DAILY DEMAND PREDICTION IN MASS FOOD PRODUCTION WITH ENSEMBLES OF TREES MODEL

Derya Yergök *[1], Mehmet Acı [2]

[1] Mersin University, Engineering Faculty, Computer Engineering Department, Mersin, Turkey, deryayergok@gmail.com

[2] Mersin University, Engineering Faculty, Computer Engineering Department, Mersin, Turkey, maci@mersin.edu.tr

ABSTRACT
In this study, it is aimed to minimize the cost and waste by estimating the number of people who eat food in places where mass food is produced. In this study, one of the regression methods, the tree ensemble model, was applied to the data obtained from the university refectory. The data set was trained and tested by the tree ensemble method and the number of people eating was estimated. The tree community regression model used includes two different models, boosted tree and bagged tree. When these models are compared, the best result is achieved in the boosted tree model.

Keywords: *Regression, Food Demand Forecast, Prediction, Ensembles of Trees.*

* Derya Yergök

1. INTRODUCTION

Mass food production is seen and important in many establishments such as refectories of universities, restaurants, cafes and hotels. The list of meals in such establishments, although it is clear how much food will come out every day, it is not possible to reach the exact number. An average number of meals comes out. This results in lack of food or too much food. Therefore, cost is the most important problem affecting companies. Every company that produces food wants to minimize its cost. In this case, the most important role in the costs is to determine the exact number of people eating. Accurate estimates of the number of people who eat will result in a reduction in cost and waste, while will greatly benefit businesses.

The aim of this study is to estimate the number of people who will eat that day according to the meal to be prepared by using the data of the university refectory. Accordingly, it is aimed to prevent the problems that may arise if too much or less food is produced. Various regression methods were used to solve the problem. MATLAB's statistics and machine learning tools are trained with tree community models to predict data using supervised machine learning with the Regression Learner App in the toolbox.

Community methods combine several decision tree classifiers to provide better predictive performance than a single decision tree classifier. The main principle behind the community model is that a group of weak learners come together to form a strong learner, thereby enhancing the accuracy of the model. When we try to estimate the target variable using any machine learning technique, the main reasons for the difference in actual and predicted values are noise, variance and bias. The community helps to reduce these factors (excluding noise, which is an irreducible error). Another way of thinking about community learning is the tales of blind men and elephants. All blind men had their own elephant identification. Although each definition was correct, it would have been better to come together and discuss their meaning before concluding. This story perfectly describes the community learning method (Shubham, 2018).

In classification, clustering and regression problems, the concept of community learning has been proposed to increase the stability and predictive accuracy of a single learning algorithm. A community learning-based model combines individually trained models and produces a single final estimate (Kapucu and Çubukçu, 2019).

Community-based tree models are better compared to single tree models. Bagging and reinforcement are the two most popular learning methods of tree communities (Panda et al., 2009).

Using techniques such as bagging and reinforcement helps reduce variance and increases the robustness of the model. Combinations of multiple classifiers reduce variance, especially in the case of unstable classifiers, and can produce a more reliable classification than a single classifier (Shubham, 2018).

Bagging creates different classifiers, only if the basic learning algorithm is unstable. The ambiguity in the basic learning algorithm means that minor changes to the training set cause major changes in the learned classifier. Breiman (1996) explores the causes of imbalance in learning algorithms and discusses a way to reduce or eliminate it.

Bagging (and to a lesser extent reinforcement) can be seen as ways to use this instability to improve classification accuracy (Dietterich, 2000).

Reinforcement is one of several techniques that aims to improve the performance of many models by adapting them to a single model and combining them for forecasting. Reinforcement is a method used to improve model accuracy. Rather than finding a single and highly accurate prediction rule, it is based on the idea that very difficult rules are easier to find (Elith et al., 2008).

The data set used in the research was divided into two groups as test and training data set. After the model was trained with training data, test data and estimates were obtained. These operations were performed in MATLAB software environment. The Regression Learner App, available in MATLAB's statistics and machine learning toolbox, proposes regression models for predicting data using supervised machine learning.

In this study, two different tree communities were designed. Principal component analysis (PCA) was used in both of the designed models. PCA is a multivariate technique that analyzes a data table in which observations are defined by several quantitative dependent variables associated with each other. Its purpose is to extract important information from the table, to show it as a new set of orthogonal variables called fundamental components, and to show the similarity model between the observations and the scores on the maps (Abdi and Williams, 2010). Among the methods, the best results were obtained with boosted trees.

The accuracy of the proposed models; Mean Square Error (MSE), Mean Absolute Error (MAE) and Multiple Correlation Coefficient (R) were determined.

2. DATASET AND MODELS

2.1. Dataset

The data set consists of 249 rows and 11 columns. Before these values were used, the text in it was converted to numeric values. In addition, a grouping operation was performed for similar data. In this way, the learning ability of the network is increased. The data set contains 11 properties. 10 of these are input data (weekdays, salary days, soups, main course, side meal, complementary meal, calories, meat amount, Ramadan, lesson / exam / holiday period) and 1 output data (number of people). Table 1 shows an example of the attribute content. An example of a dataset that is converted to numerical values is given in Table 2.

Table 1. A sample of dataset content

Attribute	Range	Description
Days of the Week (A)	1-5	Monday,Tuesday, etc.
Salary Day (B)	0-1	Salary day or not
Soups (C)	1-5	Chicken/beef, vegetables, legumes, yoghurt, pastry.
Main Courses (D)	1-5	Beef ,chicken, fish, vegetables, legumes.
Side Meal (E)	1-3	Rice, pastry, legumes.
Complementary Meal (F)	1-5	Yoghurt, fruit, salad, sugary,drink
Calorie (G)	1-3	Calories of food

Amount of Meat (H)	1-4	How many grams of meat
Ramadan (I)	0-1	Ramadan or not
Lesson/ Exam/Holiday (J)	1-3	Semester or exam week or holiday
Number of People (K)	1-9	Number of people eating at the dining hall

Table 2. A sample of dataset

A	B	C	D	E	F	G	H	I	J	K
2	0	3	1	1	1	3	3	0	3	1
3	0	4	2	1	2	3	4	0	3	1
4	0	5	1	2	3	1	2	0	3	1
5	0	3	1	1	1	2	2	0	3	1
1	0	3	1	1	2	2	2	0	2	3
2	0	1	2	1	1	3	4	0	2	3
3	0	1	1	2	1	2	1	0	2	2
4	0	4	1	1	4	3	3	0	2	2
5	0	4	2	3	1	2	4	0	2	2
1	1	3	5	1	1	3	1	0	3	1

2.2. Ensembles of Trees

Community learning is a machine learning paradigm in which multiple students are trained to solve the same problem. Unlike ordinary machine learning approaches that attempt to learn a hypothesis from educational data, community methods attempt to form a series of hypotheses and combine them to use them (Zhou, 2015). Tree communities are one of the community learning models in which more than one regression tree is combined and found the most appropriate result. A simple tree ensemble structure is shown in Fig.1.

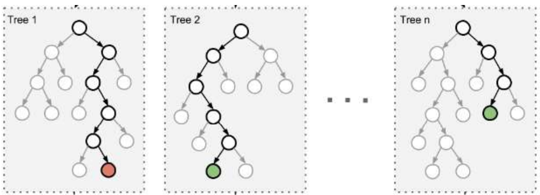

Fig. 1. Ensembles of trees model (Smolyakov,2017)

2.2.1. Boosted trees

Empowerment refers to a group of algorithms that use weighted averages to make weak learners more powerful learners. Unlike bagging, which allows each model to function independently, it then collects the outputs without preferring any model. Empowerment is all about “teamwork”. Each working model determines which features the next model will focus on (Shubham, 2018).

2.2.2. Bagged trees

Bagging is a simple and very powerful community method. Bagging is the application of the Bootstrap procedure to a high-variability machine learning algorithm, typically decision trees (Shubham, 2018).

Ensembles of trees model is recommended by L. Breiman. It is a method of retraining the basic learner by deriving new sets of training from an existing training set. Replacement is done by sampling. The training set is produced by randomly selecting a training set with n samples from bagging. Each selected sample is put back into the training set. Some examples are not included in the new training set, while others take place more than once (Güzel, 2018).

3. RESULTS AND DISCUSSIONS

All prediction models were evaluated in terms of three performances measured, (1) R value is used for measuring the correlation between target and predicted values, (2) MSE measures the average of the squares of the errors, (3) MAE measures the closeness of the predictions to the target values (Witten and Frank, 2005). Equations of these performance measures are given in equations (1), (2) and (3), respectively.

$$\boldsymbol{R} = \sqrt{\mathbf{1} - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O_m)^2}} \tag{1}$$

$$\boldsymbol{MSE} = \frac{1}{n}\left[\sum_{i=1}^{n}(O_i - P_i)^2\right] \tag{2}$$

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \tag{3}$$

where n is the number of data points used for testing, P_i is the predicted value, O_i is the observed value and O_m is the average of the observed values.

According to the results given in Table 3, the following results are obtained: In the Ensembles of Trees model, the best result was achieved with Boosted Trees. For MSE 0.5103, for MAE 0.5036 and for R 0.9591 values are achieved.

Table 3. The results obtained with the Ensembles of Trees model.

Models	MSE	MAE	R
Boosted Trees	**0.5103**	**0.5036**	**0.9591**
Bagged Trees	1.0147	0.7630	0.9165

The dataset consists of 249 rows and 11 columns. Column 11 is the result column, which contains the estimated values of the number of people who eat. Figures 2 and 3 show the performance graphs of the boosted tree and bagged tree models. The X-axis of the graph shows all data consisting of 249 lines, and the Y-axis represents the result data. The fields shown in black represent the actual values and the fields shown in gray represent the estimated values. Both methods yielded good results and boosted tree model showed better results than bagged tree model.

For the methods in the tree community model, MATLAB's current parameter values were played and high results were obtained. It was found that the parameters of the models were changed in accordance with the data set and the problem. In both methods, PCA was activated and the current value of 95 was 100. At the same time, advanced settings have been changed for both methods. For boosted tree model; the minimum leaf size was 8, the number of learners was 30, the learning rate was 0.25. For the bagged tree model; the minimum leaf size was 1, the number of learners was 23. With these settings, error rates decreased and R value increased.

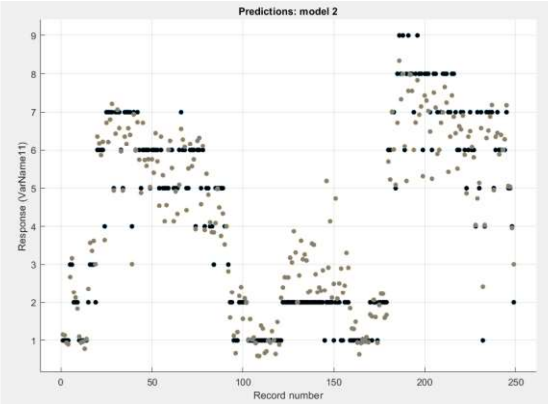

Fig. 2. Response graph of boosted tree model

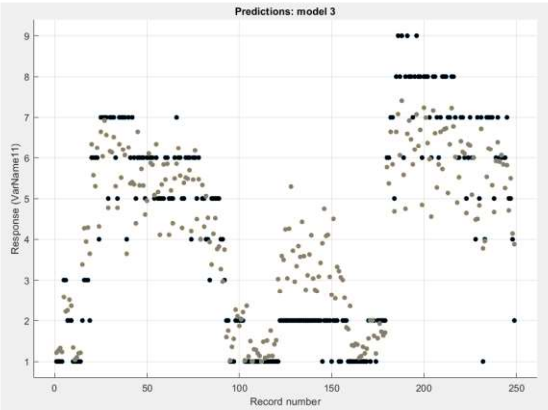

Fig. 3. Response graph of bagged tree model

4. CONCLUSION

In this study, using the data obtained from the university refectory, a regression model of the tree community was designed to predict the number of people eating food according to the food to be prepared and successful results were obtained. Two methods were modeled in the study and boosted tree model was found to reach better results than bagged tree model.

Successful results show that existing methods can be improved and better results can be obtained by using different methods and models.

ACKNOWLEDGEMENTS

This study was supported by Mersin University Scientific Research Projects Unit with Project Number 2019-1-TP2-3215.

REFERENCES

Abdi, H., and Williams, L. J. (2010). "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics*, Vol.*2,* No.4, pp.433-459.

Breiman, L. (1996). "Bagging predictors." *Machine learning,* Vol.24, No.2, pp.123-140.

Dietterich, T. G. (2000). "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization." Machine learning, Vol.40, No.2, pp. 139-157.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). "A working guide to boosted regression trees." Journal of Animal Ecology, Vol.77, No.4, pp.802-813.

Güzel, K. (2018). Kolektif Öğrenme (Ensemble Learning) ve Bagging Algoritması https://medium.com/@billmuhh/kolektif-öğrenme-ve-bagging-algoritması-e8ea3d932b72 [Accessed 22 Aug 2019]

Kapucu, C. and Çubukçu, M. (2019). "Fotovoltaik Sistemlerde Topluluk Öğrenmesi Temelli Hata Tespiti." Bilişim Teknolojileri Dergisi, Vol.12, No.2, pp.83-91.

Panda, B., Herbach, J. S., Basu, S., and Bayardo, R. J. (2009). "Planet: massively parallel learning of tree ensembles with mapreduce." Proceedings of the VLDB Endowment, pp.1426-1437.

Shubham, J.(2018). Ensemble Learning Bagging and Boosting https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e [Accessed 22 Aug 2019]

Smolyakov, V.(2017). Ensemble Learning to Improve Machine Learning Results https://blog.statsbot.co/ensemble-learning-d1dcd548e936 [Accessed 22 Aug 2019]

Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Burlingon, USA.

Zhou, Z. H. (2015). "Ensemble learning. "*Encyclopedia of biometrics*, pp.411-416.