to 20 years old surveyed passengers) and elderly people (65 to 80 years old surveyed passengers) are generally neutral or dissatisfied with the flight service nonetheless adult (20 to 60 years old surveyed passengers) are generally satisfied with the flight.

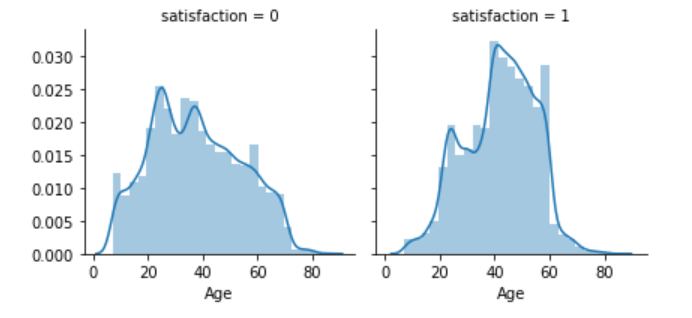


Fig. 5. Age & Satisfaction Relation Tendency

Specifically, we also find that people who are from 44 to 57 years old are most likely to have a satisfied attitude towards flight service. This is shown in Figure 6.



Fig. 6. Age & Satisfaction Correlation Graph

**4.4 Count Plot Description**

Next, to better understand the corresponding distribution between target value and variable values, several important categorical variables are selected, which are flight seat class, gate location, customer type, and check in service, and their distribution is visualized in terms of satisfaction attitude. The tables are shown below.

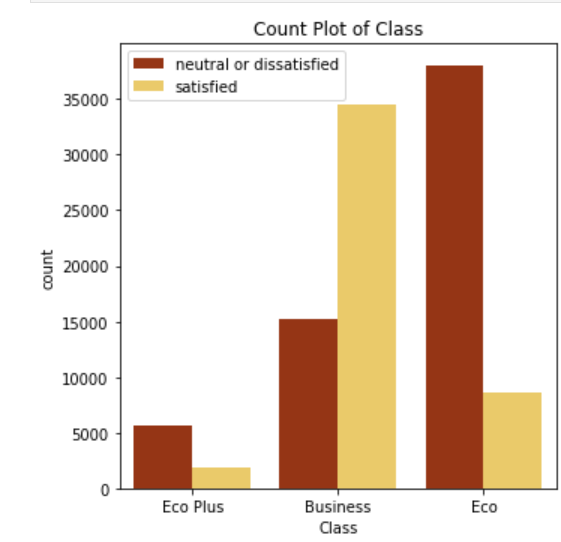


Fig. 7. Count Plot of Flight Seat Class

In figure 7, we find that people taking economic and economic plus seat are more likely to have neutral or dissatisfied attitude to flight service whilst people taking business class seat are less likely to have the same indifferent or negative sentiment.

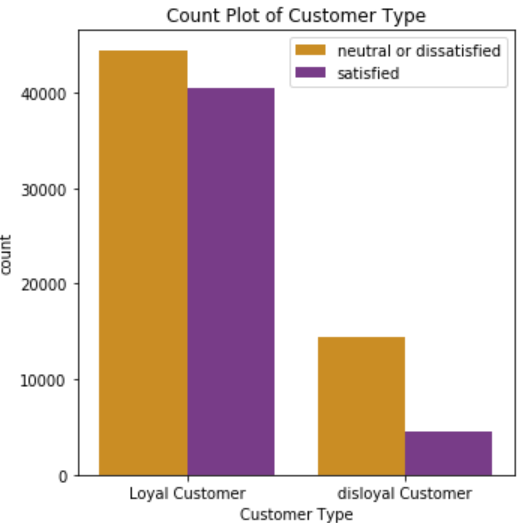


Fig. 8. Count Plot of Customer Type

In addition, in figure 8, loyal customers have the higher proportion of satisfied customers compared to disloyal customers.

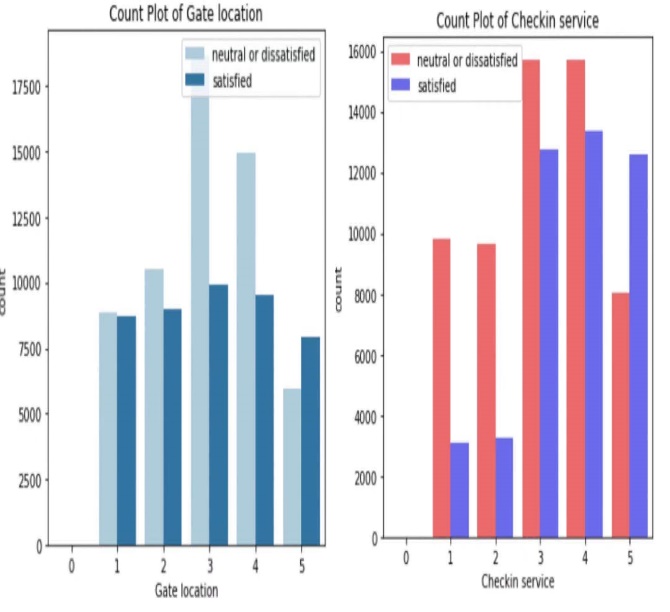


Fig. 9. Count Plot of Gate location

& Check in service

Moreover, in figure 9, the distributions of neutral or dissatisfied customers’ attitude to gate location and check in service are quite similar, whilst more satisfied customers are more pleased with check in service compared to gate location service.

**4.5 Correlation**

At the beginning, the overall correlation graph is placed below.

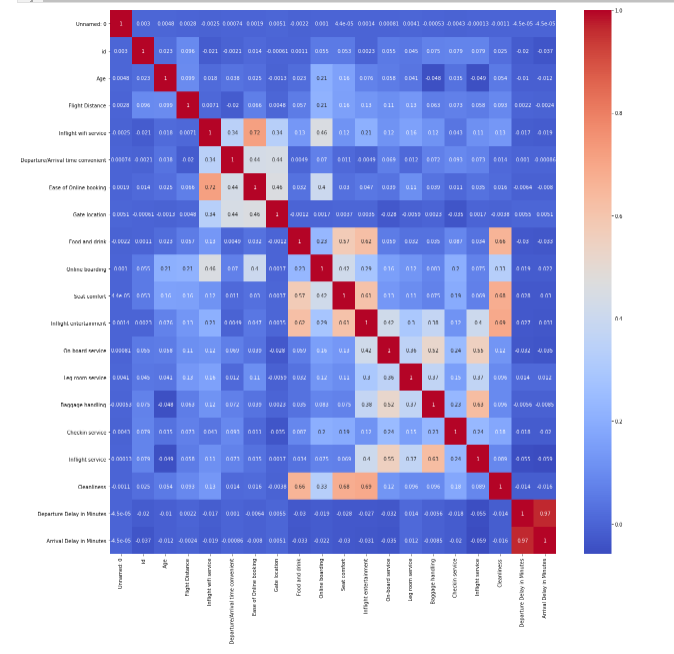


Fig. 10. General Correlation Table

Following, one hot encoding correlation test is made to find the correlation between satisfied level and other features. The table shows the visualized correlated relationship between target features and other variable features.

In general, online boarding service, seat class, and type of travel have a strong positive correlation with customers’ satisfaction attitude. In contrast, gender, departure or arrival time convenience, and arrival and departure delay have a negative correlation with customers’ satisfaction attitude. In other words, when people are unsatisfied with arrive delay and departure time convenience, they are more likely to have a negative or neutral attitude towards flight service.

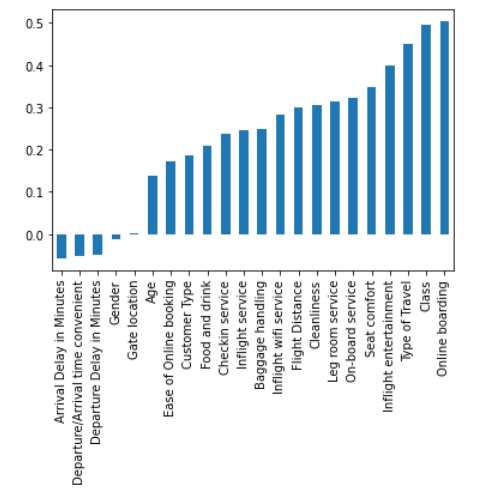


Fig. 11. The Correlation Table between Satisfied Level & Other Features

Since online boarding has a strongest positive correlation with customers’ satisfaction, we would like to check the correlation between online boarding and other variables to seek reasons of this strong positive correlation. In figure 12, it shows that inflight Wi-Fi service has the strongest correlation with online boarding. Therefore, it can be inferred that people tending to enjoy Wi-Fi service are more likely to take online boarding, and meanwhile they are more willing to have a satisfied feedback on flight service.

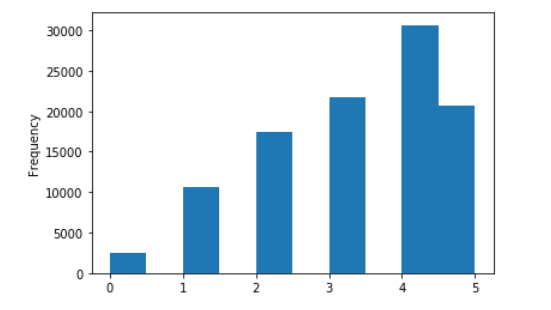
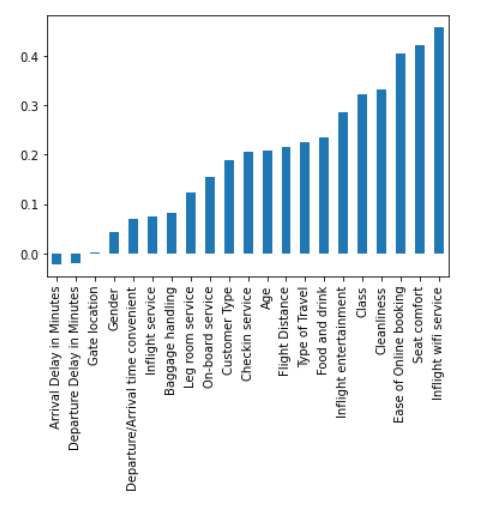


Fig. 12. The Correlation Table between Online Boarding & Other Variables

Furthermore, interestingly, in terms of the correlation between food and drink service and passenger satisfaction, 20 percent of passengers who do not like the flight food and drink service are satisfied with the flight service. The figure is shown below.

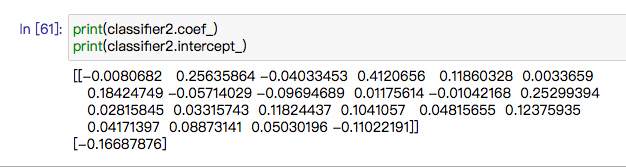
Fig. 13. The Correlation Table between Food and Drink service & Passenger Satisfaction

**5. Data Processing**

**5.1 Tran/Test Split**

The datasets have been divided into training data and testing data, two files originally when we collected from Kaggle. The proportion of train/test split is 0.25.

**5.2 Model Evaluation**

Considering machine learning methods learnt from course, five typical classifier models are chosen to train and test the datasets: Logistic Regression, Linear SVM Regression, Random Forest Regression, Decision Tree Regression and Ada-Boost Regression. The same features and target are used to train and test in these five classifiers. They will be evaluated respectively in the following sections.

**5.2.1 Logistic Regression**

The first classifier used is logistic regression which is often used for binary classification problem. The total CPU time is 482 ms and wall time is 275 ms.

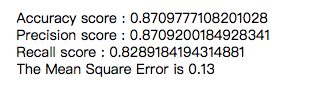
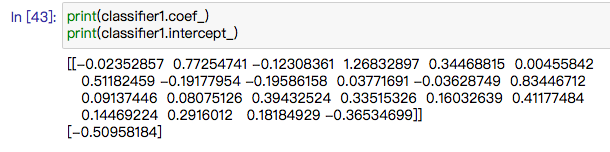
****The coefficients of features in training and intercept are shown in Figure 14.

Fig. 14. Coefficients and Intercept for Logistic Regression

The confusion matrix is shown below, the vertical axis stands for testing labels, whereas horizontal axis is predicted labels, green area shows the accurate match between test and prediction. Through confusion matrix, its accuracy score is calculated as 0.872, precision is 0.870 and recall is 0.831.

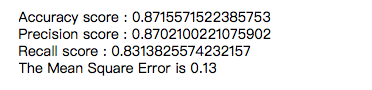


Fig.15. Evaluation Scores for Logistic Regression

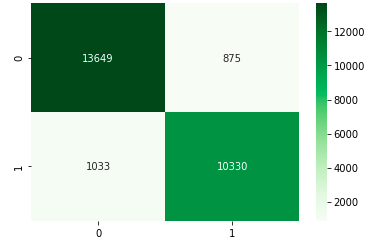


Fig. 16. Confusion Matrix for Logistic Regression

**5.2.2 Linear SVM**

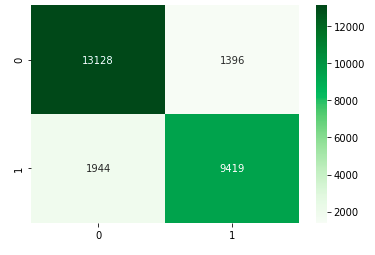
Then, we use linear SVM for training and testing datasets. Linear support vector classification is used in this case.

Fig. 17. Coefficients and Intercept for Liner SVM

Total 31.8s CPU time also wall time is 31.8s, which is much longer than other classification approaches.

Looking at its confusion matrix, accuracy score is 0.871, the same score for precision and 0.829 for recall score, that is similar with logistic regression performed. However, the prediction result for positive (satisfaction = satisfied) is worse.

Fig. 18. Evaluation Score for Linear SVM

Fig. 19. Confusion Matrix for Linear SVM

**5.2.3 Decision Tree**

Decision tree works for 732ms total CPU times and 733ms wall time.

Its Confusion Matrix and score for accuracy, precision, recall and the mean square error are shown in the below pictures.

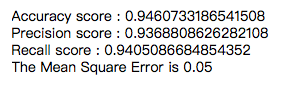


Fig. 20. Evaluation Score for Decision Tree

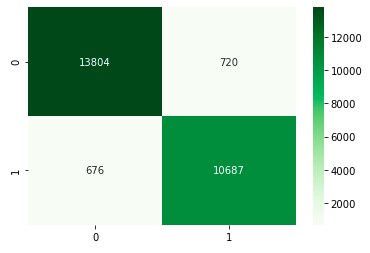


Fig. 21. Confusion Matrix for Decision Tree

**5.2.4 Random Forest**

Random forest is a kind of bagging method that is to reduce variance in model. It is constructed by many decision trees which are not disturbed with each other.

Its work spends 14.2s in CPU times 14.5s in wall time.

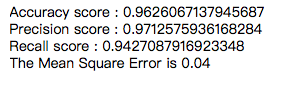


Fig. 22. Evaluation Score for Random Forest

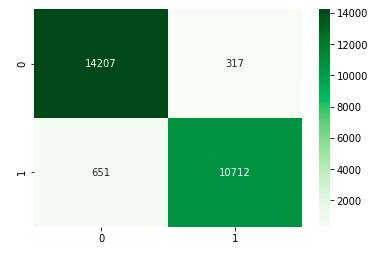


Fig. 23. Confusion Matrix for Random Forest

**5.2.5 Ada-Boost**

Finally, one of the boosting methods called Ada-Boost is used. It ran the training process using 5.63s for total CPU times, 5.79s for wall time.

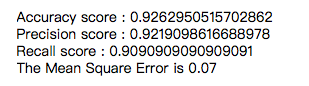


Fig. 24. Evaluation Score for Ada-Boost

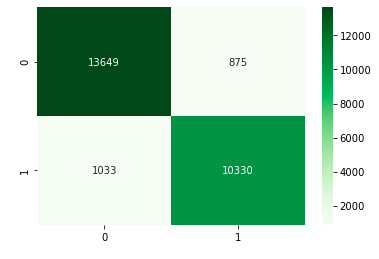


Fig. 25. Confusion Matrix for Ada-Boost

**5.2.6 Comparison & Model Choosing**

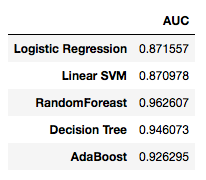
****Based on the previous model training and testing, the accuracy scores for five classifiers are concentrated in the following figure.

Fig. 26. Accuracy score for five models

It is obvious that Random Forest has the best performance in accuracy with 0.962, followed by decision tree. The worse one is linear SVM with 0.871 accuracy. Meanwhile, the mean squared error demonstrates the same story. In terms of running time, logistic regression is the fastest method, whereas linear SVM ran much slower than other classifiers. (Table 1)

|  |  |  |
| --- | --- | --- |
|  | CPU times | Wall time |
| Logistic Regression | 482 ms | 275ms |
| Linear SVM | 31.8s | 31.8s |
| Decision Tree | 732ms | 733ms |
| Random Forest | 14.2s | 14.4s |
| Ada-Boost | 5.63s | 5.79s |

Table 1. Running times for five models

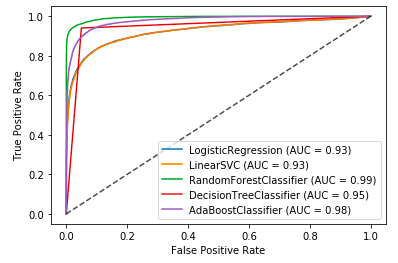
****In ROC curve, the closer curve to the left-upper, the more accurate of the model. Looking at the curve for five classifiers, random forest performs the best with the largest value of AUC and logistic regression and linear SVC have the smallest AUC value.

Fig. 27. ROC-Curve for five models

It can be concluded that random forest is suitable for the datasets with high-dimensional features and does not need the dimensionality reduction. However, the training speed for random forest is average.

Therefore, in the case of flight passengers’ satisfaction, random forest would be selected as the classifier to train and test datasets.

**6. Limitation**

There are too many variables so that some classifier models such as logistic model and decision trees, cannot visualize the relationships of variables.

**7. Conclusion**

After analysis, we found that online boarding service, seat class, and type of travel have a strong positive correlation with customers’ satisfaction attitude, which means that airlines can improve customers’ satisfaction by improving online boarding service. In contrast, gender, departure or arrival time convenience, and arrival and departure delay have a negative correlation with customers’ satisfaction attitude, which means that airlines should try their best to reduce the situation of arrival and departure delay caused by human factors. Additionally, random forest is the appropriate classifier to analyze flight passengers’ satisfaction. This research help airlines find out what factors lead to customer satisfaction. Therefore, airlines can win the loyalty of customers by improving these factors.

**References**

Asyiqin, S., 2020. Singapore Airlines Introduces No-Destination Flights For Those Who Miss Flying. [online] WORLD OF BUZZ. Available at: <https://worldofbuzz.com/singapore-airlines-introduces-no-destination-flights-for-those-who-miss-flying/> [Accessed 12 November 2020].

Qantas 7-hour flight to nowhere sells out in 10 minutes. (2020). Retrieved 12 November 2020, from https://abc30.com/flight-to-nowhere-qantas-airlines-flights-with-no-destination-flying-in-covid-19-pandemic/6454022/