Machine Learning

Prediction of the number of cafeteria users

강연범 김윤아 이다현

Yeon-Beom Kang Yun-A Kim Da-Hyun Lee

Jeonbuk National Univ.

Software Engineering

Abstract

Predicting the number of restaurant users is an important decision-making process. If the amount of food prepared is greater than the amount of food actually consumed, the quality of the food is decreased. In addition, as the amount of leftover increases, the cost of processing it is increased. In particular, in large restaurants such as cafeteria, it shall be prevented necessarily. Therefore, the purpose of this study is to prepare the appropriate amount of food by implementing a model that pre.

Keywords

institutional foodservice, meal forecasting, classification model, machine learning, big data analytics

1. Introduction

Research Background and Needs

Machine learning and application and learning algorithms are being used in various fields such as dentistry, physical education, distribution accounting, and roads, and machine learning techniques that model and predict using real data in more fields and areas will be applied in the future.

However, the application and utilization of machine learning technology is still insufficient in the restaurant industry, and user prediction at group cafeterias is mainly based on experience.

For the efficient operation of group cafeterias, large companies that operate them have been planning and ordering meals based on estimating the number of users through their own user prediction modeling. However, in many group cafeterias, except for some group cafeterias operated by large companies, systematic and scientific user predictions are not made due to lack of professionals and limited budgets, and various side effects are caused by inaccurate forecasting of manpower demand. In most food service centers, incorrect prediction of the number of people by experience leads to problems such as poor order planning, reduced order quantity and additional orders, increased food costs, inefficient human operation and efficiency, and poor food quality.

Despite this situation, in Korea, research related to user prediction is limited to research that identifies important variables that affect user prediction, and research showing user prediction accuracy is still insufficient.

In addition, theoretical studies on how to deal with food waste, such as analysis of leftover food at group cafeterias, analysis of the cause of residual

food, investigation of preference for food loss, and environmental pollution related to food waste, and related diet-based education are still insufficient.

Research Purpose

The purpose of this study is to propose a method of predicting the number of users of a group cafeteria based on data science that can reduce food waste generated in a group cafeteria, and to show the possibility of using the technique through actual data. This study aims to contribute to predicting the appropriate number of users from the stage of planning menus and ordering ingredients for cooking, and to reduce leftover food waste from public institutions and to efficiently operate the group cafeteria through data science prediction. In addition, machine learning techniques are introduced for most group cafeterias, which are currently being conducted by experiences such as nutritionists and cooks, and prediction algorithms are developed and carried out through variables for predicting drinking water in group cafeterias and user prediction modeling based on data science. Ultimately, by predicting the number of people based on data science, efficient manpower management and placement, scientific food production management, loss and cost of food ingredients due to leftover food, and prevention of food sold out, we will contribute to the efficient operation of group cafeteria.

2. Data

The data feature consists of the date, day, number of headquarters, number of vacationers, number of business trips of headquarters, number of telecommuters belonging to the current headquarters, number of overtime orders, breakfast menu, lunch menu, and dinner menu. The data hypothesis was set as follows. It is the same as "Is there an impact on meals before and after COVID-19?", "Did telecommuting occur before and after COVID-19?" and "Is there a difference in the amount of lunch and dinner depending on the day?". As a result of checking the missing data, there were no missing values in both train data and test data.

```
일자                False
요일                False
본사정원수             False
본사휴가자수            False
본사출장자수            False
본사시간외근무명령서승인건수    False
현본사소속재택근무자수       False
조식메뉴              False
중식메뉴              False
석식메뉴              False
중식계               False
석식계               False
dtype: bool
```

figure 1. Train missing data

```
일자                False
요일                False
본사정원수             False
본사휴가자수            False
본사출장자수            False
본사시간외근무명령서승인건수    False
현본사소속재택근무자수       False
조식메뉴              False
중식메뉴              False
석식메뉴              False
dtype: bool
```

figure 2. Test missing data

3. Baseline Design

After creating a data frame, two models were created and learned to predict lunch and dinner. Random forest was used to set the frequency of the most class to be accurate. Using Random Forest, the performance of the lunch baseline model was -91.67 and the performance of the dinner baseline model was -85.18.

4. EDA_Data Preprocessing

First, I modified the column name.

"일자" was used as "date",

"요일" was used as "dow",

"본사정원 수" was used as "employees",

"휴가자 수" was used as "dayoff",

"출장자 수" was used as "bustrip",

"야근자수" was used as "ovtime",

"재택근무자수" was used as a "remote",

“조식 메뉴” was used as a ”brk”,

“중식 메뉴” was used as a “ln”.

“석식 메뉴” was used as a “dn”.

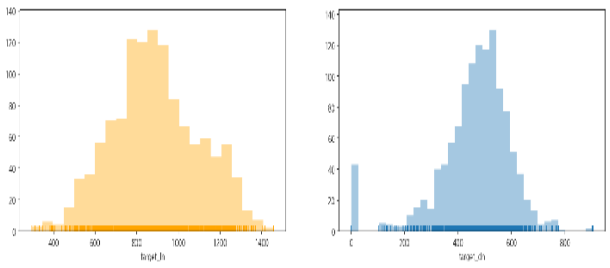

figure 3. Target Distribution

After that, the target distribution was confirmed. Figure 3 is a target distribution diagram. It was confirmed that there is a day when the number of dinner users is zero, and the number of dinner users is in the range of about 100 to 800, and that the number of dinner users is generally less than lunch.

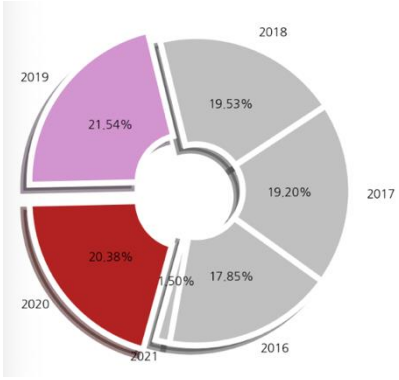

figure 4. Comparison of the number of users by year

In addition, referring to Figure 4, when comparing the number of users in 2019 and 2020, the number of lunch/dinner users did not change significantly due to COVID-19. Rather, in the case of lunch, the average number of users (890) in 2022 was higher than that of 2019 (850). Due to COVID-19, it can be said that he reduced his lunch outing and ate in-house. In Figure 3, it can be seen that there is a day with 0 evening users, and when looked at the data to find the reason, we found the following commonalities.

	date	dayoff	bustrip	ovtime	remote	dow	dn	target_dn
204	2016-11-30	68	207	0	0.0	수	*	0.0
224	2016-12-28	166	225	0	0.0	수	*	0.0
244	2017-01-25	79	203	0	0.0	수	*	0.0
262	2017-02-22	75	252	0	0.0	수	*	0.0
281	2017-03-22	53	235	0	0.0	수	*	0.0

그림 6. Target=0 table

As shown in Figure 6, you can see that the number of people on the last Wednesday evening of every

month is zero. The day was "Self-Development Day," and everyone had to leave work on time.

Therefore, the number of dinner users on the last Wednesday of every month is zero. In addition, there were days when there were 0 users even though the dinner menu was provided, and the dates of 2017.09.27 and 2018.02.14 were confirmed as the day before the holiday and the day before the Lunar New Year holiday. For this reason, the number of people on the last Wednesday in the prediction model was set to 0.

	중식계	석식계
요일	-0.731563	-0.312112
본사정원수	-0.115529	-0.173852
본사휴가자수	-0.391975	-0.316894
본사출장자수	-0.512680	-0.188164
본사시간외근무명령서승인건수	0.535611	0.571168
현본사소속재택근무자수	0.076509	-0.057534
중식계	1.000000	0.508287
석식계	0.508287	1.000000

figure 7. Target Column Correlation

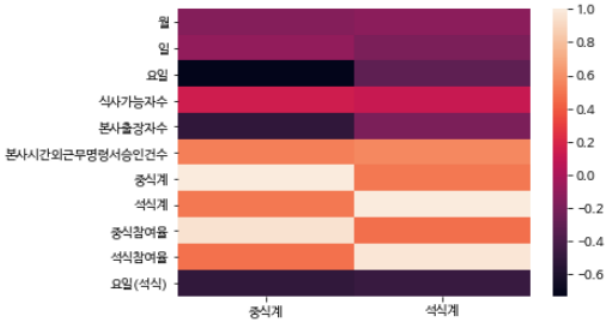

figure 8. Target Column Correlation Heatmap

Looking at Figure 7 and Figure 8, the number of days of the week, the number of approvals for overtime work orders at the headquarters, and the number of business trips at the headquarters showed a high correlation.

Dinner showed a high correlation between the number of approval of the headquarters overtime order, the day of the week, and vacationers. When the correlation was shown as a target, it was confirmed that similar columns showed a high correlation at lunch and dinner.

It can be seen that the scatterplot graph of Figure 9 shows a similar pattern to the correlation table.

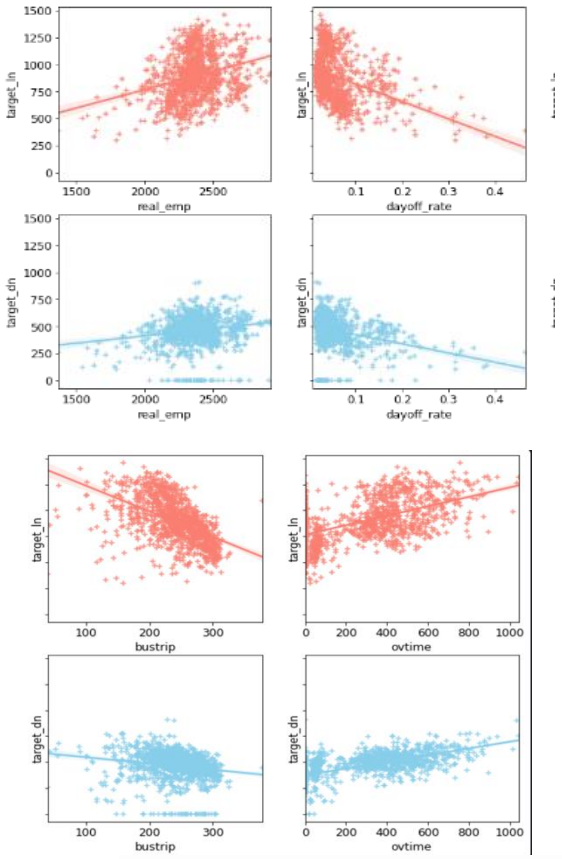

figure 9. Target Column Correlation Scatterplot Graph

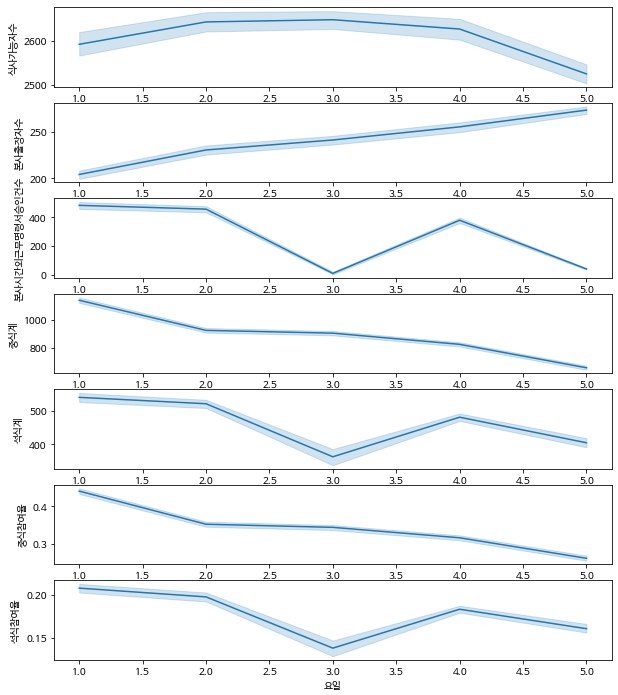

figure 10. Percentage of lunch and dinner by day graph

The process of checking whether there was a significant difference by day of the week was also performed, and the results are shown in Figure 10 above. For lunch, graphs appeared in the order of Monday, Tuesday, Wednesday, Thursday, and Friday. The graph was shown in the order of Monday, Tuesday, Thursday, Friday, and the number of dinner users at a rate similar to the ratio of the number of overtime workers.

Since there is a difference in the ratio of meals by day of the week, the date and day of the week were converted into numerical types by label encoding. Monday to Friday represented 5, converting the day of the week into a number. The number of people who can eat was subtracted from the number of people who are on leave from the headquarters and the number of telecommuters from the current headquarters, and an accurate column of participation in meals was created to convert it into a ratio. I created a new day column that mapped the days of the week to the dinner rack

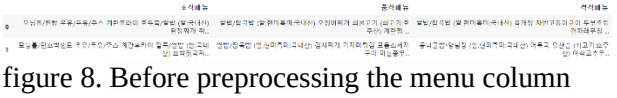

figure 8. Before preprocessing the menu column

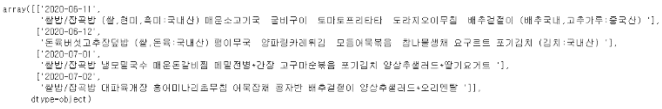

figure 9. After preprocessing the menu column

figure 10. Pre-processing of menu column reflecting abnormalities

figure 11. Combined menu data and preprocessing data

As a result of thinking a lot about how to preprocess the menu column, it was decided to divide it by menu and use it for model learning with only a specific menu. Unnecessary characters were removed from the menu, and strings were separated by spaces, and empty elements and country of origin information were deleted. Figure 8 is the data before the pre-processing of the menu column, and Figure 9 is the data after the pre-processing of the menu column.

As can be seen in Figure 9, abnormalities were found in the menu preprocessing process. The order of the side and kimchi on the menu was changed from 2020.07.01. In addition, it was found that the operation of the cafeteria was suspended from 2020.06.13 to 2020.06.30.

The menu column was preprocessed by reflecting

these abnormalities. In the lunch train data, it was divided into rice, soup, side dishes, kimchi, and side dishes. After that, the number of rice data and the number of side dish data were checked. There are too many categories of menus to combine menu data and preprocessed data, so only the main side dishes that are likely to affect the target were added to the existing data.

This can be seen in Figure 11.
Finally, after separating the target data, encoding was performed.

5. Validate machine learning model and verify performance

	BaseLine	RandomForest	LGBMRegressor	CatBoost
중식계 best score	-91.637135	-82.898094	-83.462019	-77.526155
석식계 best score	-85.303553	-72.433913	-70.673669	-67.981968

figure 12. Modeling Performance Results

We used three models: RandomForestRegressor, an ensemble method for learning multiple decision trees. LGBMRegressor, a tree-based framework with GradientBoosting. a new machine learning technique that outperforms existing boosting models, and a Catboost model focused on processing Category features. The modeling performance results are shown in Figure 12, and the first Catboost model showed high performance both at lunch and dinner. The measured value of the lunch system model is -77.526155 and the measured value of the dinner system model is -67.981968. The second highest performance model is Random Forest, followed by LGBM Regressor..

6. Machine Learning Model Analysis

We checked how the model performs on feature importance. Among the two lunch and dinner machine learning models, we investigated the model's feature importance with more diverse

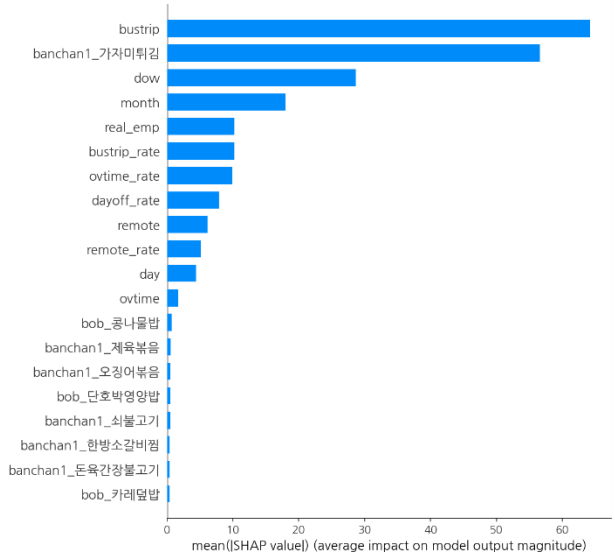

figure 13. SHAP

figure 14. SHAP Force Plot

To find out why the model predicted as shown in Figure 13, we drew the SHAP Force Plot as shown in Figure 14. As a result, the feature importance is high and the higher variable is bustrip and next, it was possible to derive an insight that the number of users of the cafeteria decreases at lunch when fried flounder is served as a side dish, which is a lower variable.

7. Conclusion

We created a model that predicts the number of cafeteria users using three models. Random Forest Regressor, LGBM Retressor, and CatBoost models were used, and the modeling performance showed high performance in the order of CatBoost, Random Forest, and LGBM Regressor.

Although a model for predicting the number of users per cafeteria was established, there was a limitation that this model could not be used as it is in all companies. Since it is a model created based on a specific company's data set, the number and distribution of data will vary from company to company. Therefore, perfect generalization is not possible on the current data.

By combining weather data, it is expected that performance will be improved by checking how the number of users at the cafeteria changes at lunch on rainy days. It will also look into public holidays in advance. We expect that the prediction performance will be improved by applying zero predictors for lunch and dinner on holidays.

The future goal of our model is to create a model that can predict the number of cafeteria users for all companies by adding data from other companies. In addition, I think that identifying the expected number of cafeteria users in advance using the predictive model can help to prepare ingredients and minimize food waste.

8. Reference

외국문헌

Guolin Ke, 2017, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”

Musa Osman, 2021, “ML-LGBM: A Machine Learning Model Based on Light Gradient Boosting Machine for the Detection of Version Number Attacks in RPL-Based Networks

S. Liu, Y.C. Shin, 2019, Additive manufacturing of Ti6Al4V alloy: A review, Mater. Des. Vol. 164, p. 107552.

A.V.S.R. Prasad, K. Ramji, G.L. Datta, 2014, An Experimental Study of Wire EDM on Ti-6Al-4V Alloy, Proc. Mater. Sci., Vol. 5,

D. Herzog, V. Seyda, E. Wycisk, C. Emmelmann, 2016, Additive manufacturing of metals, Acta Mater

P.A. Kobryn, S.L. Semiatin, 2001, The laser additive manufacture of Ti-6Al-4V, JOM

L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, "CatBoost: unbiased boosting with categorical features", Advanced in Neural Information Processing Systems 31

A. L. Lee, A Study on Author Profiling of Web-based Korean Web Text of Elementary and Middle School Students using Word2Vec and Bi-LSTM for Binary Classification, Master's thesis, Ewha Womans University, pp. 15-16, 2020.

국내 문헌

곽동경, 장미라(1993), “식단 개선을 위한 급식 평가”, 국민건강, 93(12), 2-12.

김광지, 박기용(2006), “컨조인트 분석을 통한 대학급식소의 효율적인 운영에 관 한 연구”, 한국조리학회지, 12(4), 33-45.

김광지, 조용범(2007), “대학교 학생 식당의 서비스품질, 메뉴품질, 가격이 고객 만족에 미치는 영향 - 부산지역을 중심으로 -,” 한국조리학회, 13(3), 127-136.

김정만(1997), “호텔 서비스품질이 소비자 만족에 미치는 영향에 관한 연구,” 관 광학연구, 20(2), 156-167.

박기용(2004), 외식산업경영학, 대왕사: 서울

박영숙, 이경애(1993), “대학 구내 식당에 대한 이용자의 만족도”, 순천향대학교 논문집, 16(3), 799-806.

박정영, 윤혜려(2002), “외식서비스 인카운터의 고객만족에 있어서 중요 요인에 관한 연구,” 관광학연구, 25(2), 339-360.

양기승(2001), “호텔부페 레스토랑 고객의 만족 지각요인에 관한 실증 조사”, 여 행학연구, 21, 241-263.

양일선, 신서영, 이혜영, 이소정, 채인숙(2000), “컨조인트 분석과 다차원척도법을 이용한 대학급식소의 전략적 운영 방안 모색”, 한국식생활문화학회지, 15(1), 155.

이기춘, 조희경(1996), “의료서비스에 대한 소비자만족, 불만호소행동 및 재구매 - 50 - 의도 - 종합병원을 중심으로 -”, 소비자학연구, 7, 87-108.

이유재(1991), “고객만족형성과정의 제품과 서비스간 차이에 대한 연구,” 소비자 학연구, 8(1), 28-49. 이유재(1995), “고객만족의 정의 및 측정에 관한 연구”, 경영논집, 29(1,2), 145-168.

이유재(1997), “고객만족형성과정의 제품과 서비스간 차이에 대한 연구,” 소비자 학 연구, 8(1), 101-118.

이유재(1997), “고객만족의 결과변수에 대한 이론적 연구,” 경영논집, 201-232.

이유재(2004), 서비스마케팅, 3 판, 학현사: 서울.

이유재, 라선아(2002), “브랜드 퍼스낼리티-브랜드 동일시-브랜드자산 모형: 이 용자와 비이용자간 차이에 대한 탐색적 연구”, 마케팅연구, 17(3), 1-33. 이유재, 라선아(2003), “서비스 품질의 각 차원이 CS 에 미치는 상대적 영향에 대한 연구 - 기존고객과 잠재고객의 비교를 중심으로 -”, 마케팅연구, 18(4), 67-97.

이학식, 김영(1999), “서비스품질과 서비스가치”, 한국마케팅저널, 1(2), 77-99. 이해영(2005), “대학내 신규 학생식당의 운영 모델 제안을 위한 급식서비스 속성 의 상대적 중요도 규명”, 한국식품영양과학회지, 34(7), 1028-1034.

이훈영(2006), 마케팅조사론, 2 판, 청람: 서울.

최경숙(1999), “대진대 학생들의 교내식당 이용 만족도와 급식메뉴 개발을 위한 음식기호도 조사”, 대진논집, 7, 273-288.

최창권(2003), “서비스스케이프 품질이 레스토랑 애호도에 미치는 영향,” 대구대 학교박사 학위논문.