Forecasting Customers Visiting Using Machine Learning and Characteristics Analysis with Low Forecasting Accuracy Days

Takashi Tanizaki, Yuta Hanayama, Takeshi Shimmura

Forecasting Customers Visiting using Machine Learning and Characteristics Analysis with Low Forecasting Accuracy Days

Takashi Tanizaki[1], Yuta Hanayama[1] and Takeshi Shimmura[2]

[1] Graduate School of Systems Engineering, Kindai University,
1 Takaya-Umenobe, Higashi-Hiroshima, 739-2116, Japan
tanizaki@hiro.kindai.ac.jp
[2] Ritsumeikan University,
1-1-1 Nogi-Higashi, Kusatsu 525-8577, Japan

Abstract. In this paper, the number of customers visiting restaurants is forecasted using machine learning and statistical analysis. There are some researches on forecasting the number of customers visiting restaurants using past data on the number of visitors. In this research, in addition to the above data, external data such as weather data and events existing in ubiquitous was used for forecasting. Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Random Forest Regression are used for machine learning, Stepwise is used for statistical analysis. Among above five methods, the forecasting accuracy using Bayesian Linear Regression was the highest. The forecasting accuracy did not tend to improve even if the training data period was extended. Based on these forecasting results, the characteristics of days with low forecasting accuracy are analyzed. It was found that the human psychology around the payday and the reservation customers affected the number of visitors. On the other hand, the weather data such as temperature, precipitation and wind speed did not affect the accuracy.

Keywords: Forecasting Customers Visiting, Machine Learning, Statistical Analysis.

1 Introduction

The service industry is an important industry that accounts for about 70% of Japan's GDP. It has a very high impact on the Japanese economy. However, its productivity is low compared to the manufacturing industry. In many countries, the productivity growth of the service industry is lower than that of the manufacturing industry. Among them, Japan has a big difference in productivity growth, and improving the productivity of the service industry is an important issue for the country. In order to solve such problems, it is important to change the method of business improvement in the service industry from “experience and intuition” to “engineering method”. From the above background, we are researching the improvement of the productivity of the

restaurant using the engineering method. The goal of our research is ways to improve store management by improving employee job arrangements and cocking material orders based on accurate forecasts of restaurant customer numbers. As the first step, we research a method for forecasting the number of customers visiting restaurants.

There are some research papers on forecasting the number of Customer Visiting using machine learning. In [1], a comprehensive literature review and classification of restaurant sales and consumer demand techniques are presented. In addition, the data used for the forecasting in the surveyed literature are also summarized. Multiple Regression, Autoregressive Integrated Moving Average, Artificial Neural Networks, Bayesian Network Model and other methods are used as the forecasting method. The paper reports that it is difficult for forecasters to choose the right technique for their unique situations. Data such as the past number of customers and sales, seasons, and days of the week are used for forecasting. In [2], it is proposed an approach to forecasting how many future visitors would go to a restaurant using big data and supervised learning. It is used for big data involving restaurant information, historical visits, and historical reservations in this method. In the past research, internal data, such as the past number of customers and reservations were used for forecasting, but external data were not used.

From the above, we are researching forecasting methods using internal data such as Point-of-Sales and external data in the ubiquitous environment such as weather, events, etc. in order to improve the accuracy of forecasting. In this paper, we analyze the relation between training data and forecasted results by machine learning and statistical analysis and discuss the characteristics with low forecasting accuracy days.

2 Forecasting Method

In this research, the number of customers visiting is forecasted using machine learning and statistical analysis with internal data and external data in the ubiquitous environment. Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Random Forest Regression are used for machine learning, Stepwise is used for statistical analysis. We used Azure Machine Learning and Python as a machine learning tool and SPSS as a statistical analysis tool.

(1) Bayesian Linear Regression

Bayesian Linear Regression (Bayesian) is a method of applying Bayesian network to machine learning. Bayesian network is a probabilistic model in which conditional dependencies among a plurality of random variables are represented by a graph structure and dependency relationships between the random variables are represented by conditional probabilities [3]. By using the Bayesian network, we can obtain probability distribution about unobserved variables when observing some variables, and treat the value with the highest probability value as the predicted value of that variable.

(2) Boosted Decision Tree Regression

Boosted Decision Tree Regression (Boosted) is a method machine learning using boosting. Boosting refers to a general and provably effective method of producing a very accurate forecasting rule by combining rough and moderately inaccurate rules of

thumb [4]. In this method, as a result of learning using N learning machines, learning is focused on the case by increasing the weight of the incorrectly forecasted.

(3) Decision Forest Regression

Decision Forest Regression (Decision) is a method of machine learning using Random Forest [5]. Random Forest is an ensemble learning method that constructs a forest using multiple decision trees and performs majority decision on the result of learning for each decision tree. In order to prevent extreme bias in learning of each decision tree, learning data used in each decision tree is extracted with randomness. As a result, overfitting is prevented and high generalization performance is obtained.

(4) Random Forest Regression

There are two types of methods in the random forest: “classification” and “regression” [6]. The difference between the two methods is that “classification” is used if data can be divided into classes, and “regression” is used to forecast continuous data such as time series. In this research, we use Random Forest Regression (Random) because we forecast time series data.

(5) Stepwise

Stepwise is a method of constructing a regression model by searching for a combination of objective variables that can most explain the explanatory variable by sequentially increasing or decreasing the objective variable [7]. When adding highly objective variables to regression formulas, there are variables that have already been added, which become useless due to their relevance to objective variables added later. Therefore, each time an objective variable is added, the variable that becomes insignificant for the explanatory variable is deleted from the regression formula.

3 Forecasting the Number of Customers Visiting

Using the customers visiting data of four stores from restaurant chain A of the joint research, the number of customers visiting was forecasted. The number of customers visiting from '18/5/1 to '19/4/30 was forecasted using the customer visiting result from '14/5/1 to '18/4/30. The weather data used was the data from the Japan Meteorological Agency's observation station closest to the location of each restaurant.

We compared the forecasted results with customer visiting results during the same period. Table 1 shows explanatory variables used for forecasting. The forecasting ratio α, that is ratio of the number of forecasted customers to that of actual customers, is calculated using the equations (1) and (2).

p_i: Actual number of customers visiting on i−th day

e_i: Forecasted number of customers visiting on i−th day

N: Forecasting period

α_i: Forecasting ratio on i−th day

$$\alpha_i = \sum_{i=1}^{N} \frac{p_i - |p_i - e_i|}{p_i} \tag{1}$$

$$\alpha = \frac{\sum_{i=1}^{N} \alpha_i}{N} \tag{2}$$

Table 1. Explanatory variable

Category	Explanatory variable	Definition	Category	Explanatory variable	Definition
Month	January	Jan/1-Jan/31	Event	January 1st	January 1st
	February	Feb/1-Feb/28		January 2nd	January 2nd
	March	Mar/1-Mar/31		January 3rd	January 3rd
	April	Apr/1-Apr/30		Year-end	Dec/29-Dec/31
	May	May/1-May/31		End of year party	Weekday of December
	June	Jun/1-Jun/30		Christmas eve	December 24
	July	Jul/1-Jul/31		Coming-of-age day	Second Monday in January
	August	Aug/1-Aug/31		Setsubun	February 2nd
	Septemner	Sep/1-Sep/30		Obon	Aug/13-Aug/15
	October	Oct/1-Oct/31		New year's party	Weekday till the coming-of-age day except Jan/1-Jan/3
	Novenber	Nov/1-Nov/30		Farewell party	Weekday in March
	December	Dec/1-Dec/31		Welcome party	Weekday in April
The day of the week	Monday	Weekday and the next day is weekday	Weather	Average wind speed	Average wind speed per day (m/s)
	Tuesday	Weekday and the next day is weekday		Maximum wind speed	Maximum wind speed per day (m/s)
	Wednesday	Weekday and the next day is weekday		Highest temperature	Highest temperature in a day (℃)
	Thuesday	Weekday and the next day is weekday		Lowest temperature	Lowest temperature in a day (℃)
	Fryday	Weekday and the next day is weekday		Amount of precipitation	Amount of precipitation in a day (mm)
	Saturday	Even if the target day is a holiday it is Saturday.		Maximum precipitation	Maximum amount of precipitation in ten minutes (mm)
	Sunday	Sunday and the next day is weekday Even if the target day is a holiday it is Sunday.		Maximum instantaneous	Maximum instantaneous
	Sunday during holidays	Sunday and the next day is holiday Even if the target day is a holiday it is Sunday.		wind speed	wind speed in a day (m/s)
	Holiday	Holiday and the nextday is weekday			
	Holiday during holidays	Holiday and the nextday is holiday			
	Before holiday	Weekday and the next day is holiday			
	Lastday during holidays	The last day of three or more consecutive holidays			

Table 2. Forecast results

Restaurant	Data period	Bayesian	Boosted	Decision	Random	Stepwise
W	Four years	82.7	78.3	75.2	81.1	82.4
	Three tears	83.1	79.1	79.8	81.4	82.4
	Two Years	82.7	75.1	79.0	81.3	82.2
	One Year	81.8	73.4	79.6	77.5	81.4
X	Four years	77.0	72.4	73.6	70.6	77.6
	Three tears	75.9	74.7	74.1	72.4	76.0
	Two Years	76.8	70.2	71.7	75.0	76.6
	One Year	77.0	64.6	72.8	74.6	76.2
Y	Four years	77.2	77.4	74.8	73.8	73.2
	Three tears	73.5	73.7	75.9	72.2	71.1
	Two Years	76.3	73.3	75.0	72.2	75.4
	One Year	78.4	72.4	75.3	76.3	76.9
Z	Four years	81.7	81.6	82.2	82.8	81.4
	Three tears	82.9	81.6	82.5	83.1	82.9
	Two Years	83.8	80.0	82.8	82.6	83.8
	One Year	73.0	77.5	81.5	82.4	82.4

Table 2 shows the forecast results for four restaurants in Restaurant Chain A. The highest forecasting ratio for each restaurant and the training data period is marked in yellow. Bayesian had a high forecasting ratio for all four restaurants. Fig. 1 shows a graph of actual numbers and forecasted numbers of customers visiting by Bayesian using learning data for two years at Restaurant Z, which has the highest forecasting ratio. There is a difference between the forecasted number and the actual number during the Bon holiday period in August and the welcome and farewell party period from the end of March to the beginning of April. Based on these forecasting results, we analyzed the characteristics of days with low forecasting accuracy.

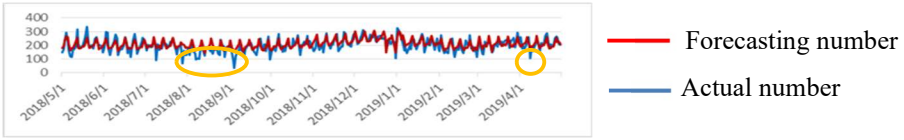

Fig. 1. Actual and forecasted numbers of customers visiting at Z (Bayesian, two years)

4 Analysis of characteristics with low forecasting accuracy days

The characteristics with low forecasting accuracy days were analyzed by focusing on the difference between the actual and forecasted numbers of customers visiting, explanatory variables such as days of the week, weather, events, and the number of res-

ervation customers. The method and the training data period that had the highest forecasting accuracy for each restaurant were selected and analyzed. We analyzed the forecasting results of Bayesian with three years training data for restaurant W, the forecasting results of Stepwise with for years training data for X, the forecasting results of Bayesian with one-year training data for Y, and the forecasting results of Bayesian with two years training data for Z.

4.1 Characteristics analysis for week and day of week

We analyzed week and day of week characteristics for the days when the forecasted numbers exceeded the error range calculated by equations (3) to (6).

q_{ij}: Forecasted number of customers visiting on i–th day using method j. J is the combination of method and training data period (j = 1, 2, …, 20)

y_j: The annual sum of the absolute value of the difference between forecasted number using method j and actual number

min(y_j): Minimum value of y_j

Y: Annual average of $min(y_j)$

$\bar{p}$: Annual average of actual values

β: Error ratio

k_i: Error range on i-th day

$$y_j = \sum_{i=1}^{365} |p_i - q_{ij}| \quad (3)$$

$$Y = \frac{\min(y_j)}{365} \quad (4)$$

$$\beta_j = \frac{Y}{\bar{p}} \quad (5)$$

$$k_i = x_i \pm \beta\bar{p} \quad (6)$$

Table 3. Results of characteristics analysis for week and day of week

Restaurant	Week	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Total
W	1st	4(33%)	7(58%)	5(42%)	5(42%)	5(42%)	5(42%)	5(42%)	36
	2nd	5(42%)	2(17%)	4(33%)	4(33%)	5(42%)	4(33%)	2(17%)	26
	3rd	1(8%)	8(67%)	4(33%)	6(50%)	5(42%)	6(50%)	6(50%)	36
	4th & 5th	8(50%)	10(63%)	10(50%)	7(44%)	9(56%)	8(50%)	6(38%)	58
	Total	18	27	23	22	24	23	19	156
X	1st	4(33%)	5(42%)	9(75%)	6(50%)	6(50%)	4(33%)	6(50%)	40
	2nd	7(58%)	5(42%)	6(50%)	4(33%)	5(42%)	5(42%)	4(33%)	36
	3rd	3(25%)	7(58%)	2(17%)	2(17%)	3(25%)	5(42%)	2(17%)	24
	4th & 5th	5(31%)	5(31%)	9(56%)	9(56%)	8(50%)	6(38%)	8(50%)	50
	Total	19	22	26	21	22	20	20	150
Y	1st	2(17%)	3(25%)	8(67%)	8(67%)	1(8%)	4(33%)	5(42%)	31
	2nd	4(33%)	4(33%)	3(25%)	6(50%)	6(50%)	5(42%)	5(42%)	33
	3rd	2(17%)	7(58%)	6(50%)	5(42%)	5(42%)	4(33%)	5(42%)	34
	4th & 5th	7(44%)	7(41%)	5(31%)	6(38%)	7(44%)	10(63%)	6(38%)	48
	Total	15	21	22	25	19	23	21	146
Z	1st	6(50%)	6(50%)	4(33%)	7(58%)	8(67%)	6(50%)	8(67%)	45
	2nd	4(33%)	7(58%)	5(42%)	3(25%)	3(25%)	6(50%)	5(42%)	33
	3rd	6(50%)	3(25%)	5(42%)	4(33%)	6(50%)	4(33%)	4(33%)	32
	4th & 5th	9(56%)	5(29%)	5(31%)	8(50%)	7(44%)	5(31%)	7(44%)	46
	Total	25	21	19	22	24	21	24	156

Table 3 shows the results. The cases that exceed more than half of the applicable annual days are marked in yellow. In restaurants W and X, there were many days outside the error range in the fourth and fifth weeks. In Japan, there are paydays in the fourth or fifth weeks. Therefore, it is considered that human psychology after salary income may affect the use of restaurants. However, in restaurants Y and Z, there is no

tendency to be low for the forecasting accuracy of the fourth and fifth weeks. Therefore, the effect of human psychology after salary income is considered to be limited.

4.2 Characteristics analysis for weather

We analyzed the weather data that could be a factor that changes human behavior. In this research, we focused on three factors: temperature, precipitation, and wind speed. Based on the definition of forecast terms by the Japan Meteorological Agency, criteria for "temperature", "precipitation" and "wind speed" were set.

(1) Temperature

If the temperature difference is extremely large compared to the normal temperature of the season, If the temperature difference is extremely large compared to the normal temperature in the season, the human's behavior may be different. If the temperature is higher than normal in summer or lower than normal in winter, you may hesitate to go out. On the other hand, if the temperature is lower than normal in summer or higher than normal in winter, you may go out positively. From this, in this research, the criterion with a large difference in temperature was ± 5 ° C of the monthly average maximum temperature and the monthly minimum temperature.

(2) Precipitation

If the precipitation is strong, you may hesitate to go out. The Japan Meteorological Agency defines strong precipitation as having an hourly precipitation of 20 mm or more [8]. From this, in this study, the criteria with strong rainfall are that both of the following conditions are satisfied.

- 10 minutes precipitation is 3.5 mm or more
- Total precipitation per day is 20mm or more

(3) Wind speed

There is a possibility that hesitating to go out in case of strong wind. The Japan Meteorological Agency defines strong wind as the average wind speed of 20 m/s or more [8]. From this, in this study, the criterion of a strong wind is that the average wind speed, the maximum wind speed, or the maximum instantaneous wind speed is 20 m/s or more.

A statistical test was performed to determine whether the weather deviated from any of the above three criteria influenced forecasting the number of customers visiting. For this purpose, we performed a x^2 test for the following null hypothesis using a cross-tabulation table.

Null hypothesis: There is no relation between the weather outside the criteria and the forecasted number of customers visiting.

Alternative hypothesis: There is some relation between the weather outside the criteria and the forecasted number of customers visiting.

Table 4 shows x^2 value for the number of weather data that deviated from each criteria value. According to Table 4, the null hypothesis cannot be rejected because the x^2 value does not exceed the one-sided 5% point of 3.84 at all four restaurants. Therefore, weather data does not affect the forecasted number of customers visiting.

Table 4. x^2 value for weather data

Restaurant	W	X	Y	Z
Temperature	0.06	0.05	0.38	3.16
Precipitation	0.18	0.26	0.15	0.15
Wind speed	1.30	0.05	1.34	0.32

4.3 Characteristics analysis for reservation customers

There are reservation customers who make reservations in advance and come to the restaurant. Since reservation customers rarely cancel reservations, it is thought that if there are many reservation customers, this may cause a difference between the actual number and the forecasted number. Therefore, we analyzed whether there was a difference in the ratio of the number of reservation customers to that of actual customers between the group where the forecasted number of visiting customers coming was within the error range and the group where that was outside the error range. For this purpose, we performed the Mann-Whitney U test for the following null hypothesis.

Null hypothesis: The average ratio of the number of reservation customers to that of actual customers is equal for groups with forecasted customers within and outside the error range.

Alternative hypothesis: The average ratio of the number of reservation customers to that of actual customers is different for groups with forecasted customers within and outside the error range.

R_i: The number of reservation customers on i-th day

γ_i: The ratio of the number of reservation customers to that of actual customers on i-th day

R_A: Sum of ranks within the error range

R_B: Sum of ranks outside the error range

m: Number of days within error range

n: Number of days outside error range

U: Either U_A or U_B

$$\gamma_i = \frac{R_i}{p_i} \tag{7}$$

$$U_A = R_A - \frac{n(n+1)}{2} \tag{8}$$

$$U_B = R_B - \frac{m(m+1)}{2} \tag{9}$$

$$Z = \frac{U - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \tag{10}$$

Table 5 shows the statistics Z calculated by equations (7) to (10) for each restaurant. The statistic Z has the property of following the standard normal distribution under the null hypothesis. The statistics Z is out of ± 1.96 which is the 5% point on both sides of the standard normal distribution in all restaurants. Therefore, the null hypothesis is rejected and the alternative hypothesis is adopted. Since the average ratio of the number of reservation customers to that of actual customers is different for groups with forecasted customers within and outside the error range, the number of reserved customers affects the forecasting the number of customers visiting.

Table 5. Statistics Z for each restaurant

W	X	Y	Z
2.26	-3.10	-2.25	-2.93

5 Conclusion

In this paper, we analyze the relation between training data and forecasted results calculated by machine learning and statistical analysis and discuss the characteristics on days with low forecasting accuracy. The following results were obtained.

(1) Forecasting results of customers visiting

Among the five methods, the forecasting accuracy using Bayesian was the highest. The forecasting accuracy did not tend to improve even if the training data period was extended.

(2) Analysis of characteristics with low forecasting accuracy days

It was found that the human psychology around the payday and the reservation customers affected the number of visitors. On the other hand, the weather data such as temperature, precipitation and wind speed did not affect the accuracy.

In future research, we improve the forecasting ratio by conducting data analysis considering human behavior patterns and restaurant locations and developing new methods and forecasting models using new machine learning or deep learning with additional explanatory variables. After further improving the forecasting ratio, we plan to study how to improve store management.

References

1. Agnieszka L., Nick C., Jim S.: Restaurant Sales and Customer Demand Forecasting: Literature Survey and Categorization of Methods, Smart City 360, Springer, pp.479-491 (2016).
2. Xu M., Yanshan T., Chu L., Yuehui Z: Predicting Future Visitors od Restaurants using Big Data, Proceedings of 2018 International Conference on Machine Learning and Cybernetics, vol.1, pp.269-274 (2018).
3. Motomura Y., Kurata T., Yamamoto Y.: Community-Based Participatory Service Engineering: Case Studies and Technologies, Global Perspectives on Service Science: Japan, Springer, pp.63-78 (2016).
4. Bernard S., Fabio R., Flavio R., Fabricio F., Jonice O.: Scholar Performance Prediction using Boosted Regression Trees Techniques, European Symposium on Artificial Neural Networks 2017 proceedings, pp.329-334 (2017).
5. Antonio C., Jamie A., Ender K.: Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning, Foundations and Trends in Computer Graphics and Vision, vol.7, nos.2-3, pp.81-227 (2012).
6. Sebastian R.: Python Machine Learning (Japanese Edition), Impress Corp, pp.86-87 (2016).
7. B.W.Boich, C.J.Huang: Applied Statistics analysis (Japanese Edition), Morikita Publishing Co. Ltd., pp.27-28, pp.167-172 (1968).
8. Japan Meteorological Agency: Rain and wind, https://www.jma.go.jp/jma/kishou/books/amekaze/amekaze.pdf (2020 3.3 access) (in Japanese)