Food Demand Prediction

T.Karan#1,V.Aditya#2, S.Pavan#3, G.Praveen#4, Y. Hemanth#5

==

***Abstract*—** **The vital aspect in the world of business is to have a proper analysis of their business outcomes. This outcome plays a major role in the development of the business. One of the expanding business spheres is food delivering companies**. **The vital factor in running such a food delivery company that is located at various branches in the city is to maintain the stock properly and prepare the food in time to deliver it to the customers. The Aim of the project is to develop a prediction model that predicts the Number of orders based on the unique id of the meal.**

**In this paper, number of order is used to forecast stock of items, using machine learning with internal and external data. In this we provide an appropriate algorithm for demand forecasting which is capable of overpowering the wastage of short life items. Proposed algorithm like Linear Regression, XGBoost, Light boost Regressor, Catboost Regressor and Random Forest algorithm are used that considerably improves the forecasting performance.**

***Keywords***— **Prediction, Analysis, Machine Learning, Data Science, XGBoost , Light Boost , Cat Boost**

1. Introduction

As in today’s competitive life even the business has become more difficult. Demand for food is increasing day by day with the increase in the population of every country. Estimation of the demand in food consumption plays a vital role in supplying or generating resources to produce the required amount of food. To meet this challenge, we need to predict the demand in food consumption for future so that the hunger of everyone can be satisfied. We will analyse all the previous year’s data on how the food demand has been throughout the restaurants.

The major task that one would face in this will be predicting the quantity of products to be bought and prepared. It is very difficult to predict the number of orders in a given restaurant on a given day. A wrong prediction may end up purchasing and preparing less amount of food which will cause shortage or purchasing and preparing more which will lead to wastage of food. So, predicting the exact demand is a challenge because of uncertainty and fluctuations in consumer demand.

1. LITERATURE REVIEW

**K.Aishwarya, Aishwarya.N.Rao, Nikita Kumari, Akshit Mishra, Mrs.Rashmi M R , June 2020 [1]**

The success of a restaurant not only depends on taste, ambience but also on service. The most important part among the services is serving fresh food. In order to provide this, the restaurants need to prepare food daily, this requires buying some of fresh self life food products every day. The major task that one would face in this will be predicting the quantity of products to be bought and prepared. It is very difficult to predict the number of orders in a given restaurant on a given day. A wrong prediction may end up purchasing and preparing less amount of food which will cause shortage or purchasing and preparing more which will lead to wastage of food. So, predicting the exact demand is a challenge because of uncertainty and fluctuations in consumer demand. These variations ad fluctuations in demand may be because of price change, promotions, change in customer’s preferences and weather changes.

**N. de P. Barbosa, E.da S.Christo, and K. A. Costa , September 2019 [2]**

The supply chain activities planning and control depends of accurate estimates of the volumes of products and services to be processed and the estimates come as forecasts (Ballou, 2007). One of the biggest challenges of food and beverage manufacturers is adjust the production and the stocks to minimize the loss of products due to its short perishability.

Time series analysis is very important in a wide range of applications, especially when it comes to forecasting, and it encloses many different forecasting models. However, it is necessary to determine which model best suits each situation (de Oliveira Silva et al. 2014). T

here are countless models to develop forecasts presentedin the bibliography. From market research to the most complexmethods computational. The reference (Ballou, 2007) presents the three main groups of these methods, they are: qualitative, historicalprojections and causal.

Depending on the series and the desiredtime to be forecasted it is possible to choose a technique thatbest fits. Besides choosing the best technique, the forecasting tobe generated by the model chosen should be as close to real aspossible (Junior and Filho, 2012). In other words, the errors of forecasting should beminimized, so the production managers plan the production inattention to the market and minimizing the costs.

**Takashi Tanizakia, Tomohiro Hoshinoa ,Takeshi Shimmurab ,Takeshi Takenaka, July 2018[3]**

The service industry is an important industry accounting for about 70% of Japan's GDP. However, since the labor productivity of the service industry is lower than that of the manufacturing industry, its improvement is an important policy issue of Japan. Especially in the labor-intensive service industry labor productivity is low because service goods are consumed at the same time as they are provided.

Among the laborintensive service industries, restaurant industry, which is integrated production and sales industry, has improved its inventory possibilities by separating service production functions from sales by introducing a central kitchen.

In this context, the main challenge in modelling and analysis is now not only to cope with single products, a limited product range or existing product families, but also to be able to analyze and to compare products to define new product families. It can be observed that classical existing product families are regrouped in function of clients.

**Claudimar Pereira Da Veiga, Cássia Rita Pereira Da Veiga, Anderson Catapan, Wesley Vieira da Silva , November 2017 [4]**

Companies were driven by the many changes that, over the past few decades, Brazil experienced in both the economic and political spheres to seek offered product solutions which promoted profit generation, productive process efficiency and improved quality. In addition challenges of economic nature, in light of current restrictions as to the availability of non-renewable resources, companies were additionally shifted into a position whereby rethinking their future strategies became mandatory so as to ensure the sustainability of their very operation. Despite initial resistance to change, migrating to a sustainable supply chain condition became a core strategic competitive factor to ensure the continuation of the business in itself.

Within an industrial segment characterized by rapid and regular changes, companies are required to develop processes that ensure greater awareness as to future scenarios and possible outcomes to in turn enable their maintaining of sustainable competitive advantages (Teece, Pisano & Shuen, 1997).

Demand predictions thus take on a prime role before the planning of sustainable operation within the chosen marketplace, whether from a macro or micro-economic standpoint. However, merely having a demand forecasting routine or system in place within the organization does not necessarily address end objectives. What truly ensures a given organization is capable of obtaining improved environmental economic and social performances is the quality of the information that routines and systems generate and thus offer as input to the process itself.

Kuo and Xue (1999) clearly state that the key success factor, i.e., core point of attention, ensuring quality of decision processes is precisely the construction of an accurate demand forecast. Predictions attempt to calculate and foresee future events, conditions and contexts, offering the best possible assessment of commercial and available market information.

**Ahmet Selman Bozkir,** **Ebru Akcapinar Sezer, 2020 [5]**

Fluctuations and unpredictability in food demand generally cause problems in economic point of view in public food courts. In this study, to overcome this problem and predict actual consumption demand for a specified menu in a selected date, three decision tree methods (CART, CHAID and Microsoft Decision Trees) are utilized.

A two year period dataset which is gathered from food courts of Hacettepe University in Turkey is used during the analyses. As a result, prediction accuracies up to 0.83 in R2are achieved. By this study, it’s shown that decision tree methodology is suitable for food consumption prediction.

Data mining, on the other hand, is discovering hidden relationships, correlations, trends and associations in data with the help of smart algorithms over automatic or semi-automatic approaches. Besides, data mining involves various techniques such as statistics, neural networks, decision tree, genetic algorithm, and visualization techniques that have been developed over the years.

In general, data mining methodologies are classified as predictive and descriptive methods. While the purpose of descriptive methods is defining the nature of data, making future predictions based on current observations is the main goal of predictive ones. Therefore, predictive data mining approaches are suitable solution candidates for the main problem defined in this study.

1. Proposed Methodology
2. *Dataset Description*

There are total 3 datasets used here in our project , info data set. Train.csv data set includes all the attributes, the id’s of the fulfilment centers and meals, and also the num\_orders which is needed to be predicted this is used to train the models, the fulfilment\_center\_info.csv file provides us with the details of the food centres which are providing the food. The meal\_info.csv file contains all the extra and particular information about the means which are indicated with an id in the train.csv data set. We have merged the train, fulfilment center and the meal info datasets to form one train dataset and used that to build the model.

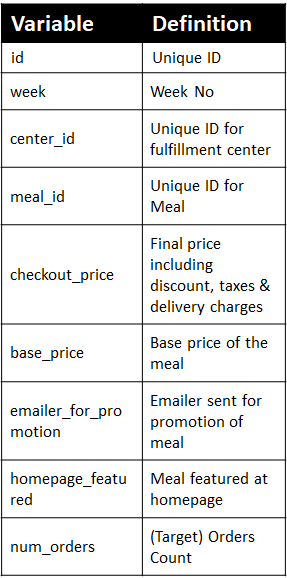


Fig 1.1 Train Dataset

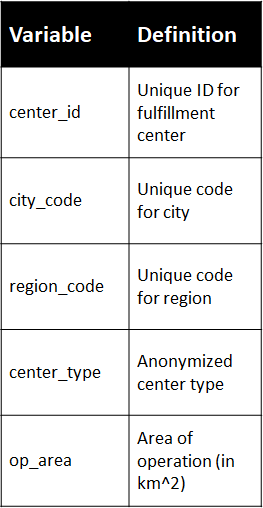


Fig 1.2 Fulfilment Dataset

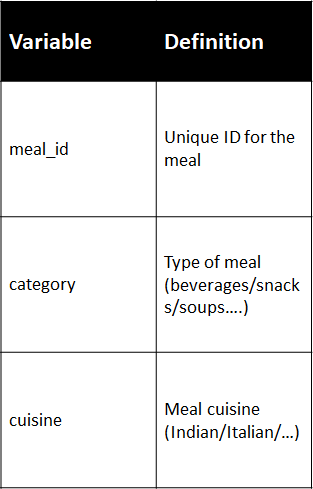


Fig 1.3 Meal Info Dataset

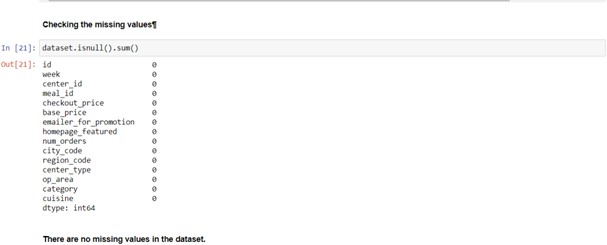
1. *Requirements*

The implementation of the predictive modelling of Food Demand Forecasting undergoes several steps to give the accurate prediction. We used the RMSE (Root-Mean square error) as the evaluation metric of the model. Linear Regression, XGBoost, Light Boost regressor, Catboost Regressor and Random Forest Algorithms are used to build the model. The less the RMSE values, the more accurate is the prediction model.

We use train and test data sets to tarin the models and test them with our predictions. Sample submission data set is used to check weather our predicted file is in the same format as the sample submission file.

1. *Data Pre-Processing*

We have imported the datasets. We have come to the dimensions and the complete information about the datasets. We have checked the presence of duplicate values. We have assigned a random constant value to the target variable i.e. number of orders as we have to predict that column after training the model. we have merged the meal info, fulfillment Centre, test datasets with train dataset to form one train dataset. We used the

data.isnull().sum() to find out the missing values and found that there are no missing values in the dataset.

The most common problems you can find with raw data can be divided into 3 groups:

**Missing data:** you can also see this as inaccurate data since the information that isn’t there creates gaps that might be relevant to the final analysis. Missing data often appears when there’s a problem in the collection phase, such as a glitch that caused a system’s downtime, mistakes in data entry, or issues with biometrics use, among others.

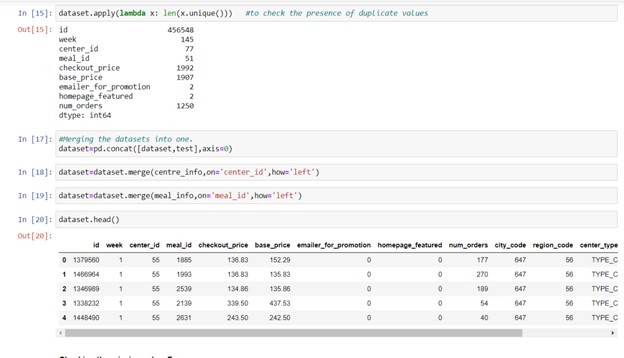
**Noisy data:** this group encompasses erroneous data and outliers that you can find in the data set but that is just meaningless information. Here you can see noise made of human mistakes, rare exceptions, mislabels, and other issues during data gathering.

**Inconsistent data:** inconsistencies happen when you keep files with similar data in different formats and files. Duplicates in different formats, mistakes in codes of names, or the absence of data constraints often lead to inconsistent data that introduces deviations that you have to deal with before analysis.

Since mistakes, redundancies, missing values, and inconsistencies all compromise the integrity of the set, you need to fix all those issues for a more accurate outcome. This can be achieved through data preprocessing.

Checking the duplicate values and merging all the datasets into train dataset is observed

Checking the missing values



1. *Feature Engineering*

We extract some useful features from existing attributes that helps in improving the performance of prediction model. We created some new features that are as follows, discount amount, 'discount percent', compare\_week\_price, compare\_week\_price y/n.

1. *Encoding Categorical Variables*

The encoding of the categorical variables is essential as the machine learning models gives better prediction result with continuous and numerical variables. The complete removal of categorical variables would lead to loss of information. Here we used get\_dummies() method to convert each category of a variable to a number

1. *Predictive Modelling*

The most crucial stage and the heart of the project is the predictive modeling using machine learning algorithms. we used the Linear Regression Model, Xgboost, LightBoost regressor, Catboost Regressor and Random forest Algorithms implementing the project. We have scikit learn package and did the test train split to train the model and tested them on test dataset. The algorithms and the models build using them are briefed in Module description session. Evaluation metric is the vital part of building a effective model as we get feedback from this metrics to improve the model further.

Here, we used the **RMSE** (Root-Mean Squared Error) as evaluation metric:

https://s3.amazonaws.com/thinkific/file_uploads/118220/images/b9d/893/657/1549265342189.jpg

1. *Proposed System*

The implementation of the predictive modelling of Food Demand Forecasting undergoes several steps to give the accurate prediction. We used the RMSE (Root-Mean square error) as the evaluation metric of the model. Linear Regression, Xgboost, LightBoost regressor, Catboost Regressor and Random Forest Algorithms are used to build the model. The less the RMSE values, the more accurate is the prediction model.

1. MODULE DESCRIPTIONS AND METHODS

In this research, the number of customers is forecasted using machine learning and statistical analysis method with internal data and external data in the ubiquitous environment. Linear Regression , XG Boost , Cat Boost , Light Boost and Random Forest Algorithms are used for machine learning, Stepwise method is used for statistical analysis method. We used Jupyter Notebook as a machine learning tool.

*1. Linear Regression*

Linear Regression is the simplest algorithm and it is most widely used algorithm for predictive modeling. Here the target variable is dependent on the attributes of the dataset. It basically gives us an equation, where we have our features as independent variables, on which our target variable is dependent upon.

In this paper, we have developed the 3 models using Linear Regression. We have imported the Linear Regression from sklearn\_linear\_model package, split the dataset into train and test and performed linear regression directly on the dataset. In that case we got a huge value of RMSE value.

In the second Model we performed the Standard Scaling and log transformation on the target variable and then the RMSE value obtained is good compared to first model. For the third model we have created some attributes namely quarter and year, cleared the outliers using quantile method, did the log transformation on target variable and performed the linear regression in which we got a less RMSE value compared to all previous models.

*2. XG Boost*

XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. The algorithm's implementation was designed to maximize computation time and memory resources. One of the design goals was to make the most of the resources available to train the model. The two reasons to use XGBoost are Execution speed and Model performance. XGBoost dominates structured or tabular datasets on classification and regression predictive modelling problems. In general, XGBoost is quick. When compared to other gradient boosting implementations, this one is lightning fast.

In this paper, we implemented XG Boost in two models ensuring with same parameters and with different attributes. In model-1, we used some of the attributes from dataset where these two models are used to build training data sets. In model-2, we had included some more attributes from feature engineering for showing the different RSME values with different attributes. After comparing the RSME values of both models, we can conclude that model-2 having with good RSME value compared to model-1.

*3. Light Boost*

Light boost Algorithm is high performance gradient boosting framework based on decision tree algorithm, The leaf wise algorithm can reduce more loss than the level wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms. Important features of the algorithm are **num**\_**leaves**: number of leaves in one tree; default = 31; type =int, **max\_depth**: Specify the max depth to which tree will grow. This parameter is used to deal with overfitting. **min\_data\_in\_leaf**: Min amount of data in one leaf. **feature\_fraction**: default=1; specifies the fraction of features to be taken for each iteration. **bagging\_fraction**: default=1; specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting.

In this paper, we have developed the 4-models using the Light boost Regressor. In Model-1, we have used only some hyper parameters and developed the model. In Model-2 we have taken all the hyper parameters into consideration. In Model- 3, we have dropped some more input attributes from the train dataset. In Model- 4, We have dropped input attributes such as base price and discount amount and obtained an RMSE Model-3. Compared with XGBoost Algorithm the RMSE values in four models of Light boost regressor are less.

*4. Cat Boost*

Cat Boost algorithm can be used to solve a wide range of problems and can deal with many types of data with unique parameters which gives us accurate results. Categorical

data in particularly yields better accuracy when cat boost algorithm is applied. With many parameters to tune with we can easily arrive at the best result out of all. Cat Boost algorithm also uses the gradient boosting algorithm which can yield very good result even with very small data. This algorithm can be used without any pre-processing which is required in many other algorithms. It doesn’t require any hyper parameter tuning even with all the various parameters to work. The other qualities include robustness, more accuracy and easy to implement. Most of the datasets can be processed by the default settings of the parameters without tuning them.

The implementation of cat boost in this project is processed with 2 models. Each model has different tuning parameters. These models are used to build the training data sets. The variables that are used in both the models are num\_oders, week, discount\_amount, city\_code, Quarter\_Q2, base price, with these models fixed the tuning parameters learning rate and the depth are varied between 0.1-0.3 and 6-9

These training models have yielded the best RMSE values out of all the remaining models which we have generated. With the ame parameters the prediction file is generated using test data set

*5. Random Forest*

Random forest algorithm is one of the ensembles learning algorithm. It is a algorithm with decision tree (CART) Model, in which the sub trees are learned so that the resulting predictions from all the sub-trees have less correlation. Here it tries to build multiple CART models with different samples and initial variables. It will repeat the same process and then make a final prediction on each observation and the final prediction is result of each prediction. This final prediction can simply be mean of each prediction. For example, say we have 1000 observations in the complete population with 10 variables. It takes a random sample of 200 observations and 5 randomly chosen initial variables to build CART model. It will repeat the process say 10 to 12 times and then make a final prediction on each observation.

Random forest gives more accurate predictions Here it uses only a subset of features that are selected random out of total population. It has a benefit of handling large data. In this project we have given the n\_estimators as 200 in the model and build the model using the algorithm. We have got an RMSE value of 0.5289 that is almost similar to the value obtained in the XGBoost Model-1.