ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse951022021.pdf
https://doi.org/10.30534/ijatcse/2021/951022021

Supermarket Sales Prediction Using Regression

[1]Melvin Tom, [2]Nayana Raju, [3]Asha Issac, [4]Jeswin James,[5]Rani Saritha R
[1]PG student Saintgits College of Engineering,Pathamuttom,melvintom63@gmail.com
[2]PG student Saintgits College of Engineering Pathamuttom,nayanaraju1@gmail.com
[3]PG student Saintgits College of Engineering Pathamuttom,ashaissac77@gmail.com
[4]PG student Saintgits College of Engineering Pathamuttom,jeswinjames555@gmail.com
[5]Asst.Professor, Saintgits College of Engineering Pathamuttom,rani.saritha@saintgits.org

ABSTRACT
Sales forecasting is an important when it comes to companies who are engaged in retailing, logistics, manufacturing, marketing and wholesaling. It allows companies to allocate resources efficiently, to estimate revenue of the sales and to plan strategies which are better for company's future. In this paper, predicting product sales from a particular store is done in a way that produces better performance compared to any machine learning algorithms. The dataset used for this project is Big Mart Sales data of the 2013.Nowadays shopping malls and Supermarkets keep track of the sales data of the each and every individual item for predicting the future demand of the customer. It contains large amount of customer data and the item attributes.
Further, the frequent patterns are detected by mining the data from the data warehouse. Then the data can be used for predicting the sales of the future with the help of several machine learning techniques (algorithms) for the companies like Big Mart. In this project, we propose a model using the Xgboost algorithm for predicting sales of companies like Big Mart and founded that it produces better performance compared to other existing models. An analysis of this model with other models in terms of their performance metrics is made in this project. Big Mart is an online marketplace where people can buy or sell or advertise your merchandise at low cost. The goal of the paper is to make Big Mart the shopping paradise for the buyers and a marketing solutions for the sellers as well. The ultimate aim is the complete satisfaction of the customers. The project "SUPERMARKET SALES PREDICTION" builds a predictive model and finds out the sales of each of the product at a particular store. The Big Mart use this model to under the properties of the products which plays a major role in increasing the sales. This can also be done on the basis hypothesis that should be done before looking at the data.

Key words: Regression, Sales, Prediction, Data Exploration, Bigmart, XGBoost, Ridge Regression.

1. INTRODUCTION

Supply and demand are two fundamental concepts of sellers and customers. Predicting the demand accurately is critical for the organizations in order to be able to formulate the plans. Big Mart is an online stop marketplace, where we can buy, sell or advertise your merchandise at a very low cost. The major goal is to make Big Mart the shopping paradise for the buyers and the marketing solutions for the sellers.

The ultimate goal is to multiply with the customers. The project "SUPERMARKET SALES PREDICTION" aims to build a predictive model and find out the sales of each of the products at a particular store. The Big Mart can use this model to understand the properties of the products which plays a key role in increasing the sales. This can also be done on the basis hypothesis that should be done before looking at the data.

In this project, we propose a predictive model using XG Boost technique for predicting the sales of company's like Big Mart and found that the model produces better performance as compared to existing models. The major aim of this machine learning project is to build a predictive model and also to search out sales of each of the products at a particular selected store. Using this machine learning model, the Supermarket sales prediction tries to understand the properties of the products and stores which plays a key role in increasing the sales of products. Sales Prediction is used to predict the sales of different products sold at various outlets in different cities of a Big Mart Company. Using this model, we will try to understand the properties of the products and stores which play a major role in increasing sales. Here python is used as programming language and Jupyter Notebook is used as tools. To build this application, machine learning aspects such as Supervised Learning task, Regression task are used. This is mainly done in order to predict the sales of a company stores in the future.
The Various processes used are: Exploring the data and Data Pre-processing, Feature Engineering, Creating Model, Evaluation. Supervised learning helps you to understand the flow of the data and knowing the sale prices, etc. The regression task uses several different algorithms to predict the sales prices. It also includes task such as data visualization, cleaning and transformation. Various Algorithms used are: Linear Regression, Multiple Linear Regression, Decision Tree Regression, XG Boost Regression, Random Forest Regression.
In this paper, we have designed a predictive model by as XG Boost technique and experimented on the Big Mart 2013 dataset for predicting sales of the product from the particular outlet.

2. RELATED WORK

There are various regression models are implemented in crime predictions health sectors house predictions, sales prediction etc. In cardiovascular risk prediction based on XG Boost. Sales Forecast is used to predict sales of products getting sold in various stores of Big Mart Company. As the volume of the products increases, growing areas become more and more predictable by hand predicting them become more difficult. Here python is used as a programming language and Jupyter Notebook is used as a tool. To build this app, machine learning features such as supervised learning function, Regression function is used. This is mainly done to predict the future sales of the company's store products.

The various methods used are: Data Processing and Data Processing, Engineering Feature, Model Design, Testing. The regression function uses several algorithms to predict prices. It involves work such as data detection, cleaning and transformation. The profits made by the company are directly proportional to the accurate sales forecasts, Big Marts wants an accurate prediction algorithm so that the company does not lose anything. Experiments support that our strategies produces more accurate and precise forecasts compared to other available methods such as decision trees, community retreats etc. It is found to be 88% accurate.

In the house price prediction using regression techniques, People are cautious when they are trying to buy a new home with their budgets and the market strategies. The purpose of this paper is to predict the corresponding housing prices for the homeless according to their financial conditions and preferences. With the above sales analysis, ride distances and development warnings, estimated prices will be estimated. This paper includes predictions using various Regression techniques such as Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting and Ada Boost Regression. House price forecasts on a set of data are made using all of the methods mentioned above to find the best among them.Some of the cost-related factors were also taken into account such as physical condition, mind and location etc.The multiple linear regression is found to be 91.77% accurate.

In crime prediction using K-Nearest Neighboring Algorithm a developing country like India, it is not uncommon for people to hear about crime on a regular basis. With the rapidity of urbanization, we must always be aware of our surroundings. To avoid misfortune, we will try to monitor crime rates through KNN predictability. It will predict, on average, the type of crime, when, where and when. This data will provide a criminal record in an area that can assist in criminal investigations. It will also provide us with the most committed crimes in a particular region. In this, we will use the nearby k-neighborhood algorithm for machine learning. It is found to be 99.51% accurate.In the paper prediction of the sales value in online shopping using linear regression, the purpose is to analyze the sale of a supermarket, and to predict their future sales by helping them to increase their profits and make their product better and more competitive in terms of market trends by generating customer satisfaction. The method used to predict sales is the Linear Regression Algorithm, which is a popular algorithm in the field of Machine Learning. Sales data from 2011-13 and 2014 data for the year have been made. Subsequently, the 2014 real-time data are also captured and the 2014 real-time of data are compared with predicted data to calculate accuracy of forecast. This is done to ensure that our results are real. This will help them a lot to take the necessary steps to increase their sales. In this paper, we will explain how to deal with such data and predict supermarket sales for years to come from available tools such as machine learning. A brief description of the processes involved in the implementation It is found to be 84% accurate and is comparatively higher as compared to other algorithms. The Prediction of the values are nearly accurate to the original values that the original that would have been got.

3. PROPOSED SYSTEM

After pre-processing and filling the missing values, we used ensemble classifier using Random Forest, Decision trees, Ridge regression, Xgboost and Linear regression. The details of the proposed method are explained in the following section.

Dataset Description of Bigmart Sale

The name of the dataset is "BigMart" Dataset and dataset consists of 12 attributes. Out of these 12 attributes, the response variable is the Item Outlet Sales and remaining are mostly used as the predictor variables. This data-set consists of 8523 products across the different cities. After considering all, a dataset is formed and finally the dataset is divided into two sets, training dataset and testing dataset in the ratio 80: 20.

3.1 Data Exploration

In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. There are 1559 different products and also have 10 unique outlets which are present in the dataset. The Item type contains 16 unique values. Whereas two types of Item Fat Content are there but some of them are misspelled as regular instead of 'Regular' and low fat, LF instead of Low Fat.

3.2 Data Cleaning

It was observed from the previous section that the attributes Outlet_Size and Item_Weight has missing values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight the missing values are replaced by mean of their attribute value. The missing values are numerical where the replacing them by mean and mode reduces the correlation between the imputed attributes. For our model we are assuming that there will be no relationship between the measured attribute and imputed attribute.

3.3 Feature Engineering

Some nuances have been observed in the data-set during the data exploration phase. So this phase is used in resolving all the nuances found in the dataset and make them ready for building appropriate model. During this it was seen that the value of Item visibility had a zero value, so mean value of the item visibility of that product is used in place of zero values attribute. All categorical attributes discrepancies are solved by modifying all categorical attributes into an appropriate one. To avoid this, we create a third category named Item fat content. The Item Identifier attribute has the unique ID starting with either DR, FD or NC. So, we create an attribute named Item Type New having values like Foods, Drinks and Non-consumables. Finally, for determining how old an outlet is, an additional attribute Year is added to the dataset.

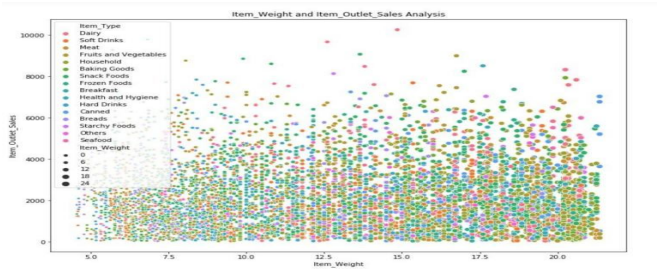

Figure 1: Item weight and item outlet sales analysis

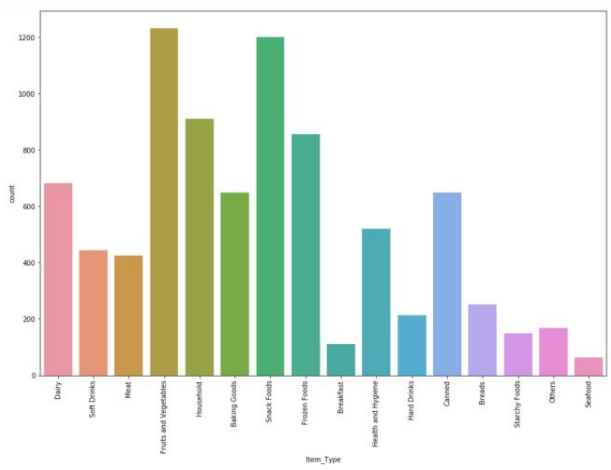

Figure 2: Item Type

3.4 Model Building

Model Building after completing all the previous phases, this dataset is ready to build model. In this paper, we propose a model using the Xgboost algorithm and also we compare it with the other machine learning techniques such as Ridge regression, Linear regression, Decision tree etc.

4. ALGORITHMS USED

4.1 Linear Regression

It is a supervised algorithm in which the predicted value is continuous and also have a constant slope. It is a model that is used for any data that suffers from multicollinearity. It's used to predict the values within a continuous range, rather than classifying them into categories. It is a statistical method that is used for a predictiveanalysis. Linear regression algorithm shows a linear relationship between a dependent variable and an independent (y) variables, hence it is called as linear regression. Since it shows the linear relationship, ie it shows how the dependent variable is changing in accordance to independent variable's value.
The linear regression model provides a straight line representing the relationship between the variables.

Linear Regression can also be represented as:
$y= a0+a1x+ \varepsilon$ Here,

Y= Target Variable
X= predictor Variable
a0= intercept of the line
a1 = coefficient of Linear regression
ε = error
x and y variables are datasets values for Linear Regression model

4.2 Random Forest Algorithm

Random forest is one of the most flexible and easy machine learning algorithms that produces even without any parameter setting a great result most of the time it is used. And it is one of the most widely used algorithms, because it is simple and unique. The bagging method is used to train the forest that is the ensemble of the decision tress. The learning models are combined so that the overall result will be high as compared to others. This method uses several decision tress and then combine them as a single one. Classification and regression problems both use random forest methods.

4.3 Ridge Regression Algorithm

Tikhonov Regularization or ridge regression, is a commonly used regression technique for the approximation of an answer for an equation having any unique solutions. This is very common type of problem in machine learning problems, where the "required" solution must be chosen from limited set of data. Specifically, for A x=b which has no unique solution for x, ridge minimizes
□□A· x−b| | 2+| | Γ· x| | 2 for find a solution, where Γ is the user defined matrix.
The data is divided into three datasets:
1.Training dataset
2. Validation dataset
3. Test dataset
Training dataset:
A training set is used to train the model so that the model can be used to produce an output when certain inputs are provided.
Validation dataset:
A validation set is used to verify the skill of the model and is held back from testing and training.
Test dataset:
Test dataset is used to test the model which the trained using the machine learning algorithms.

4.4 XGBoost Algorithm

XGboost is the one of the most popular and one of the highest accuracy providing machine learning algorithm used in the present day and is implemented regardless on the type of prediction task at hand ie whether it is regression or classification. It is an implementation of decision trees which are gradient boosted which are designed for performance and speed that is competitive machine learning. This algorithm is well known for providing better solutions as compared to other machine learning algorithms. In fact, since it has been developed, it has become the "state-of-the-art" machine learning algorithm to deal well-structured data. It is a library for distributed gradient boosting which is

optimized. It is a software library that we can download and install on from internet and install in our system, then access them through a wide variety of interfaces. Specifically, it supports the following interfaces such as:

Command Line Interface (CLI).
- C++ (the language in which the library is written).
- Python interface as well as a model in scikit-learn.
- R interface as well as a model in the caret package.
- Java and JVM languages like Scala and platforms like Hadoop.

4.5 Decision Tree Algorithm

The decision tree algorithm is a machine learning algorithm with tree structure like the flow chart. The internal node of a decision tree represents features of the model, the branch of the tree represents a decision tree rule, and individual leaf node represents the outcome or label. The top most node of the decision tree is known by the name root node of the tree. It divides the tree into two halves on the basis of an attribute value. The tree is divided in a recursive manner and is known as recursive partitioning. This structure of the tree helps in decision making. The flow chart diagram helps to mimic the human thinking. Because of this reason it is easy to read and intercept. It follows a SOP or Sum of Product representation. It is also known as Disjunctive Normal Form.

4.6 Lasso Regression

It stands for Least Absolute Shrinkage Selection Operator.

Linear regression is a standard regression type, which always assumes that there is a linear relationship between the inputs and the output variables.Lasso Regression is a famous type of linear regression that has an L1 penalty. This shrinks the coefficients for those input variables that doesn't contribute to the prediction task. The L1 penalty allows some of the coefficient values to go to the value of zero, which allows the input variables to be removed effectively from the model, providing an automatic feature selection.

Mathematical equation of Lasso Regression is:

Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients) lasso regression

Where,

λ denotes the amount of shrinkage.

- $\lambda = 0$ implies all features are considered and it is equivalent to the linear regression where only the sum of squares is considered to build a model
- $\lambda = \infty$ implies no feature is considered i.e., infinity it eliminates more and more features
- The bias increases with increase in λ
- variance increases with decrease in λ

Linear regression refers to a model that assumes a linear relationship between input variables and the target variable.

5. RESULTS

5.1 Algorithms Used and Accuracy

ALGORITHMS	ACCURACY
Linear Regression Algorithm	56.0
Ridge Regression Algorithm	46.0
Lasso Regression Algorithm	55.0
Decision Tree Algorithm	62.0
Random Forest Algorithm	61.0
XGBoost Algorithm	88.51

Table 1: Algorithm and Accuracy

6. CONCLUSION

In this project, basics of machine learning and the associated data processing and modelling algorithms are described, and their application in predicting sales of different Big Mart shopping outlets. The implementation, show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales. Multiple instances parameters and various other factors can also be used for predicting the sales more innovatively and successfully. Accuracy plays a major role in prediction systems, can be significantly increased when the parameters used are increased. Also how the sub-models work also can lead to improving the productivity of the system.

As the profit made is directly proportional to the sales predictions made accurately, the Big marts aim accurate predictions so that the company will not suffer any losses. In this paper, we have designed a model by Xgboost technique, linear regression, random forest etc. and experimented it on the Big Mart 2013 dataset for sales prediction of the product of a particular outlet. Experiments support that our technique produce more accurate prediction compared to than other available techniques like decision trees, ridge regression etc.

REFERENCES

1. Crime Prediction using K Nearest Neighbour Algorithm by Akash Kumar, Aniket Verma,Gandhali Shinde,YashSukhdeve,Nidhi Lal published by 2020 International Conference on Emerging Trends in Information Technology and Engineering.

2. Cardiovascular Risk Prediction based on XGBoost by Nitten S, Rajliwall, Rachel Davey, GirijaChetty published by 2018 5th Asia Pacific World Congress on Computer Science and Engineering.

3. House Price Prediction Using Regressio Techniques: A Comparative Study by CH.Raga Madhuri, Anuradha G, M.VaniPujitha published by IEEE 6th International Conference on smart structures and systems 2019.

4. Prediction of Sales Value in online shopping using Linear Regression by Gopalakrishnan T, Ritesh Choudhary Sarada Prasad published by 4th International Conference on Computing Communication and Automation.

5. A comparative study of big mart sales prediction by Gopal Behera, Neeta Nain published by 4th International Conference on Computer Vision and Image Processing at: MNIT Jaipur.

6. Sales prediction using machine learning algorithms by Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, Dr. Nikhilkumar Shardoor published by International Research Journal of Engineering and Technology June 2020.