**BIG MART SALES DATASET**

**ABSTRACT**

Everybody wants to know how to buy goods cheaper or how to advertise them at low cost. Here is the answer. That is Big Mart. Big Mart is online one stop marketplace where you can buy or sell or advertise your merchandise at low cost. The goal is to make Big Mart the shopping paradise for buyers and the marketing solutions for the sellers. The ultimate goal is to prosper with customers. The project “**BIGMART SALES DATASET”** aims to build a predictive model and find out the sales of each product at a particular store. Big Mart will **use this model to understand the properties of products and stores which play a key role in increasing sales.** This can also be done based on the hypothesis that should be done before looking at the data.

**INTRODUCTION**

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.

**MODULES**

**Predictive Modeling:**

In order to find a decent model to predict sales we performed an extensive search of various machine learning models available in R, in particular of those accessible through the caret wrapper. In the end, however, models from the h2o package yielded the best results for the task. In particular, deep learning neural networks h2o.deeplearning and gradient boosting regression trees h2o.gbm performed particularly well. An ensemble of various such models, constructed in h2oEnsemble.R forms the basis of our submission. Here, we used only the 12 most important predictors to avoid over-fitting. To include some features we may have missed with this rather small sub set of predictors we supplemented the ensemble with a deep learning neural net using 23 predictors.

Following algorithms are used:

1. Linear Regression Model
2. Ridge Regression Model
3. Decision Tree Model
4. Random Forest Model

**EXISTING SYSTEM**

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose.

**PROPOSED SYSTEM**

The data scientists at Big Mart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, Big Mart will try to understand the properties of products and stores which play a key role in increasing sales.

**Advantages:**

* This is an easily scalable model to provide detailed information and accurate predictions for sales volume for different types of products as there is a lot of data out there.
* It is the percentage of display space in a store given to that particular item. Looking at the average visibility of items given in each store type and outlet.

**Goals:**

* Replacing the Nans, identifying outliers, feature selection and normalization – for both training and testing data.
* Building the regression models: linear and decision tree. Predicting the sales, cross validating the scores, calculating the R^2.
* Classifying the training data with a decision tree and a random forest and calculating the accuracy score and the R^2.

**CONCLUSION**

The ML algorithm that perform the best was XGBoost with RMSE = 1041 which got me in the first 25%. The next step will be looking at Hyperparameter Tuning and Ensembling.

Hence, we propose a software tool for forecasting future sales volume based on the historical sales data. Using this tool, the accuracy of prediction for multiple linear regressions and random forests can be determined.

**BIBLIOGRAPHY**

[1] H. M. Al-Hamadi “Long-Term Electric Power Load Forecasting Using Fuzzy Linear Regression Technique” ,IEEE Mar.2011

[2] Yanming Yang “Prediction and Analysis of Aero-Material Consumption Based on Multivariate Linear Regression Model” , 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis