International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

A TWO-LEVEL STATISTICAL MODEL FOR BIG MART SALES PREDICTION

DR. M. CHAITANYA KISHORE REDDY[1], T. P. V. S. SANJAY[2], R. GEETHA MEGHANA[3], T. LAKSHMI BHAVANI[4]

- 1 PROFESSOR Dept of IT, NRI Institute of technology, A.P, India - 521212
- 2,3,4 UG Scholar, Dept of IT, NRI Institute of Technology, A.P, India - 521212

Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Big Mart, a one-stop-shopping centre, has been discussed to predict the sales of different types of items and for understanding the effects of different factors on the items' sales. Taking various aspects of a dataset collected for Big Mart, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to take decisions to improve sales.

Keywords: Machine Learning, Sales Prediction, Big Mart, Random Forest, Linear Regression

Introduction

In today's modern world, huge shopping centres such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management. The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task's data. This can then further be used for forecasting future sales by means of employing machine learning algorithms such as the random forests and simple or multiple linear regression model.

Machine Learning

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analysed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life [1]. As the technology

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects. In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results [2]. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key [11].

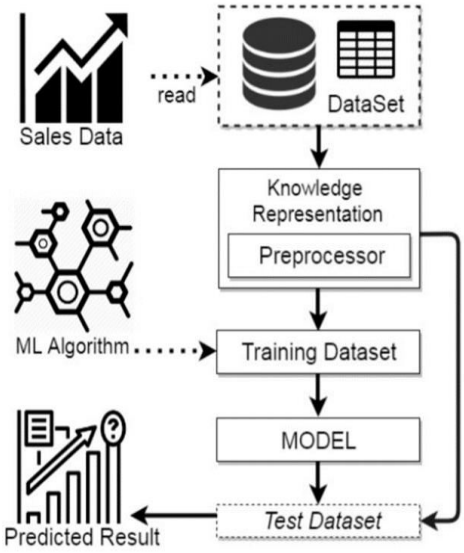
ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way. This is followed by explaining the methodology ensued and the prediction results observed on implementation. Various machine learning algorithms include [3]:

Linear Regression: It can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on basis of data set.

K-Nearest Neighbours (KNN): It is a learning algorithm which is based on instances and knowledge gained through them [4]. Unlike mining in data stream scenarios, cases where every sample can simultaneously belong to multiple classes in hierarchical multi-label classification problems, k-NN is being proposed to be applied to predict outputs in structured form [5].

Decision tree: It is an intuitive model having low bias and it can be adopted to build a classification tree with root node being the first to be taken into account in a top-down manner. It is a classic model for machine learning [6].

Architecture:



Implementation

1. Training and Testing Head
2. Removing Unwanted Columns & Features
3. Complete Description of Dataset
4. Making Correction in Item Fat Content Column
5. Dimensionality Reduction
6. Multi-Linear Regression
7. Random Forest Model
8. Polynomial Regression
9. Let us make Prediction on Test Set

Predictive Modelling:

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

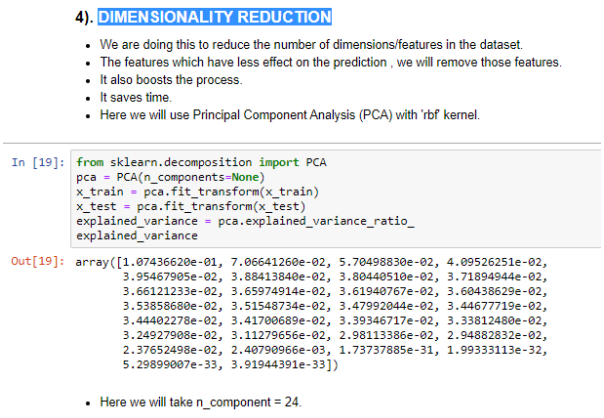
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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 inh2oEnsemble.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

Results



4). DIMENSIONALITY REDUCTION

- We are doing this to reduce the number of dimensions/features in the dataset.
- The features which have less effect on the prediction , we will remove those features.
- It also boosts the process.
- It saves time.
- Here we will use Principal Component Analysis (PCA) with 'rbf' kernel.

```
In [19]: from sklearn.decomposition import PCA
         pca = PCA(n_components=None)
         x_train = pca.fit_transform(x_train)
         x_test = pca.fit_transform(x_test)
         explained_variance = pca.explained_variance_ratio_
         explained_variance

Out[19]: array([1.07436620e-01, 7.06641260e-02, 5.70498830e-02, 4.09526251e-02,
                3.95467905e-02, 3.88413840e-02, 3.80440510e-02, 3.71894944e-02,
                3.66121233e-02, 3.65974914e-02, 3.61940767e-02, 3.60438629e-02,
                3.53858680e-02, 3.51548734e-02, 3.47992044e-02, 3.44677719e-02,
                3.44402278e-02, 3.41700689e-02, 3.39346717e-02, 3.33812480e-02,
                3.24927908e-02, 3.11279656e-02, 2.98113386e-02, 2.94882832e-02,
                2.37652498e-02, 2.40790966e-03, 1.73737885e-31, 1.99333113e-32,
                5.29899007e-33, 3.91944391e-33])
```

- Here we will take n_component = 24.

Machine learning and theassociated dataprocessing and modelling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centres at different locations. On implementation, the prediction results 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 factors can be used to make this sales prediction more innovative and successful. Accuracy, which plays a key role in prediction-based systems, can be significantly increased as the number of parameters used are increased. Also, a look into how the sub-models work can lead to increase in productivity of system. The project can be further collaborated in a web-based application or in any device supported with an in-built intelligence by virtue of Internet of Things (IoT), to be more feasible for use. Various stakeholders concerned with sales information can also provide more inputs to help in hypothesis generation and more instances can be taken into consideration such that more precise results that are closer to real world situations are generated. When combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overall development of corporation's tasks on the whole. One of the main highlights is more expressive regression outputs, which are more understandable bounded with some of accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stage of regression model-building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

Conclusion

In this paper, basics of machine learning and the associated data processing and modelling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centres at different locations. On implementation, the prediction results show the correlation among different

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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 factors can be used to make this sales prediction more innovative and successful. Accuracy, which plays a key role in prediction-based systems, can be significantly increased as the number of parameters used are increased. Also, a look into how the sub-models work can lead to increase in productivity of system. The project can be further collaborated in a web-based application or in any device supported with an in-built intelligence by virtue of Internet of Things (IoT), to be more feasible for use. Various stakeholders concerned with sales information can also provide more inputs to help in hypothesis generation and more instances can be taken into consideration such that more precise results that are closer to real world situations are generated. When combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overall development of corporation's tasks on the whole. One of the main highlights is more expressive regression outputs, which are more understandable bounded with some of accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stage of regression model building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

References

[1] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University, UK, 32, 34.

[2] Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publications.

[3] Shashua, A. (2009). Introduction to machine learning: Class notes 67577. arXiv preprint arxiv:0904.3664.

[4] MacKay, D. J., & Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

[5] Daumé III, H. (2012). A course in machine learning. Publisher, ciml. info, 5, 69.

[6] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[7] Cerrada, M., & Aguilar, J. (2008). Reinforcement learning in system identification. In Reinforcement Learning. IntechOpen.

[8] Welling, M. (2011). A first encounter with Machine Learning. Irvine, CA.: University of California, 12.

[9] Learning, M. (1994). Neural and Statistical Classification. Editors D. Mitchie et. al, 350.

[10] Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30-36.

[11] Downey, A. B. (2011). Think stats. " O'Reilly Media, Inc.".

[12] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.