Airline Customer Satisfaction Prediction

Machine Learning ICA Report

Anatolii Kolesov

B1159927

Teesside University

2022

Table of Contents

Abstract
This report describes development of a machine learning solution that would allow to predict satisfaction of customers of an airline based on individual satisfaction scores and the customers' and flights' features. Three machine learning algorithms, such as naive bayes, decision tree and logistic regression were utilised. The data for training and testing of the models was acquired from a large publicly available dataset of an American airline company. All three algorithms performed well, with the decision tree algorithm achieving 95% testing accuracy. The potential of machine learning applications in customer relations as well as directions for future research are discussed.

Introduction and background
In the modern world with growing service economy sector (Buera & Kaboski, 2012) customer satisfaction is one of the key business metrics as it allows to see how many customers are likely to buy the company's products or services again (Ennew et al., 2015). Repetitive purchases or use of company's services by loyal customers generate the company's income.

Customer satisfaction also defines the image of a company in the market and the reviews existing customers leave of the company which in turn can influence if new customers will be attracted or not (Kumar & Zymbler, 2019; Ye et al., 2011). For instance, a study investigating the economic effects of online ratings of restaurants (Ghose & Ipeirotis, 2011) found that positive reviews are associated with better business.

In highly competitive airline industry, customer satisfaction is crucial performance indicator as there are multiple service providers and customers can fly with a different company next time if they are not satisfied with the service. There are also multiple websites that aggregate the reviews of airlines by customers such as Skytrax and OAG which increases the influence of customers' opinion even more. Hence, there is a need to be able to track and possibly predict customer satisfaction to ensure that the company stays afloat.

Machine Learning applications can be extremely helpful when tackling such problems (Lee & Shin, 2020). Models that are able to automatically identify if a customer is satisfied or not can allow to make better marketing decisions (e.g. offering additional services or special discounts for satisfied customers) maximising the profits of the company (Aka et al., 2016; Yeung & Ennew, 2000).

The goal of this study is to develop an algorithm that will predict the general satisfaction of airline customers with the service they were provided based on the scores they evaluate the service with as well as details of their journey (type of travel, flight distance) and customers themselves (age and gender).

Data exploration and feature selection
For this study the data were obtained from Kaggle (2020) website. The two tables already separated for training and testing were merged and shuffled to ensure more randomised results.

The original dataset contains 23 columns and 129880 rows. Approximately 400 rows had missing values in them – since this is a very small proportion compared to the overall size of the dataset, these rows were simply removed. A detailed description of each column is provided in table 1.

Table 1. Variable description (original dataset)

Variable name	Description	Type
Id	A unique ID of each customer	Numeric
Gender	Customer's gender	Categorical (Male, Female)
Age	Customer's age	Numeric
Type of Travel	The reason why the passenger is travelling	Categorical (Personal Travel, Business Travel)
Class	Travel class of the passenger	Categorical (Business, Eco, Eco Plus)
Flight distance	The distance of the flight in kilometres.	Numeric
Inflight wifi service	Satisfaction with the wifi service onboard	Categorical 0 – Not Applicable, 1-5 – Satisfaction level. Will be treated as numeric.
Departure/Arrival time convenient	Satisfaction with the arrival and departure times	Categorical (1-5 – Satisfaction level). Will be treated as numeric.
Ease of Online booking	Satisfaction with online booking service	
Gate location	Satisfaction with the gate location	
Food and drink	Satisfaction with food and drink onboard	
Online boarding	Satisfaction with the online boarding	
Seat comfort	Satisfaction with the seat comfort	
Inflight entertainment	Satisfaction with the entertainment onboard	
On-board service	Satisfaction with the onboard service	
Leg room service	Satisfaction with the leg room available	
Baggage handling	Satisfaction with the baggage handling	
Check-in service	Satisfaction with check-in service	
Inflight service	Satisfaction with inflight service	
Cleanliness	Satisfaction with how clean the aeroplane was	
Departure Delay in Minutes	How many minutes the departure time was delayed by	
Arrival Delay in Minutes	How many minutes the arrival time was delayed by	
Satisfaction	Overall satisfaction level with the airline	Categorical (2 levels – "satisfied" and "neutral or dissatisfied")

As the ID column does not contain any meaningful information it was removed. All other variables are related to the overall customer satisfaction outcome. The descriptive statistics of these variables are presented in Tables 2a and 2b. Individual satisfaction scores, although being categorical in nature, will be treated as numeric in the model development process.

Table 2a. Descriptive statistics of numeric variables in the dataset.

Variable name	Descriptive statistics
Age	Min 7 Mean 39.43 Std 15.12 Max 85
Flight distance	Min 31 Mean 1190.21 Std 997.56 Max 4983
Inflight wifi service	Min 0 Mean 2.73 Std 1.33 Max 5
Departure/Arrival time convenient	Min 0 Mean 3.06 Std 1.53 Max 5
Ease of Online booking	Min 0 Mean 2.76 Std 1.40 Max 5
Gate location	Min 0 Mean 2.98 Std 1.28 Max 5
Food and drink	Min 0 Mean 3.20 Std 1.33 Max 5
Online boarding	Min 0 Mean 3.25 Std 1.35 Max 5
Seat comfort	Min 0 Mean 3.44 Std 1.32 Max 5
Inflight entertainment	Min 0 Mean 3.36 Std 1.34 Max 5
On-board service	Min 0 Mean 3.38

	Std 1.29 Max 5
Leg room service	Min 0 Mean 3.35 Std 1.32 Max 5
Baggage handling	Min 0 Mean 3.63 Std 1.18 Max 5
Check-in service	Min 0 Mean 3.30 Std 1.27 Max 5
Inflight service	Min 0 Mean 3.64 Std 1.17 Max 5
Cleanliness	Min 0 Mean 3.29 Std 1.31 Max 5
Departure Delay in Minutes	Min 0 Mean 14.64 Std 37.93 Max 1592
Arrival Delay in Minutes	Min 0 Mean 15.09 Std 38.47 Max 1584

Table 2b. Descriptive statistics of categorical variables in the dataset.

Variable name	**Number of instances in each category**
Gender	Female – 51% Male – 49%
Type of Travel	Business travel – 69% Personal Travel – 31%
Class	Business – 48% Eco – 45% Eco Plus – 7%
Customer Type	Loyal – 83% Disloyal – 17%
Satisfaction (target variable)	Neutral of Dissatisfied – 56% Satisfied – 44%

The histograms showing the distributions of the numeric variables are shown in Figure 1. From the pictures and the tables, it can be seen that age ranges from 7 to 85, being relatively normally distributed. Most of the arrival and departure delays are under 1 hour with a long tail of

extremely rare longer delays. The distribution of flight distance is also skewed to the right, averaging at 1200 kilometres.

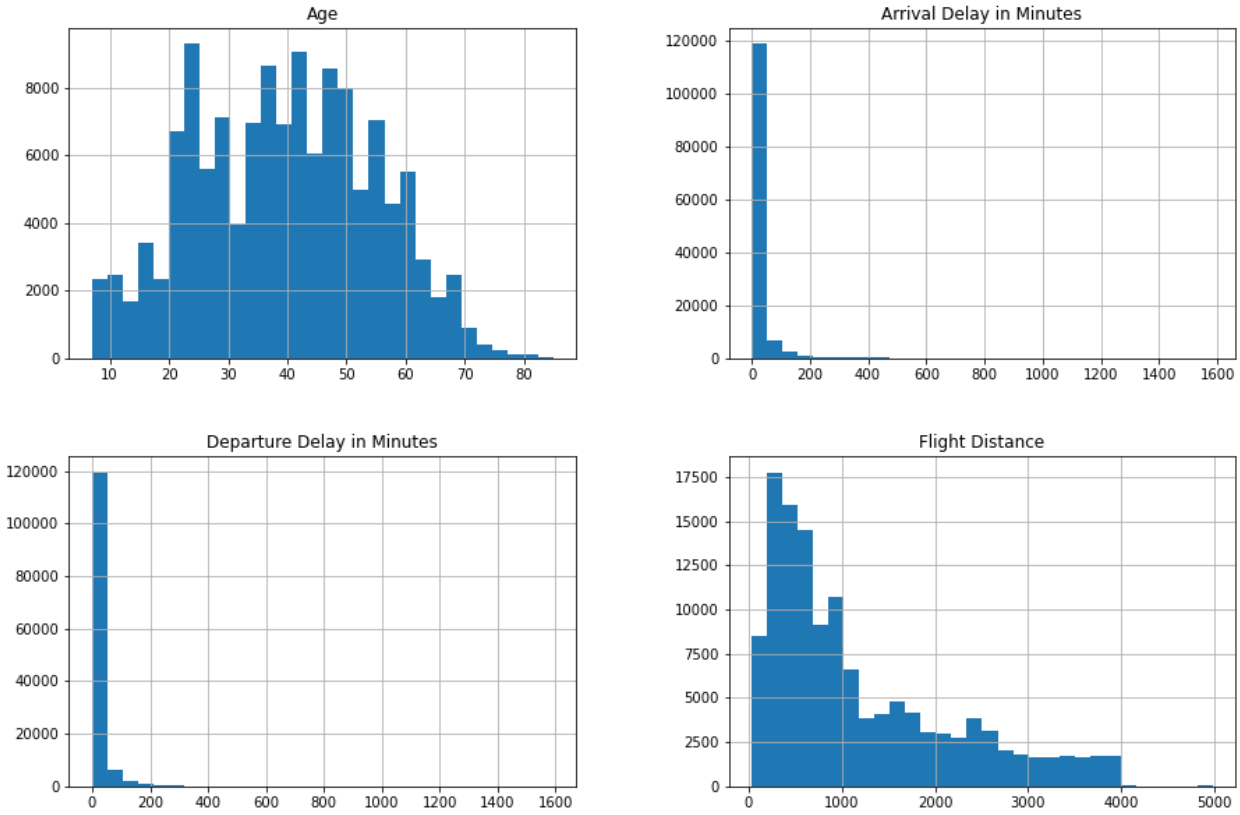

Figure 1. Distributions of numeric variables in the dataset

The distributions of all 14 satisfaction scores are presented in Figure 2. For most of the scores the values are closer to 3-4, with inflight wifi service and ease of online booking averaging below other scores at 2.7 – perhaps these two services require some improvement.

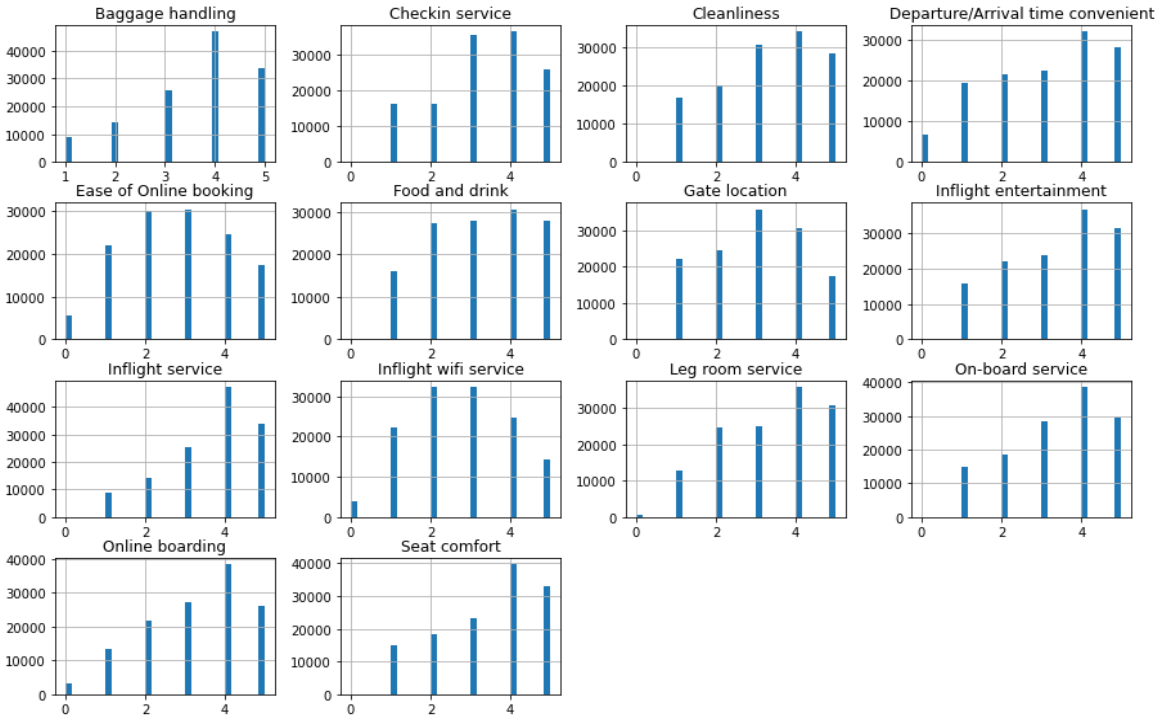

Figure 2. Distributions of satisfaction scores in the dataset

The proportions of categorical variables in the dataset are shown in Figure 3. The proportions of genders are almost equal, there is much more loyal customers, 2/3 of them travel for business purposes and the proportion of business and eco class is almost equal as well.

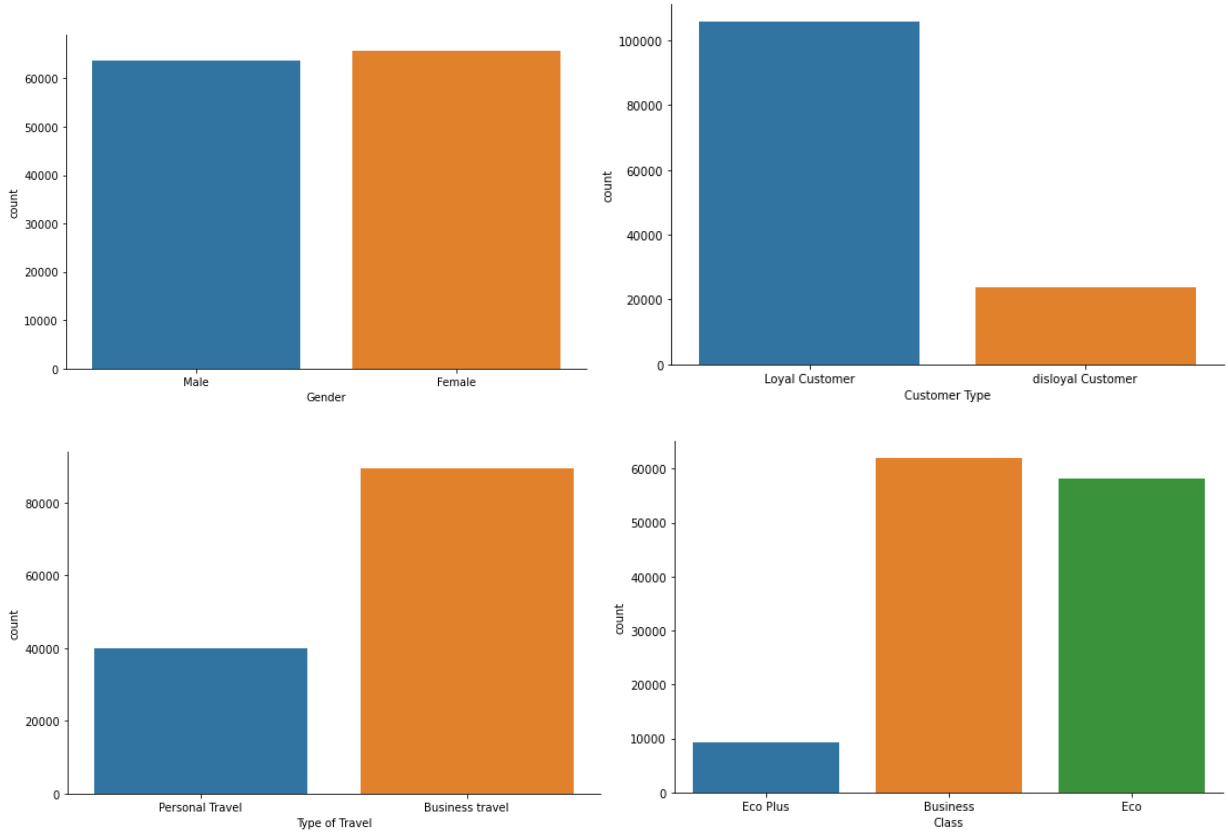

Figure 3. Proportions of categorical variables in the dataset

The proportion of satisfied customers is shown in Figure 4. The proportions are rather balanced, although there is slightly more neutral or dissatisfied (will from now on be treated as dissatisfied) than satisfied. On one hand the data are balanced, on the other hand there is a large proportion of dissatisfied customers, which confirms the need for a solution that would be able to predict customer satisfaction.

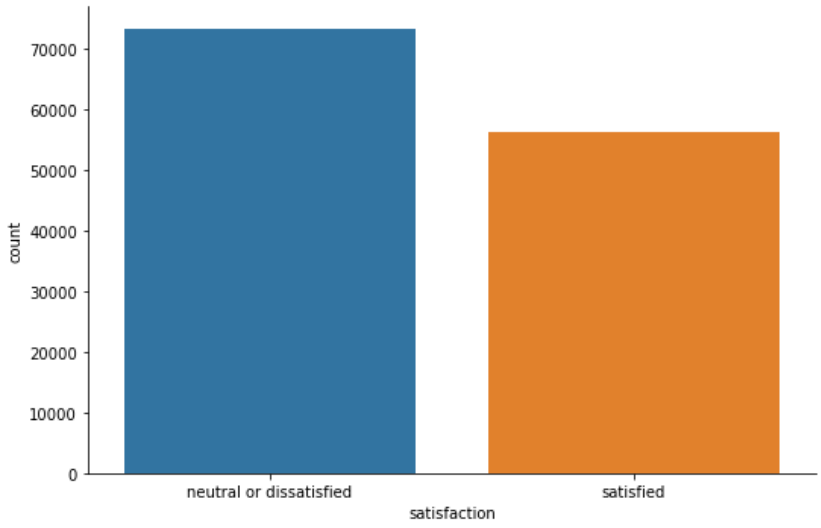

Figure 4. Proportion of satisfied customers

After all pre-processing steps, categorical variables (gender, class, etc.) were recoded into dummy variables and were later treated as numeric in the model development process. All the variables were standardised prior to model development.

The correlation between the variables was explored using a correlation matrix and variance inflation factor for each variable (Figures 5 and 6).

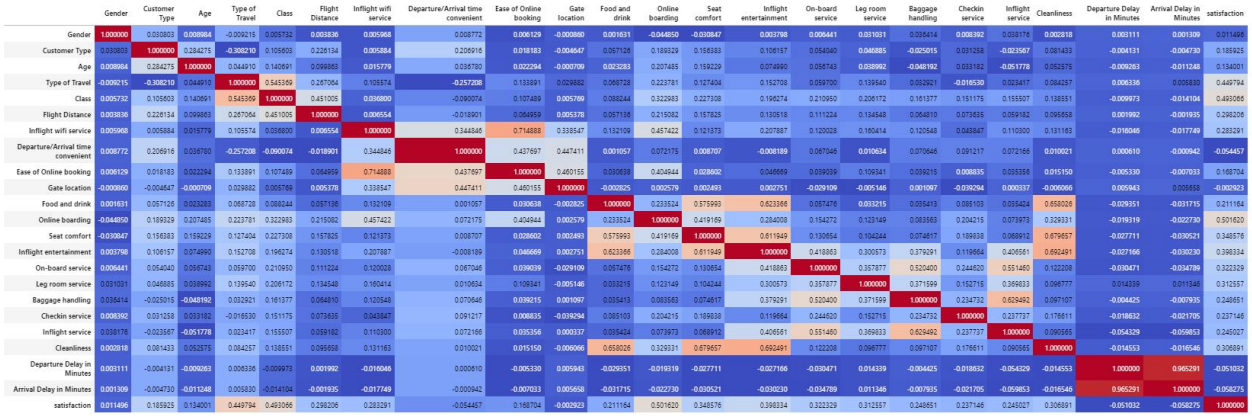

Figure 5. Correlation matrix

	feature	VIF
0	Gender	1.902693
1	Customer Type	2.833467
2	Age	1.115047
3	Type of Travel	3.909823
4	Class	4.219434
5	Flight Distance	1.214850
6	Inflight wifi service	2.435016
7	Departure/Arrival time convenient	1.667838
8	Ease of Online booking	2.695322
9	Gate location	1.506366
10	Food and drink	2.171993
11	Online boarding	1.934397
12	Seat comfort	2.394158
13	Inflight entertainment	3.835170
14	On-board service	1.771479
15	Leg room service	1.307681
16	Baggage handling	1.906549
17	Checkin service	1.222506
18	Inflight service	2.074780
19	Cleanliness	2.857515
20	Departure Delay in Minutes	14.669587
21	Arrival Delay in Minutes	14.682873

Figure 6. Variance Inflation Factor (VIF) table

Both metrics show that there is a high correlation between Departure and Arrival Delay which is essential since the arrival time depends on departure time. In order to avoid multicollinearity, Arrival Delay in Minutes will be removed, as the departure delay already provides the information about the change between the planned and actual departure and arrival times.

There is a correlation of 0.7 between the Ease of Online Booking and Inflight Wifi Service variables, however since it is not >0.7 it can be tolerated (Dormann et al., 2013).

After all these steps, three machine learning models were developed.

Model Development & Analysis

The purpose of this study is to develop an algorithm for binary classification of customers into satisfied and not satisfied based on multiple numeric features. For this problem three binary classification methods were used – Naive Bayes (NB), Decision Trees (DT) and Logistic Regression (LR). The same dataset, split into training (70%) and test (30%) subsets was used to train and test all three algorithms. The specific model tuning procedures and resulting accuracy for each algorithm are described further.

Naive Bayes

Naive Bayes is a classification algorithm that is based on Bayes Theorem and relies on the assumption that all of the features are independent from each other (Tang et al., 2016).

This algorithm has been applied in a study predicting satisfaction of hospitality services' clients through text classification of customer reviews (Sánchez-Franco et al., 2019). In this study Bayesian classifier was applied directly to predict if a customer is satisfied or not.

A Gaussian Bayesian classifier was developed, the confusion matrix with the predictions made by this classifier is presented in Figure 5. The number of false positives and false negatives is similar, the overall accuracy of the Bayesian classifier was 86%.

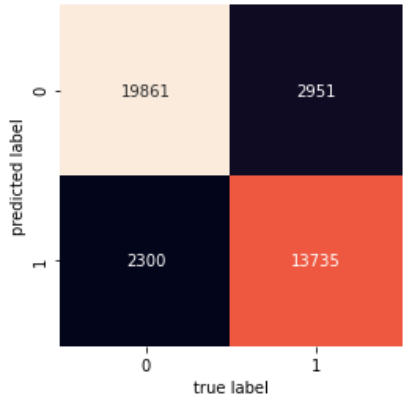

Figure 5. Confusion matrix of Gaussian Bayesian classifier

Decision Tree

Decision tree is an algorithm that continuously splits the data based on a certain parameter that most likely predicts the outcome until all the data are split to each individual observation or until the point of pre-defined maximum splits is reached (Charbuty & Abdulazeez, 2021). The main advantage of this algorithm is its high interpretability as every split can be formulated as a logical rule (e.g. "age > 30", "distance travelled < 100 km" etc).

A decision tree model was developed to predict customer satisfaction, the confusion matrix showing the predictions made with the decision tree is presented in Figure 6. The initial accuracy of the model was already 94.6%.

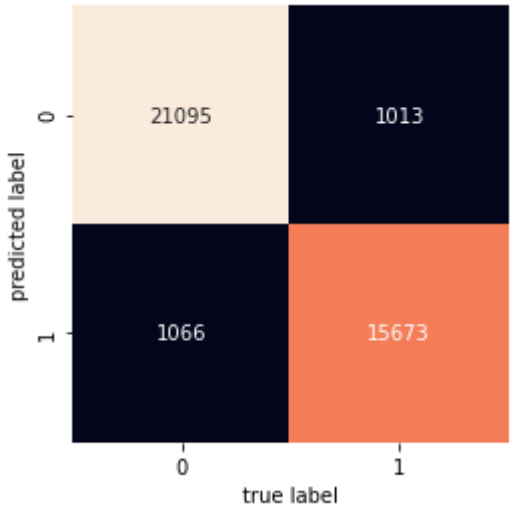

Figure 6. Confusion matrix of the decision tree

The model was further tuned by trying both Gini and entropy information criteria and by adjusting the maximum depth parameter of the tree. Gini criterion yielded higher accuracy and restricting the maximum depth to 20 only increased the accuracy by 0.4%, with lower numbers of maximum depth resulting in ever smaller accuracy increases.

Logistic regression

Logistic Regression is a classification algorithm that outputs the probability of an observation belonging to a particular class based on its features (Nick & Campbell, 2007). A simple logistic regression model was developed, the confusion matrix with the predictions made by this model is presented in Figure 8. The accuracy of this model is 82.6%.

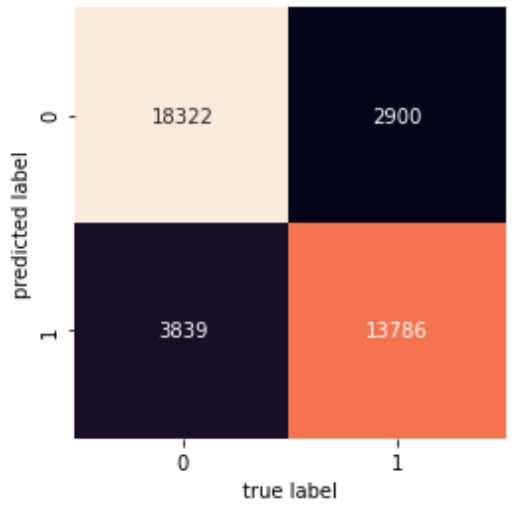

Figure 7. Confusion matrix of the logistic regression

The optimal logistic regression model was found by creating several models with different solvers and different values of C (from 100 to 0.01). The highest accuracy was achieved by the model using C of 0.1, l2 penalty, and liblinear solver and was equal to 87.3%.

Results

The highest accuracy of 95% and the best overall results were achieved by the decision tree algorithm. Given the high interpretability of the decision tree algorithm (Charbuty & Abdulazeez, 2021) this solution seems to be the best for this particular problem. It could be further developed to formulate a set of logical rules that lead to a satisfied customer and then target those particular parameters to increase the satisfaction rate among customers.

Other algorithms performed decently as well, there could be potential in these algorithms for similar problems.

Discussion, Conclusions and Future Work

Machine learning algorithms have great potential in assisting businesses in tackling their problems, making predictions about the future, and tracking the most important business indicators to make better decisions.

The developed solution using decision trees achieves the optimal mixture of accuracy and interpretability (Lee & Shin, 2020), providing an interpretable tool to predict satisfaction of airline passengers with the company's services. Naive Bayes and Logistic Regression models also demonstrated commendable performance and could be considered in future studies for similar problems.

There are other classification algorithms that were not used in this study but could also be tested in application to this problem. As the data that was used for model development in this study describes only one airline company, future studies, exploring either data from a different company or multiple companies are needed before general conclusions about the industry could be made. Additionally, future research could focus on analysing the text of the reviews, as was already done in other studies (Sánchez-Franco et al., 2019) as numeric data with different satisfaction scores is harder to obtain as customers need to fill in surveys and they might not always be willing to do so.

References

Aka, D. O., Kehinde, O. J., & Ogunnaike, O. O. (2016). Relationship Marketing and Customer Satisfaction: A Conceptual Perspective. *Binus Business Review*, *7*(2), 185. https://doi.org/10.21512/bbr.v7i2.1502

Buera, F. J., & Kaboski, J. P. (2012). The Rise of the Service Economy. *American Economic Review*, *102*(6), 2540–2569. https://doi.org/10.1257/aer.102.6.2540

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Ennew, C. T., Binks, M. R., & Chiplin, B. (2015). Customer Satisfaction and Customer Retention: An Examination of Small Businesses and Their Banks in the UK. In E. J. Wilson & W. C. Black (Eds.), *Proceedings of the 1994 Academy of Marketing Science (AMS) Annual Conference* (pp. 188–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-13162-7_49

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188

Kaggle (2020). Airline Passenger Satisfaction. Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction (Accessed on 02.05.2022)

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, *6*(1), 62. https://doi.org/10.1186/s40537-019-0224-1

Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), 157–170. https://doi.org/10.1016/j.bushor.2019.10.005

Nick, T. G., & Campbell, K. M. (2007). Logistic Regression. In W. T. Ambrosius (Ed.), *Topics in Biostatistics* (Vol. 404, pp. 273–301). Humana Press. https://doi.org/10.1007/978-1-59745-530-5_14

Sánchez-Franco, M. J., Navarro-García, A., & Rondán-Cataluña, F. J. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research*, *101*, 499–506. https://doi.org/10.1016/j.jbusres.2018.12.051

Tang, B., Kay, S., & He, H. (2016). Toward Optimal Feature Selection in Naive Bayes for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2508–2521. https://doi.org/10.1109/TKDE.2016.2563436

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, *27*(2), 634–639. https://doi.org/10.1016/j.chb.2010.04.014

Yeung, M. C. H., & Ennew, C. T. (2000). From customer satisfaction to profitability. *Journal of Strategic Marketing*, *8*(4), 313–326. https://doi.org/10.1080/09652540010003663