Sales Prediction Feature Analysis in Big Mart Data based on Univariate, Bivariate and XGBoost Methods

T K Thivakaran[1], Dr M Ramesh[2]
[1]Research Scholar, Annamalai University, Tamil Nadu, India.
[2]Professor, Annamalai University, Tamil Nadu, India.

Abstract

Sale prediction is required for supermarkets to understand the requirements of customers and to increase the sales of products. Feature selection is an important process in sales prediction and this improves the performance of overall prediction. In this research, the XGBoost method (Bayesian optimization) is applied with univariate and bivariate to analyse the feature importance of the dataset. The BigMart datasets of 2013 and 2018 were used to test the performance of the developed method. The univariate method analyses the individual features related to output variables. The bivariate method analyses the feature correlation to the output variables to find the feature importance. The XGBoost method is applied with important features to improve the performance of classification. The univariate analysis shows that fruit & vegetables, snacks and food items have 14 % sales that are higher than other item types. The bivariate analysis shows that item types and item outlet sales have a high correlation. So, the item types have higher feature importance in the sales prediction in the datasets. The proposed XGBoost with the Bayesian optimization method has 0.311 MAE and the existing two-level statistical model [11] has 0.3917 MSE.

Keywords: BigMart datasets, Bivariate, Sale prediction, Univariate and XGBoost.

1 Introduction

A chain store such as ZARA, Starbucks, and McDonald's have a group of stores in various cities with standardized business models, central management, and the same brands. Moreover, the chain-store model is popular in various service categories such as fashion markets and food markets [1]. Building a sales prediction model is useful for the business to plan the product. A sales prediction model is one of the important parts of business intelligence in the store. Sales prediction provides the insights such as resources, cash flow, and workforce (labour) of the company. A sales prediction model acts as a tool that uses past and current sales data to predict sales of the company. Future sales estimation is an important factor for financial planning for any business model [2]. The E-commerce industry chain uses supply chain collaboration to improve customers logistic service experience. This model helps to plan for stock of the commodities in advance for local warehouses located in various locations [3]. This is essentially required in supermarket or pharmaceutical distributors to provide the forecast of the need for products or medicines. This also helps to avoid excessive inventory cost as well as loss of customers due to stock outages, short-term shelf-life and the need to control stocks [4]. Accurate prediction is one of the key aspects to successfully managing supermarket stocks and staff. Properly predicting future sales of supermarket items helps to precise storing of food stock and increases the profit [5].

Accurate prediction of sales volume of agricultural products helps to solve core precise marketing of online sales of agricultural products. Extracting useful and

understandable knowledge from a large amount of sales data to provide effective sales forecast results. Sales regions and consumer groups help the buyers to adjust sales plans in real-time and are also helpful for agricultural e-commerce companies to get the best promotion channels based on prediction results [6, 7]. Marketing motion pictures based on high financial stakes provide the importance of decisions and accurate box office are measured [8]. The organized large-scale retail sector is gradually placed around the world and exponentially increases the activities in the pandemic period. Modern sales system based on machine learning methods provides more accurate information to increase profit [9, 10].

This paper is organized as follows: a literature review is provided in Section 2 and the proposed method is explained in Section 3. The simulation result is given in section 4 and the results and discussion are given in Section 5. The conclusion of this research work is given in Section 6.

2 Literature Review

Feature engineering improves machine learning prediction and classification performance. Sales prediction models based on machine learning techniques with feature selection helps to improve the efficiency of the model. Various methods involved in applying the feature selection and machine learning methods in sales prediction were reviewed in this section.

Punam, *et al*. [11] proposed a two-level prediction model to predict the sales of a particular outlet than the single predictive model. The Big Mart sales data of 2013 was used to test the performance of the developed method. A feature engineering method was applied to improve the performance of the prediction process. Linear regression, Support Vector Machine (SVM) and cubist were used in the stacking ensemble method for two-level prediction model. Single models such as linear regression, support vector machine and k-nearest neighbour were compared with the two-level approach method. The result shows that the developed method has higher performance compared to existing methods. The developed method has lower performance in feature engineering and the SVM model has an imbalance data problem.

Chen, *et al*. [12] developed a framework of TADA based on encoder-decoder Recurrent Neural Network (RNN) for sales prediction. An online learning model based on reservoir similarity is applied to enhance the performance of the developed method. The developed model selects ‘hard’ data samples to mine apparent dynamic patterns in model construction. Two real-world datasets were used to test the performance of the developed model in the sales prediction. The LSTM based model has the limitations of vanishing gradient problem that affects the performance of the classification. The developed model also suffers from the limitation of the cold-start problem in the classification.

Huang, *et al*. [13] applied the Dependency SCOR-topic sentiments (DSTS) model to improve the sales prediction. The auto-regressive method was applied in the developed DSTS model to test the tea performance prediction. The developed method distribution, review text analysis, and topic probability increase the performance of the sales prediction. The study improves the performance of the sales prediction using sentiment analysis. The efficiency of sentiment analysis was less and classification performance needs to be improved.

Xia, *et al*. [14] proposed the ForeXGBoost method to improve the performance of the sales prediction. The carefully-designed data filling algorithm was used in the ForeXGBoost method to missing values and improve the quality of data. The ForeXGBoost method applied the sliding window method to analyse production data features and historical sales to improve the performance of sales prediction. The data correlation and information gain were applied to analyse the importance of the different attributes. The XGBoost prediction model improves the performance of sales prediction and short computation time for vehicle sales prediction. This analysis shows that the XGBoost method has a lower overhead in the sale prediction process compared to the existing method. The XGBoost method has the limitation of overfitting problem that affects the performance of prediction.

Behera and Nain, [15] proposed Grid Search Optimization (GSO) method to optimize the parameter and select the best hyperparameter for XGBoost techniques for sales forecasting. The Big mart dataset was used to test the performance of the developed method for sales prediction. The XGBoost method with GSO provides higher performance in the prediction of sales. The overfitting problem in XGBoost classifier affects the performance of the prediction and has lower efficiency in prediction.

3 Proposed Method

3.1 Univariate Analysis

A generalization of a linear model is represented in a GLM and suitable transformations are used to transform nonlinear data into learning form to expand its scope. The linear regression limitation is handled by GLMs and assumes the linear relationship between input and output. Adding a step of transforming data part into another domain to provide a non-linear relationship between input and output. This process is called the basis function. Logistic regression is one of the widely used basis functions and a logistic function is applied to transform non-linearity into linear. The output is mapped between a range of [0, 1] in logistic function and equivalent to a probability density function.

The linear regression output is added with an exponential function in logistic regression $y_i \in R$ and constrain into $y_i \in [0, 1]$. The input and predicted output relationship is calculated using equation (1).

$$\hat{y_i} = \sigma(\sum_{j=1}^{n} x_{i,j} \times w_j + w_0) \quad (1)$$

Logistic regression output represents asymmetrical distribution between $[-\infty, \infty]$ and this is suitable for classification problems.

3.2 XGBoost Algorithm

The XGBoost method is a machine learning method and gradient boosting method, which consists of weak predictors' sequences. This method is based on the. Gradient boosting is iterative tree estimation, residuals obtained at each step and adaptive estimates updates. Gradient descent technique is used in Gradient boosting method and this splits the favours to reduce the point of the objective function.

The XGBoost optimization is compared with gradient boosting due to regularization to avoid bias and overfitting, missing values management, tree pruning operations, parallel and distribution computing use, and its scalability.

The variables x_i is a set of values in input data and predict the variable y_i, as given in equation (2).

$$\{(x_i - y_i)\}_{i=1}^{n} \quad (2)$$

This consists of a training dataset, the model predicts the variable value y_i based on variable x_i to characterize multiple features. The predicted value is $\hat{y}_i = \sum_j \theta_j x_{ij}$ is used in a linear regression problem, where the weight of x_j is denoted as θ_j. The model parameters are denoted as θ in a generic problem.

The objective function measures the model ability to fit training data that consists of two terms, as given in equation (3).

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (3)$$

Where the regularization term is denoted as $\Omega(\theta)$ and the training loss function is denoted as $L(\theta)$. The prediction is evaluated using the differentiable function of a loss function. The regularization term helps to control model complexity and avoid overfitting.

The loss function of Taylor expansion is used in XGBoost to design an objective function, as given in equation (4).

$$Obj(\theta) = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (4)$$

Where $g_i = \partial_{\hat{y}_i^{t-1}} L(y_i, \hat{y}_i^{t-1})$, while $h_i = \partial^2_{\hat{y}^{t-1}} L(y_i, \hat{y}_i^{t-1})$. The following quantities are defined, as given in equations (5) to (7).

$$G_j = \sum_{i \in I_j} g_i \quad (5)$$

$$H_j = \sum_{i \in I_j} h_i \quad (6)$$

$$I_j = \{i | q(x_i) = j\} \quad (7)$$

The j^{th} leaf optimal weight value is denoted as $\theta_j = -\left(\frac{G_j}{H_j + \lambda}\right)$ that returns the leaf index itself. The j^{th} leaf instance set is denoted as I_j and mapping function of data instance into tree leaf. The model optimizes based on objective function is given in equation (8).

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (8)$$

Where characterize of tree leaves is denoted as T.

The algorithm computation cost is due to all tree training simultaneous. The split candidate evaluation based on gain function is given as in equation (9).

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_P^2}{H_P + \lambda} \quad (9)$$

Where left nodes (subscript L) is contributed based on the first term, the right nodes (subscript R) is contributed based on the second term, the parent leaf node (subscript P) is contributed by the last term. The greatest gain of the split condition is selected and the pruning method is used to optimize a tree level to reduce overfitting.

Bayesian optimization could find the optimal value through only a small number of samples. Compared with traditional optimization methods, it does not need the explicit expression of the function. Hence, Bayesian optimization is appropriate for tuning hyperparameters. In this section, the Bayesian optimization algorithm is applied to optimize hyperparameters for three widely used machine learning models.

4 Simulation Setup

Sales prediction is a challenging task that involves analysing the features related to sales of the product. This research involves applying the XGBoost classifier to analyse the features and provide the prediction. This section provides the details of datasets and system configuration in the implementation of the proposed method.

Dataset: The big mart sales 2013 and 2018 datasets were used to test the performance of the proposed method. The BigMart has data of 1559 products across 10 stores in different cities and also has certain attributes of products and stores. The dataset has 8523 instances in the train set and the test data has 5681 instances. The variables and descriptions of the dataset are shown in Table 1.

Table 1. The variables and description of the dataset

Variables	Description
Item_Identifier	Unique Product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_visibility	The % of a total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs.
Item_MRP	Maximum retail price(list price) of the product.
Outlet_Identifier	Unique store ID.
Outlet_Establishment_Year	The year in which the store was established.
Outlet_Size	The size of the store in terms of ground area covered.
Outlet_Location_Type	The type of city in which the store is located.
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Item Weight, Item Visibility, Item MRP, and Item Outlet_Sales(Target Variable) are numerical features. Item Identifier, Item Fat Content (Ordinal Feature), Item Type, Outlet Itemtifier, Outlet Establishment Year, Outlet Size (Ordinal Feature), Outlet Location Type (Ordinal Feature), and Outlet Type (Ordinal Feature) are categorical features. The dataset consists of 4 float type variables, 1 integer type and 7 object type features. The Item Establishment Year is a categorical feature because it contains the fixed value and does not convert its data types. Ordinal features are Item fat content, outlet size, outlet location type and outlet type due to its values being arranged in some order.

System Configuration: The proposed method is implemented in the system consisting of an Intel i9 processor, 128 GB of RAM, 22 GB graphics card and Windows 10 OS. The Python tool is used to implement the proposed method.

5 Result and Discussion

The sale prediction model helps to improve the profit of the company and to understand the requirement of customers. Feature analysis is an important process in the sale prediction model and feature analysis improves efficiency. In this research, the XGBoost model with univariate and bivariate analysis was applied to analyse the important features in the dataset. The univariate model analyses the feature importance of individual features related to the output variable. The bivariate method applies two features to analyse the correlation related to the output variable.

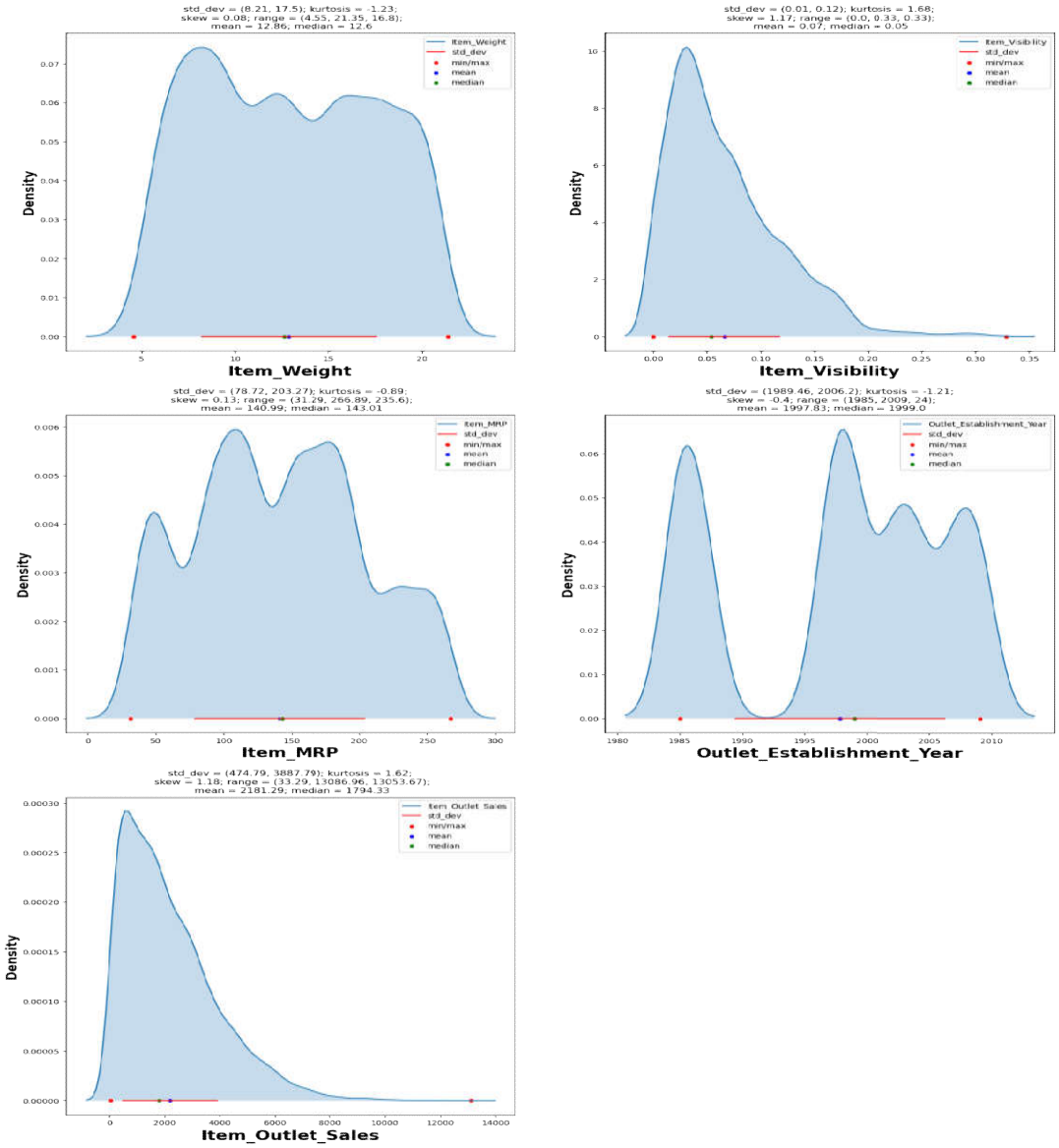

Figure 1. The univariate analysis of features

The univariate analysis of features is shown in Figure 1. The total count of item weight is 7060 that is less than the length of a dataset and this also has missing values. The average weight of all items is 12 kg and the maximum weight of items is 21 kg. The stores are not selling heavyweight items. The maximum price of an item is Rs 266 and this shows that stores do not sell costly items like laptops, mobile phones and TV etc. The recent store was established in 2009 and the oldest store is established in 1985. The average sale of the

item is Rs 2181 and the maximum sale is Rs. 13,086. The weight of items is present in the range of 4 – 22 and the average weight of items is 12. Some items are not visible and the maximum visibility of the item is 33 %. The price of the item is in the range of Rs 31 – 265. The most expensive item in the stores is Rs 266.89. Most of the stores are established in 1985 – 2000. This shows 1990 – 1995 is not a good period to open the stores and no stores are established in this period. Most of the stores have a maximum sales in between 450 to 3900 and only a few stores have more than 6000 sales.

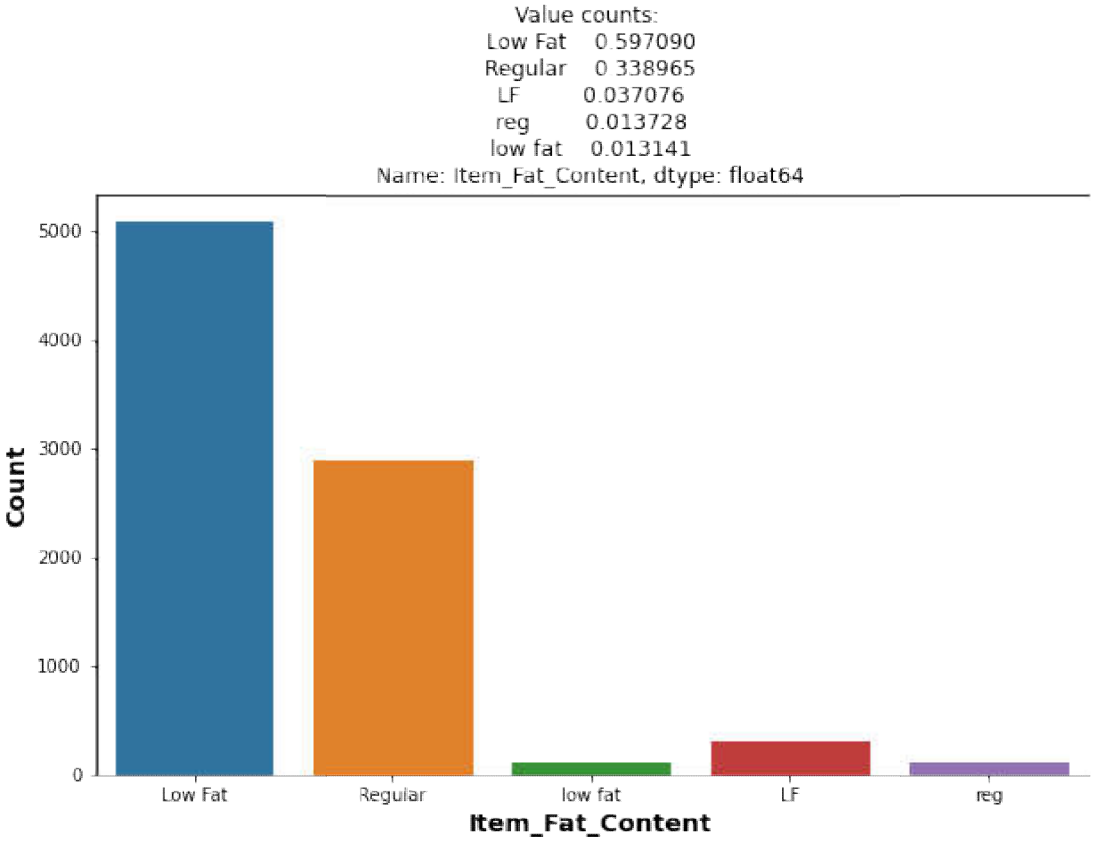

Figure 2. The count of item fat content

The count of item fat content is shown in Figure 2. This shows that around 64 % of total items have low fat and the remaining have regular fat. The 'Low Fat', 'low fat, and 'LF' needs to be renamed as 'low fat'. The 'Regular' and 'reg' need to rename as 'Regular fat' in pre-processing.

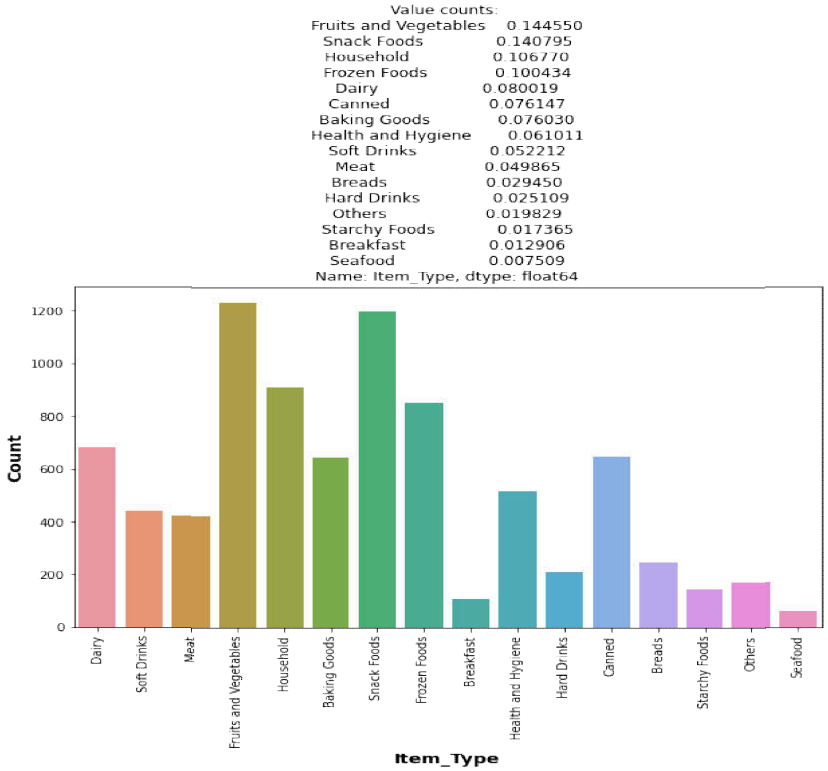

Figure 3. The count of item type sales

The count of item type sales is shown in Figure 3 and this shows the number of items sold for the respective type. More than 14 % of items (more than 1200 items) are fruits & vegetables, snacks and foods. This shows that sales of seafood and breakfast types of items are very less.

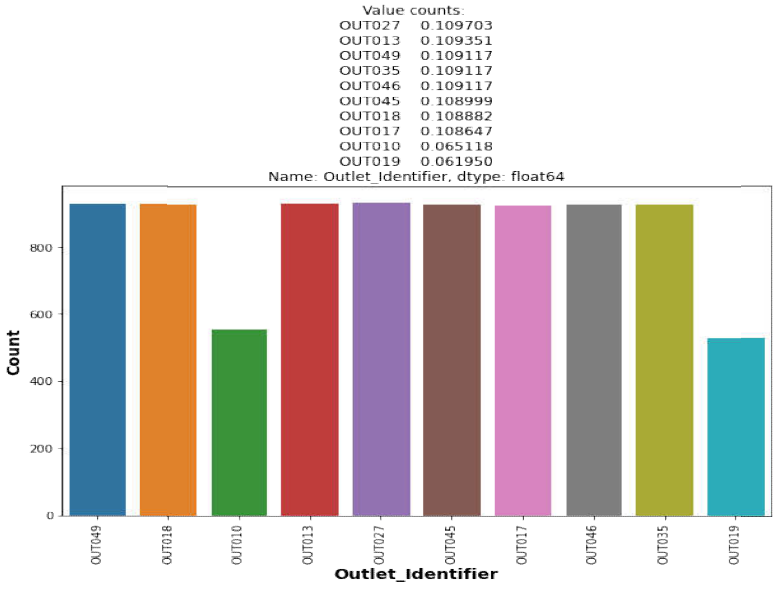

Figure 4. Outlet identifiers of 10 stores

The outlet identifiers of 10 stores are shown in Figure 4 and this shows the number of sales of each store. Most of the stores are almost sold the same number of items except OUT010 and OUT019 stores.

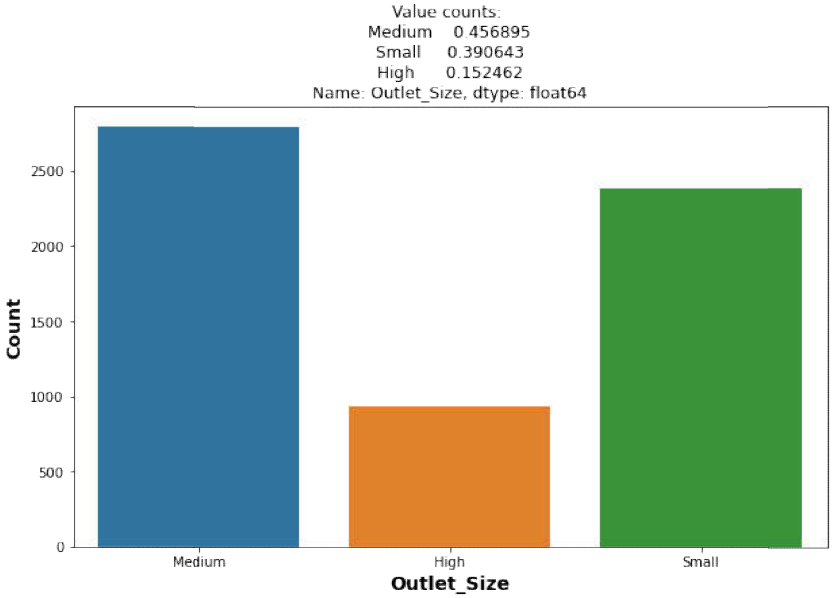

Figure 5. Sales count based on the size of stores

The sales count based on the size of stores is shown in Figure 5 and this shows three categorical values of small, medium and High size stores. This shows that medium-size stores have high sales values of 45 %, small size stores have 39 % and high size stores have 15 %.

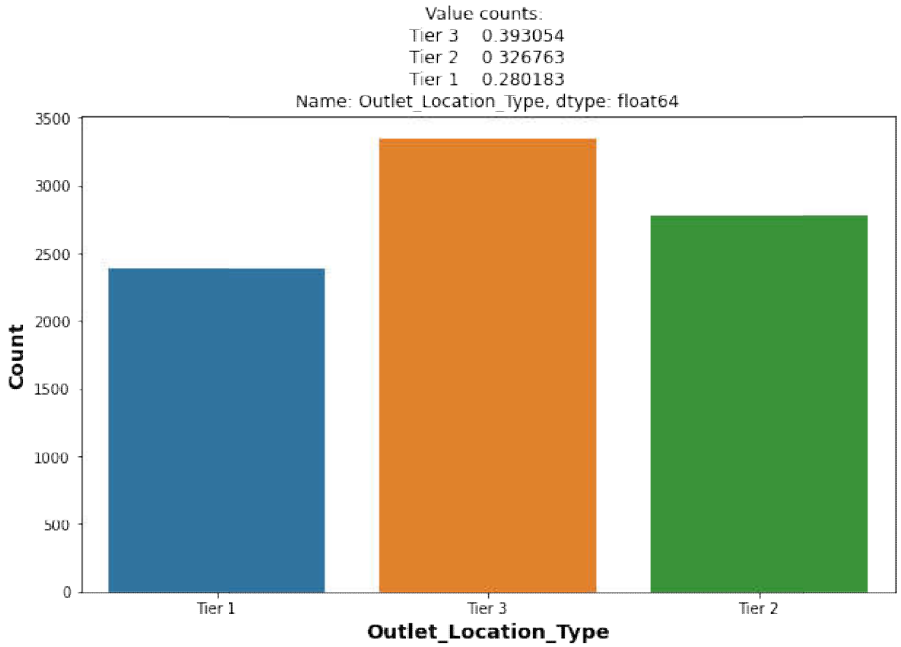

Figure 6. The sales based on Tier cities

The count of sales based on Tier cities is shown in Figure 6 and shows the sales from Tier 1, 2 and 3 cities. This shows that Tier 3 cities have higher sales compared to Tier 1 and Tier 2 cities. Tier 1 cities have 28 % sales, Tier 2 cities have 32 % sales and Tier 3 cities have 39 % sales.

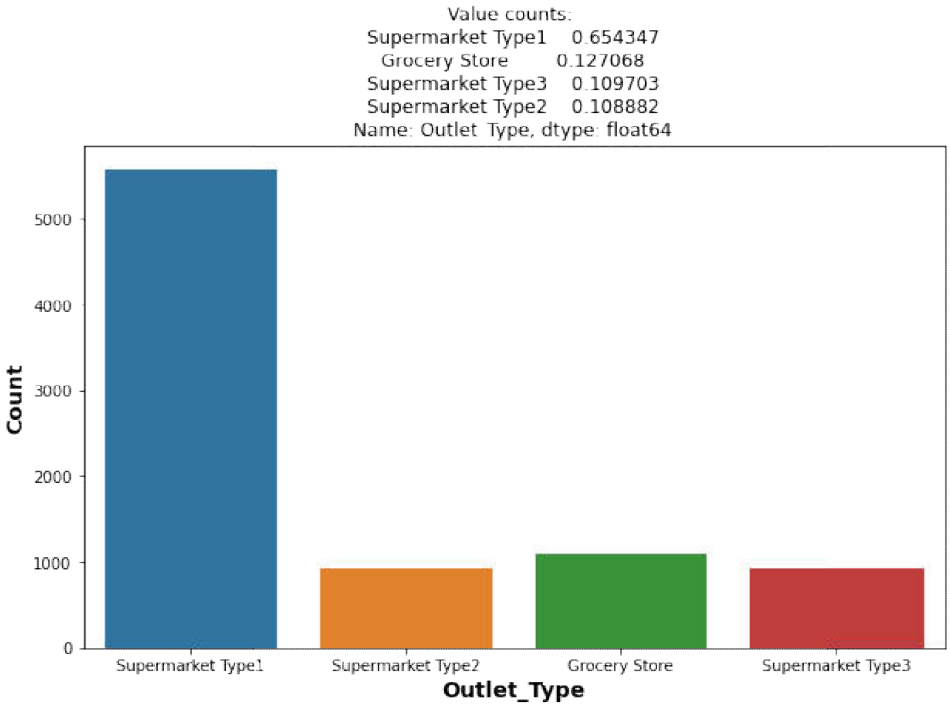

Figure 7. The sales count based on supermarket types

The sales count based on the types of supermarket is shown in Figure 7 and this shows the number of sales from each supermarket type. The supermarket type 1 has 65 % of sales and this is twice of other types of stores. The supermarket type 2 has 10.8 %, supermarket type 3 has 10.9 %, and Grocery store has 12.7 % sales.

The Item weight feature has 17.16 % of missing values and outlet size has 28.27 % of missing values. The percentage of missing is high so it shouldn't be dropped and need to be replaced.

Numerical analysis: The weight of the item is in the range of 4 to 22 and the average weight is 12 %. Some of the items are not visible and have a maximum of 33 % visibility. The price of the item range is Rs 31 to Rs 265 and the most expensive item in the store is Rs 266.89. Most of the store is established in the year of 1985-1990 and 1995 to 2000. Most of the stores have maximum sales of 450-3900 and a few stores have 6000 sales.

Categorical analysis: Around 64% of the total item has low fat and the remaining contain regular fat. More than 14% of items (more than 1200 items) are fruits & vegetables, snacks and foods. Seafood and breakfast type item sales are very less. Except for OUT010 and OUT019 stores, other stores in the datasets have the same number of sales. Medium size stores have 45 % of the total number of items sold and 15 % of sales are from high size. Tier 3 cities have 39 % of items sold, Tier 2 cities stores have 32 % items sold and Tier 1 cities have 28 % items sold. The Supermarket Type 1 stores have 65 % of the item sold and other types have less sales.

5.1 Bivariate analysis

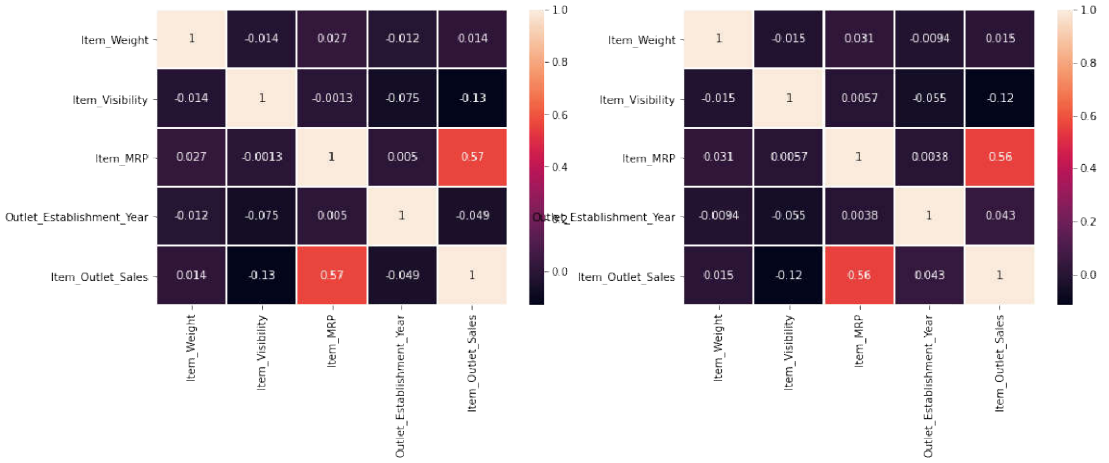

Figure 8. The correlation analysis of features in datasets

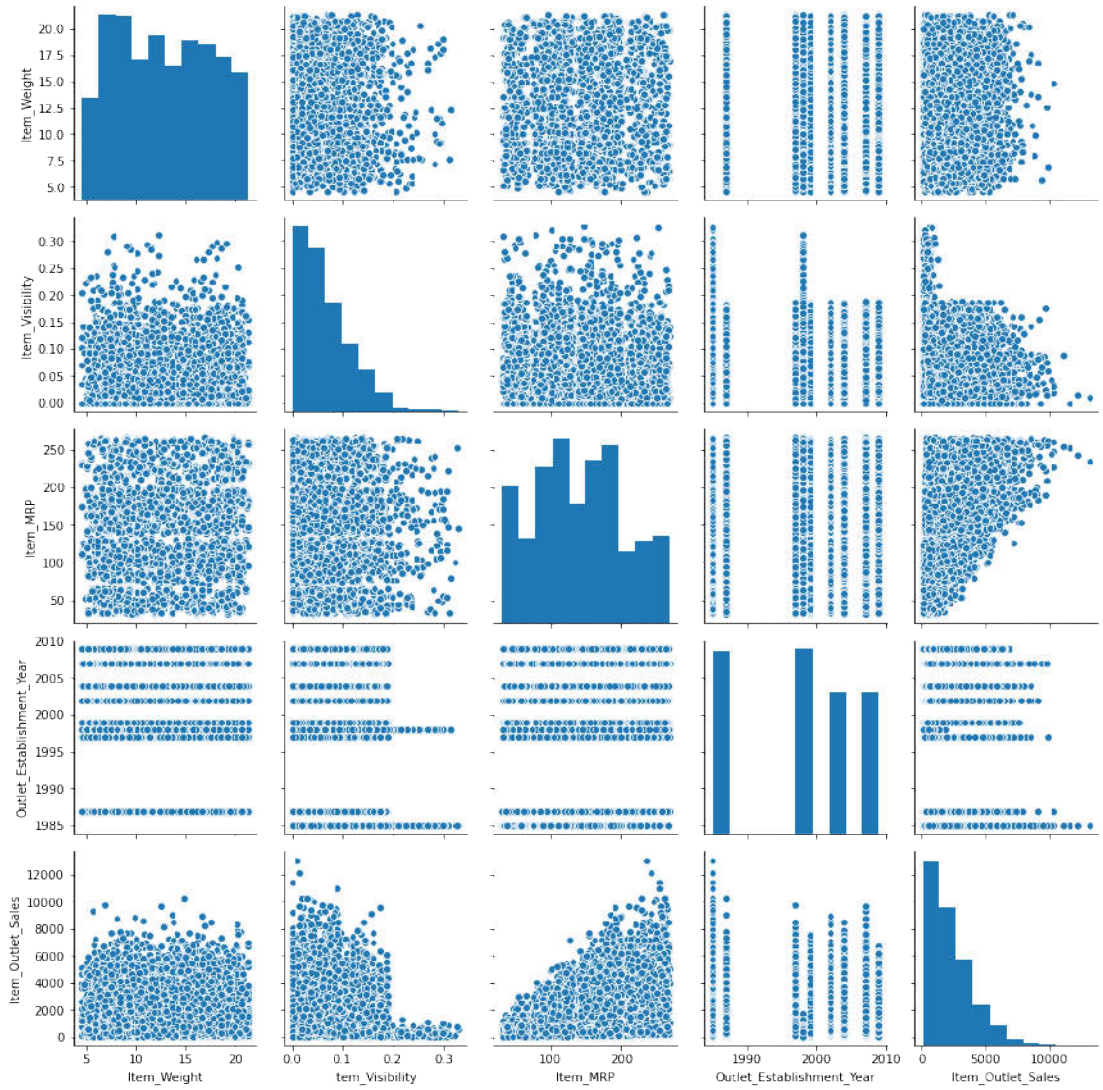

Figure 9. A joint plot of features in a dataset

The correlation heat map and joint plot of features in a dataset are shown in Figure 8 and Figure 9. The item MRP and item outlet features have a high correlation compared to other features in the datasets. The increase in item visibility decreases the item outlet sales due to it having a negative correlation. The item weight and item establishment year have a lower correlation with item outlet sales.

Item weight is not an important feature for the item outlet sales and there is no significant change in item outlet sales and item weight. The item fat content distribution is slightly right skew. The significant difference in item outlet sales of various item types. The item types feature is an important feature in the dataset based on correlation. Dairy products have higher item outlet sales compared to others. A significant difference is present between item outlet sales of various outlet sizes. Medium size stores have more item outlet sales compared to other and small size stores have low item outlet sales. The 'Medium' outlet size of mean item outlet sales is more than 2500, 'high' is less than 2500, and 'small' is less than 200. The significant difference between the stores' item outlet sales of various location types. Tier 2 cities stores have more sales and Tier 1 cities have less sales. The average sale of Tier 2 cities is 2324 and Tier 1 has 2279 sales. The supermarket type 3 has more sales than others and the average sales of supermarket type 3 are 3694 sales. Grocery stores have less item outlet sales. There is no significant difference in item outlet sales for various stores.

5.2 Comparative Analysis

The XGBoost method with Bayesian optimization has the advantage of applying the probability model to find the optimal parameter for the classifier. The probability model reduces the error value due to its adaptive learning of data. The Grid search optimization has the limitation of applying specific values to search for optimal parameters and doesn't adaptive change the parameter value. The XGBoost method has the advantage of developing a tree structure to classify the data.

Table 2. Comparative Analysis of proposed and existing methods

Methods	**MAE**	**MSE**	**RMSE**
Two-level Statistical Model [11]	0.3917	-	-
XGBoost [15]	-	134.08	180.2
XGBoost (Grid Search Optimization) [15]	-	129.9	178.7
XGBoost (Bayesian Optimization)	0.311	34.37	5.86

The proposed XGBoost with the Bayesian optimization method is compared with existing methods in sales prediction, as shown in Table 2. The proposed XGBoost with the Bayesian optimization method has a lower error value compared to existing methods. The proposed XGBoost with Bayesian optimization has 0.311 MAE, and the existing two-level statistical model [11] has a 0.3917 MAE value.

6 Conclusion

Sales prediction is a challenging task due to the presence of features in the dataset related to output variables. Feature selection helps to increase the efficiency of the sales prediction model. In this research, the XGBoost method (Bayesian optimization) with univariate and bivariate are proposed to increase the feature analysis of the sales prediction.

The univariate method analyses the individual feature importance in the study. The bivariate method analyses the correlation between the two features related to output variables to find the important features. The univariate method shows that item types of food & vegetables, snacks and foods have higher sales of 14 % compared to other items. The bivariate method shows that item types have a higher correlation with sales item outlets and item types have higher feature importance in the model. The XGBoost with Bayesian optimization method has 34.37 MSE and the existing XGBoost with Grid search optimization method has 129.9 MSE. The future work of this method involves applying neural network-based models to improve the performance of sales prediction.

References

[1] Zhao, J., Du, B., Sun, L., Lv, W., Liu, Y. and Xiong, H., 2021. Deep multi-task learning with relational attention for business success prediction. Pattern Recognition, 110, p.107469.

[2] Kohli, S., Godwin, G.T. and Urolagin, S., 2021. Sales Prediction Using Linear and KNN Regression. In Advances in Machine Learning and Computational Intelligence (pp. 321-329). Springer, Singapore.

[3] Ji, S., Wang, X., Zhao, W. and Guo, D., 2019. An application of a three-stage XGBoost-based model to sales forecasting of a cross-border E-commerce enterprise. Mathematical Problems in Engineering, 2019.

[4] Ribeiro, A., Seruca, I. and Durão, N., 2017. Improving organizational decision support: Detection of outliers and sales prediction for a pharmaceutical distribution company. Procedia computer science, 121, pp.282-290.

[5] Posch, K., Truden, C., Hungerländer, P. and Pilz, J., 2021. A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants. International Journal of Forecasting.

[6] Mu, S., Wang, Y., Wang, F. and Ogiela, L., 2021. Transformative computing for products sales forecast based on SCIM. Applied Soft Computing, p.107520.

[7] Branda, F., Marozzo, F. and Talia, D., 2020. Ticket Sales Prediction and Dynamic Pricing Strategies in Public Transport. Big Data and Cognitive Computing, 4(4), p.36.

[8] Bogaert, M., Ballings, M., Van den Poel, D. and Oztekin, A., 2021. Box office sales and social media: A cross-platform comparison of predictive ability and mechanisms. Decision Support Systems, p.113517.

[9] Massaro, A., Panarese, A., Giannone, D. and Galiano, A., 2021. Augmented Data and XGBoost Improvement for Sales Forecasting in the Large-Scale Retail Sector. Applied Sciences, 11(17), p.7793.

[10] Khalil Zadeh, N., Sepehri, M.M. and Farvaresh, H., 2014. Intelligent sales prediction for pharmaceutical distribution companies: A data mining based approach. Mathematical Problems in Engineering, 2014.

[11] Punam, K., Pamula, R. and Jain, P.K., 2018, September. A two-level statistical model for big mart sales prediction. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 617-620). IEEE.

[12] Chen, T., Yin, H., Chen, H., Wang, H., Zhou, X. and Li, X., 2020. Online sales prediction via trend alignment-based multitask recurrent neural networks. Knowledge and Information Systems, 62(6), pp.2139-2167.

[13] Huang, L., Dou, Z., Hu, Y. and Huang, R., 2019. Online Sales Prediction: An Analysis With Dependency SCOR-Topic Sentiment Model. IEEE Access, 7, pp.79791-79797.

[14] Xia, Z., Xue, S., Wu, L., Sun, J., Chen, Y. and Zhang, R., 2020. ForeXGBoost: passenger car sales prediction based on XGBoost. Distributed and Parallel Databases, 38, pp.713-738.

[15] Behera, G. and Nain, N., 2019, November. Grid Search Optimization (GSO) Based Future Sales Prediction For Big Mart. In 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 172-178). IEEE.