Food Demand Forecast for Online Food Delivery Service Using CatBoost Model

Ansh Pujara, V. Pattabiraman, and R. Parvathi

Abstract Online food ordering has been proven a great source for businesses from a wide range of sectors. By using an online food ordering system, you can get your food to be delivered to your door without consuming much time. For businesses operating in the food industry including restaurants, agriculture, and many others, accurate forecast is of crucial importance because of the unpredictable demand pattern. In several studies, the choice of an appropriate forecasting model remains a concerning point. In this context, this research aims to analyse the performance of the CatBoost Gradient boosting model for the prediction of the amount of raw materials required for a meal delivery company that operates in multiple cities having multiple centres.

Keywords Food demand · CatBoost · Demand forecast

1 Introduction

Demand forecasting is an integral component of any rising online business. Without proper demand forecasting techniques in place, it can be nearly impossible to have the required amount of stock available at any given time [1]. A major facet of businesses today is the notion of supply chain integration, whereby resources are combined to provide value to the end consumer and where all the upstream firms realize the importance of integration.

An online food delivery service has to face several difficulties like dealing with a lot of perishable raw materials, which makes it perhaps more essential for such a company to accurately forecast daily and weekly demand [2]. The task is to predict the demand of raw materials for 10 weeks in future using the historical data for a product-centre combination of the past 145 weeks.

A. Pujara · V. Pattabiraman · R. Parvathi (✉)
Vellore Institute of Technology, Chennai, Tamil Nadu, India
e-mail: ansh.pujara2018@vitstudent.ac.in; pattabiraman.v@vit.ac.in; parvathi.r@vit.ac.in

A. Haldorai et al. (eds.), *3rd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-78750-9_9

Prediction of the type of raw material required at a particular delivery centre/branch depends upon numerous factors such as the type of meal ordered at the centre the most, the type of raw materials available nearest to a centre, etc. This chapter will mainly focus on one factor, that is, the type of meal ordered from a particular centre. The results will determine the demand for raw materials for the centre-meal combination [2, 3].

Many forecasting models have been developed in recent years, and almost all of them share the same basic idea, but follow paradigms from different fields [1]. But the present chapter aims at the performance of the CatBoost Regressor. In the next section, the theoretical structure will be discussed in order to support the model proposed in this study.

2 Literature Survey

Online food ordering is the way of ordering food through an Internet site or a mobile phone application, generally handled by a particular restaurant or a mediator service (such as Zomato, Swiggy, Dunzo, etc.). The product may either be a ready-to-eat food product (delivered directly from a licenced home kitchen, restaurant, or a ghost kitchen) or food not explicitly prepared for direct consumption (like vegetables directly from a farm, frozen meats, etc) [3, 4]. The history dates back to 1994, when the famous pizza chain, Pizza Hut started accepting orders online. Since then, restaurants, hotels, groceries, etc. are majorly accepting orders through online mode. In online food orders placed at restaurants/chains, the restaurants prefer to use their own website/ app, or opt to employ a delivery vendor. Usually, food delivery orders are on demand and meant to be consumed immediately, including hot meals already prepared. Ordering food for delivery usually includes contacting a local restaurant or chain online or by telephone. For deliveries in large cities, it is not possible to deliver the meal on time and fresh [5]. Thus, restaurants have their own chains/franchises in major landmarks, which helps them to attract customers from the whole city.

According to a 2019 market analysis on food delivery services, it was found that the global online ordered food delivery market was valued at ₹7.1755 trillion and is expected to rise approximately 9% a year, approaching ₹10.267 trillion by 2023 [5]. The study defined the market as follows:

1. Meals ordered online that the restaurants deliver directly, either ordered via a portal (e.g. Zomato) or through its website.
2. Online meal orders and deliveries all driven by a platform (like Swiggy).
3. Online orders for takeaway in the restaurant by the customer. Excluding orders made via phone call (Fig. 1).

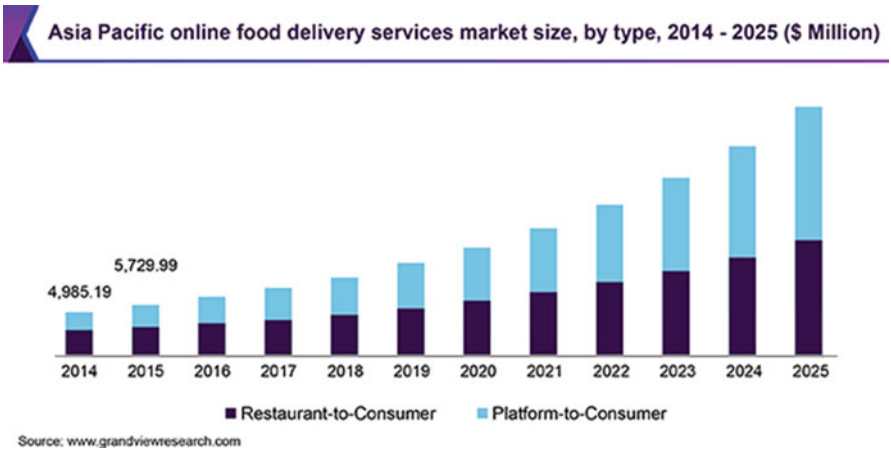

Fig. 1 Market analysis of online food delivery system [5]

3 Methodology

3.1 CatBoost Regressor

CatBoost is based on the Gradient Boosting library. Gradient boosting is an effective machine-learning technique that is extensively used in various types of business challenges like fraud detection, recommendations, demand/ supply forecast, and it achieves state-of-the-art results. Also, decision trees have been found as the base predictors when it comes to the implementation of gradient boosting algorithms. While dealing with pure numerical data, it is suitable to use decision trees, but during implementation many datasets contain categorical features that sometimes act as most important features for prediction [6].

A standard solution for dealing with categorical features that are not necessarily equivalent to one another is to substitute each feature with a number at the time of pre-processing [6, 7].

Gradient Boosting Algorithms like XGBoost, LightGBM, etc. deals with categorical data through one-hot encoding of such features before training the model. Unlike other algorithms, CatBoost does not require one-hot encoding of categorical features during model building.

CatBoost, similar to the various Gradient Boosting strategies, fabricates new trees to surmise the angle of the present tree. CatBoost utilizes a more efficient procedure that reduces overfitting and permits to utilize the entire dataset for preparation [6]. To be specific, a random permutation of the dataset in use is carried out, and an average label value for each example is computed. CatBoost comprises two boosting models, Ordered and Plain. Here, Ordered boosting is used [7].

In CatBoost, oblivious trees are used as base predictors. In such trees, the same criterion of splitting is used over a whole level of the tree point. Such trees are

balanced and less likely to undergo overfitting [6]. CatBoost model evaluator makes use of these facts: It first converts all used one-hot encoded features, float features and statistics into binary and then uses those binary features to compute model predictions [6, 7].

Assuming a dataset having $\{(x_k, y_k)\}_{k=1}^{n}$, , where x_k is a vector of features chosen randomly and y_k is a target, initially CatBoost generates $s + 1$ independent random permutations of the dataset $(\sigma_0, \ldots, \sigma_s)$ [6, 7]. Now, permutations $\sigma_1, \ldots, \sigma_s$ will be used for evaluation of splits, while σ_0 will be used for choosing the leaf values of the trees obtained [7]. In ordered boosting, we have several supporting models $M_1, \ldots M_n$such that each M_i model will act as the current prediction for i the sample into the permutation σ_r [7]. In order to prevent "prediction shift", CatBoost makes use of permutations such that $\sigma_1 = \sigma_2$. This technique prevents the use of target for training of model for "Target Statistic" calculation or "gradient estimation".

Algorithm 1: Ordered Boosting

input: $\{(x_k, y_k)\}_{k=1}^{n}, I;$

$\sigma \leftarrow random\ permutation\ of\ [1, n]; M_i \leftarrow 0\ for\ i = 1..n;$

for $t \leftarrow 1$ **to** I **do**

  for $i \leftarrow 1$ **to** n **do**

    $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_i);$

  for $i \leftarrow 1$ **to** n **do**

    $\Delta M \leftarrow LearnModel((x_j, r_j): \sigma(j) \leq i);$

    $M_i \leftarrow M_i + \Delta M;$

return M_n

Note: Originally published in "*CatBoost: unbiased boosting with categorical features support*" [4]

How CatBoost works?
1. Initially, one categorical feature is chosen to start with, let us say x.
2. Next, randomly a row is chosen from the training set, and a random level of this categorical feature is exchanged with a number.
3. The number chosen is generally based on the target variable depending on the category level. To put it another way, the target number is based on the estimated output variable.
4. A splitting function is used to construct two different sets of training data: One set with all the features will have a greater target variable, and the other set with smaller target variables.

4 Implementation

Here, Python3 is used to implement the model available in the CatBoost library by Yandex. To implement the CatBoost regressor, Online Food Delivery System's dataset is used (available online). The dataset contains historical data from 77 distinct delivery centres each of either Type A, Type B or Type C. The time frame of the data provided is 145 weeks. It contains several categorical features that have a direct impact on the prediction like operating area, type of cuisine, category of cuisine, etc. The aim is to perform regression analysis on the number of orders for the upcoming 10 weeks, which will tell about the required raw materials. Figure 2 is a glimpse of the dataset used in this study.

An exploratory analysis of the dataset was done and is shown below to observe the relationship between the variables (Figs. 3, 4, 5, 6, 7 and 8).

CatBoost offers a versatile interface for parameter selection and tuning. The parameters are selected after observing different values through grid search, which builds a model for every combination of hyperparameters specified and evaluates each model. The considered values for the parameters were: learning rate = [0.02, 0.03, 0.1]; tree depth = [4, 6, 8]; l2 regularization = [1, 3, 6, 9, 10]. The below table shows the best combinations where the resulting RMSE values were best (Table 1 and Fig. 9).

As RMSE (Root Mean Squared Error) is the loss function used in this analysis, the data needs to be converted into log values. Some of the features were even transformed for better understanding of the model. In this dataset, almost all the features were equally important for the prediction, but to achieve a better result, some extra features were also added like discount on base price, discount ratio, difference between price, certain *lag* features and *ewm* (exponentially weighted mean) features.

Below are shown some of the features participating in the forecast (Fig. 10).

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). It is the square root of MSE (Mean Squared Error) [8, 10]. In case of unbiased estimators, RMSE is just the square root of variance, which is actually Standard Deviation [2, 8]. In case of RMSLE (Root Mean Squared Logarithmic Error), the log of the predictions and actual values are taken. So basically, changes occur in the variance that is being measured [2].

Table 1 RMSE values for different combinations of hyperparameters

Learning rate	Tree depth	L2 regularization	**RMSE** (evaluation metric)
0.02	4	6	0.4547964942
0.03	6	6	0.443667502
0.1	6	9	0.4441142703
0.02	8	10	0.4396277843

	id	week	center_id	meal_id	checkout_price	base_price	\
0	1379560	1	55	1885	136.83	152.29	
1	1018704	2	55	1885	135.83	152.29	
2	1196273	3	55	1885	132.92	133.92	
3	1116527	4	55	1885	135.86	134.86	
4	1343872	5	55	1885	146.50	147.50	
5	1493612	6	55	1885	146.53	146.53	
6	1110832	7	55	1885	145.53	146.53	
7	1461167	8	55	1885	146.53	145.53	
8	1102364	9	55	1885	134.83	134.83	
9	1018130	10	55	1885	144.56	143.56	

	emailer_for_promotion	homepage_featured	num_orders	city_code_x	...	\
0	0	0	177	647	...	
1	0	0	323	647	...	
2	0	0	96	647	...	
3	0	0	163	647	...	
4	0	0	215	647	...	
5	0	0	285	647	...	
6	0	0	148	647	...	
7	0	0	135	647	...	
8	0	0	175	647	...	
9	0	0	175	647	...	

	center_type_y	op_area_y	category_y	cuisine_y	city_code	region_code	\
0	TYPE_C	2.0	Beverages	Thai	647	56	
1	TYPE_C	2.0	Beverages	Thai	647	56	
2	TYPE_C	2.0	Beverages	Thai	647	56	
3	TYPE_C	2.0	Beverages	Thai	647	56	
4	TYPE_C	2.0	Beverages	Thai	647	56	
5	TYPE_C	2.0	Beverages	Thai	647	56	
6	TYPE_C	2.0	Beverages	Thai	647	56	
7	TYPE_C	2.0	Beverages	Thai	647	56	
8	TYPE_C	2.0	Beverages	Thai	647	56	
9	TYPE_C	2.0	Beverages	Thai	647	56	

	center_type	op_area	category	cuisine
0	TYPE_C	2.0	Beverages	Thai
1	TYPE_C	2.0	Beverages	Thai
2	TYPE_C	2.0	Beverages	Thai
3	TYPE_C	2.0	Beverages	Thai
4	TYPE_C	2.0	Beverages	Thai
5	TYPE_C	2.0	Beverages	Thai
6	TYPE_C	2.0	Beverages	Thai
7	TYPE_C	2.0	Beverages	Thai
8	TYPE_C	2.0	Beverages	Thai
9	TYPE_C	2.0	Beverages	Thai

[10 rows x 27 columns]

Fig. 2 Training dataset (first 10 rows)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{1}$$

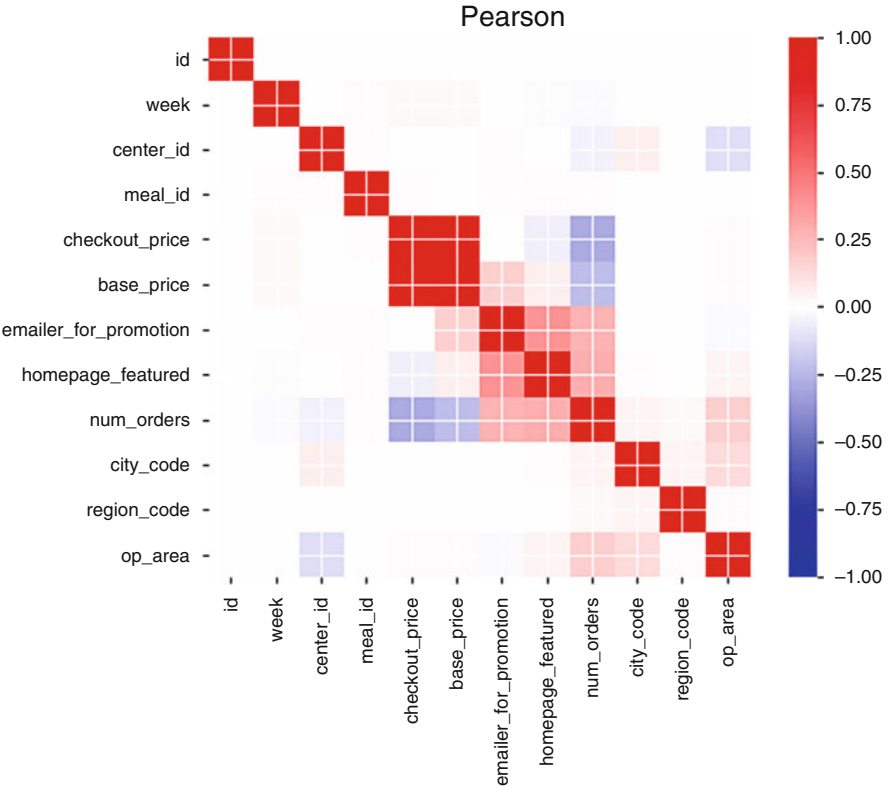

Fig. 3 Pearson correlation analysis between variables

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\log\ (p_i + 1) - \log\,(a_i + 1))^2} \tag{2}$$

Once the features are decided, the data is split into train, validation and test. But here, the train data is divided for validation using K-Fold Cross Validation. And finally, the regression model is applied to the test dataset resulting in the number of orders for the next 10 weeks data.

5 Analyses and Results

This research on the demand forecasting of an Online Food Delivery service has brought some insightful points to notice. The evaluation metric, Root Mean Squared Logarithmic Error value obtained in this study was 0.0887.

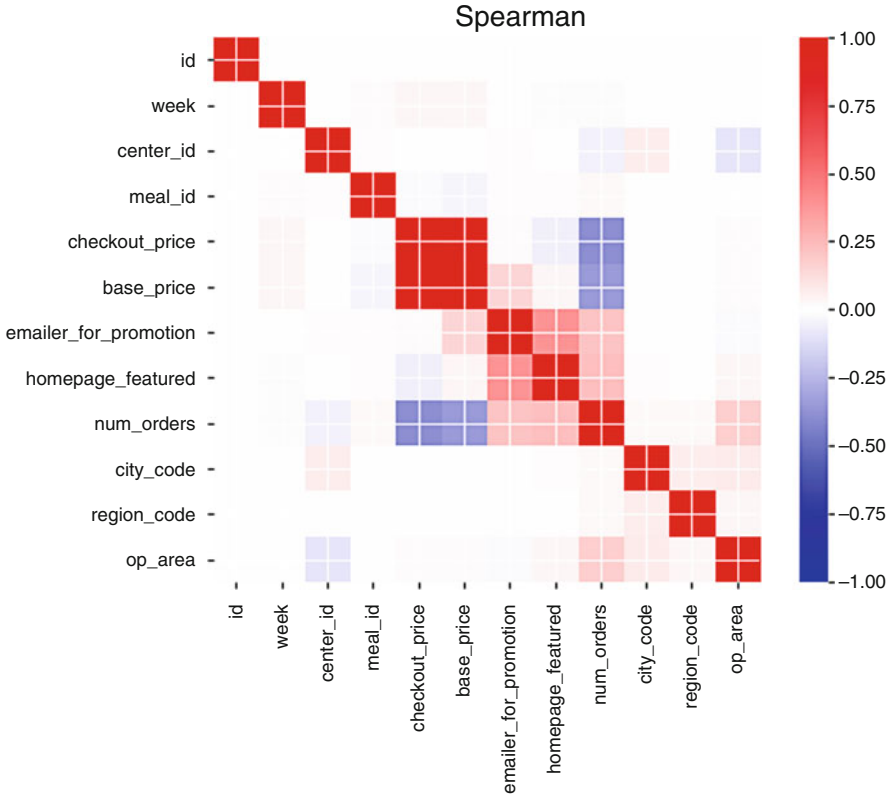

Fig. 4 Spearman correlation analysis between variables

When compared to other studies, the following results we obtained (Table 2).

The fit of the model can be observed in the below figure (Fig. 11), which depicts the sensitivity analysis of the model.

For a precise output of a forecasting model, adjustments must be made to the parameters for the technique applied. CatBoost, just like other machine learning packages, provides with a way to compare the feature (variable) importance [6]. This can help in building a better model, and allows for easier interpretation of results. Figure 12 is a plot showing the Feature Importance.

It is observed that avg_orders (average orders) have the highest importance among all variables. Evidently, CatBoost has shown some state-of-the-art results with most of the popular datasets containing categorical features. In this chapter, it shows more than expected results.

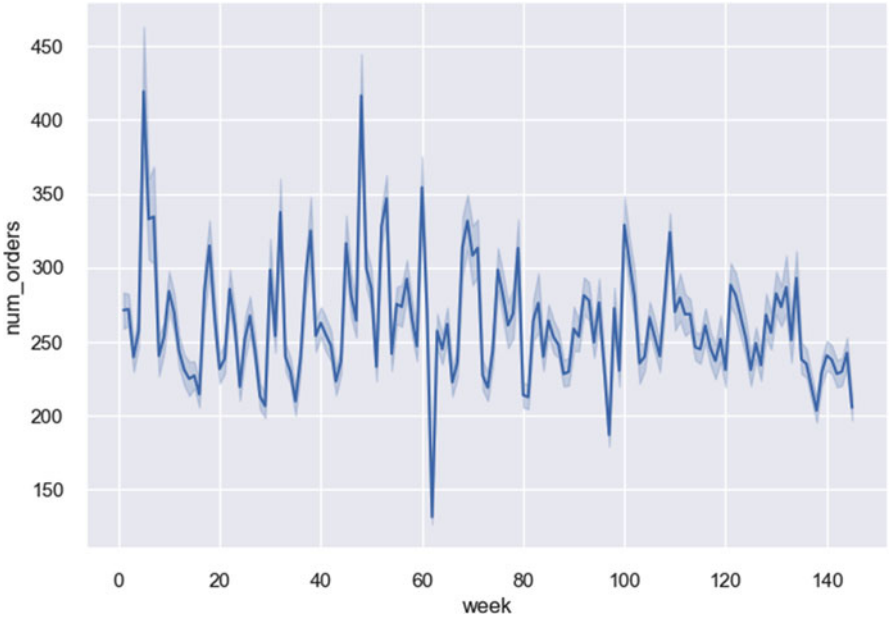

Fig. 5 Weeks vs number of orders

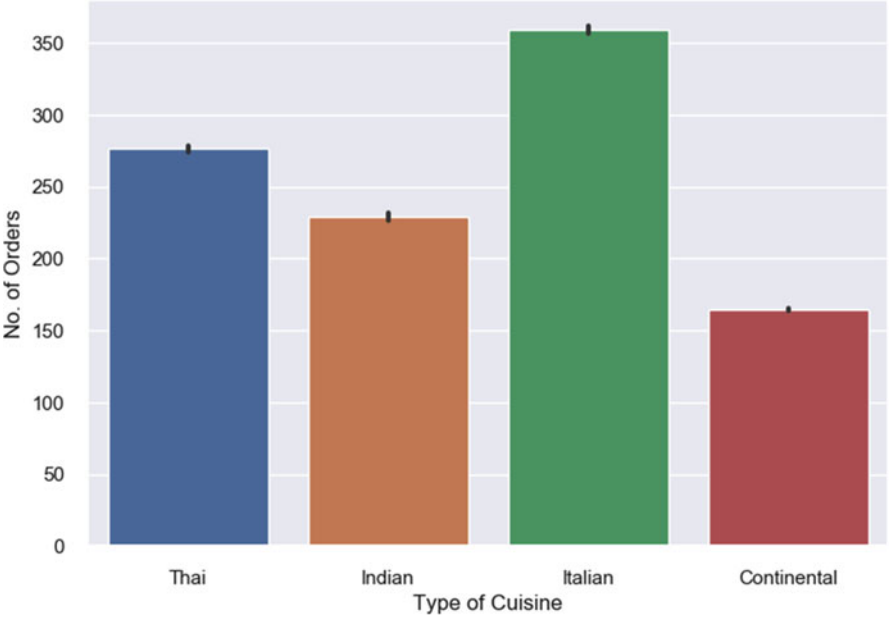

Fig. 6 Types of cuisine vs number of orders

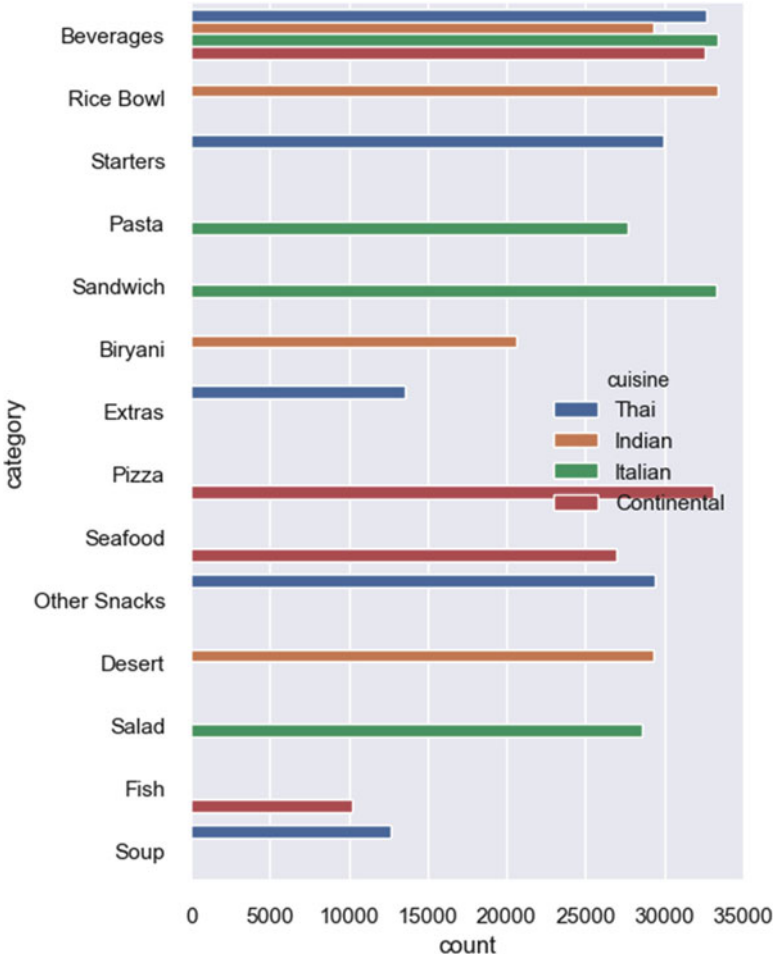

Fig. 7 Count vs categories

Also, applications for the Internet of Things (IoT) have evolved immensely, producing a vast amount of data needed for intelligent data processing. To optimize asset utilization as well as customer value, IoT supply chain management tools mine valuable data inside and outside the business [9].

IoT devices communicate with each other to collect and share data [9]. The results obtained through the study can provide insightful results for the data collected through the IoT devices and help in forecasting the demand of raw materials as well as automate the overall process of supply chain management (Fig. 13).

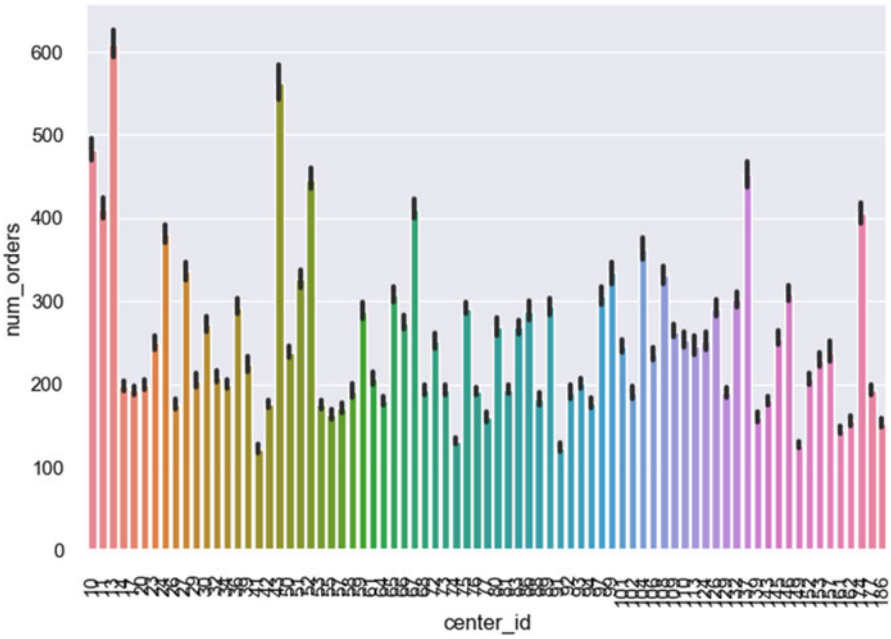

Fig. 8 Centre ID vs number of orders

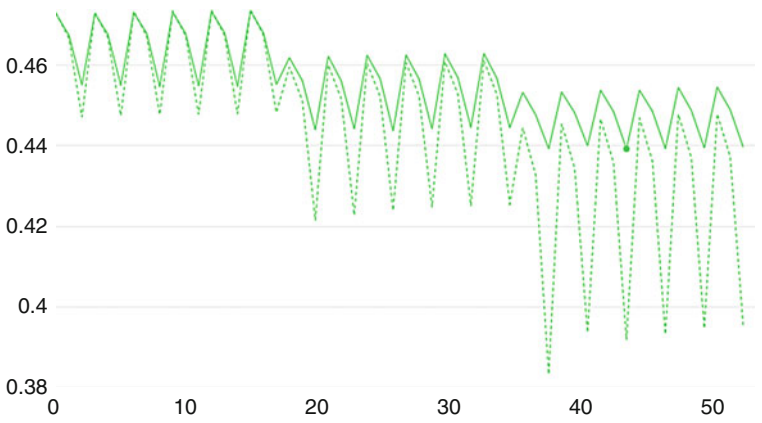

Fig. 9 Plot showing the variation in RMSE value based on the hyperparameter values

Fig. 10 Features
Note: There are other features responsible that are not present in the above list

```
['week',
 'center_id',
 'meal_id',
 'checkout_price',
 'base_price',
 'emailer_for_promotion',
 'homepage_featured',
 'city_code',
 'region_code',
 'center_type',
 'op_area',
 'category',
 'cuisine',
 'train_or_test',
 'discount_on_base',
 'discount_ratio',
 'price_last_curr_diff']
```

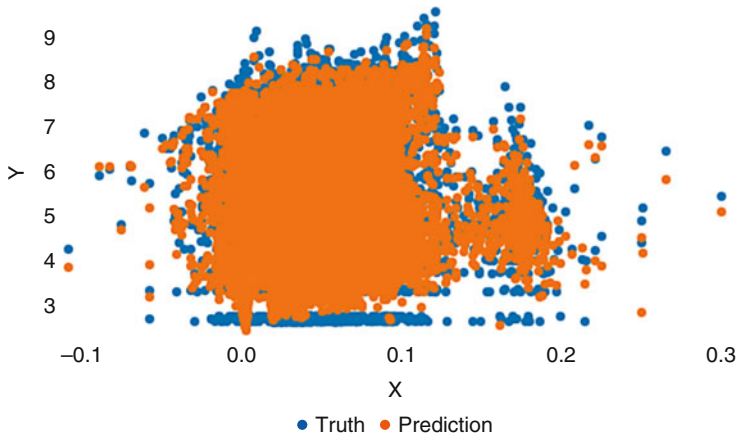

Fig. 11 True values vs predicted values

Table 2 Performance of different models on the food demand dataset

	LightGBM	XGBoost	**CatBoost**
RMSE	0.583	0.813	**0.446**
RMSLE	0.218	0.146	**0.0887**

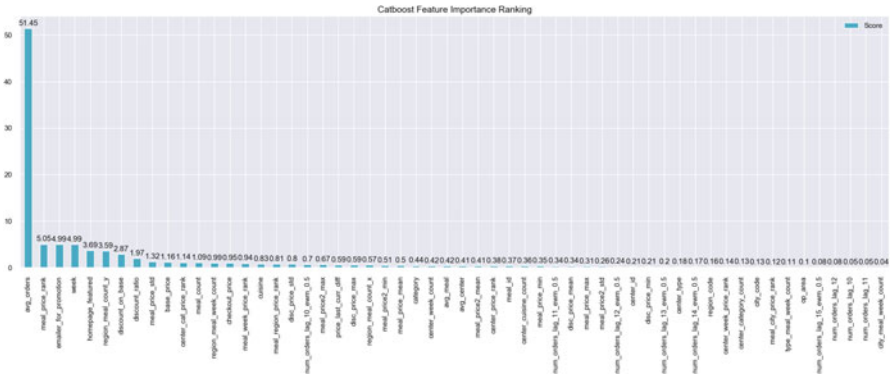

Fig. 12 Feature importance Score

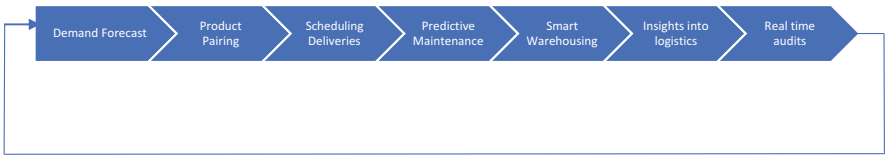

Fig. 13 How IoT enhances efficiency in the food supply chain

6 Conclusion

From the results obtained from this analysis, the following conclusions can be drawn:

1. The CatBoost model applies shift and ewm (exponentially weighted mean) features to the dataset as the data used contains historical data values.
2. CatBoost proves to offer improved performance as compared to other regression models that also deals with categorical data.
3. The historical data has a great and major impact on predicting the future values.
4. For any retailer, manufacturer or supplier, forecasting product demand is essential for its business' growth. Demand forecasting helps determine the quantities that should be purchased, produced and shipped.

In general, accurate demand forecasts help the manufacturers/suppliers obtain a correct idea of the raw materials requirements in future, which leads to high levels of customer services. On the other hand, erroneous forecasts can lead to high-cost operations and a poor customer service.

References

1. Zhang, Y., Xu, H., & Zheng, Y. (2009). Chinese residents cold chain logistics demand forecasting based on GM (1, 1) model. *African Journal of Business Management, 6*(14), 5136–5141.
2. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? –arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250.
3. Pigatto, G., Machado, J. G. D. C. F., dos Santos Negreti, A., & Machado, L. M. (2017). Have you chosen your request? Analysis of online food delivery companies in Brazil. *British Food Journal, 119*, 639.
4. Keeble, M., Adams, J., Sacks, G., Vanderlee, L., White, C. M., Hammond, D., & Burgoine, T. (2020). Use of online food delivery services to order food prepared away-from-home and associated sociodemographic characteristics: A cross-sectional, multi-country analysis. *International Journal of Environmental Research and Public Health, 17*(14), 5190.
5. Zulkarnain, K., Ahasanul, H., & Selim, A. (2015). Key success factors of online food ordering services: An empirical study. *Malaysian Institute of Management, 50*(2), 19–36.
6. Dorogush, A.V., Ershov, V. & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.
7. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in neural information processing systems* (pp. 6638–6648).
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
9. Adi, E., Anwar, A., Baig, Z., & Zeadally, S. (2020). Machine learning and data analytics for the IoT. In *Neural computing and applications* (pp. 1–29).
10. Da Veiga, C. P., Da Veiga, C. R. P., Catapan, A., Tortato, U., & Da Silva, W. V. (2014). Demand forecasting in food retail: A comparison between the Holt-winters and ARIMA models. *WSEAS Transactions on Business and Economics, 11*(1), 608–614.