|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group # | Following Instructions (11 pts) | Writing Quality (11 pts) | Abstract (12 pts) | Data Collection / Cleaning / Exploration (11 pts) | Data Exploration Insights (11 pts) | Methodology (11 pts) | Predictions (11 pts) | Inference (11 pts) | Conclusion (11 pts) | Additional Discretionary  Points | Numeric Grade | Letter Grade | Notes |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 15 | 11 | 11 | 12 | 11 | 11 | 11 | 11 | 11 | 11 | 0 | 100 | A | Very nice work. Easy to read, easy to follow, things make sense. |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Analysis and Prediction of**

**Airline Passenger Satisfaction**

Group 15

Qian Zhang, Yeling Cai, Yunhan Zhang, Zekai Wei

**Table of Contents**

[1. Abstract 4](#_Toc121674729)

[1.1. Project Overview 4](#_Toc121674730)

[1.2. Data Overview 4](#_Toc121674731)

[1.3. The Link of Dataset 4](#_Toc121674732)

[1.4. Interesting Things About the Dataset 4](#_Toc121674733)

[1.5. Predictions 4](#_Toc121674734)

[1.6. Inferences 5](#_Toc121674735)

[1.7. A brief conclusion 5](#_Toc121674736)

[2. Data Collecting/Cleaning 5](#_Toc121674737)

[2.1. Data Sampling 5](#_Toc121674738)

[2.2. Data Cleaning 5](#_Toc121674739)

[3. Data Exploration Insights Using Standard Statistical Techniques 6](#_Toc121674740)

[3.1. Correlation between Each Variables 6](#_Toc121674741)

[3.2. Exploratory Data Analysis 7](#_Toc121674742)

[4. Methodology 9](#_Toc121674743)

[5. Model Prediction 9](#_Toc121674744)

[5.1. Prediction of Overall Satisfaction (with all variables) 9](#_Toc121674745)

[5.2. Prediction of Overall Satisfaction (only with sub-satisfaction variables) 10](#_Toc121674746)

[5.3. Prediction of Customer Type 11](#_Toc121674747)

[6. Model Inference 11](#_Toc121674748)

[6.1. Model Inference of Predicting Overall satisfaction 11](#_Toc121674749)

[6.2. Model Inference of Predicting Overall satisfaction (with sub-satisfaction) 11](#_Toc121674750)

[6.3. Model Inference of Predicting Customer Type 12](#_Toc121674751)

[7. Conclusion 14](#_Toc121674752)

[7.1. Prediction Summary 14](#_Toc121674753)

[7.2. Inference Summary 14](#_Toc121674754)

[8. Appendix 15](#_Toc121674755)

# Abstract

## Project Overview

Our main purpose is to predict passenger satisfaction, and determine the influencing factors of satisfaction, to improve the service quality of the airline. We also want to have a picture of what a loyal member will look like, to help the airline identify potential loyal customers.

## Data Overview

The dataset records 129,880 American Airlines passenger satisfaction surveys, with one column for satisfaction ratings and 22 for explanatory variables which contain customer type and service surveys on the flight. The raw data set contains 129,880 rows, and 24 cols.

## The Link of Dataset

<https://www.kaggle.com/datasets/johndddddd/customer-satisfaction>

## Interesting Things About the Dataset

* This dataset describes nearly every aspect of the experience of flying with a passenger. Through this data set, more detailed factors can be provided to help airlines improve their services in a targeted manner.
* The average score of seat comfort is the lowest, 2.84, which means many passengers are not satisfied with seats on the plane. Therefore, maybe the quality of seats is the most important problem to be concerned about.
* 50% of Departure Delay in Minutes and Arrival Delay in Minutes are 0, and 75% of both are around 12, but the average delay is around 15 minutes. This shows that the flight took off normally in most cases, but because some extreme delays occurred, the data showed this trend.

## Predictions

* We created models to predict passengers’ overall satisfaction with all variables to help airline leaders make decisions and improve service quality.
* We created models to predict passengers’ overall satisfaction only with sub-satisfaction variables to get a list of the importance of features.
* We created models to predict the Customer Type and identify potential loyal customers for the airline.

## Inferences

* We want to compare AUC scores of all models to identify the best model in every prediction task.
* We want to find the relationship between sub-satisfaction and overall satisfaction, then get a list of the importance of metrics.
* We want to identify some common features of potential loyal customers for the airline and give some advice to the airline.

## A brief conclusion

We used 11 models in total to predict overall satisfaction and customer type successfully. Then we compared all models and got the best model in each task. We also get a list of the importance of sub-satisfaction, which are the aspects the airline should pay more attention to. In addition, we get a list of importance of features of potential loyal customers. Then, we come up with some advice to the airline according to our analysis.

# Data Collecting/Cleaning

## Data Sampling

Since the raw data is large, we took sample data to analyze. To avoid bias in the final prediction results, we extracted data from the original data according to the Satisfied ratio. After sampling, our data has a total of 10035 rows.

## Data Cleaning

In this data set, there is fewer missing data, and there are only a small amount of null values in Arrival Delay in Minutes. Considering that null values will affect the machine learning results, for the accuracy of the prediction, we delete the entire row which contained null values instead of filling null with values.

# Data Exploration Insights Using Standard Statistical Techniques

## Correlation between Each Variables

电脑萤幕画面

中度可信度描述已自动生成We did a preliminary exploration of the data, and in the satisfaction survey, there were nearly 55% of passengers (71,087 people) who were satisfied with the flight experience. Then, we made a Heatmap to visualize the correlation between variables. According to *Figure 3.1*, it can be found that Inflight Entertainment has the greatest impact on customer satisfaction with 0.52, followed by Ease of Online Booking, Online Support, On-board Service, and Online Boarding. The above variables are positively correlated with customer satisfaction. In Customer Type, we found that the Type of Travel had a large correlation of -0.31. That said, customers with more personal travel are generally more loyal.

Figure 3.1 Airline Satisfaction Correlation Heatmap

## Exploratory Data Analysis

According to *Figure 3.1*, we made plots to see the relationship between specific variables. *Figure 3.2* shows the number of ratings of customers from different classes. According to the plot, there are more customers from the economy and the economy class plus feel neutral or dissatisfied than satisfied. However, business class passengers are more than twice as likely to be satisfied as dissatisfied. Our guess is that business class customers are satisfied with the service they pay more for. But if customers do not pay extra for the service, accordingly airlines will not provide the service that most people would be happy with. We believe that the premise of widening the service gap between different classes is to ensure that the service with the lowest consumption can still be satisfied by most people.

图表, 条形图

描述已自动生成

Figure 3.2 Satisfaction from Different Classes

图表, 直方图

描述已自动生成

Figure 3.3 Flight Distance of Different Customer Types

The information we hope to get from *Figure 3.3* is to understand whether passengers with different loyalty to airlines will choose airplane as a travel mode when traveling at different distances. Clearly, the distribution of passengers is similar for medium and long-distance travel. However, for short-distance travel, almost only loyal customers choose to travel by airplane.

图表, 折线图

描述已自动生成

Figure 3.4 Percentage of Satisfaction for Delay of Departure & Arrival

In *Figure 3.4*, we can find out the correlations between satisfaction percentage and delay time (Departure and Arrival). By looking at the trend, we found that the percentage of passenger satisfaction decreases as the delay time increases. Customers are quite satisfied with delay time less than 30 minutes and on the other side, customers have become very unsatisfied with higher than 90 minutes delay time.

# Methodology

We are going to build three different models for different objectives. For model 1, we want to predict passengers’ overall satisfaction to help airline leaders decide how to improve service quality. For model 2, we want to predict passengers’ overall satisfaction to find the relationship between sub-satisfaction and overall satisfaction, then get the importance of each metric. For model 3, we want to predict the customer type and identify potential loyal customers for the airline.

Firstly, we need to convert the column of satisfaction into a binary variable. Therefore, we transfer satisfied into 1 and others (neutral and unsatisfied) into 0. Secondly, we applied String Encoding to the existing variables: gender; customer\_type; type\_of\_travel and class since these variables are all binary. And by doing this will make our models and classification much easier to do. Lastly, to make the model predictions more rigorous we separated our dataset into training; validation and testing with a ratio of 0.6 / 0.2 / 0.2.

# Model Prediction

## Prediction of Overall Satisfaction (with all variables)

**Goal:** We want to predict passengers’ overall satisfaction to help airline leaders decide how to improve service quality. So, we used all variables to predict the overall satisfaction.

**Predictors:** Gender, Customer Type, Age, Travel Type, Class, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes, and 14 Sub-satisfaction Scores (Seat comfort, Wifi service, Food and drink, etc.)

We used four model types: logic regression, Random Forest[1], GBT, and Deep Learning[2]. *Table 5.1* shows the result of all models. We used grid search, and each model used AUC as a scoring metric because the dependent variable, overall satisfaction, is binary. In logic regression and deep learning models, we also used 3 cross validations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Value being predicted | Model type | Scoring metric | Number of folds (Cross validation) | Resulting Model Score | Best Model Parameter |
| Overall Satisfaction | Logic Regression | AUC | 3 | 0.9130 | ElasticNetParam=0  RegParam = 0.01 |
| Overall Satisfaction | Random Forest | AUC | / | 0.9814  (highest) | numTrees = 30  maxDepth = 20 |
| Overall Satisfaction | GBT | AUC | / | 0.9314 | maxIter = 50  maxDepth = 5 |
| Overall Satisfaction | Deep Learning | AUC | 3 | 0.5229 | StepSize = 0.1  MaxIter = 200 |

Table 5.1 Prediction of overall satisfaction with all variables

## Prediction of Overall Satisfaction (only with sub-satisfaction variables)

**Goal:** We want to predict passengers’ overall satisfaction to find the relationship between sub-satisfaction and overall satisfaction, then get the importance of each metric. So, the airline managers can decide which aspects they can do more.

**Predictors:** 14 Sub-satisfaction Scores (Seat comfort, Wifi service, Food and drink, etc.)

We used three model types: logic regression, Random Forest, GBT. *Table 5.2* shows the result of all models. It is worth noting that, compared to the previous model, we only used 14 sub-satisfaction scores to predict overall satisfaction. And these variables are in the same unit, so we can compare their coefficient later.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Value being predicted | Model type | Scoring metric | Number of folds (Cross validation) | Resulting Model Score | Best Model Parameter |
| Overall Satisfaction | Logic Regression | AUC | 3 | 0.8902 | ElasticNetParam=0.01  RegParam = 0.001 |
| Overall Satisfaction | Random Forest | AUC | / | 0.9772  (highest) | numTrees = 30  maxDepth = 20 |
| Overall Satisfaction | GBT | AUC | / | 0.9165 | maxIter = 50  maxDepth = 8 |

Table 5.2 Prediction of overall satisfaction with sub-satisfaction variables

## Prediction of Customer Type

**Goal:** We want to create models to predict the Customer Type and identify potential loyal customers for the airline.

**Predictors:** Gender, Age, Travel Type, Class, Flight Distance, and Overall Satisfaction Score

**Extra Feature Engine:**

Because every variable is not in the same unit, we standardize features to compare coefficients of features.

We used four model types: logic regression, Random Forest, GBT, and Deep Learning. *Table 5.3* shows the result of all models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Value being predicted | Model type | Scoring metric | Number of folds (Cross validation) | Resulting Model Score | Best Model Parameter |
| Customer Type | Logic Regression | AUC | 3 | 0.8995 | ElasticNetParam=0.1  RegParam = 0.001 |
| Customer Type | Random Forest | AUC | / | 0.9420  (highest) | numTrees = 50  maxDepth = 10 |
| Customer Type | GBT | AUC | / | 0.8319 | maxIter = 50  maxDepth = 8 |
| Customer Type | Deep Learning | AUC | 3 | 0.4807 | StepSize = 0.5  MaxIter = 1000 |

Table 5.3 Prediction of customer type

# Model Inference

## Model Inference of Predicting Overall satisfaction

From Part 5.1, we can see the AUC score of random forest is highest, higher than 90%. Therefore, random forest is the best model in the task of predicting overall satisfaction. However, the AUC score of deep learning is just around 50%, which means deep learning is not suitable for this task.

## Model Inference of Predicting Overall satisfaction (with sub-satisfaction)

According to Part 5.2, we used sub-satisfaction variables to predict overall satisfaction. We used logic regression, random forest, and GBT to predict. The AUC score of random forest is highest. Therefore, the following are results and inferences of random forest.

**Feature Name**:

'Seat comfort', 'Departure/Arrival time convenient', 'Food and drink', 'Gate location', 'Inflight wifi service', 'Inflight entertainment', 'Online support', 'Ease of Online booking', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Cleanliness', 'Online boarding'

**Inference Goal**:

To get the importance of each metric(sub-satisfaction) and help the airline decide how to improve their service.

**Feature Importance**:

表格

中度可信度描述已自动生成

Table 6.1 Feature Importance

**Insights**:

We noticed that “Inflight entertainment”, “seat comfort”, and “Ease of Online booking” are the top 3 metrics that are most important to overall satisfaction. Therefore, the airline is supposed to pay more attention to improve these three aspects to improve the service of the airline.

## Model Inference of Predicting Customer Type

According to part 5.3, the AUC score of random forest is highest.

**Feature Name**:

'Age', 'satisfaction', 'encoded\_Gender', 'encoded\_Type\_of\_Travel', 'Flight Distance', 'encoded\_Class'

**Inference Goal**:

To get a list of feature importance and get a picture of what a loyal member will look like. Then, identify potential loyal customers for the airline and targeted advertising to potential loyal customers.

**Feature Importance**:

表格

描述已自动生成

Table 6.2 Feature Importance

We noticed that “age”, “travel type”, and “flight distance” are top 3 features when predicting the customer type.

**Further analysis**:

We wanted to analyze the top 2 features, “age” and “travel type”. So, we get a plot of the percentage of loyal members in different ages and a plot of the percentage of loyal members in different travel types.

图表, 折线图

描述已自动生成图表, 条形图

描述已自动生成

Figure 6.1 Figure 6.2

From *Figure 6.1*, we noticed that the percentage of loyal members for passengers between 20 and 30 are low, the percentages are much higher for passengers between 50 and 70. From *Figure 6.2*, we noticed that the percentage of loyal members is much higher for passengers traveling for personal purposes.

**Insights**:

Passengers between 50 and 70 years old and usually traveling for personal purposes should be target customers when the airline tries to find the potential loyal members. Besides, the airline could cooperate with more companies to improve the percentage of loyal members for passengers traveling for business travel. Therefore, the airline will get more clients.

# Conclusion

## Prediction Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Value being predicted | Predictors | Best Model Type | Resulting Model Score (validation data) | Resulting Model Score (testing data) |
| Overall Satisfaction | all variables | Random Forest | 0.9814 | 0.9803 |
| Overall Satisfaction | 14 sub-satisfaction variables | Random Forest | 0.9772 | 0.9738 |
| Customer Type | Gender, Age, Travel Type, Class, Flight Distance, and Overall Satisfaction Score | Random Forest | 0.9420 | 0.9382 |

Table 7.1 Prediction Summary of best models

## Inference Summary

* Decision tree is the best model to predict the overall satisfaction and customer type.
* *Table 7.2* shows the feature importance of passengers’ satisfaction. “Inflight entertainment”, “seat comfort” and “Ease of Online booking” are three aspects that the airline should pay more attention to and improve.

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Importance |
| 1 | Inflight entertainment | 0.251763 |
| 2 | Seat comfort | 0.138475 |
| 3 | Ease of Online booking | 0.103454 |
| 4 | Online support | 0.076307 |
| 5 | Leg room service | 0.060950 |

Table 7.2 Feature Importance of predicting Satisfaction

* *Table 7.3* shows the feature importance when predicting customer type. Passengers in 50-70 and with the purpose of personal travel are potential clients becoming loyal members. The airline can push targeted, personalized ads to them and work harder to cooperate with companies.

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Importance |
| 1 | Age | 0.297430 |
| 2 | Type of Travel | 0.249437 |
| 3 | Flight Distance | 0.200656 |
| 4 | Overall Satisfaction | 0.147066 |
| 5 | Class | 0.085535 |

Table 7.3 Feature Importance of predicting Customer Type

# Appendix

[1] *Classification and regression - Spark 3.3.1 Documentation*

<https://spark.apache.org/docs/latest/ml-classification-regression.html#decision-tree-classifier>

[2] *MultilayerPerceptronClassifier — PySpark 3.3.1 documentation*

<https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.classification.MultilayerPerceptronClassifier.html>