Food Inventory Predictor

Mohamed Yilmaz Ibrahim
Purdue University
West Lafayette, Indiana, USA
PUID: 0033795176
ibrahi35@purdue.edu

Rahul Senguttuvan
Purdue University
West Lafayette, Indiana, USA
PUID : 0033583173
rsengutt@purdue.edu

Vignesh Somasundaram
Purdue University
West Lafayette, Indiana, USA
PUID : 0033886171
vsomasu@purdue.edu

ABSTRACT

As an organization in the food supply chain, estimating the food demands has always been a challenging problem to solve. Poor estimation of meal requirements may lead to excess or shortage of food. It, therefore, becomes essential to keep track of food consumption to meet the customers' demands. A system that predicts food demand can help organizations make informed decisions concerning the purchase of raw materials and avoiding wastage. In this study, we propose a food inventory prediction system that aids users in keeping track of their stock and make wise purchasing and preparation decisions. Our study involves (i) A literature review to formulate the design requirements, (ii) An iterative design process (iii) A comparison study with three Machine Learning algorithms. We plan to evaluate the tool by running the prediction model against the test data collected from the data set to predict the food demands for up to 10 weeks.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by regression**; **Classification and regression trees**.

KEYWORDS

Data set, Neural networks, Data mining, Artificial Intelligence, Machine Learning, Supervised Learning

ACM Reference Format:
Mohamed Yilmaz Ibrahim, Rahul Senguttuvan, and Vignesh Somasundaram. 2021. Food Inventory Predictor. In *CS 57300 Data Mining Fall 2021 : Purdue University WL*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 MOTIVATION

With the increase in food consumption, restaurants face a challenge when catering to the needs of their customers. A restaurant will have to provide quality food and make sure they have sufficient stock of ingredients available to provide such quality food. Accessibility to a variety of options has made managing food quantities of every individual option difficult. This would lead either to an excess or a shortage in production of the varieties. Being able to predict the quantity needed 7-10 weeks in advance would not only manage production efficiently but will also have a ripple effect in reducing the prices of " in-demand " groceries. Food Organization would be able to change the prices of the stocked items based on availability, to increase their profit.

This application is helpful not only in restaurants but also in places like Universities. With the increase in the student population, the university dorms can better plan their meals and produce food only according to their needs. The university can reduce their spending by using our application and also ensure the students needs are satisfied.

2 RELATED WORK

The food demand prediction system proposed in this paper is a time series forecasting task. Making use of a machine learning approach to deal with a time series forecasting task is more powerful and flexible than traditional algorithms ([7]). It can be considered powerful because it allows the use of modern state-of-the-art supervised learning algorithms such as support vector machines for regression and decision trees ([5]).

The paper [4] by Hasmin et al. proposes a double exponential smoothing technique for forecasting the stock in an inventory. Exponential smoothing computes the average of data in a particular period and predicts the next stage's value by providing weights exponentially to the previous periods. They use double exponential smoothing where they have the smoothing parameter value to be between 0 and 1. The validation of the method is done by the Mean Absolute Percentage Error method and also black-box testing. The system is evaluated to have low error rates while forecasting. While forecasting using this method is good, it needs to be assessed with larger datasets and different locations. Also, it does not provide a methodology to suggest prices to increase the number of orders.

Artificial intelligence forecasting methods have been receiving much attention lately to solve problems that are hardly solved by the use of traditional methods. They have been cited to have the ability to learn like humans, by accumulating knowledge through repetitive learning activities. Therefore in this study, we focus on using artificial intelligence techniques for predicting food demands.

Liang, in her paper ([6]) proposed the use of Analytical Hierarchical Process (AHP) to deconstruct complex decision-making problems to well defined layers. Then, for each of these well-defined layers, the relative weights are derived through pairwise comparisons. After identifying the key factors using this model, analysis of sequential pattern are included to predict the forthcoming input. Sequential Analysis helps find repeated patterns even if no obvious relationship exists among different events. This paper produces a high

Data Mining Fall '21, August - December, 2021, Purdue University, IN

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

accuracy result but takes into consideration the data is continuous and sequential. The work we propose in our paper requires the usage of discrete values through time series mining.

Agarwal et al. in his paper ([2]) introduces the usage of the model, SARIMAX (Seasonal Autoregressive Integrated Moving Average) which is a modification of the existing SARIMA model that takes Multiple Linear Regressions into account. The model also considers holiday effect which increases the accuracy of the prediction. The dataset has been acquired through the usage of Smart Dustbins and object detection using Convolutional Neural Networks. Once the image discarded is detected, the value is then used to populate the dataset. Once the dataset has been acquired, the data is split into training and test data. The training data is then used to find the auto-regressive moving average stock for the model. Although this model gives an accurate estimate of the wastage of individual raw materials, our tool does not rely on a set of images for the data and hence we would not be using a CNN for object detection.

The work in [9] by Cetinkaya et al. discusses the prediction of daily food demand at Kirikkale University using Artificial Neural Networks. In this work, the various factors affecting food consumption are identified initially and then normalized. The data set is split into two classes, one for the students and one for the staff. The two datasets are then passed to the ANN model, and the accurate number of hidden layers and the proper activation function are determined by trial and error. The results were evaluated using Root Mean Squared Error(RMSE) and Mean Absolute Percent Error(MAPE). Although the model yielded promising results, it considered only the data of a particular location, classified it into sets, and worked separately on each of them. Our work requires the prediction to be performed given any location and type of meal.

Naseri et.al [3] have proposed the use of Recurrent Neural Networks (RNN) in demand forecasting tasks. The proposed RNN model comprised of four layers: an input layer, a hidden layer, a context layer and an output layer. In this paper, datasets of around thirty diffrent types of spare parts from Arak petrochemical company in Iran and three performance measures: Percentage Best (PB), Adjusted Mean Absolute Percentage Error (A-MAPE) and Mean Absolute Scaled Error (MASE) was proposed. Their results show that Neural Networks can be used with promising results in comparison with traditional forecasting methods.

Afifi, in his paper [1] proposed the application of data mining to forecast the demand of short life cycle products. The steps toward the forecast after the data pre-processing includes using an incremental k-means algorithm to identify the sales profiles of historical products by grouping similar data. The clustering results was measured by a distortion error value with a lower value indicating a better clustering result. The next step involves using the RULES-6 algorithm to extract if-then rules from the training data set which is used to build a classification model that maps the descriptive attributes to identified historical products. The results of the Clustering indicate that the sales of behavior of identified products have been accurately represented through the graph and the RULES-6 algorithm has generated lesser Rules than data points in the training set. However, on application to the test data, the actual sales associated with a cluster vary for certain weeks. This is due to the failure of descriptive attributes used to build the rule set. While the accuracy of the forecast model is high, inaccurate profiles could be a result of inaccurate descriptive attributes chosen.

Yang et al. in their paper [8], present a forecasting model for predicting the sales of bakery products in a shop. They investigated the idea of using the sales in the first few hours of a day to predict the sales for the rest of the day. The work used regression analysis and neural networks with 50 hidden nodes to find the model that best suited this use case. They used an adaptive learning rate method called ADADELTA and 5-fold cross-validation to test the models' performance and accuracy. The work made several assumptions, such as the products being replenished at given times and human interventions made during that time. The paper concluded that Neural Networks perform better than the regression technique in most cases because the former can learn the patterns in the data quickly. The models presented have several shortcomings, such as it uses the sales at the start of the day for predicting the sales for the rest of the day, which may not be efficient in all the scenarios. Also, the work concentrates on short-term patterns, and predicting long-term inventory management using that is a challenge.

Our work differs from the above mentioned literature in that our tool is a mixed-initiative approach that involves an intelligent agent that it provides suggestions on the prices of the meals depending on demand. Unlike the literature studies, a simulator component is added where the users can visualize the price trends helping them make informed decisions. A detailed design of our tool is proposed in the following sections.

3 IMPLEMENTATION

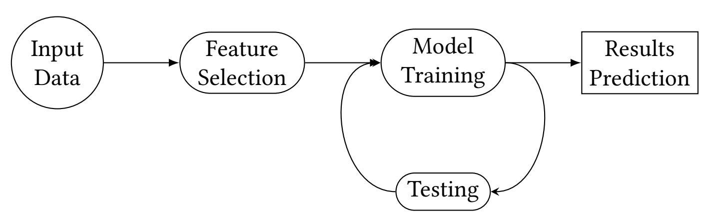

Fig 1: Flow diagram of the Food Inventory Prediction System

In our research, the food inventory prediction is done through machine learning and statistical analysis methods. We plan on using Regression techniques, Artificial Neural Networks for our implementation. From the literature survey, we learnt that Artificial Neural Networks are able to learn about time-series data-set better than other machine learning models. We plan on training the model with techniques like Support Vector Regressor, Long-Short term memory, identify the correct weight for the parameters and compare the results. The language used in our implementation is Python3 and we are making use of Jupyter notebook for the project.

4 DATA ANALYSIS

Exploratory Data Analysis refers to the process of performing initial investigations on the data set so as to discover patterns, detect any anomalies, test hypothesis and check assumptions with the help of summary statistics and graphical representations. We have performed Exploratory Data Analysis to identify patterns in the data. The following are the results:

Our dataset contains 456548 instances for training. The training data contains information on:
Center ID - A unique ID denoting the center for which the prediction has to be made
Meal ID - The unique ID for a meal
Checkout price - The price of the food item during checkout
Base Price - The price of the food item
Number of orders - Number of orders in a particular center for a meal

Each center has 4 unique characteristics representing it apart from its ID. They are denoted by city code, region code, center type and op area. And each meal has 3 characteristics associated with it. It being meal cuisine and category (Such as Beverages, Desserts etc.) Our dataset is cleansed and has no null values. Each entry in the dataset belongs to the dtype object and contains only integer, float and string values.
Each entry in the data is unique with no elements have frequency greater than 1.

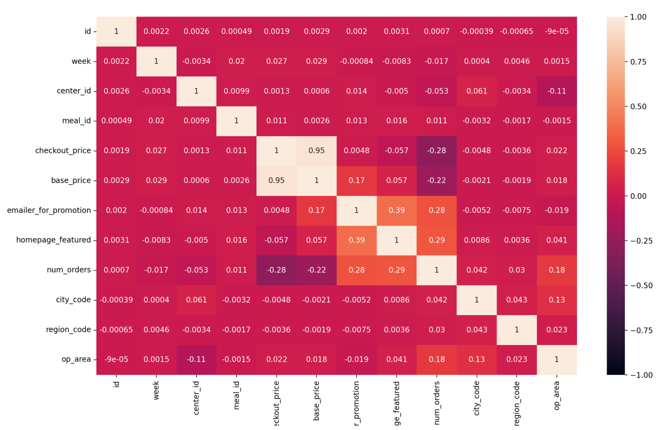

Fig 2: Correlation Heatmap

To provide information on the correlation between the features, we have plotted a correlation heatmap. The above heatmap (Figure 2) provides us information on the correlation between the independent features in our dataset. From the heatmap, we can infer that there exists minimal correlation between the features in the dataset. Thus the features are entirely independent in the dataset. We can see that the checkout price has strong positive correlation with the base price.

From the graph, we can also infer that the most of the other features have less correlation and so are independent features. We can drop the features which have zero correlation against the other features.

For four of our common features (Week, Checkout price, Base price and Number of Orders), let us plot an histogram to view the frequency of values across the training instances. From these graphs, we can learn what values to concentrate the most during the training of the model.

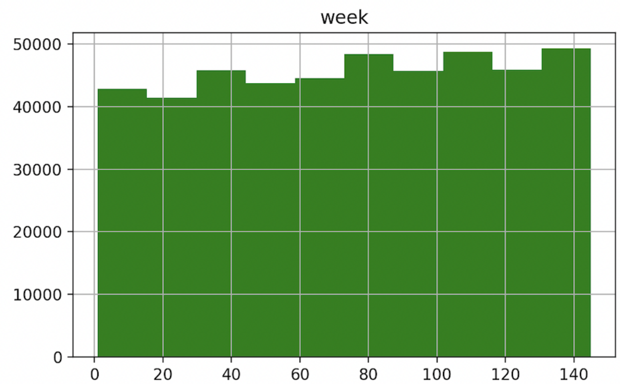

Fig 3: Week distribution

The above figure provides information on the distribution of instances pertaining weeks. The instances are spread across different weeks and we have a uniform distribution. The dataset has equal distribution across all the weeks so we do not have any bias in the dataset with respect to the week feature.

The below figure provides information on the base price distribution. We can infer that most of the food items have base prices amounting to around 300. We can also infer that we have less instances for amounts 100, 600 and 800. Thus, we can focus more on the items ranging around 300 for our prediction model.

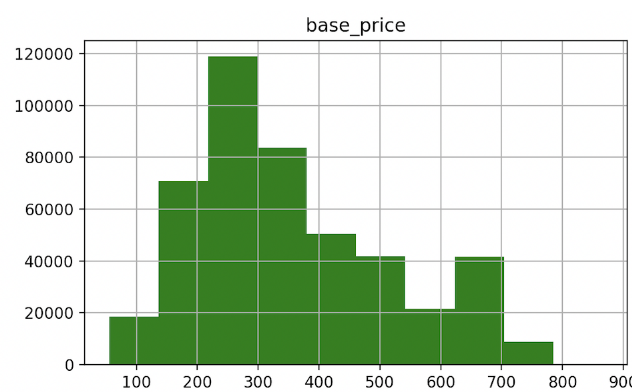

Fig 4: Base Price feature distribution

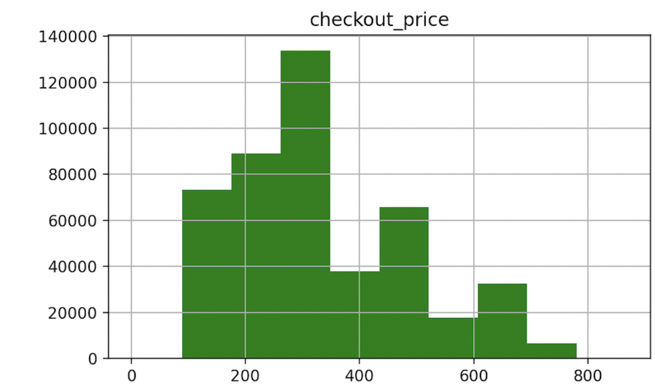

Fig 5: Checkout Price feature distribution

The above figure provides information on the checkout price distribution. The checkout price is also quite similar to the base price distribution of the items. Around 100k instances have checkout prices around 300. Even the base price distribution had results similar to this.

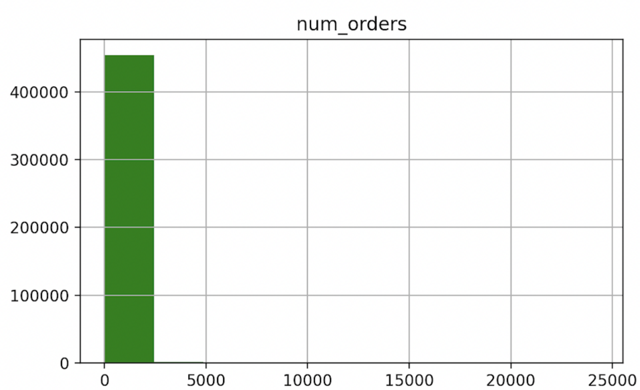

Fig 6: Number of orders distribution

From the Order count distribution, we can see that the number of orders for food items is less than the 5000 range. All 400k instances in the dataset have less order counts.
Thus for the bar graph distributions across the four common features, we get to understand how the instances are spread across in the dataset.
Our next step would be to further analyze the data and remove information which deviate from our intended tasks ie. we will have to remove the outlier information from the dataset.

Removing Outliers: To remove outliers from the dataset, we can make use of a whisker plot to detect outliers if any.

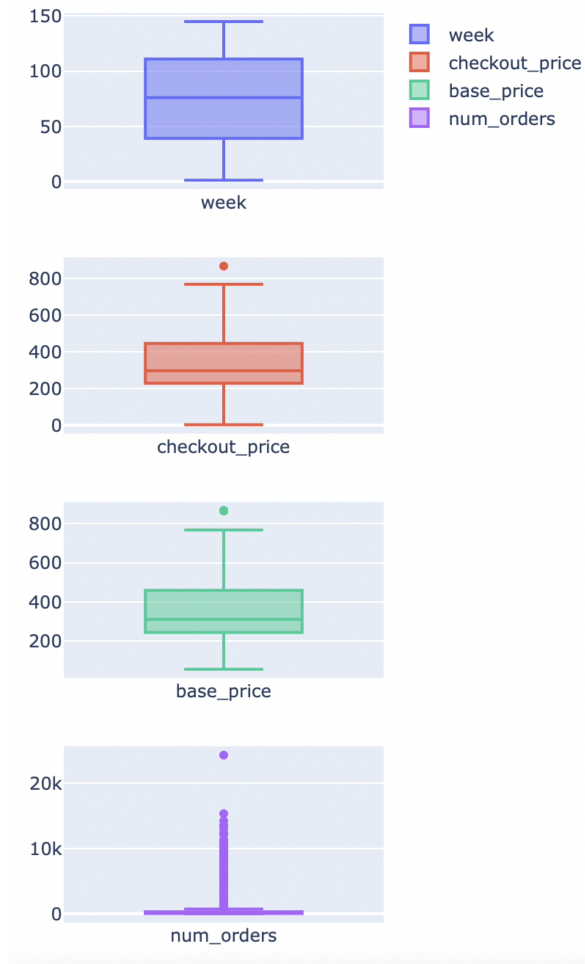

Fig 7: Detection of outliers across different features

From the above whisker plots, we can observe one outlier in each of checkout price, base price and number of orders. With the removal of outliers, we will have less variance in the dataset and the regression models should be able to predict better. We have removed the outliers from the dataset for the model to perform better.

5 TIMELINE

We have conducted the literature survey on the various models that have been proposed previously for predicting time series data sets. Pre-processing for the data set is conducted during this phase. We have performed Exploratory Data Analysis on the data set. We have merged the different data sets given, identified the patterns in the data and plotted them to understand the data to choose the right model. We have also figured out the outliers in the data set and removed them. The next step in this process involves training the models with Artificial Neural Network (ANN) techniques like long short term memory (lstm) and regression techniques. We will identify the best weights of the parameters for each of the models and compare their performance based on mean-squared error.

6 EXPERIMENTAL RESULTS

Initially, we used the Linear Regression model, one of the first regression models taught in the lecture, on our dataset. We applied one-hot encoding for the categorical data present in the data set. Initially there were three columns : center_type, category and cuisine. We sorted them in a lexicographical order with the last value as the reference vector for each of the column. We then split the data set into 80 percent of training and 20 percent of testing. After training the model, the R-squared value of the training and test set model is found to be around 41%. This shows the linear regression model does not perform well in the given dataset which has non-linear and time series data.

7 CHALLENGES

- Identifying the correct preprocessing steps for cleansing data was a challenge. Since it is a large dataset comprising more than 450k instances, identifying the features of importance contributing to the result was a challenge as well.
- Our linear regression performed poorly in the dataset. We are looking forward to using advanced models that may help yield better prediction results.
- We had to spend a good amount of time expanding our learning curve to gain insights on Exploratory Data Analysis.

8 CONCLUSION

Presently, the management of raw materials to meet the food demands at restaurants, grocery shops and university food courts may not be optimal which can lead to either an excess or shortage of food produced. With the help of our application, the organizations can predict the demand for upcoming weeks and efficiently manage their food production and increase their revenue. The users can make use of this tool to aid their decision-making process with respect to food management.

REFERENCES

[1] Ashraf A. Afifi. 2020. Demand Forecasting of Short Life Cycle Products Using Data Mining Techniques. Springer International Publishing, 151–162. https://doi.org/10.1007/978-3-030-49161-1_14

[2] Harsh Agarwal, Bhavya Ahir, Pramod Bide, Somil Jain, and Harshit Barot. 2020. Minimization of Food Waste in Retail Sector using Time-Series Analysis and Object Detection Algorithm. In *2020 International Conference for Emerging Technology (INCET)*. 1–7. https://doi.org/10.1109/INCET49848.2020.9154156

[3] M. R. Amin-Naseri and Bahman ROSTAMI-TABAR. 2008. Neural network approach to lumpy demand forecasting for spare parts in process industries. 1378 – 1382. https://doi.org/10.1109/ICCCE.2008.4580831

[4] Erfan Hasmin and Nurul Aini. 2020. Data Mining For Inventory Forecasting Using Double Exponential Smoothing Method. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*. 1–5. https://doi.org/10.1109/ICORIS50180.2020.9320765

[5] Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic Model Trees. *Machine Learning* 59 (02 2005), 161–205. https://doi.org/10.1007/s10994-005-0466-3

[6] Chih-Chin Liang. 2013. Smart Inventory Management System of Food-Processing-and- Distribution Industry. *Procedia Computer Science* 17 (12 2013), 373,Äì378. https://doi.org/10.1016/j.procs.2013.05.048

[7] Grigorios Tsoumakas. 2019. A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review* 52 (06 2019). https://doi.org/10.1007/s10462-018-9637-z

[8] Chao-Lung Yang and Hendri Sutrisno. 2018. Short-Term Sales Forecast of Perishable Goods for Franchise Business. In *2018 10th International Conference on Knowledge and Smart Technology (KST)*. 101–105. https://doi.org/10.1109/KST.2018.8426091

[9] Zeynep Çetinkaya and Erdal Erdal. 2019. Daily Food Demand Forecast with Artificial Neural Networks: Kırıkkale University Case. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*. 1–6. https://doi.org/10.1109/UBMK.2019.8907105