Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

Research Paper

An Ensemble Based Predictive Modeling in Forecasting Sales of Big Mart

Statistic

KEYWORDS : Statistical models, Machine Learning, Ensemble Learning.

T. Leo Alexander — Associate Professor, Department of Statistics Loyola College Chennai – 34

D. Delwin Christopher — Research Scholar, Department of Statistics Loyola College Chennai – 34

ABSTRACT

In this paper we present the analysis of various products of BigMart presented by their unique identification codes with a view of developing a predictive model for the sales of each of those products. The analysis is performed on a data set for which the past sales is provided and the prediction is made on another data set. The data has been obtained from ANALYTICS VIDHYA (http://www.analyticsvidhya.com/).

We have used some basic statistical predictive models like general linear model, principle component analysis based model and other machine learning techniques like random forest, support vector machine and neural network to develop the ensemble based predictive model.

1. INTRODUCTION

The statistical learning of the ensemble modeling to predict the sales of BigMart can be beneficially adopted for the wholesale and retail vendor joints in India. It helps in understanding the factors that influence the sales of similar products in a better manner. The machine learning techniques adopted aims at reducing the variability to the maximum in predicting the sales.

In every technique that has been adopted, all the related issues has been addressed. Accuracy of the prediction is determined by fitting the development model on the valuation data.

OBJECTIVE

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.

Please note that the data have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly.

2. BIGMART

Big Mart is online one stop marketplace where we can buy or sell or advertise our merchandise at low cost. The goal is to make Big Mart the shopping paradise for buyers and the marketing solution for sellers. The data is extracted from **ANALYTICS VIDHYA** (http://www.analyticsvidhya.com/). It is a website where information on big data and machine learning is shared. Online competitions are also conducted and the winners are being rewarded for good results. The data set is related with retail domain and challenge is to predict sales of product across multiple stores. We have train (8523) and test (5681) data set, train data set has both input and output variable(s). We need to predict the sales for test data set.

2.1 EVALUATION METRIC

The model performance is evaluated on the basis of the prediction of the sales for the test data (test.csv), which contains similar data-points as train except for the sales to be predicted.

The experts at Analytics Vidhya, have the actual sales for the test dataset, against which the predictions will be evaluated. The metric being used is the Root Mean Square Error (RMSE) value to judge our response.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\Pr edicted_i - Actual_i)^2}{N}}$$

where N : Total number of observations, Predicted : The predicted values of sales, Actual : The actual values of sales.

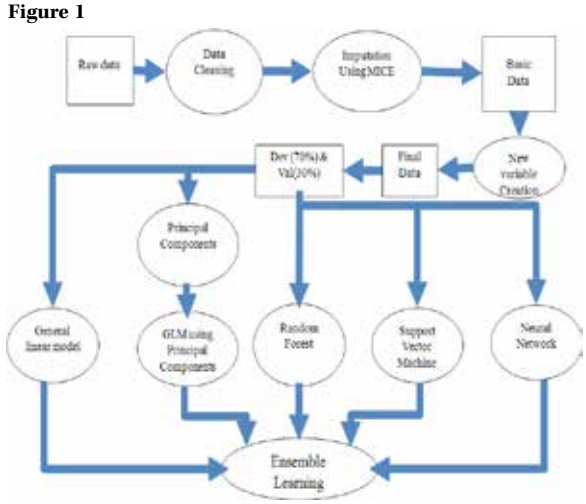
3. ANALYTICAL PROCEDURE

The analytical procedures involves building predictive models by all mentioned techniques, comparing them and arriving at the best predictive models for the ensemble learning.

3.1 PROCEDURAL BREAKDOWN

The flowchart below gives an outline of the analytical procedures that are done for the fulfillment of the study.

Figure 1



3.2 MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS

Multiple imputation is the method of choice for complex incomplete data problems. Missing data that occur in more than one variable presents a special challenge *(Karin Groothuis-Oudshroon 2011)*. Two general approaches for imputing multivariate data have emerged: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). ***Imputation*** of missing data is done using ***MICE.***

3.3 SEGMENTATION

In order to build a regression model, ***segmentation*** of the vari-

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

ables is been done. The presence of categorical variables with too many categories makes the iterative procedure complex and merely impossible in building a linear model. Thus segmentation of categories for the categorical variables is essential. The procedure is such that the categories are listed in the decreasing order and then the segmentation is done in an appropriate business way such that the number of categories is under a tolerable limit.

3.4 LINEAR REGRESSION MODELS

A linear regression model assumes that the regression function $E(Y|X)$ is linear in the inputs $X_1,...,X_p$. Linear models were largely developed in the pre-computer age of statistics, but even in today's computer era there are still good reasons to study and use them. They are simple and often provide an adequate and interpretable description of how the inputs affect the output. For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data. Finally, linear methods can be applied to transformations of the inputs and this considerably expands their scope.

Since the continuous variables are not normally distributed, the regression model is built with transformed variables. It is made obvious by plotting the residuals against the variables. From the model summary it is observed that only the variables Item MRP, Outlet Identifier, Outlet establishment year, Outlet size, Outlet location type and Outlet type are significant at 5% level of significance.

In the Table 1, testing for Multicollinearity, we observe that none of the variables suffers from multicollinearity since all the VIF values are lesser than 5.

Table 1

	Collinearity Statistics	
	Tolerance	VIF
(Constant)		
Item weight	0.977	1.023
Item fat content	0.97	1.031
Item visibility	0.914	1.094
Item MRP	0.994	1.006
Outlet identifier	0.231	4.323
Outlet establishment year	0.237	4.214
OS1	0.431	2.32
OS2	0.372	2.686
OLT1	0.312	3.206
OLT2	0.338	2.96
Outlet type	0.225	4.447
Item type 1	0.797	1.255
Item type 2	0.802	1.247

Remark 1: Since most of the variables do not follow normality, we intend to generate the principal components as an alternative procedure to obtain predictive model.

PRINCIPAL COMPONENT ANALYSIS

Principal Component analysis is done and the first 2 principal components which contain the maximum amount of information is identified. The principal components of a set of data in $\mathbf{R}^p$ provide a sequence of best linear approximations to that data, of all ranks $q \leq p$.

By observing the scree plot we observe that only the first two out of the 30 generated principal components contain around 98% of the total information.

Using the principal components a new regression model is built considering all the 30 components that were generated and the RMSE is recorded.

Remark 2: To obtain a much more precise predictive model we try to build a predictive model using the machine learning technique random forest.

RANDOM FOREST

Two well-known methods are boosting and bagging of classification trees. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees - each is independently constructed. (*Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani;* 2015)

Using the Random Forest, prediction of the sales is made. Care is taken in fixing the optimum number of trees .

FIXING THE OPTIMUM NUMBER OF TREES

From table 4, we observe that the random forest with ntree=1000 and mtry=5 is optimum.

Table 2

METHOD	RMSE for val
RF with ntree=100 mtry=3	1130.906
RF with ntree=500 mtry=3	1127.151
RF with ntree=500 mtry=5	1124.319
RF with ntree=1000 mtry=5	1123.194
RF with ntree=2000 mtry=5	1123.387

IMPORTANCE OF VARIABLES

The percentage increase in MSE helps to identify the sparsely distributed variables like item MRP, outlet establishment year and outlet type from the Table 3.

Table 3

Variables	%IncMSE
Item_Weight	19.2885011
Item_Fat_Content	5.2037534
Item_Visibility	12.2855373
Item_Type1	-0.2143772
Item_Type2	3.3053666
Item_MRP	369.6293884
Outlet_Identifier	31.9631645
Outlet_Establishment_Year	93.5076146
OS1	22.3598953
OS2	46.0831689
OLT1	39.2464472
OLT2	17.5602835
Outlet_Type	51.5608179

Remark 3: To obtain a much more precise predictive model we

try to build a predictive model using the machine learning technique support vector machine.

3.7 SUPPORT VECTOR MACHINE

We describe generalizations of linear decision boundaries for regression. Optimal separating hyper-planes are introduced for the case when two classes are linearly separable. Here the extensions to the non separable case, where the classes overlap. These techniques are then generalized to what is known as the support vector machine, which produces nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space. The second set of methods generalize Fisher's linear discriminant analysis (LDA). *(Reference David L.Olson, Dursun Delen;* 2008). We build predictive models using support vector machine and record the RMSE.

Remark 4: To obtain a much more precise predictive model we try to build a predictive model using the machine learning technique neural network.

3.8 NEURAL NETWORK

We describe a class of learning methods that was developed separately in different fields-statistics and artificial intelligence-based on essentially identical models. The central idea is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. The result is a powerful learning method, with widespread applications in many fields. *(Ian H.Witten, Eibe Franc;* 2005)

Thus we build predictive models using neural network and record the RMSE.

Remark 5: To obtain a more precise model model, we concentrate in building an ensemble based model by merging the best of models.

3.9 ENSEMBLE LEARNING

The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models. Linear model, random forest, support vector machine and neural network are ensemble methods for regression, where a *committee* of trees each cast a vote for the predicted values.

3.9.1 COMPARISON OF PREDICTIVE MODELS BY RMSE

Table 4

METHOD	RMSE for val
GLM with power 6	1324.356
GLM with PCA	1229.603
RF with ntree=1000 mtry=5	1123.194
SVM	1200.793
ANN with size = 10	2776.58

COMPARISON OF PREDICTIVE MODELS BY CORRELATION

Table 5
Correlation

	Linear model with power 6	Random Forest	Linear model with PCA	Support Vector Machine	Neural Network
Linear model with power 6	1	0.817	0.124	0.88	0.8
Random Forest		1	0.928	0.084	0.981
Linear model with PCA			1	0.087	0.083
Support Vector Machine				1	0.912
Neural Network					1

By observing both the Table 4 and Table 5, we notice that the RMSE for general linear model using PCR and prediction using random forest techniques are comparatively less. The correlation between the same is also 0.928, showing high correlation. The p-value is also 0<0.05, showing that there is a significant correlation between the two. Hence we concentrate in building a new model having the predicted values of the two corresponding methods as the independent variables on the valuation data set and observe to obtain a much convincing RMSE. The established models are used to predict for the actual 'test' data set. The ensemble technique is then repeated for the new data again supposing that the RMSE value will be comparatively in this case too as in the validation data set case.

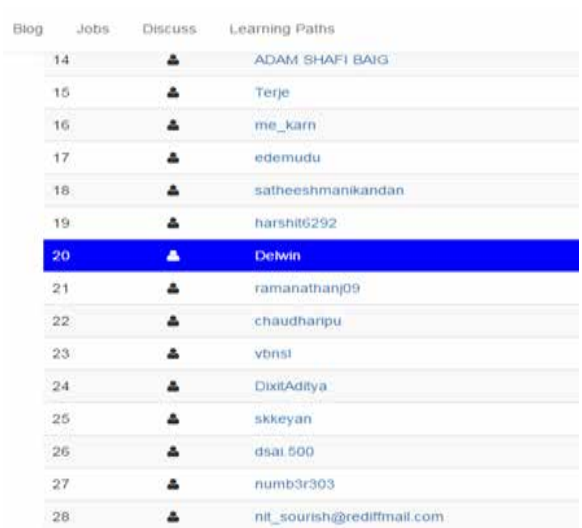
4. CONCLUSION

The technique of 'Multivariate Imputation by Chained Equations' is understood. Descriptive statistics is studied using pictorial representation. Building of general linear model and it is studied giving importance to multi collinearity issues. Building a linear model using principal component analysis is done. Data mining technique random forest is used as a method to predict the sales, including the decision on fixing the optimal number of trees.

Other machine learning techniques including support vector machine and neural network methods were also studied and used to predict the sales. It was found that the general linear model using the principal component analysis and the random forest techniques produce better results which is been decided by the RMSE values. Ensemble learning is presumed with a view of enhancing the RMSE value for the final test data set. Minor data analysis techniques such as dividing the data as for development and validation is done. The final ranking after uploading the predicted values for the train data set fetched 20th rank where more than 1000 contestants had participated. The final RMSE value achieved is 1171.800

Remark 6: The final ranking after uploading the predicted values for the train data set fetched 20th rank for the co-author Delwin Christopher where more than 1000 contestants had participated as shown in figure 5.

Figure 2
Achieved rank



Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

REFERENCE

1. *David L.Olson, Dursun Delen;* (2008). Advanced Data Mining Techniques; Springer series, New York
2. *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani;* (2015). An Introduction to Statistical Learning; Springer series, New York
3. *Ian H.Witten, Eibe Franc;* (2005). Data Mining; Morgan Kaufmann Publishers, San Francisco
4. *Karin Groothuis-Oudshroon* (2011). Journal of statistical Software;
5. *Trevor Hastie, Robert Tibshirani, Jerome Friedman*; (2008). The Elements of Statistical Learning; Springer series, New York