School of Computer Science and Engineering
VIT Chennai
Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: COMPUTER SCIENCE AND ENGINEERING (SCOPE)

Course: CSE2004 (Database management system)

Slot: D2 SLOT

Faculty: Dr. M. Premalatha

Component: J component

Title:

Big-Mart-Sales-Prediction

Team Member(s): [Min 1, Max-3 in a team]

- Mallapalli Kusuma Sai – 20BCE1553
- Penta Revanth- 20BPS1161
- Mandla Sheshi Kiran Reddy-20BAI1061

Abstract

The word forecasting means the process of making predictions based on past and present data and most commonly by analysis of trends which is important in many fields to prevent loss or damage. One of the domain where we have use this forecasting is the sales. In the technical world today, the companies which are involved in marketing, wholesale, retail and the logistics give importance to the sales forecasting. It allows companies to work efficiently and allocate their capital and resources as per the proper prediction and to evaluate achievable sales revenue in a better manner and to have a strategy for the future of the company. The data is mined for the future patterns of the data. Data exploration, Data cleaning and the feature engineering plays and important role.The feature engineering which is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms and this feature engineering is very important in the prediction of sales. The tree-based algorithms fall under the feature engineering which makes the work easy for us.

Keywords: **Data mining, Feature engineering, Data exploration, Data cleaning, Tree based algorithms, Sales prediction.**

Introduction

With the lively and the fast development of the companies which are involved in the supermarkets or store chains and the increase in the customers mostly paying through the medium of electronic payments there is a serious competition increasing among the companies which are actually leading to the invention of new methods in data mining/Machine learning. Every company is trying to make the customer more comfortable and also more satisfied with their offers and the personalised recommendations based on the prediction Our main objective is to find out the properties of the product and to store the product which impacts the sale**.** In this project, we are having the forecast for the sales dataset of big mart in a number of big mart stores across various location types of the year 2013. According to the characteristics of the data, we can use the method of multiple linear regression analysis decision tree algorithm and the random forest to forecast the sales of the company. Every different type of algorithm at ground level operates differently and makes different predictions. All aspects of the dataset are analysed and then the final prediction is made

1. Data Set Description:

ANSWER:

Big Mart is an International Retail Corporation. Big Mart Sales data of the year 2013 has been used as the dataset for the proposed work. Big Mart sales data has 12 attributes and rest of the attributes play the role of predictor variables.

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

The dataset has 8523 observations. We have train has 8523 rows and 13 columns depicted as (8523,13) and test dataset has has 5681 rows and 12 columns depicted as (5681,12). The train data set has both input and output variables. We need to predict the sales for test data set. The pre-processing steps involved are 1) Acquire the dataset.

2) Import all the crucial libraries.

3) Import the dataset.

4) Data Cleaning

5) Filling the missing values

6) Numerical and One-Hot Coding of Categorical variables

7) Exporting Data

8) Model Building

URL:(Link For the Dataset)

Test dataset: (Github) (Text format)

- https://github.com/shrikant-temburwar/Big-Mart-Sales-Prediction/blob/master/test.csv
- https://github.com/shrikant-temburwar/Big-Mart-Sales-Prediction/blob/master/test_modified.csv **(modified test datset)**

Train dataset:(Github) (Text format)

- https://github.com/shrikant-temburwar/Big-Mart-Sales-Prediction/blob/master/train.csv
- https://github.com/shrikant-temburwar/Big-Mart-Sales-Prediction/blob/master/test_modified.csv **(modified train dataset)**

Database (Kaggle) (Tables and relations format for same database in the github)

- https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data

2. Consider the schema alone and normalize it till BCBF using schema decomposition

ANSWER:

- In our code we, will try to remove the insertion , update , anamaly using the data pre-processing , feature engineering and feature transformation.
- Now we will take the schema and will take a small part from our test data and will noramlize it.

ORIGINAL SCHEMA:

Item_ Identifier	Item_ Weight	Item_ Fat_Content	Item_ Visibility	Item_ Type	Item_ MRP	Outlet_ Identifier	Outlet_ Establishment_ Year	Outlet_ Size	Outlet_ Location_ Type	Outlet_ Type

Minimal Cover

Item_Identifier → Item_Weight
Item_Identifier → Item_Fat_Content
Item_Identifier → Item_Type
Item_Identifier → Item_Visibility , Item_MRP
Item_Identifier → Item_Visibility , Outlet_Identifier
Outlet_Identifier → Outlet_Establishment_Year
Outlet_Establishment_Year → Outlet_Type

Candidate Keys

{Item_Identifier, Item_Visibility ,Outlet_Size,Outlet_Location_Type}

Normalize to 2NF

Attributes

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Type

Functional Dependencies

Item_Identifier ➔ Item_Weight, Item_Fat_Content , Item_Type

Attributes

Item_Identifier	Item_Visibility	Outlet_Identifier	Item_MRP	Outlet_Establishment_Year	Outlet_Type

Functional Dependencies

Outlet_Identifier ➔ *Outlet_Establishment_Year*

Item_Identifier ➔ *Item_Visibility ,Outlet_Identifier Item_MRP*

Outlet_Establishment_Year ➔ *Outlet_Type*

Attributes

Item_Identifier	Item_Visibility	Outlet_Size	Outlet_Location_Type

Functional Dependencies

No functional dependancies

Normalize to 3NF & BCNF:

Attributes

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Type

Functional Dependencies

Item_Identifier ➡ Item_Weight, Item_Fat_Content , Item_Type

Attributes

Item_Identifier	Item_Visibility	Outlet_Identifier	Item_MRP

Functional Dependencies

Item_Identifier ➡ *Item_Visibility*, Outlet_Identifier , Item_MRP

Attributes

Outlet_Identifier	*Outlet_establishment_year*

Functional Dependencies

Outlet_Identifier ➡ Outlet_establishment_year

Attributes

Outlet_Identifier	*Outlet_Type*

Functional Dependencies

Outlet_Identifier ➡ Outlet_Type

Attributes

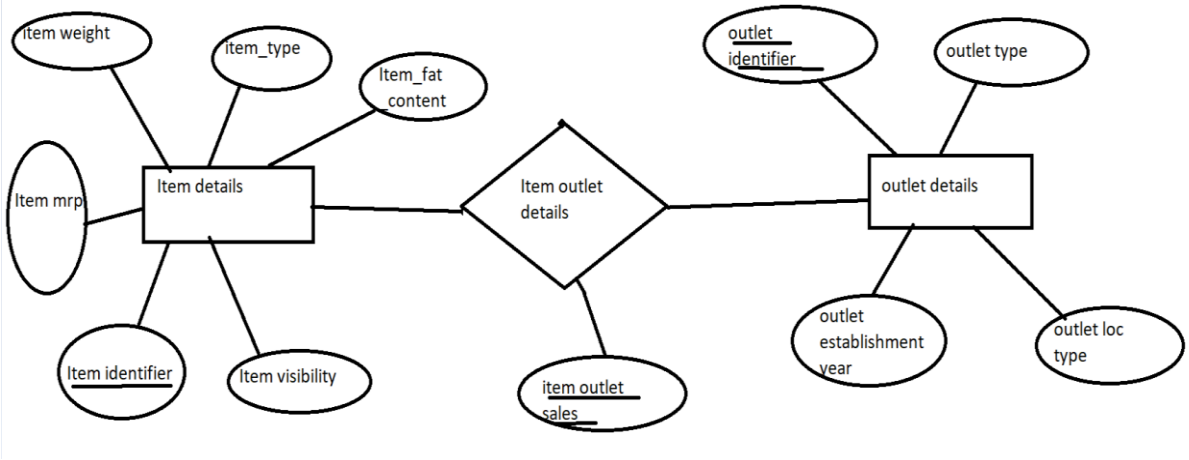
Item_Identifier	Item_Visibility	Outlet_Size	Outlet_Location_Type

Functional Dependencies

No functional dependancies

3. Draw the ER diagram for the final decomposed schema stating the key attributes, mapping cardinalities, participation constraint, and so on

ANSWER:



4. **Methodology and Algorithm used:**

ANSWER:

The method we are going to use is the **"Prediction Methodology".** Predictive analytics is the area of data science focused on interpreting existing data in order to make informed predictions about future events. It includes a variety of statistics techniques .In the data mining case prediction is nothing but looking for patterns and relationships in large stores of data. Using it allows executives to predict upcoming challenges, identify opportunities for growth, and optimize their internal operations. There isn't a single way to do predictive analytics, though; depending on the goal, different methods provide the best results. Since in the sales analysis our objective is to find the properties of the product and to store and analyse them, the prediction methodology will do the best business while implementing the algorithm. We also use this method in the sales called the Cross-validation which refers to a set of methods for measuring the performance of a given predictive model on new test data sets. The training set, used to train (i.e. build) the model; and the testing set (or validation set), used to test (i.e. validate) the model by estimating the prediction error.
The two main algorithms used are the

1) **RANDOM FOREST MODEL:** We will be going to use this algorithm because there are very few constraints in the random forest algorithm thus very easy to use it and we will not be introducing the biases from our end and because it is
simple ,it is highly interpretable. It is the collection of decision trees which helps to give correct output by making use of bagging mechanism. Random Forest Algorithm is used to incorporate predictions from

multiple decision trees into a single model. This algorithm uses bagging mechanism to create a forest of decision trees. It the incorporates the predictions from multiple decision trees to give very accurate predictions.

2) **DECISION TREE MODEL:** We will also use the decision tree algorithm Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. he goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). The sales need the use of training set and the predicition as mentioned above in such case the decision tree model will do its best in the prediction domain of problemes.
In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.Using this model Bigmart will understand the properties of products and their impacts on sales. We will also try to execute the **Linear regression model** algorithm in the level of execution and will find out the best working model for the **BIGMART SALES**.

5. Implementation

ANSWER:

- We will be using the required packages and will be implementing the code like the panda , numpy , snsplot , matplotlib etc.
- First we will be doing the analysis of the train data set separately which helps in the prediction.
- We will be calculating the skewness, correlation and will be drawing the bar-graphs , histograms for comparing the relation between the attributes in the data analysis part
- We will be checking the datatypes of the attributes also for the data analysis
- Next step is the data pre-processing . Here we will be checking for the missing values and null values.
- So we will be concatenating the train and the test dataset into the "data" dataset and will be filling the values with the mean or mode according to the nature of the attributes which we studied in the data analysis part.

- Next we will be doing the feature engineering and feature transformations steps to divide the attributes into categories and sub divisions and decomposition of schema to clean the data .
- We will be exporting the cleaned and analysed data into the train_modified and test_modified datasets for implementing the algorithms.
- Decision tree algorithm:

  By using the modified datasets we will be implementing the decision tree algorithm.

  We will use the decisiontreeregressor package from the sklearn.tree and will predict the results of the test dataset in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes and will return the array of prediction values.

  We will calculate the root mean squared error and the regressor score and will be importing the resultant dataset into our system.
- Linear regression:

  By using the modified datasets we will be implementing the linear regression algorithm. We will import the linearregression package from the sklearn.linear and will predict the results of test dataset by finding linear relationship between target and one or more predictors. We will calculate the root mean squared error and the regressor score and will be importing the resultant dataset into our system.
- Random forest:

  By using the modified datasets we will be implementing the random forest algorithm. We will import the randomforest package from the sklearn.ensemble and will predict the results of test dataset.It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. We will calculate the root mean squared error and the regressor score and will be importing the resultant dataset into our system.
- In this way we will implement the algorithm and predict the results.

6. Results and Discussion

ANSWER:

- Where in the cross-validation stage the dataset is divided randomly into 20 subsets with roughly equal sizes. Out of the 20 subsets, 19 subsets are used as training data and the remaining subset forms the test data also called leave-one-out cross validation.
- Every models is first trained by using the training data and then used to predict accuracy by using test data and this continues until each subset is tested once.
- From data visualization, it was observed that lowest sales were produced in smallest locations. However, in some cases it was found that medium size location produced highest sales though it was type-3 (there are three type of super market e.g. super market type-1, type-2 and type-3) super market instead of largest size location to increase the product sales of Big mart in a particular outlet, more locations should be switched to Type 3 Super-markets.

Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general-purpose error metric for numerical predictions.

RMSE is a good measure of accuracy, but only to compare prediction errors of different models or model configurations for a particular variable and not between variables, as it is scale-dependent.

After the implementation the results are:

ALGORITHM	RMSE (root mean squared error)	OUTLET_SALES_PREDICTIONS
Decision-tree algorithm:	1095	`array([1556.65096825, 1349.5129 0433, 471.30684669, ...])`
Linear regression:	1127	`array([ [1852.99962636, 1579.6253 0278, 1888.44660091, ..., ]`
Random forest:	1061	`array([1659.36233923, 1359.1374 969 , 587.17212218, ...,])`

7. Conclusion:

ANSWER:

- RMSE is used to measure the performance of two different types of supervised problems

- The RMSE value is low for the random forest so the random forest method can be applied for this prediction and analysis.
- Low RMSE value shows that there are less errors around the regression line.
- As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring more accurate prediction algorithm so that the company will not suffer any losses
- **So we will be choosing the Random forest algorithm among the Decision tree, linear regression algorithms.**

THE END