


PRICE PREDICTION MODEL FOR RESTAURANTS IN ISTANBUL BY USING
MACHINE LEARNING ALGORITHMS

Kevser Şahinbaş* 

Sending Date: 25.07.2022

Acceptance Date: 16.08.2022

Araştırma Makalesi/ Research Article

Doi: <https://doi.org/10.38009/ekimad.1148216>

Abstract

Today, companies have created new products based on data and accelerated the digitalization processes of businesses with the concept of data science. In this study, a price prediction model is proposed with machine learning algorithms by collecting the data of businesses in the food and beverage sector in Istanbul. In this study, different machine learning modeling algorithms such as XGBoost, Random Forest, Artificial Neural Network, K-Nearest Neighbor, Multi Linear Regression and CatBoost were used for restaurant price prediction. Classification algorithms were tested for price prediction, and as a result of the evaluation, it was observed that XGBoost algorithm achieve the highest performance with 0.023236 RMSE and 0.0005399 MSE error rates. By this study, business owners will be able to understand how new developments they will make in their businesses will benefit in terms of price and customer feedback. It will enable entrepreneurs to have information about what features a new business should have and the average price they will offer to their customers according to these features. In addition, entrepreneurs who want to open a restaurant will learn how much they should cost, provide price performance, and increase their profitability by selling more products because they will sell their products at affordable prices. Accurate pricing is one of the four important concepts of marketing. The company needs to make the right pricing in order to hold on and create customer loyalty.

Keywords: Data Science, Price Strategy, Data Management, Prediction Model, Classification Algorithms.

Jel Classification: C02, C60, D40

İSTANBUL'DAKİ RESTORANLAR İÇİN MAKİNE ÖĞRENME Sİ ALGORİTMALARI KULLANILARAK
FİYAT TAHMİN MODELİ

Öz

Günümüzde veri bilimi kavramıyla birlikte firmalar veriye dayalı yeni ürünler ortaya çıkarmış ve işletmelerin dijitalleşme süreçlerini hızlandırmıştır. Bu çalışmada İstanbul'da bulunan yeme-içme sektöründeki işletmelerin verileri toplanarak makine öğrenmesi algoritmaları ile bir fiyat tahmin modeli önerilmiştir. Bu çalışmada restoran fiyat tahmini için XGBoost, Random Forest, Artificial Neural Network, K-Nearest Neighbor, Multi Linear Regression ve CatBoost gibi farklı makine öğrenmesi modelleme algoritmaları kullanılmıştır. Fiyat tahmini için sınıflandırma algoritmaları test edilmiş ve değerlendirme sonucunda XGBoost algoritmasının 0.023236 RMSE ve 0.0005399 MSE hata oranları ile en iyi algoritma olduğu gözlemlenmiştir. Bu çalışma sayesinde işletme sahipleri işletmelerinde yapacakları yeni geliştirmelerin fiyat ve müşteri geri bildirimleri açısından ne derece fayda sağlayacağını anlayabileceklerdir. Girişimcilerin yeni kuracakları bir işletmenin hangi özelliklere sahip olması gerektiğini ve bu özelliklere göre müşterilerine sunacakları ortalama fiyat konusunda bilgi sahibi olmasını sağlayacaktır. Ayrıca restoran açmak isteyen girişimcilerin neye ne kadar maliyet koyması gerektiğini öğrenecek, fiyat performansı sağlayacak, ürünlerini makul fiyata satacağı için daha çok ürün satıp karlılığını arttıracaktır. Doğru fiyatlandırma pazarlamanın dört önemli kavramlarından biridir. Firmanın tutunması, müşteri sadakati oluşturması için doğru fiyatlandırma yapması gerekmektedir.

Anahtar Kelimeler: Veri Bilimi, Fiyat Stratejisi, Veri Yönetimi, Tahmin Modellemesi, Sınıflandırma Algoritmaları

Jel Sınıflandırması: C02, C60, D40

* Asst. Prof., Istanbul Medipol University, Faculty of Business Administration and Management Sciences, Department of Management Information Systems, ksahinbas@medipol.edu.tr

1. Introduction

Machine learning (ML) methods have been used in different areas such as suggesting a suitable place to start a business or proposing the expected score of a business from the features a restaurant has. In this area, studies have been conducted to personalize food recommendations for users by using fuzzy logic. User satisfaction and helping users were the main concerns in the studies. Risk assessment is an important factor for entrepreneurs before starting a business (Osman, 2016). Traditionally, businesses have focused on how users rate their service to evaluate whether they like it or not. For this reason, it is aimed to help entrepreneurs who will invest in the restaurant business to make an optimal decision. The lack of work by entrepreneurs in decision making inspired the creation of a model to assist in the expected rating of restaurant features. A supervised machine learning model is proposed for this purpose.

A restaurant's rating depends on many factors, including the average cost of food, restaurant type, number of ratings, cuisines served, location, types of facilities available, number of reviews, and views. For this reason, managing these factors can be difficult. In this study, a price prediction model is proposed that reveals what customers value most in a restaurant.

Firstly, data on restaurants in Istanbul were obtained by web scraping technology and exported to a file for price prediction of the dataset by Zomato (zomato.com). Zomato was founded in India in 2008 for food reviews and now operates in 24 different countries. Zomato and other dedicated Online Opinion Platforms (OOPs) enable customers to share their thoughts and experiences with accommodation service providers (Litvin, Goldsmith, & Pan, 2018). OOPs is a community-based platform that allows customers to publish restaurant evaluations and potential customers to visit and read those reviews to aid in their decision-making (Li et al., 2013).

Customers who have recently visited a restaurant can leave a comment and a 1 to 5-star rating on the Zomato website to share their experience with others. The reviewer will then go to OOP, read the review, and evaluate both the normative (like star ratings) and informative (like recommendation frameworks) parts of argument quality (Eslami et al., 2018). Zomato has revolutionized the way consumers search for restaurants. It also assisted customers in finding affordable dining options.

In this study, it is aimed to make a price prediction for restaurant management. When entrepreneurs want to start a business, the proposed model will use machine learning to calculate the expected average price of a restaurant depending on the characteristics of the restaurants and will pave the way for access to suitable features. The survey helps entrepreneurs to decide while establishing a restaurant about prediction of price.

The study is organized as follows. Section 2 includes literature survey about prediction models about restaurant. Section 3 consists of four subsections. First subsection presents detailed of the dataset. Second subsection explains different machine learning methods such as XGBoost, Random Forest, Artificial Neural Network, K-Nearest Neighbor, Multi Linear Regression and CatBoost. Third subsection defines evaluation metrics. Fourth subsection presents the overall system architecture. Section 4 presents information about dataset and analysis. In Section 5, results are shown in detail and the performance metrics are compared. Section 6 provides feature importance. At a result, section 7 concludes the paper.

2. Literature Survey

This section summarizes the main contributions of the surveys in the literature. A brief explanation is given for each study and the difference of this study from other studies are explained.

The study investigated which contextual and descriptive features of restaurant reviews influence the reviewer's decision to think a review valuable and hence "Like" it using Dual Process Theory and Social Impact Theory. Both qualitative and quantitative approaches were used to examine a large data

sample of 58,468 restaurant reviews on Zomato. The findings proved that a reviewer's "Like" formation was influenced by the knowledge component of the positive recommendation framework as well as the normative factors of high argument quality and moderate recommendation ratings. Their research emphasized the crucial filtering feature that a heuristic could provide to potential customers, as well as the OOP's increased societal impact (Meek, 2021). Kulkarni et al. (2019) studied to learn about restaurants where people want to go and to determine the rating of the restaurant. In their study, different prediction models such as Support Vector Machine (SVM), Random Forest and Linear Regression, XGBoost, Decision Tree were used, and 83% score was obtained with ADABOOST. Wang (2016) et al. proposed a model using the yelp dataset to predict the success and rating of the new restaurant. They performed the Chi-square test and stochastic gradient descent to classify the restaurant features as having the greatest weight. They used classification algorithms such as SVM, Random Forest, Logistic Regression and Multilayer Neural Networks. Random Forest achieved 56% and Multilayer Neural Networks obtained 60% accuracy. They then did sentiment analysis on restaurant reviews, and using clustering algorithms, accuracy increased to 85%. Shihab et al. (2018) proposed a viable place to start a restaurant business based on available data from Yelp, where 75 features were extracted for supervised machine learning. The model calculated the expected score a restaurant receive based on the characteristics the restaurant has. Several machine learning algorithms (Support Vector Machine, Decision Tree, Logistic Regression and pre-order Decision Tree) used and selected the appropriate one. Because Yelp's review is genuine and maintained regularly, they considered a business's rating as a point of recommendation. Comparative analysis of these algorithms was applied and an algorithm that gave the best results was searched. Lunkad (2015) predicted a restaurant's rating using review data from data from Yelp. They used a support vector machine, linear regression, and a pure bias model. A better result was obtained with the linear regression value of 53.13%. Schmid et al. (2022) presented approaches to solving the sales forecasting problem in small and medium restaurants. The study examined a wide variety of ML techniques based on real datasets. LSTM and GRU neural networks were developed to assist with the gradient problem. Vicario et al. (2020) built on previous restaurant business bankruptcy prediction models. In their study, it was emphasized that many of the studies applied belonged to American companies. Deep Recurrent Convolutional Neural Network (DRCNN) was used in the study. A 10-year analysis period was chosen. Logistic Regression gave better results in studies with multiple discriminant analysis. The database used in the study was the SABI database, which included Spanish and Portuguese Companies. Unlike other studies, this study helped to predict a few years before bankruptcy. The DRCNN calculation technique provided the best success, and the technique showed a high level of classification accuracy, with a one-year accuracy of 93.50%, a two-year accuracy of 89.60%, and a three-year accuracy of 85.60%. Kim et al. (2014) focused on key factors of financial distress for US restaurants from 1988 to 2010 by applying AdaBoosted Decision Trees. The AdaBoosted Decision Tree model is recommended for the early warning system because it showed the highest prediction performance with the least error as a result. The overall accuracy with AdaBoosted Decision Tree was higher than with Decision Tree. AdaBoosted Decision Tree also outperformed Decision Tree in the kappa coefficient. Tsoumakas (2019) reviewed current ML approaches for food sales forecasting. A daily sales forecast was required for products with a short shelf life, and a weekly sales forecast was required for products with a long shelf life. Lagged variables as input variables were the main mechanism by which propositional learning algorithms capture the relationship between the past and present values of a series. It was emphasized that sequential lagging variables could be averaged over a single field to reduce the number of input variables, as a large number of input variables might have a negative effect on some learning algorithms. Gu et al. (2002) attempted to identify the financial features that distinguished bankrupt restaurant companies from non-bankrupt restaurants. A model of bankruptcy classification based on multiple discriminant analysis (MDA) has been transferred. The classification variable was taken as the properties of the failed restaurants. Paired sampling was adopted to develop the MDA model. This model classified firms within the sample quite accurately,

with an accuracy of 92 percent 1 year before bankruptcy.

The main contribution of this paper is to provide a model that presents a price prediction model using a dataset containing very wide parameters of all popular restaurants in Istanbul. While other studies make score predictions, this study aims to fill the price prediction gap and enable entrepreneurs who will open restaurants to make accurate pricing strategy.

3. Material and Method

3.1. Dataset

A model is proposed using data collected from Zomato ([zomato.com/Istanbul](https://www.zomato.com/Istanbul))¹. From the Zomato platform, information about a total of 7.932 restaurants in popular localities in and around İstanbul, including 24 features of the restaurants was obtained. Restaurants data in Taksim, Kadıköy, Beşiktaş, Karaköy, Nişantaşı, Moda, Asmalımescit, Etiler, Levent, Bebek, Ataşehir, Galata, Caddebostan, Ortaköy, Kuruçeşme, Çengelköy, Eminönü, Balat, Şenlikköy, Cihangir, Akaretler, Üsküdar, Bakırköy, Şişli, Koşuyolu, Kalamış, Arnavutköy, Süleymaniye, Ataköy are collected for the analysis. In this study, data was extracted with web-scraping technology and preprocessed for price prediction. Features are indicated in the Table 1.

Table 1: Types of attributes found in the dataset and descriptions

Variables	Data Type	Definition	Coding
Distance	Numeric	The distance between the restaurant and the town center	
Price	Numeric	The average cost for two people in the venue	
Comments	Numeric	Comments about restaurant.	
Score	Numeric	Rate of satisfaction.	
Alcohol_Available	Categorical	Does restaurant have alcohol option?	1: Yes 0: No
Home_Delivery	Categorical	Is your food order delivered right to your door?	1: Yes 0: No
Self_Service	Categorical	Does the restaurant have a self-serve?	1: Yes 0: No
Brunch	Categorical	Is there a late morning meal at the restaurant?	1: Yes 0: No
Sports_Broadcast	Categorical	Does it have sports option?	1: Yes 0: No
Parking_Lot	Categorical	Does it have private parking area?	1: Yes 0: No
View	Categorical	Does it have landscape view?	1: Yes 0: No
Outdoor_Seating	Categorical	Does restaurant have outdoor seating area?	1: Yes 0: No
Internet_Phone_Charge	Categorical	Does restaurant have wifi Internet, phone charge options?	1: Yes 0: No
Desserts and Bakes	Categorical	Does restaurant have desserts and Bakes options?	1: Yes 0: No
Smoking_Area	Categorical	Does it has smoking area?	1: Yes 0: No
Non_Alcohol_Available	Categorical	Is the restaurant non-alcohol?	1: Yes 0: No
Pet_Friendly	Categorical	Is it pet friendly place?	1: Yes 0: No
Additional_Feature	Categorical	Does restaurant have the other options?	1: Yes 0: No
Music	Categorical	Does restaurant have music option?	1: Yes 0: No
Buffet	Categorical	Does the restaurant have a buffet option?	1: Yes 0: No
Luxury_Dining	Categorical	Does the restaurant have a luxury dining option?	1: Yes 0: No
Organic_Vegan_Vegetarian	Categorical	Does the restaurant is organic, vegan or vegetarian options?	1: Yes 0: No
Good_for_Working	Categorical	Is the restaurant good for working?	1: Yes 0: No
World_Cuisine	Categorical	Does restaurant have world cuisine option?	1: Yes 0: No
Turkish_Cuisine	Categorical	Does restaurant have Turkish cuisine option?	1: Yes 0: No

¹ <https://www.zomato.com/istanbul>

3.2. Algorithms

In this section machine learning algorithms are explained in detail.

3.2.1. Random Forest (RF)

The random forest algorithm is a supervised learning algorithm used for classification and regression problems. As we can predict on the name, the random forest algorithm consists of decision trees trained with the bagging technique. In the bagging method, base learners are randomly trained with subsets in the training set. As the number of trees in the random forest algorithm increases, the algorithm gives more precise results. (Breiman, 2001) The biggest advantage of the random forest algorithm is to perform well in datasets with missing data. At this point, random forest algorithms are frequently used for both large datasets and small datasets. Another advantage of the random forest algorithm is that it provides a deeper exploration of the dataset by establishing various models on the dataset.

3.2.2. Artificial Neural Networks (ANN)

Artificial neural networks are an algorithm inspired by the working principles and functions of the human brain. The working principle of artificial neural networks is the same as the working of neural networks in the brain. Learning in biological systems is provided by synaptic connections between neurons. Information from people's sense organs updates synaptic connections. In artificial neural networks, on the other hand, samples represent information coming from sense organs. Learning occurs as a result of using examples and associating them with results. Training, on the other hand, refers to the process that continues until the determination of connection weights using examples and obtaining the best results (Kukreja, N, S, & S, 2016).

A true neuron cell consists of dendrite, soma, axon, and synapsis. Dendrites are the transmitters that transmit information to the soma, that is, the nucleus. At this point, there is a direct communication between the dendrites and the soma. Likewise, the soma has the ability to filter information from dendrites. After filtering the information from the soma dendrites, it transmits it to the synapses via the axon. The task of synapsis is to convert the signals coming from the axon to a certain threshold value and transfer them to other cells. Like the way neurons work, activation functions in artificial neural networks do the transfer process together with the conversion process, just like synapses. The relationship between inputs and transfer functions here represents the relationship between dendrites and soma. Weights are parameters in the filtering process. Artificial neural networks have many advantages. There are many cells in artificial neural networks and these cells have the feature of working simultaneously. In addition, it can work with different learning algorithms of artificial neural networks and can process real-time information. Neural networks can complete missing patterns by recognizing and classifying patterns. Neural networks have fault tolerance. This feature allows them to work with datasets with incomplete or ambiguous information.

3.2.3. K-Nearest Neighbor (KNN)

K-Nearest Neighbor algorithm is used in classification and regression modeling and is considered the easiest supervised learning algorithm compared to other algorithms. This algorithm first emerged for the solution of classification problems, then it began to be used for solving regression problems. K-Nearest Neighbors algorithm is one of the other supervised learning algorithms. Unlike, it does not have a training stage. In the approach of this algorithm, training and testing are considered the same operation. The absence of a training set makes this algorithm very easy to implement, but it is not recommended to be used on large data sets because it has low performance compared to other algorithms. (Cover & Hart, 1967)

In the K Nearest Neighbor algorithm, predictions are made based on observation similarity. It is a non-parametric supervised learning algorithm compared to other algorithms. The algorithm is

implemented in 4 steps. First, the number of neighbors is determined. The number of neighbors is expressed as the “k” value. The distance between the unknown point and all other points is calculated. The distances are listed in ascending order and the closest observation is selected. The predictive value is most often shown as the class in classification and the mean value in regression.

3.2.4. Extreme Gradient Boost (XGBoost)

XGboost algorithm is a decision tree-based machine learning algorithm. This algorithm is known as the best among decision tree-based algorithms. The reason for this is that it has been developed with various optimization and software techniques in order to make better predictions with less resource usage. This algorithm was found by Tiangi Chen and Carlos Guestrin in 2016. The biggest advantages of this algorithm are that it has high predictive power, can overcome overfit situations, can manage empty data and do all these functional operations quickly. According to the founders of the algorithm, the XGboost algorithm works 10 times faster than other decision tree-based and popular algorithms (Chen & Guestrin, 2016). The XGboost algorithm is used to construct classification and regression models. In this study, Python language was used to establish the XGboost model and the model parameters `colsample_bytree`, `learning_rate`, `max_depth` and `n_estimators` parameters were optimized.

3.2.5. Catboost Algorithm

Catboost algorithm is a gradient boosting based machine learning algorithm and developed by yandex in 2017. The algorithm was first introduced with the article “Catboost: unbiased boosting with categorical features” published in April 2017. The name of the Catboost algorithm is formed by combining the words “Category” and “Boosting” (Alshari, Saleh, & Odabas, 2021). The Catboost algorithm has a high learning speed ability. With this ability, it can work quickly with categorical, numerical and textual data. In addition, unlike other algorithms, having visualization options and GPU support provides ease of use in various fields (Alshari, Saleh, & Odabas, 2021). Catboost algorithm does not need coding while preparing data. In this way, the data preparation process takes less time and enables the algorithm to work with high performance, especially in categorical data.

In the study, the Catboost algorithm was used for price estimation in regression modeling. The algorithm was implemented with the `CatBoostRegressor` method by loading the Catboost algorithm in the Jupyter Lab environment in the Python programming language. The parameters used in the model are `iterations`, `learning_rate` and `depth`, and these parameters are optimized in the model tuning section.

3.2.6. Multi Linear Regression (MLR)

Multiple linear regression is a method used to see how independent features are related to the feature affected by other features expected to be explained in a data set (Marill, 2004). In order to determine the order of importance of the independent variables, MLP determines the effect of the independent variables in the regression equation on the dependent variable. Simple linear regression looks at whether only one feature explains the dependent feature, while multiple linear regression looks at the status of more than one feature (Marill, 2004).

3.3. Model Evaluation Metrics

3.3.1. Mean Squared Error (MSE)

MSE is an evaluation method that measures the performance of the estimator in regression models. In this evaluation metric, the estimated values are subtracted from the real values and squared. Then result of all values are added together and divided by the number of observations (Kervancı & Akay, 2020).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (1)$$

The result in the MSE value is always greater than zero, and the value close to zero is considered more successful.

3.3.2. Root Mean Squared Error (RMSE)

RMSE is a metric that calculates the distance between predicted and actual values and measures the magnitude of the error. In simpler words, it is the standard deviation of the predicted values. RMSE is calculated by squaring the value resulting from the operations (Cinaroglu, 2017).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

The RMSE value can range from 0 to ∞ . Algorithms with a lower value on this metric are considered to perform better.

3.3.3. Mean Absolute Error (MAE)

MAE is an error metric created by summing the absolute error values. In this metric, the first estimated values are subtracted from the actual values. These values are then summed up in their absolute value and divided by the number of observations (Cinaroglu, 2017). MAE value is always positive and can take an infinite value. Algorithms with lower values in this evaluation metric are considered to perform better.

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_i - y_j| \quad (3)$$

3.4. Proposed Method

Data for price prediction is obtained from the Zomato website. The diagram of the overall architecture is indicated in Figure 1. Firstly, data is extracted by scraping technology. Then data is sent to a csv file to make analysis. Bagging Algorithms, ANN, SVM, XGBoost, KNN and RF are applied to data. As a result, the model generates the predicted outputs.

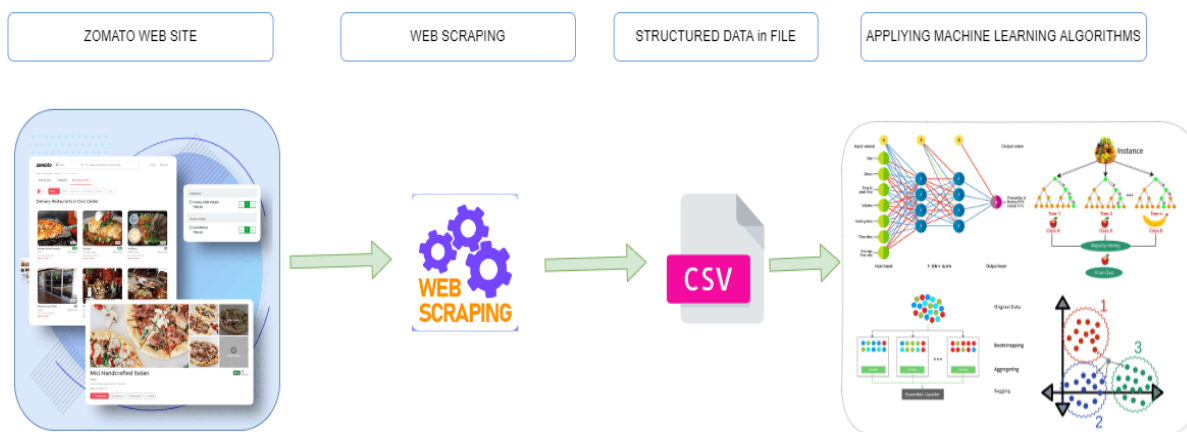


Figure 1: System Overview

4. Experimental Results

In this section, a model is proposed on the restaurant data of XGBoost, ANN, KNN, RF, Multiple Linear Regression and CatBoost classification algorithms and the analysis results are shown in detail. Python is used in the implementation of the algorithms.

4.1. Correlation Analysis

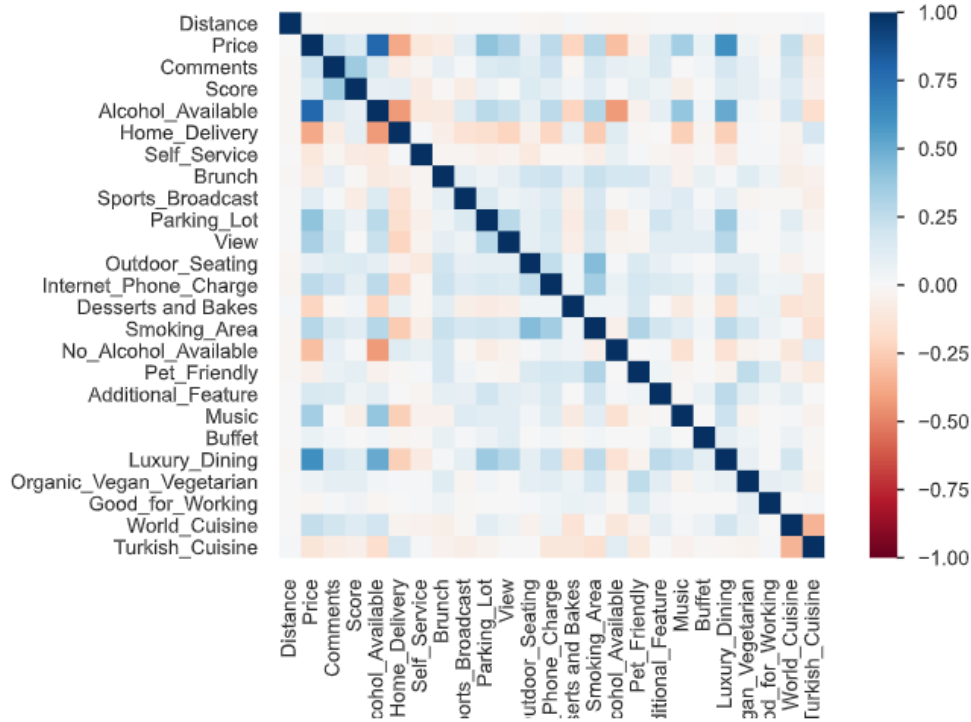


Figure 2: Correlation Analysis

In Figure 2, correlations among variables are presented. Figure 2 indicates that there is a high and positive relationship between the price variable and the alcohol available variable. It can be stated that the presence of the alcohol available variable also increases the price. Likewise, a high and positive relationship is observed between the price variable and luxury dining. Luxury dining variable increases the price. There are moderate and weak relationships among other variables.

4.2. Sample from Dataset

	Distance	Price	Comments	Score	Alcohol_Available	Home_Delivery	Self_Service	Brunch	Sports_Broadcast	Parking_Lot	...	No_Alcohol_Available	F
0	191	450.0	1737.0	4.2	1	0	0	0	0	0	0	0	0
1	183	420.0	1385.0	4.8	1	0	0	0	0	0	0	0	0
2	208	220.0	1891.0	4.1	1	0	0	1	0	0	0	0	0
3	308	260.0	511.0	4.4	1	0	0	0	0	0	0	0	0
4	112	280.0	603.0	4.7	1	0	0	0	0	0	0	0	0
...
7927	1200	65.0	8.0	3.6	0	1	0	0	0	0	0	0	0
7928	1300	30.0	40.0	4.0	0	1	0	0	0	0	0	0	0
7929	1400	70.0	10.0	2.8	0	0	0	1	0	0	0	0	1
7930	925	90.0	15.0	3.2	0	1	0	0	0	0	0	0	1
7931	473	60.0	64.0	4.3	0	1	0	0	0	0	0	0	0

Pet_Friendly	Additional_Feature	Music	Buffet	Luxury_Dining	Organic_Vegan_Vegetarian	Good_for_Working	World_Cuisine	Turkish_Cuisine
0	0	1	0	1	0	0	1	1
0	0	0	0	1	0	0	1	1
1	1	0	0	1	0	0	0	0
0	0	0	0	1	0	0	1	0
0	0	0	0	1	1	0	1	0
...
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0

Figure 3: Sample from Dataset

Sample from dataset are presented in Figure 3.

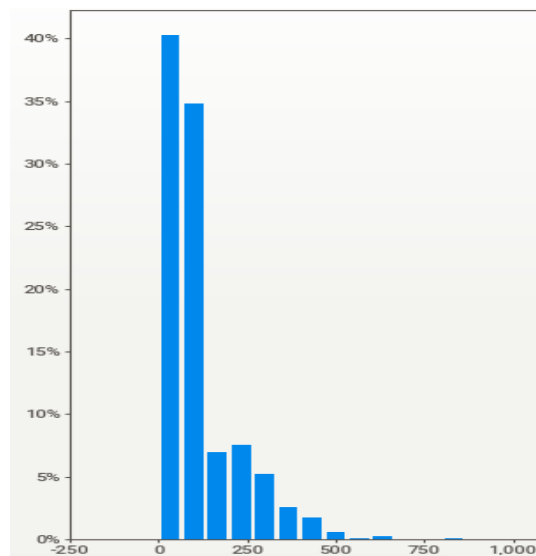


Figure 4: Price of Restaurants

Figure 4 presents price of restaurants in Istanbul.

4.3. Parameters of the Machine Learning Models

The models use many parameters. The values of these hyperparameters can change the performance of the models. The hyperparameters used in the models are explained below:

- Dataset is normalized.
- 70% dataset is reserved for train and 30% for testing.
- MSE, RMSE, MAE and R^2 values are used to calculate the error rate.
- In ANN model, hidden layers changes from 50 to 150, Optimizer is rmsprop, batch size is 1024 and epoch is 5000 cycles by trial and error and selu is used for activation function.
- In XGBoost model, colsample_bytree is 0.5, learning_rate is 0.02, max_depth is 4 and n_estimators is 500.
- For Random Forest model, parameters are tuned according to the following: max_depth is 9, max_feature is 5 and n_estimators is 500.
- In CatBoost model, iterations are 1000, learning_rate is 0.01 and depth is 8.

5. Results

In this study, the performances of XGBoost, Artificial Neural Networks, Multi Linear Regression, Random Forest, CatBoost and K-Nearest Neighbor based models are presented in terms of RMSE, MSE and MAE model evaluation metrics.

Table 2: Performance of the models in terms of RMSE, MSE and MAE

Model	RMSE	MSE	MAE
ANN	0.026035	0.0006778	0.01748
KNN	0.024085	0.0005800	0.01487
RF	0.023258	0.0005409	0.01463
XGBoost	0.023236	0.0005399	0.01468
CatBoost	0.023954	0.000573	0.014681
MLR	0.024168	0.00058	0.015861

The proposed model predicted price for restaurants. The performances of the models were calculated, and the results are shown in Table 2. When the restaurant dataset was analyzed with Random Forest and CatBoost models, it was determined that the most important parameter in price estimation for restaurants was the Score feature. The findings from Table 2 show that XGBoost algorithm achieved the highest performance rate among the other algorithms. The RMSE, MSE, and MAE values of the data analysed with XGBoost are 0.023236, 0.0005399 and 0.01468 respectively. Low scores indicate better for evaluation. It was observed that the XGBoost model has obvious advantage for price prediction for restaurant.

6. Feature Importance

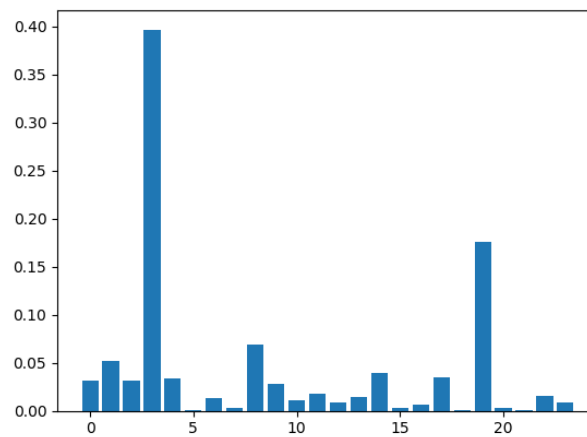


Figure 5: Feature Importance Values

Figure 5 indicates the feature importance with RF algorithm. The results from figure show that Score has the most impact on price prediction with 0.39668. The second important feature is Luxury Dining. The attributes Home Delivery and Organic, Vegan or Vegetarian features provide lowest contribution to analysis with 0.00046 and 0.00059 respectively.

7. Result and Discussion

Pricing is one of the 4Ps of marketing strategy, which is a model for improving the components of the marketing mix, i.e. the way a new product or service is brought to market. A price prediction model is proposed for entrepreneurs who want to open a restaurant by data science. According to the ML models that are applied in this study, Score is the most important parameter for prediction of restaurant price; however, Home Delivery and Organic, Vegan or Vegetarian features have no effect on price prediction model. In this study, a model is proposed for restaurant price prediction using various machine learning algorithms such as XGBoost, Random Forest, Artificial Neural Network, K-Nearest Neighbor, Multi Linear Regression and CatBoost. According to the findings, XGBoost gave the best prediction result among other algorithms. By this study, entrepreneurs who tend to open a restaurant will be aware about what and how much they should cost and will increase their profitability by selling more products as it will provide price performance. It is imperative for the restaurant to make the right pricing in order to hold on and build customer loyalty.

For the future study, it is planned to increase the scope and dataset by obtaining the dataset of the restaurants in the other three largest cities of Turkey.

AUTHOR CONTRIBUTION

The author contributed to the entire study.

STATEMENT OF CONFLICT OF INTEREST

There is no financial conflict of interest with any institution, organization, or person.

REFERENCES

- Alshari, H., Saleh, A. Y., & Odabas, A. (2021). Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance. *Erciyes University Journal of Institute Of Science and Technology*, 160-161.
- Breiman, L. (2001). Random Forest. *University of California*, 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- Cinaroglu, S. (2017). Comparison of Machine Learning Regression Methods to Predict Health Expenditures. *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, 189.
- Cover, T., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE*, 21-27.
- Ding, S., & Chen, L. (2010). Intelligent Optimization Methods for High-Dimensional Data Classification for Support Vector Machines. *Intelligent Information Management*, 1-12.
- Eslami, S. P., Ghasemaghahi, M., & Hassanein, K. (2018). Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems*, 113, 32-42.
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 1189-1232.
- Gu, Z. (2002). Analyzing bankruptcy in the restaurant industry: A multiple discriminant model. *International Journal of Hospitality Management*, 21(1), 25-42.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer*, 211.
- Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, 36, 354-362.
- Kervancı, I. S., & Akay, M. (2020). Review on Bitcoin Price Prediction Using Machine Learning and Statistical Methods. *Sakarya University Journal of Computer and Information Sciences*, 273-281.
- Kukreja, H., N, B., S, S. C., & S, K. (2016). AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK. *Journal of Electrical & Electronics Engineering, School of Engineering & Technology, Jain University*, 27-29.
- Kulkarni, A., Bhandari, D., & Bhoite, S. (2019). Restaurants Rating Prediction using Machine Learning Algorithms. *International Journal of Computer Applications Technology and Research*.
- Li, M., Huang, L., Tan, C. H., & Wei, K. K. (2013). Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, 17(4), 101-136.
- Litvin, S. W., Goldsmith, R. E., & Pan, B. (2018). A retrospective view of electronic word-of-mouth in hospitality and tourism management. *International Journal of Contemporary Hospitality Management*.
- Lunkad, K. (2015). Prediction of Yelp Rating using Yelp Reviews.

- Meek, S., Wilk, V., & Lambert, C. (2021). A big data exploration of the informational and normative influences on the helpfulness of online restaurant reviews. *Journal of business research*, 125, 354-367.
- Osman, T., Mahjabeen, M., Psyche, S. S., Urmi, A. I., Ferdous, J. S., & Rahman, R. M. (2016, June). Adaptive food suggestion engine by fuzzy logic. *In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* (pp. 1-6). IEEE.
- Schmidt, A., Kabir, M. W. U., & Hoque, M. T. (2022). Machine Learning Based Restaurant Sales Forecasting. *Machine Learning and Knowledge Extraction*, 4(1), 105-130.
- Shihab, I. F., Oishi, M. M., Islam, S., Banik, K., & Arif, H. (2018, December). A machine learning approach to suggest ideal geographical location for new restaurant establishment. *In 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 1-5). IEEE.
- Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, 52(1), 441-447.
- Vicario- Becerra, R., Alaminos, D., Aranda, E., & Fernández-Gómez, M. A. (2020). Deep recurrent convolutional neural network for bankruptcy prediction: A case of the restaurant industry. *Sustainability*, 12(12), 5180.
- Wang, A., Zeng, W., & Zhang, J. (2016). Predicting New Restaurant Success and Rating with Yelp. ser. CS221: Artificial Intelligence: *Principles and Techniques*, Stanford University.