International Journal of Research in Engineering, Science and Management
Volume 5, Issue 7, July 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Prediction of Online Product Sales using Machine Learning

K. P. Aldrin Neal John[1*], K. Adarsh Nag[2]
[1,2]*Department of Computer Science Engineering, Albertian Institute of Science and Technology, Kalamassery, India*

***Abstract*: Product sales prediction is a major aspect of purchasing management. One of the key challenges faced nowadays by organizations the dynamic, international and unpredictable business environment in which they operate. With growing customer expectations for price and quality, manufacturers today can no longer rely only on cost advantage that they have over their rivals. Forecasting the sales are crucial in determining inventory stock levels and accurately estimating the future demand for goods has been an ongoing challenge in industries If goods are not readily available or if goods availability is more than demand overall profit can be compromised. As a result, sales prediction for goods can be significant to ensure that loss is minimized. Depending on this study, our project is creating a prediction model using machine learning algorithms for accurately predicting online product sales. Our project aims to use upto date data which includes online reviews, online ratings, online promotional strategies and sentiments and various other parameters for predicting product sales.**

***Keywords*: Clustering, Machine Learning.**

1. Introduction

Data is growing in massive amount on internet and time plays very important role in every person's life. It is impossible for a single person to read whole data daily. Retailers nowadays understand his well and attempt to make use of it in an effort to gain an edge in a highly competitive market. This is specially done in an effort to make purchasing more likely, in addition to balancing the scalability and profit in setting the selling price of a product. Companies frequently introduce additional elements to the offer which are aimed at increasing the perceived value of the purchase to the customer. Sometimes decision regarding whether or not to make a purchase is dependent on price but in many cases the purchasing decision is more complex. An important aspect of managing supply chain efficiently is to have better prediction of sales such that manufacturer will not over or under purchase production products. An important aspect of managing supply chain efficiently is to have better prediction of sales such that manufacturer will not over or under purchase production products. An emerging area in prediction of sales is in big data and user-generated content on the sales of product. Given that user-generated content plays an important role in influencing the purchasing decisions of consumers, its role in helping organizations to understand and predict product demand can be investigated.

2. Literature Survey

A. *Forecasting of Bigmart Sales*

This approach was proposed by Deven Ketkar. In this methodology raw data collected at big-mart was pre-processed for missing anomalies and outliers. Then an algorithm was trained on this data to create a model. Algorithms used were Random forests and multiple Linear Regression. ETL that is Extract, Transform and Load tool was used in this methodology to get data from one database and transform it into suitable format. Data was transformed from sample raw data into understandable format. The model was used for final results. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. Different machine learning algorithms like linear regression analysis, random forest, etc. are used for prediction or forecasting of sales volume. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex. Always a better prediction is helpful, to develop as well as to enhance the strategies of business about the marketplace which is also helpful to improve the knowledge of marketplace.

B. *Sales Time Forecasting*

This approach was proposed by Bohdan M. Pavlyshenko. This methodology is a stacking approach for building regression. Ensemble of single models was studied for implementation. The algorithms used were Random Forest and regression. The results showed that using stacking techniques we can improve performance of predicting model. One of the approaches is Time Series Approach which means, A time series is a sequence of data points taken at successive, equally-spaced points in time that can be used to predict the future. A time series analysis model involves using historical data to forecast the future. It looks in the dataset for features such as trends, cyclical fluctuations, seasonality and behavourial patterns.

3. Proposed System

The study helped to design a model which can facilitate future business researches for predicting product sales in an online environment. The main objective of the project is to show that product demands can be predicted through the

*Corresponding author: aldrinneal83@gmail.com

comparative influence of promotional marketing strategies. In this we use regression algorithm for classification of datas. This study will then use a Multiple Linear Regression to predict product sales, as well as to predict the effects of the online sentiments on the same so as to design effective promotional strategies and sales tactics. It requires various parameters related to the product in-order to predict the sales demand and the output sales.

4. Methodology

The required big mart dataset is collected and products are classified into different types according the various parameters related to the product. And here Regression algorithm is used for classification for product data. For the prediction of the output sales multiple linear regression algorithms are used which is a statistical technique that uses various explanatory variables to predict the outcome of response variable. Formula used in this is y=b0+b1*x1+b2*x2+.....bn*xn, where y = dependent variable and x = independent variables Parameters. Random forest algorithms are used for getting accurate predictions for small datasets.

5. Experimental Setup

The experimental setup is as follows:

The big mart dataset contains parameters related to product which are used for predicting the sales and to analyse the product demand in various organizations. In these libraries such as pandas, numpy, mat-plotlib, seaborn, sklearn are used. For both data visualization and data model building regression algorithm are used. In this regression models such as linear regression, decision tree regression, multi-linear regression and random forest regression algorithms are used.

A. Data Cleaning

Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect. This is generally data that can have a negative impact on the model or algorithm it is fed into by reinforcing a wrong notion. It is a key step before any form of analysis can be made on it. Datasets in pipelines are often collected in small groups and merged before being fed into a model. Merging multiple datasets means that redundancies and duplicates are formed in the data, which then need to be removed. Models trained on raw datasets are forced to take in noise as information and this can lead to accurate predictions when the noise is uniform within the training and testing set only to fail when new, cleaner data is shown to it. Hence it is considered.

B. Null Value Filling

In this method we train an ML Model, Regression or Classification for Numerical or Categorical Missing Data Column respectively and then let the model predict the missing values. One of the most favourable algorithms to implement this method is KNN because it takes distance between two data points in n-dimentional vector space into account. This method is also referred to as "nearest neighbour imputation". *Multiple inputting:* This method is like Bagging based ensemble of Regression/Classification Imputation method, what I mean by that is, Regression/Classification Imputation is used Multiple times instead of a Single time and mean or voting methods is applied respectively to generalize the results.

C. User Interface

The project is run on the user's local host. In the main page of sales prediction tab, it contains several empty rows which needs to be filled with the parameters related to the product. Once after inputting them the total product sales is predicted at the end. Libraries like pandas, numpy, matplotlib, seaborn etc can be used for preprocessing and visualization purposes. Library like sklearn can be used for splitting dataset, model building purposes, score calculation purposes. We are using R2 score, mean squared error (MSE), root mean squared error (RMSE) for model evaluation purposes.

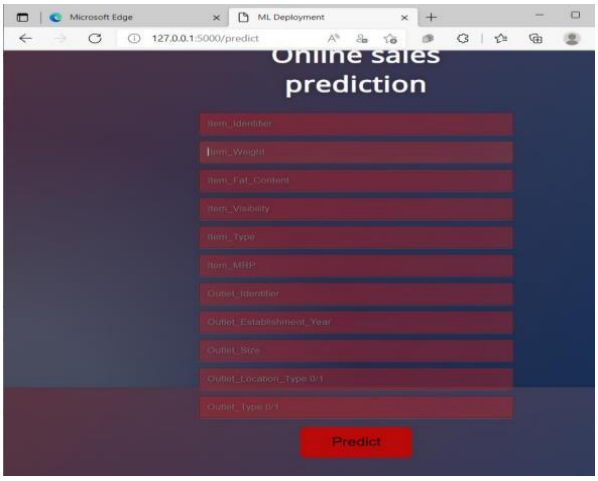

Fig. 1. Online sales prediction

6. Result

The proposed system's results demonstrate how well the model can predict sales output and helps organizations to understand product demand in the society. The proposed system's results demonstrate how well the model can predict sales output and helps organizations to understand product demand in the society.

Fig. 2. Dataset used for predicting the product sales in those particular year

In dataset certain columns indicates sales of product in

specific outlet, it can be point value not only whole numbers because it is regression problem. The strings are converted into corresponding numeric values and then for further process.

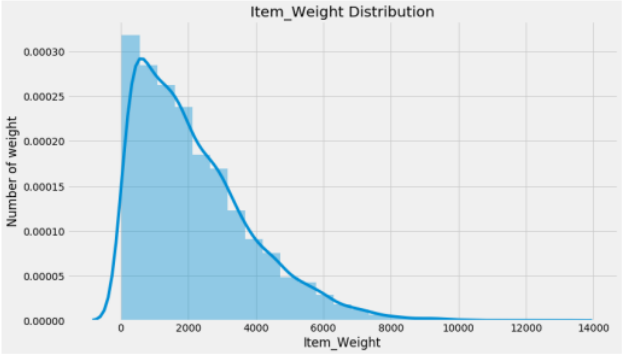

Fig. 3. Item-weight distribution graph

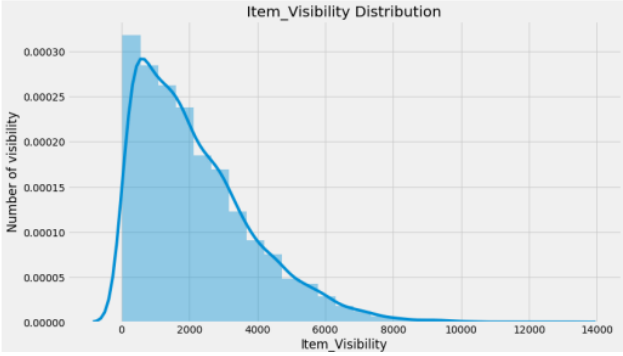

Fig. 4. Item-visibility graph

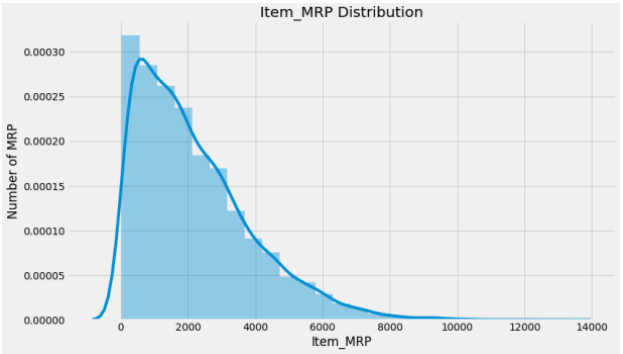

Fig. 5. Item-MRP distribution graph

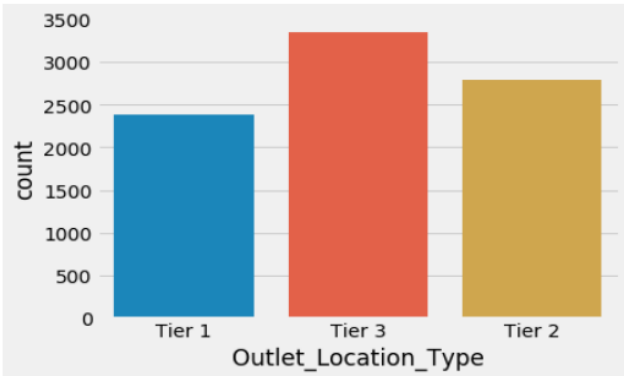

Fig. 6. Distribution of outlet types graph

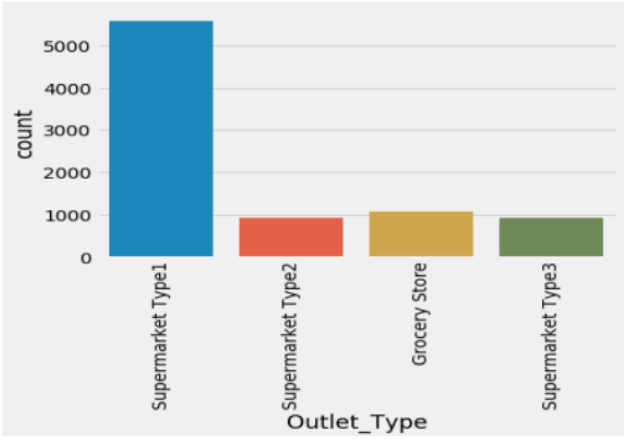

Fig. 7. Distribution of outlet size type graph

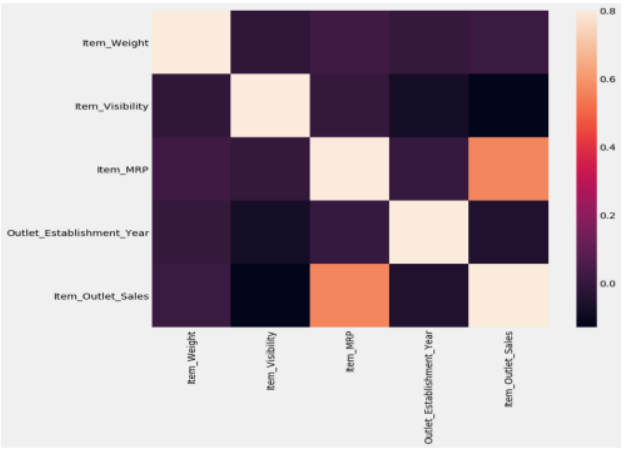

Fig. 8. Heat map of numeric features

7. Conclusion

This paper proposes of how a product's sales demand can be calculated by assigning various parameters such as establishment year, price, fat content, etc., to rows they are assigned at. It helps in finding the best model for building the best model for the project. The Dataset is allowed test and train with values in-order to undergo model evaluation, convert a model if into pickle format. On predicting the average occurred in that particular year is found and the output indicates the number of units sold in that year. The decimal value is occurred since the average of the products sold is predicted.

References

[1] East, R. Hammond, K. and Lomax W., "Measuring the impact of positive and negative word of mount on brand purchase probability", International Journal of research in Marketing, vol. 25, no. 3, pp. 215-224, 2008.
[2] Cui, G., Lui, H. And Guo, X, "The effect of online consumer reviews on new product sales", International Journal of Electronics, 2012.
[3] Bohdan M. Pavlyshenko, "Machine Learning models for sales time series forecasting," 2018.
[4] Deven Ketkar, "A Forecast for Big Data Sales based on Random Forests and Multiple Linear Regression", IJEDR, vol. 6, no. 4, 2018.
[5] https://www.tutorialspoint.com/matplotlib/index.html
[6] https://www.investopedia.com/terms/m/mlr.asp