International Journal of Internet, Broadcasting and Communication Vol.12 No.4 26-32 (2020)
http://dx.doi.org/10.7236/IJIBC.2020.12.4.26

IJIBC 20-4-4

Correlation Analysis of Airline Customer Satisfaction using Random Forest with Deep Neural Network and Support Vector Machine Model

[1]Sang Hoon Hong, [2] Bumsu Kim, [3*]Yong Gyu Jung

[1]*Student, Dept. of Medical IT, Eulji University, Korea*
[2]*Director, Div. of Customer &Media, Korea Telecomm, Korea*
[3*]*Professor, Dept. of Medical IT, Eulji University, Korea (Corresponding Author)*
[1]hshjjang10@gmail.com, [2]*ben.kim@kt.com*, [3*]ygjung@eulji.ac.kr

Abstract

There are many airline customer evaluation data, but they are insufficient in terms of predicting customer satisfaction in practice. In particular, they are generally insufficient in case of verification of data value and development of a customer satisfaction prediction model based on customer evaluation data. In this paper, airline customer satisfaction analysis is conducted through an experiment of correlation analysis between customer evaluation data provided by Google's Kaggle. The difference in accuracy varied according to the three types, which are the overall variables, the top 4 and top 8 variables with the highest correlation. To build an airline customer satisfaction prediction model, they are applied to three classification algorithms of Random Forest, SVM, DNN and conduct a classification experiment. They are divided into training data and verification data by 7:3. As a result, the DNN model showed the lowest accuracy at 86.4%, while the SVM model at 89% and the Random Forest model at 95.7% showed the highest accuracy and performance.

Keywords: *Random Forest, Support Vector Machine, SVM, Deep Neural Network, DNN, Correlation Analysis, Airline Customer Satisfaction, Kaggle*

1. Introduction

According to the 2019 Air Passenger Statistics data released by the Ministry of Land and Transport, Infrastructure and Transport, the number of air passengers in 2019 increased by 5% compared to last year, reaching a record high of 12.37 million people annually, and international flights increased by 5.2% compared to the previous year to 9,039 million. The number of passengers is recorded, and air passenger performance, excluding Japan routes, continued to increase for three years from 2016 [1]. In order to attract customers, airlines around the world are striving to improve service quality to increase customer satisfaction, which is directly related to re-use intention as well as price competition [2]. However, by using customer evaluation data on aviation service, customer satisfaction has been improved. The use of data for evaluation and prediction is insufficient. In this paper, we statistically analyzed the correlation between input data and output data using customer evaluation data for airlines, and presented a customer satisfaction prediction model through learning. The predictive model performed pre-processing (Label Encoding) on the collected data, and performed the variable selection process through correlation coefficient

Manuscript Received: August. 3, 2020 / Revised: August. 8, 2020 / Accepted: August. 14, 2020
+Corresponding Author: ygjung@eulji.ac.kr
Tel: +82-31-740-7198
Professor, Dept. of Medical IT, Eulji University, Korea

analysis. Using Random Forest, Deep Neural Network (DNN), and Support Vector Machine (SVM), an airline customer satisfaction prediction model is constructed and the accuracy of the model is compared.

2. Related Literature

Satisfaction is a consumer's subjective evaluation of the extent to which desires and demands arising from acquiring or consuming a provided product or service are satisfied, and reusability is the possibility that the customer will continue to use the product or service repeatedly after purchasing [2,3]. To prove the correlation between customer satisfaction and reuse intention, a high correlation between variables is demonstrated by conducting a correlation analysis between customer satisfaction and reuse intention for fast food businesses, pest control businesses, banks, and laundry businesses overseas. In Korea, through various studies [5,6], it is proved that customer satisfaction has a positive effect on loyalty and reuse intention, and the results are shown in Figure 1 [7].

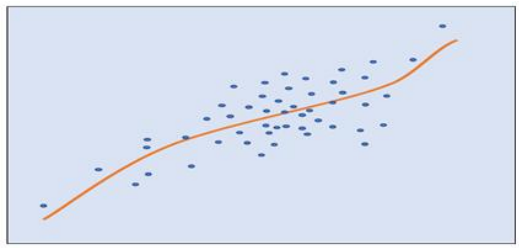

Figure 1. Non-linear Scatter plot chart of customer satisfaction(X) and repurchase intention (Y)

3. Experimental Process

The data on this study are airline customer satisfaction data, which are including gender, age, food and beverages provided on board, seat comfort, etc. They are collected by Google's Kaggle [8]. Airline customer satisfaction prediction model implementation, and performance evaluation are conducted with pre-processing.

3.1 Airline customer evaluation data

Table 1. List of input data

No	Type	Variable
0	Categorical	Gender
1		Customer Type
2		Type of Travel
3		Class
4	Numerical	Age
5		Flight Distance(knot)
6		Departure Delay in Minutes
7		Arrival Delay in Minutes
8	Grade (0~5)	Seat comfort
9		Departure/Arrival time convenient
10		Food and drink
11		Gate location
12		Inflight WIFI service
13		Inflight entertainment
14		Online support
15		Ease of Online booking
16		On-board service
17		Leg room service
18		Baggage handling
19		Check-in service
20		Cleanliness
21		Online boarding

.The calculation result of the predictive model developed in this paper is customer satisfaction, and character-type categorical data consisting of gender, customer type, reason for travel, and seat type, and numeric data consisting of age, flight distance, arrival delay, and departure delay. 22 grade-type (0~5) numeric data consisting of service and convenience are used as input data, as shown in Table 1 above.

3.2 Data preprocessing

Table 2. Label Encoding Data

Variable	Label Encoding
Gender	1 (Male)
	0 (Female)
Customer Type	1 (Loyal Customer)
	0 (disloyal Customer)
Type of Travel	1 (Business travel)
	0 (Personal Travel)
Class	2 (Business)
	1 (Eco Plus)
	0 (Eco)
Satisfaction	1 (satisfied)
	0 (dissatisfied)

In this study, 129,487 data are used by deleting all 393 missing values found in the 'arrival delay time' from the evaluation data of 129,880 passengers. Label Encoding is performed to convert four character-type characteristics (gender, customer type, travel purpose, boarding seat) into real-type categorical data, as shown in Table 2 above.

Table 3. Variables and rankings with positive (+) correlation

Rank	Variable	R (+)
1	Inflight entertainment	0.52
2	Ease of Online booking	0.43
3	Online support	0.39
4	On-board service	0.35
5	Online boarding	0.34
6	Leg room service	0.31
	Class	0.31
7	Customer Type	0.29
8	Check-in service	0.27
9	Baggage handling	0.26
	Cleanliness	0.26
10	Seat comfort	0.24
11	Inflight wi-fi service	0.23
12	Age	0.12
	Food and drink	0.12
13	Type of Travel	0.11

Since the accuracy of the prediction model may vary depending on the number and characteristics of variables to be used as input data, in this study, input data is selected according to the results of the correlation analysis used to find out only the correlation between two variables. As a result of the correlation analysis with the output data, the variable for which the positive (+) correlation (0.11 to 0.52) is calculated is calculated as a higher value than the negative (-) correlation (-0.01 to -0.21), so the positive (+) correlation Input data is used for this calculated variable, which is shown in Table 3 above.

Table 4. Dataset for building prediction model

Class	Size of Sample
1 (satisfied)	70,882
0 (dissatisfied)	58,605

In this paper, the prediction accuracy is compared by constructing a model that selected the top 4 and 8 variables with the highest correlation (including the top 4) and a model using all 22 variables, and the preprocessed data are total 129,487. Of the output data (whether customer satisfaction or not), there are 70,882 satisfactory data and 58,605 dissatisfied data, which are shown in Table 4 above.

3.3 classification algorithm

3.3.1 Support Vector Machine (SVM)

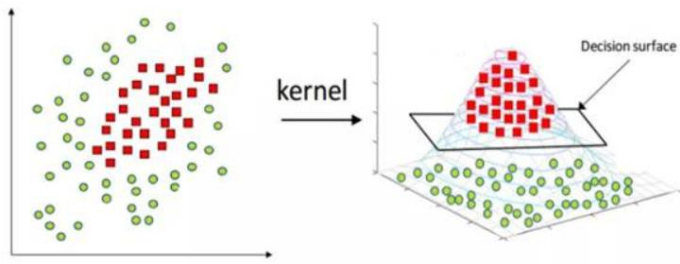

Figure 2. SVM model

SVM is to divide the N-dimensional space into an N-1 dimensional hyperplane by setting a hyperplane to separate the boundaries of the data set. In this paper, a model is constructed with RBF (Radial Basis Function) for kernel, 1.0 for cost and 1/N_features (4, 8, 22) for Gamma as parameters.

3.3.2 Random Forest

Random Forest is a classifier that composes an ensemble through bagging (Bootstrap Aggregation) of multiple decision trees, and two parameters (number of trees, number of randomly selected variables) must be set. In this paper, the number of trees is set to 500, and the designated variable is used as the input data variable, as shown in Figure 3 below.

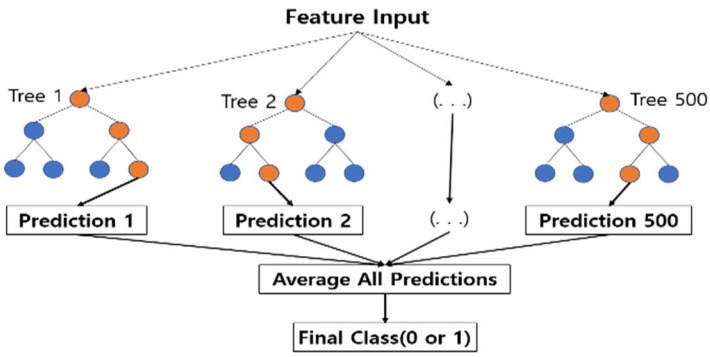

Figure 3. Random Forest Model

3.3.3 Deep Neural Network (DNN)

DNN uses more layers than existing artificial neural networks, and is used for learning pattern recognition and inference through classification and clustering, and the DNN model constructed in this paper is shown in Figure 4.

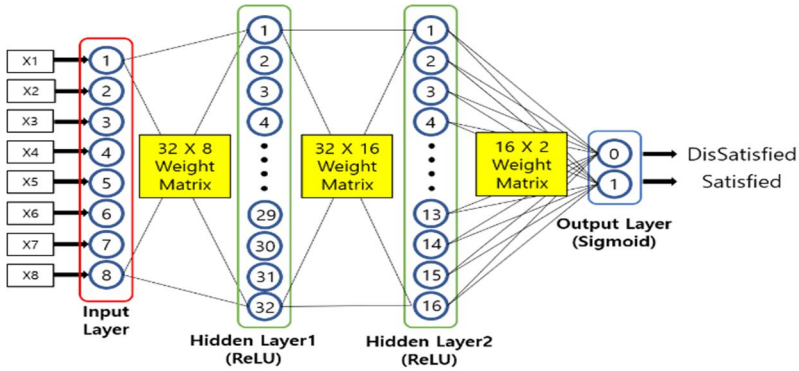

Figure 4. DNN model

4. Evaluation and Discussion

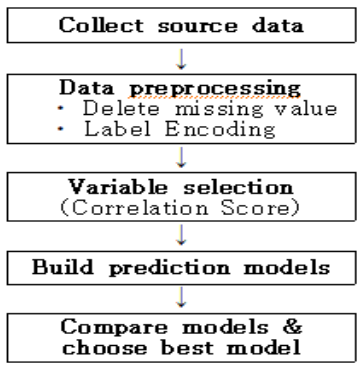

Figure 5. Prediction model construction flowchart

In this paper, in order to prevent model overfitting cases, an experiment is conducted by classifying the data set into training and verification data at a ratio of 7:3 using the 'Hold Out' method, one of the cross-validation methods. The accuracy calculated through each customer satisfaction prediction experiment is compared and analyzed, and

the prediction experiment procedure is shown in Figure 5 above.

Table 5. Accuracy of each prediction model according to variable selection

	No of variable	Random Forest	SVM	DNN
All variable	22	0.957	0.687	0.586
Correlation coefficient	4	0.859	0.858	0.824
	8	0.891	0.890	0.864

When all variables (22) are used in the Random Forest, SVM, and DNN models, accuracy of 95.7%, 68.7% and 58.6% are calculated, respectively, and 85.9% and 85.8% respectively. When the four variables with the highest correlation are selected, 82.4%. When the 8 variables with the highest correlation (including the top 4) are selected, the accuracy of 89.1%, 89%, and 86.4% respectively. When the random forest model is constructed using highly correlated variables, an average of 8.2% lower accuracy is calculated than the model constructed using all variables, and 8 (top 4) are calculated in all models than the 4 highly correlated variables. Including), an average of 3.4% higher accuracy is calculated, as shown in Table 5 above.

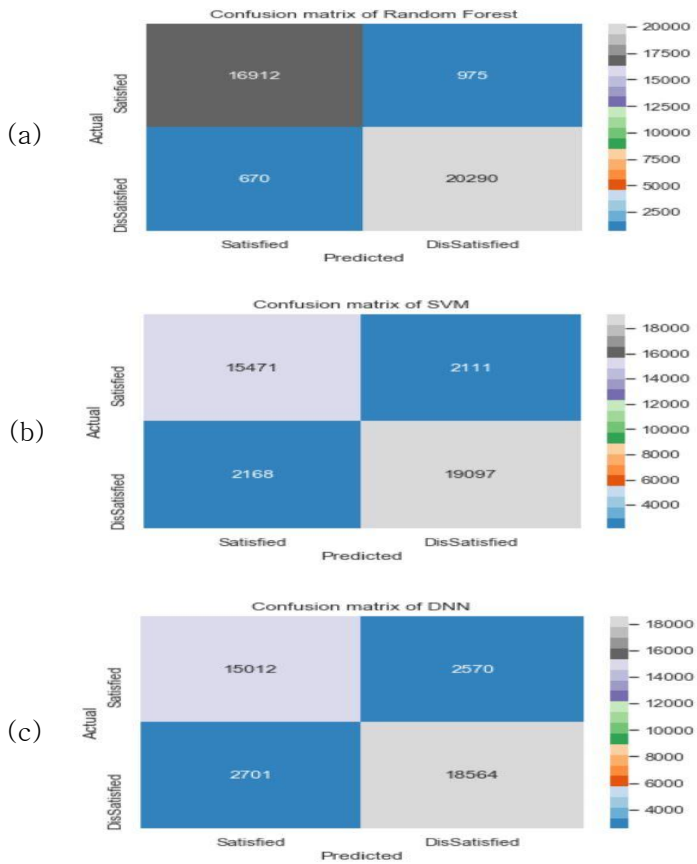

Figure 6. Confusion Matrix of Random Forest (a), SVM (b) and DNN (c)

The confusion matrix is a visualization tool used to see at a glance the performance of supervised learning in the

field of artificial intelligence. If true is said to be true, it is true positive (TP), if true is said to be false is negative, false (NP), if it is said to be true is false positive (FP), and finally false is said to be false, it is expressed as a true negative (TN). In this study, the highest accuracy is calculated from the random forest model among the three classification algorithms, and the airline customer satisfaction prediction model. Confusion matrix is shown in Figure 6 above.

As a result of confirming the Confusion Matrix, it is calculated accuracy in the three classification algorithms, which are the Random Forest, SVM, and DNN prediction model. It is shown as high performance in order, which is consistent with the ranking of accuracy.

5. Conclusion

In this paper, the customer satisfaction prediction model is proposed using customer evaluation data for airlines, and a model is constructed from selected data through a process of variable selection through data preprocessing and correlation analysis. As a result of the experiment, the accuracy of 95.7% is calculated in the random forest model with the highest accuracy and performance. through that experiment result, it is confirmed that the possibility of a customer satisfaction prediction model. In the case of random forest model, the importance of the features used in the experiment can be confirmed. it can also create marketing strategy about customers. Also top 8 with correlation in all prediction models, it is confirmed that an average of 3.4% higher accuracy is calculated than top 4 case. In the future, it plans to experiment using from domestic airlines. And if an airline establishes service management and marketing strategies by further analyzing customer evaluation data and improving the airline customer satisfaction prediction model, it will be useful to help customers re-use services and to gain a relative advantage in the fiercely competitive market.

References

[1] Kyungdoo NAM, Thomas SCHAEFER, “Forecasting international airline passenger traffic using neural networks”, The Logistics and Transportation Review, 1995, 31.3: 239-252.

[2] Insil Park, “Influence of Customer Satisfaction and Reuse Intention on Service Quality of Airline Outsourcing: Focusing on National Airlines”, Tourism Management Research, Vol. 13, No. 39, pp.27-60, 2009.

[3] Czepiel, J. A., Rosenberg, L. J., Akerele, “Perspectives on consumer satisfaction”, AMA Conference Proceedings, pp.119-123, 1997.

[4] Cronin, J, Joseph, Jr. and Steven A, Taylor, "SERVPERF Versus SERVQUAL: Reconciling Performance -Based and Perceptions-Minus- Expectations Measurement of Service Quality", Journal of Marketing, 1994, p.127. DOI: https://doi.org/10.1177/002224299405800110

[5] Seongsuk Ahn, “The Influence of Service Integrity on Customer Satisfaction, Word of Mouth and Reuse Intention: Focused on Airline Service”, Korean Society for Aviation Management, Vol. 16, No. 1, pp.91-106, 2018.

[6] Pan-ho Choi, “A Study on the Influence of Selecting Attributes of Airline on Customer Satisfaction and Loyalty”, Korea Data Analysis Society, Vol. 21, No. 1, pp.305-317, 2019.

[7] Hyeon Mi Yoo, “A Study of the Nonlinear Relationship between Customer Satisfaction and Repurchase Intension”, Journal of Channel and Retailing, 2017, 22(3): 19-38.

[8] NELLER, Todd W. “AI education matters: lessons from a kaggle click-through rate prediction competition”. AI Matters, 2018, 4.2: 5-7. DOI: https://doi.org/10.1145/3236644.3236646

[9] MANGAL, Ankita; KUMAR, Nishant. “Using big data to enhance the bosch production line performance: A kaggle challenge”. 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016. p. 2029-2035. DOI: https://doi.org/10.1109/bigdata.2016.7840826

[10] WAN, Yulai, et al. “Airlines’ reaction to high-speed rail entries: Empirical study of the Northeast Asian market”, Transportation Research Part A: Policy and Practice, 2016, 94: 532-557. DOI: https://doi.org/10.1016/j.tra.2016.10.014