**A comparative analysis of machine learning algorithms for predicting airline customer satisfaction**

Zhibo Liu

210164972

Word count: 2583

# Abstract

Customer satisfaction has long been one of the most valued types of feedback in the service industry. Especially in the airline industry, satisfaction affects a company's revenue, so it is important to be able to predict satisfaction. The purpose of this study is to use support vector machines and random forest algorithms to predict satisfaction. Use precision and ROC curves. Results The random forest model using the information gain ratio kernel yielded the best accuracy of 91.243%.

Content

[Abstract 1](#_Toc100422892)

[1. Introduction 3](#_Toc100422893)

[2. Data Mining Theory 4](#_Toc100422894)

[2.1 literature review 4](#_Toc100422895)

[2.2 support vector machine 5](#_Toc100422896)

[2.3 random forest 5](#_Toc100422897)

[2.4 evaluation 6](#_Toc100422898)

[3. Data Exploration and Preparation 6](#_Toc100422899)

[3.1 data exploration 6](#_Toc100422900)

[3.2 data preparation 7](#_Toc100422901)

[4. Experiment setup 8](#_Toc100422902)

[4.1 SVM model 10](#_Toc100422903)

[4.2 random forest model 12](#_Toc100422904)

[5. Results and Discussion 13](#_Toc100422905)

[5.1 chi-square result 13](#_Toc100422906)

[5.2 SVM model result 14](#_Toc100422907)

[5.3 random forest model result 16](#_Toc100422908)

[5.3.1 Gini Index kernel 16](#_Toc100422909)

[5.3.2 Information Gain kernel 17](#_Toc100422910)

[5.3.3 Information Gain Ratio kernel 18](#_Toc100422911)

[5.4 discussion 19](#_Toc100422912)

[6. Conclusion and reflections 20](#_Toc100422913)

[Reference: 21](#_Toc100422914)

[Appendix 24](#_Toc100422915)

# 1. Introduction

Various service industries have been focusing on how to further improve their service quality, and customer satisfaction has always been a key factor in measuring their service quality. Therefore, researchers in the past have tried to investigate the key factors affecting customer satisfaction and whether customer satisfaction can be estimated and predicted. For example, Hwang et al. (2020) pointed out in their research that the rate of repeat customers can represent the quality of service. Thus, they estimated the frequency of customer return visits in their research and made predictions by using a variety of machine learning methods. Machine learning is also used to estimate customer satisfaction in Imtiaz and Ben (2020), but the study focuses on the smartphone industry. Comments related to five popular brands in five industries were extracted on Twitter and modelled using five machine learning methods. In addition, many scholars have conducted research on customer satisfaction prediction in various industries and proved the high efficiency of machine learning methods in prediction (Kumar & Zymbler, 2019; Siebert et al., 2021; Tian et al. al., 2021; Lee, Kim & Park, 2021; Oh et al., 2022).

The data mining task in this study is to use two machine learning algorithms on the KNIME platform to predict customer satisfaction and to evaluate and compare the models built by the two algorithms. The dataset AirlineSatisfactionData.csv used contains 23 variables, and the target predictor in this study is the Satisfaction variable. It is a binary variable that indicates whether the customer is satisfied with the flight. Two machine learning algorithms will be used to make predictions, and support vector machines and random forests. Once the results are obtained, they will be evaluated and compared together using the accuracy and ROC curves.

To solve this research goal, the following parts of this paper will be divided into an introduction to data mining theory, data exploration and preparation, experimental setup on KNIME, results and discussion, and finally the conclusion.

# 2. Data Mining Theory

## 2.1 literature review

In this study, support vector machines and random forest algorithms will be used for model building to predict customer satisfaction. This is because scholars have demonstrated the efficiency of these two algorithms in past research (Kim et al., 2019; Tiwari et al.,2019; Kumar & Zymbler, 2019; Imatiaz & Ben,2020; Hwang et al., 2020; Khaild et al., 2019; Lee, Kim & Park, 2021). For example, Imatiaz and Ben's (2020) study on smartphone user satisfaction used multiple machine learning algorithms and compared them. Their research problem is also a binary class classification problem, intending to predict whether a customer's review is positive or negative. Support Vector Machines, MLP Neural Networks, Naive Bayes, Decision Trees, and Random Forests are used. The results say that SVM is the best performing model with an accuracy of 66.7%. Similarly, in the study by Kumar and Zymbler (2019), they wanted to predict airline customer satisfaction, like the objective of this study. And the target features they hope to predict in the data obtained from the Twitter platform are binary as in this study. They show that for binary class classification problems, support vector machines using margin thresholds or hyperplanes for classification are popular and effective. Therefore, this means that support vector machines may work well for binary class classification problems.

In addition to it, more machine learning algorithms have been used in previous research on predicting flight customer satisfaction, of which the random forest algorithm has shown remarkable performance (Hwang et al., 2020). Support vector machines and other algorithms such as decision trees, logistic regression, etc. are also used in their research. These algorithms are commonly used for binary classification tasks (Kim et al., 2019). Finally, the random forest algorithm has the highest accuracy in multiple dimensions. Research by Tiwari et al. (2019) and Lee, Kim, and Park (2021) also supports this conclusion, with random forests working better than multiple algorithms in predicting satisfaction. Even the accuracy of the random forest model in Lee, Kim and Park's study is as high as 96.09%.

In conclusion, support vector machines and random forests generally have better performance in predicting binary classification problems like this study based on the results of past research.

## 2.2 support vector machine

Support vector machine is a powerful supervised algorithm that usually works very well in classification problems (Saini, 2021). Support vector machines handle classification problems by finding a hyperplane between two classes in the data that best distinguishes the two classes. It finds the maximum distance between two classes by finding the maximum margin between the hyperplanes. Use kernels when dealing with nonlinear datasets, such as polynomial kernels, sigmoid kernels, and the most used RBF kernels in classification.

## 2.3 random forest

The decision tree is consistent with all forms of tree representation, starting from the starting node root node through each internal node to perform tests on features, and finally reaching the termination node. While classifying features at internal nodes, various impurity measures such as Gini and entropy can be used (Kelleher, Namee & D'Arcy, 2015). The random forest algorithm that will be used in this study is a decision tree-based bagging ensemble algorithm. It randomly generates multiple decision trees, and then the decision tree uses a random subset of features to get the result. Finally, the result of a random forest consists of the results of each decision tree (Sharma, 2020).

## 2.4 evaluation

Both accuracy and ROC curves will be used when evaluating the model. The confusion matrix is obtained after the conclusion is drawn, and the accuracy is the ratio of all true values to all values. The value of this relative ratio is important when evaluating due to different criteria (Sitara, 2021). The false-positive rate on the x-axis of the ROC curve is the ratio of FP to the sum of FP and TN in the confusion matrix, and the y-axis is the ratio of the true positive rate, TP to the sum of TP and FN (Hand, 2009). The larger the area between the curve and the X-axis, the better the model.

# 3. Data Exploration and Preparation

## 3.1 data exploration

At the beginning of all work, this study will first survey the flight satisfaction dataset. Therefore, observe first after reading and writing the data set, as shown in Figure 1.

Figure 1 read dataset

图形用户界面, 应用程序

描述已自动生成

The dataset has a total of 23 variables and 5196 rows of data. In addition to the five nominal class string format variables, the other 17 are integer variables. The distribution of each variable can be seen in Figures 26 and 27. And the range of values of multiple variables is quite different from other variables, for example, the value of the flight distance variable ranges from 56 to 4983. The target feature in this study is satisfaction, it is a binary variable, two values are neutral or dissatisfied and satisfied, respectively.

## 3.2 data preparation

From the missing value statistics in Figure 3, there are four variables with missing values. The online check\_in variable has the most missing values ​​with 1531 missing values. Such many missing values ​​can have a huge impact on the model. Therefore, following the "80% rule" proposed by Bijlsma et al. (2006), the number of online check\_in variables that do not contain missing values ​​only accounts for 70.53% of the total, which is lower than 80%, so this variable can be deleted without affecting model effect. The other three variables with missing values ​​were 747 missing in Flight Distance, 747 missing in Departure Delay in Minutes, and 15 missing in Arrival Delay in Minutes. The missing values ​​of these three variables are filled by calculating the rounded mean.

As mentioned earlier, the ranges of multiple continuous features in the dataset are very different. As pointed out by Kelleher, Namee, and D'Arcy (2015) this may hinder machine learning algorithms from building accurate models, and normalization can be used to solve this problem. This changes continuous features within a specified range while maintaining the relationship between feature values. Therefore, this study uses Min-Max Normalization to control the value of the feature between 0 and 1.

# 4. Experiment setup

In this section, we will introduce the various nodes used in kNIME. First, use the CSV Reader and the Statistics node to read and write the data file and review the descriptive analysis mentioned above. The Missing Value and Column Filter nodes were then used to fill in the missing values ​​of the three variables with the rounded mean and to remove the Online check-in variables. As shown in figure 2.

Figure 2 Missing Value node

图形用户界面, 应用程序

描述已自动生成

As mentioned in the previous section, the feature range is too large. Here, the Normalizer node is used to normalize the data, and the statistic node is used to observe the normalized data, as shown in figure 3.

Figure 3 normalized data

图形用户界面, 应用程序, Excel

描述已自动生成

Next, to select the most suitable features for building a better model and avoid the feature variables with high correlation affecting the model, the Linear Correlation and Correlation filter nodes are used to first obtain the matrix of the correlation of each feature, then remove features with a correlation higher than 0.8, as shown in Figure 4.

.

Figure 4 correlation

图片包含 散点图

描述已自动生成

Apart from that, here the author wants to further investigate whether there is any relationship between the individual features and target feature satisfaction. Therefore, the Crosstab node is used here, and the chi-square test can detect whether there is a direct and significant relationship between the respective variables and the target variable (Suresh, 2019). The null hypothesis is that the respective variables have no significant relationship with the satisfaction variable. After the inspection, use the Column Filter to filter out the features that have no significant relationship and then use the CSV writer to export the data file to create the next step.

## 4.1 SVM model

Use the CSV Reader node to read the data file just exported before building the model. In machine learning algorithms, the adjustment of hyperparameters is a very critical step, and changes in hyperparameters may affect the performance of the model. Therefore, in this study, the model learning process is placed between the Parameter Optimization Loop Start and Parameter Optimization Loop End nodes. When building the support vector machine model, set the range of the hyperparameter sigma in the Parameter Optimization Loop Start so that the algorithm uses each parameter to build the model, and finally obtains a parameter that can produce the best model, as shown in the following figure 5.

Figure 5 best parameter



The X-Partitioner and X-Aggregator nodes are used in the Parameter Optimization Loop to use the cross-validation method. Many features can lead to overfitting, so using cross validation ensures that the model generalizes better on the data, while also helping to find the best model (Goyal, 2021). As shown in figure 6 below, 5-fold cross validation is set up in the node, and stratified sampling is used to ensure uniform sub-data.

Figure 6 cross-validation

图形用户界面, 文本, 应用程序

描述已自动生成

The model is then built using the SVM Learner and SVM Predictor nodes in cross validation. Choose to use the RBF kernel in SVM, which is rarely used due to the inefficiency of other kernels compared to the RBF kernel (Saini, 2021). And the kernel\_param\_sigma parameter is used in Parameter sigma parameter control set in Optimization Loop Start. As Figures 7 and 8 are shown below. Finally, use the ROC Curve node and Score node to evaluate the performance of the model.

Figure 7 sigma

图形用户界面

中度可信度描述已自动生成

Figure 8 kernel selection

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

## 4.2 random forest model

The establishment process of the random forest model is the same as that of SVM except for the hyperparameter range settings and machine learning algorithm nodes. The hyperparameter range settings are shown in Figure 9:

Figure 9 random forest parameter

图形用户界面, 文本, 应用程序

描述已自动生成

For a better model, the most appropriate split criterion method should be selected in the Random Forest Learner node. Kelleher, Namee, and D'Arcy (2015) pointed out that the effectiveness of feature selection methods may be affected by the data, so to find the best model one should try different split criteria. Therefore, this study uses three split criteria to build the model, and finally compares the model performance.

# 5. Results and Discussion

## 5.1 chi-square result

The results of the chi-square test are shown in the figure below. The test results of the four variables Gender, id, Flight Distance and Departure Delay In Minutes indicate that the P-value is much greater than 0.05, so the null hypothesis cannot be rejected for these variables, that is, these four variables are satisfactory with the target feature. degree is not statistically significant. Therefore, these four variables are not used when building the model.

Table 1 Chi\_square test result

|  |  |  |  |
| --- | --- | --- | --- |
| Chi-square | DF | Value | Prob |
| Gender | 1 | 2.669 | 0.1023 |
| Customer Loyalty | 1 | 146.0653 | 1.26E-33 |
| Age | 73 | 394.727 | 1.16E-45 |
| Type of Travel | 1 | 1,066.9476 | 5.04E-234 |
| Class | 2 | 1,152.5229 | 5.41E-251 |
| Departure/Arrival time convenient | 5 | 35.9271 | 9.82E-7 |
| Ease of Online booking | 5 | 468.1516 | 5.96E-99 |
| Gate location | 4 | 123.3481 | 1.03E-25 |
| Food and drink | 5 | 259.0252 | 6.36E-54 |
| Seat comfort | 4 | 780.2506 | 1.46E-167 |
| Inflight entertainment | 4 | 927.4484 | 1.88E-199 |
| On-board service | 4 | 592.0553 | 8.12E-127 |
| Leg room service | 5 | 522.3608 | 1.19E-110 |
| Baggage handling | 4 | 460.6974 | 2.11E-98 |
| Checkin service | 4 | 367.8368 | 2.47E-78 |
| Inflight service | 4 | 407.5593 | 6.47E-87 |
| Cleanliness | 4 | 584.7353 | 3.12E-125 |
| Departure Delay in Minutes | 186 | 136.8086 | 0.9973 |
| id | 5195 | 2,639.6632 | 1 |
| Flight Distance | 1643 | 1,161.5192 | 1 |

## 5.2 SVM model result

After the parameter optimization loop ends, the parameters that can generate the best support vector machine model are obtained, as shown in Figure 10. The results of the best model produced when sigma is set to 0.6 are shown in Figure 11. There are 4642 correct values for all predictions, with an accuracy of 89.338%. Figures 12 and 13 show the results of the ROC curves. The area under the ROC curve is 0.953, which is very close to 1, which means that the model generated by the support vector machine can produce a good prediction effect.

Figure 10 SVM best parameter

图形用户界面, 文本, 应用程序

描述已自动生成

Figure 11 SVM confusion matrix

表格

描述已自动生成

Figure 12 SVM ROC curve

图片包含 图表

描述已自动生成

Figure 13 area under the curve



## 5.3 random forest model result

### 5.3.1 Gini Index kernel

The model uses the Split criteria of the Gini Index, and the parameter optimization loop obtains the parameters that can produce the best model, as shown in Figure 14. The results of the best model produced when the number of models was set to 60 are shown in Figure 15. Combining Figure 16 and Figure 17, the accuracy of the model is 90.916%, and the area between the ROC curve and the X-axis is 0.97, which all verify the superior performance of the model.

Figure 14



Figure 15

表格

中度可信度描述已自动生成

Figure 16

图片包含 图表

描述已自动生成

Figure 17



### 5.3.2 Information Gain kernel

This model uses the Split criteria of Information Gain, the parameters of the best model are shown in Figure 18, and the results are shown in Figure 19. Combining Figure 20 and Figure 21, the accuracy of the model is 91.109%, and the area between the ROC curve and the X-axis is 0.969, which is also a high-performance model.

Figure 18



Figure 19

表格

描述已自动生成

Figure 20

图片包含 图表

描述已自动生成

Figure 21



### 5.3.3 Information Gain Ratio kernel

The model uses the Information Gain ratio, and the optimal parameters are shown in Figure 22. Figures 23, 24 and 25 show that the accuracy of the model is 91.243%, and the area between the ROC curve and the X-axis is 0.972, which is the best model.

Figure 22



Figure 23

图形用户界面, 应用程序

描述已自动生成

Figure 24

图片包含 图表

描述已自动生成

Figure 25



## 5.4 discussion

From the results of these models, among the models using the three split criteria methods, the model using the information gain ratio produced the best results with the highest accuracy but was not far behind the other two models. Comparing the models generated by the two algorithms, all random forest models perform better than support vector machine models, which may be because random forests are better at handling multi-featured high-latitude data. However, random forests produce a better performance at the cost of a much longer runtime than SVM models and models that are too complex to interpret.

# 6. Conclusion and reflections

In this study, four models were established using SVM and random forests with three kernels. The results showed that the random forest model using the Information Gain Ratio kernel had the best performance, with an accuracy of 91.243%. Although the effect of the model is very objective, it can be further improved. The features have been screened by correlation and chi-square test before the model is built in this study, but the high-latitude pair data can still be reduced by PCA (Nasution, Sitompul & Ramli, 2018), which can further remove irrelevant data, allowing the algorithm to classify better to improve accuracy.

# Reference:

Bijlsma, Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., van Ommen, B., & Smilde, A. K. (2006). Large-Scale Human Metabolomics Studies:  A Strategy for Data (Pre-) Processing and Validation. Analytical Chemistry (Washington), 78(2), 567–574. https://doi.org/10.1021/ac051495j

Cavalcante Siebert, Bianchi Filho, J. F., Silva Júnior, E. J. da, Kazumi Yamakawa, E., & Catapan, A. (2021). Predicting customer satisfaction for distribution companies using machine learning. International Journal of Energy Sector Management, 15(4), 743–764. <https://doi.org/10.1108/IJESM-10-2018-0007>

Goyal, C. (2021). Importance of Cross Validation: Are Evaluation Metrics enough? *Analytics Vidhya.* Retrieved from: <https://www.analyticsvidhya.com/blog/2021/05/importance-of-cross-validation-are-evaluation-metrics-enough/>

Hand. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hwang, Kim, J., Park, E., & Kwon, S. J. (2020). Who will be your next customer: A machine learning approach to customer return visits in airline services. Journal of Business Research, 121, 121–126. <https://doi.org/10.1016/j.jbusres.2020.08.025>

Imtiaz, & Ben Islam, M. K. (2020). Identifying Significance of Product Features on Customer Satisfaction Recognizing Public Sentiment Polarity: Analysis of Smart Phone Industry Using Machine-Learning Approaches. Applied Artificial Intelligence, 34(11), 832–848. <https://doi.org/10.1080/08839514.2020.1787676>

Kelleher, Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies. The MIT Press.

Khalid, Ashraf, I., Mehmood, A., Ullah, S., Ahmad, M., & Choi, G. S. (2020). GBSVM: Sentiment classification from unstructured reviews using ensemble classifier. Applied Sciences, 10(8), 2788. <https://doi.org/10.3390/APP10082788>

Kim, J., Bae, K., Park, E., & del Pobil, A. P. (2019). Who will subscribe to my streaming channel? The case of twitch. In Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (pp. 247-251).

Kumar, & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(1), 1–16. <https://doi.org/10.1186/s40537-019-0224-1>

Lee, Ji, H., Kim, J., & Park, E. (2021). What books will be your bestseller? A machine learning approach with Amazon Kindle. Electronic Library, 39(1), 137–151. https://doi.org/10.1108/EL-08-2020-0234

Nasution, Sitompul, O. S., & Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4.5 classification. Journal of Physics. Conference Series, 978(1), 12058. <https://doi.org/10.1088/1742-6596/978/1/012058>

Oh, Ji, H., Kim, J., Park, E., & del Pobil, A. P. (2022). Deep learning model based on expectation-confirmation theory to predict customer satisfaction in hospitality service. Information Technology & Tourism, 24(1), 109–126. <https://doi.org/10.1007/s40558-022-00222-z>

Sitara, A. (2021). Interpretation of Performance Measures to Evaluate Models. Retrieved from: [https://www.analyticsvidhya.com/blog/2021/03/interpretation-of-performance-measures-to-evaluate-model/#](https://www.analyticsvidhya.com/blog/2021/03/interpretation-of-performance-measures-to-evaluate-model/)

Saini, A. (2021). Support Vector Machine (SVM): A Complete guide for beginners. Retrieved from: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/#h2\_15

Sharma, A. (2020). Decision Tree vs. Random Forest – Which Algorithm Should you Use?. *Analytics Vidhya.* Retrieved from: <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#h2_4>

Suresh, A. (2019). What is the Chi-Square Test and How Does it Work? An Intuitive Explanation with R Code. *Analytics Vidhya.* Retrieved from: <https://www.analyticsvidhya.com/blog/2019/11/what-is-chi-square-test-how-it-works/>

Tian, Zhong, R. Y., Vatankhah Barenji, A., Wang, Y. T., Li, Z., & Rong, Y. (2021). A blockchain-based evaluation approach for customer delivery satisfaction in sustainable urban logistics. International Journal of Production Research, 59(7), 2229–2249. <https://doi.org/10.1080/00207543.2020.1809733>

Tiwari, Pandey, H. M., Khamparia, A., & Kumar, S. (2019). Twitter-based opinion mining for flight service utilizing machine learning. Informatica (Ljubljana), 43(3), 381–386. <https://doi.org/10.31449/inf.v43i3.2615>

# Appendix

Figure 26 Numeric Statistics

表格

低可信度描述已自动生成

Figure 27 Nominal Statistics

表格

描述已自动生成