**CPSC 531**

Advance Database Systems

**Store Sales Estimation**

****

Submitted By:

**Ajinkya Bijwe**

**CWID: 802332916**

**California State University, Fullerton**

**ajinkyabijwe@csu.fullerton.edu**

Under the Guidance of

**Kyong-Seop Shin**

**Department of Computer Science**

**California State University, Fullerton**

**Table of Contents**

|  |  |
| --- | --- |
| Topics | Page Number |
| 1. Introduction | 1 |
| 1. Digital Forensics History | 3 |
| 1. Computer Forensics Policies for Privacy Protection | 5 |
| 1. Digital Forensics Data Encryption 2. Tools for data Encryption 3. Problems Related to encryption | 8  9  9 |
| 1. Digital Privacy Current Practices in Digital Forensics | 10 |
| 1. Conclusion | 11 |
| References | 12 |

**Abstract**

Knowledge Discovery in Database (KDD) is process to retrieve the useful data from the collection of datasets. This technique includes the data selection, data cleansing and data preparation. The application of the KDD include the fraud detection, marketing and telecommunication. So we used this mining techniques in our project to retrieve and forecast the data. The store sales estimation is the way to predict the amount of sales on the certain date based on the certain constraints such as temperature, holidays and historical data of the previous years about the sales. The sales forecasting help us to predict the future steps to be taken to maximize the profit. This ultimately leads to better decisions in the business. This project is about writing the logic to forecast the sales on the holidays on the future dates where external datasets are provided by stores. This logic is used to estimate the data of the store and its departments regarding the sales. This logic uses various algorithms of data mining to calculate the final result in an external portable human readable file.

1. **Introduction**

The Knowledge discovery is the technique of the data mining in which we discover the knowledge from the collection of the data. In the past the knowledge discovery was performed manually. But today as we are generating terabytes of data each day it is now impossible to perform the knowledge discovery manually. Moreover, for thriving any business and finding the pattern the data is very essential. So several software logic and tools are developed to make the suitable assumptions and discover the invisible data. The KDD has reached its optimum level in the past decade. It now contains many techniques which comprises of knowledge acquisition, Bayesian statistics and query optimization. The primary objective of the KDD is to retrieve the high level knowledge from the low-level data. (Technopedia, 2013)

The Store Sales Estimation is the study and assessment of the sales data during the holiday season in the respective year. From the archival data of past years we can determine and predict the sales of the forthcoming years.

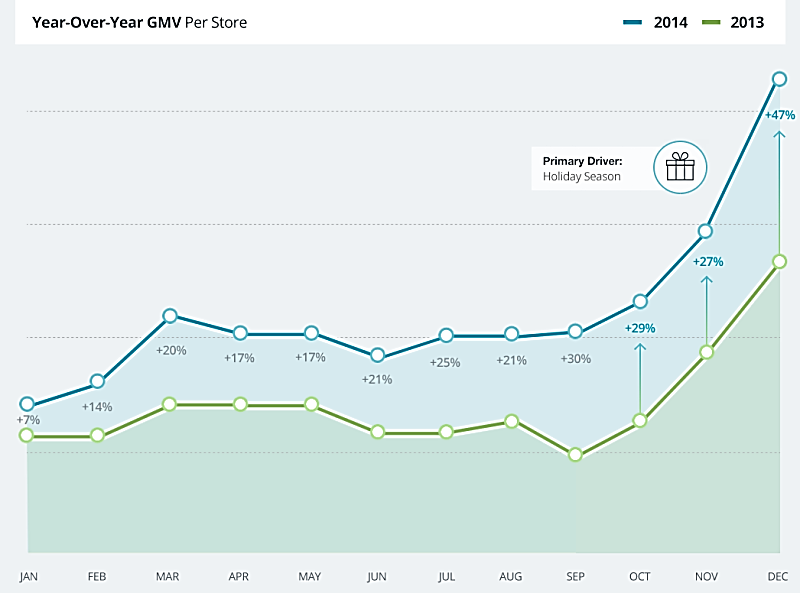


Figure 1 Increase in Store Sales During Holidays For Year 2013-2014 (Marketing Land, 2015)

In the above Figure 1 we can see there is the substantial increase in the amount of sales during the months of October, November and December for the year 2014 as compared to sales of 2013. The amount of sale increased is maximum for the December as compared to previous sales for the December of the year 2013. From the above figure we can also determine the sales increases at the end of the year 2013 and 2014. From this statistics we can determine that there is the need for the appropriate and useful model to assess the amount of sale to be happen for the current year so as to maximize the outcome from the business. So we created a logic complete based on the proximity algorithms model to envision the sales for the store with its departments.

1. **Problem Statement**

As we all know the sales of the stores depend upon the prices of the commodities in that store and also on the holiday seasons. If the store do not proper manage their sales it can result in heavy losses. So there is the need of the proper model to estimate the future sales in that store. This can be done using retrieving the useful historical data of the past years related to sales and different elements such as prices, holidays etc. So we created a model which can analyze and retrieve data to provide the suitable output for the user. This model uses Random Forest algorithm to show the final data. In this project we have taken into consideration the holiday, sales of the previous year, date, store id and department id of that store to calculate the final outcome. The most important constraints in this project are previous year’s sales data and the holiday markdown sales. The final result will show the future sales and holiday sales.

1. **Literature Review**

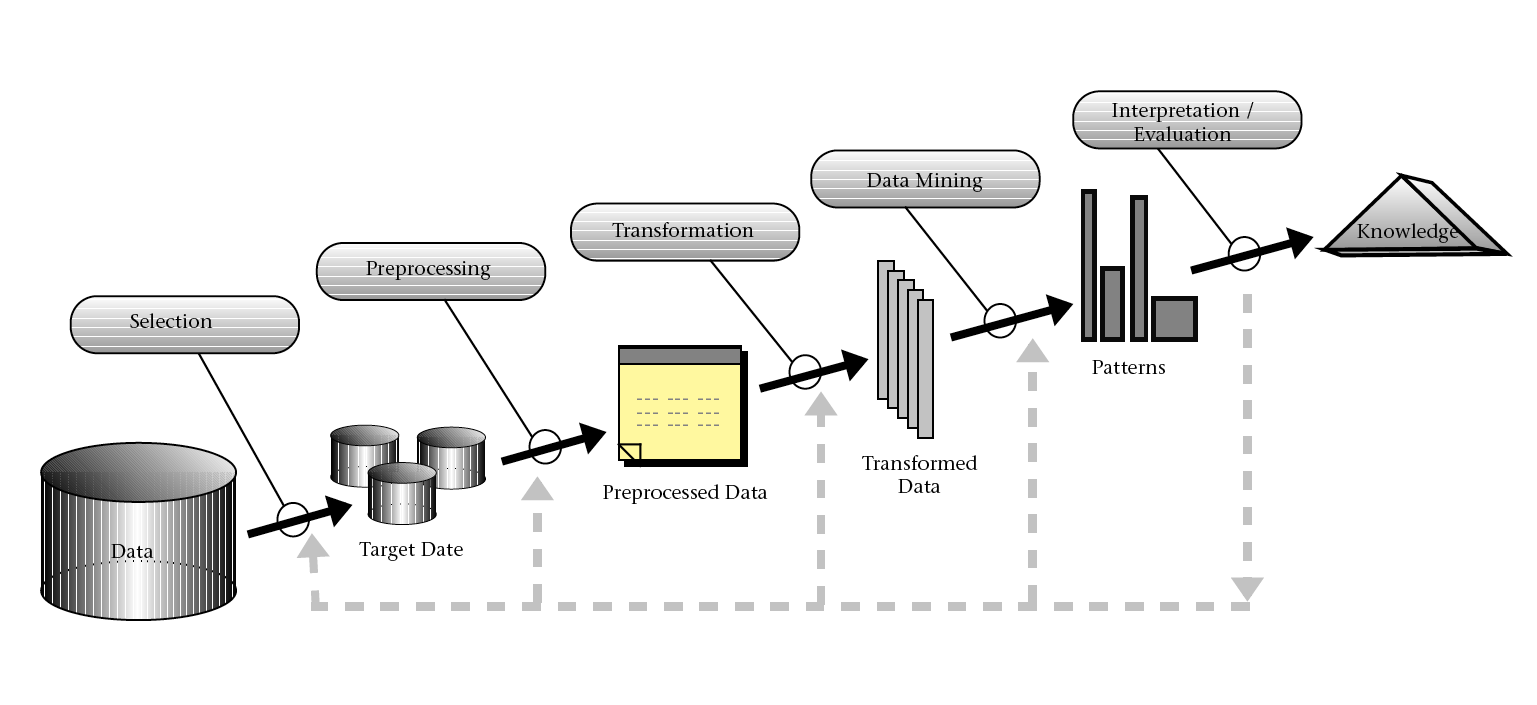
The KDD’s goal is to extract knowledge from the large datasets. It uses data mining algorithms and also uses subsampling and transformation of the database. (Overview of the KDD Process)

Figure 2 KDD Process (Fayyad, 1996)

As stated in the above figure the KDD process is divided into various phases. The first phase is the **selection phase** where the relevant data is either get selected or created. Only the information related to our goal is selected and also the data which contains some background knowledge. Occasionally the information from the ubiquitous sources are selected but before doing this the compatibility of the data is checked. From this process the targeted data sources are created for the next phase of the preprocessing. (Deutsch, 2015)

In the **preprocessing phase** important elements related to our goal are taken out and filtered. In this phase it is effectively determined that the retrieved knowledge is good without applying any type of the data mining tools but by just doing the preprocessing of the data. The less bad data or noise is in data the effective and efficient the data mining will be. This process include the cleaning of the invalid data, the treatment of the missing values in the data and the creation of the new attributes of the data. (Deutsch, 2015)

In the **Transformation phase** the preprocessed data from the preprocessing phase is converted to the transformed data. The data in this phase should be in the data mining competent format. This phase may result in the number of different format of the data since variable data mining may need different variable data formats. The data in this phase can be manually reduced or automatically reduced. The reduction of data in this phase takes place due to lossless aggregation or loss of the full selection of the most needed variables in dataset. A sample selection can be used to find the final solution of the whole data. (Deutsch, 2015)

In the **Data Mining Phase** the function of the data mining is used. Existing or the new techniques can be used in this phase. The output of this phase is converted to the pattern of data. In this phase data mining task is chosen to see whether the goal of the KDD is to classification clustering or regression. It is also decided that which parameters and models may be appropriate for the model of the data. Identical data mining algorithm is matched with the whole criteria of the KDD process. (Deutsch, 2015)

In the final **interpretation and evaluation phase** it is determined that the retrieved pattern is interesting or not. The primary task of this phase is to represent the result in the proper way so as to examine it completely. In this phase it is also evaluated that it contains the knowledge which is required. If the proper data is not retrieved the procedure falls back to the previous phase until the final knowledge is not retrieved. This phase contains the important final data which is required to conclude the procedures and plan. This whole phase are also called as Decision Support system (DDS) into the field of the directed and marketing automation fields. (Deutsch, 2015)

In our logic the data is retrieved from the datasets and selected in so as to get the targeted data. In the second phase the data is cleaned and missing data is replaced with the help of the aggregation then that data is transformed and converted so as to get the data ready for the data mining phase. We then apply the random forest algorithm to get patterns and the final data which is need to make decisions. The final outputted .csv files will contains the sorted and predicted data for the sales of the store and its departments.

1. **Methodology**

The problem solution is determined by the data mining algorithm technique called Random Forest. This algorithm can be found as a package in the R programming language. This language is used for the statistical analysis and graphics of the data. The R programming language was created by Ross Ihaka and Robert Gentlemen. The source code for R is written in C and FORTRAN. In this language user can create their own packages and upload onto the R server. It contains many wide range of varieties of statistical and graphical techniques. It includes linear and nonlinear modeling, clustering etc. techniques. It is extensible through functions and the extensions. We have used the R studio to edit and display the UI related to the code. The R studio can install any package simply through its UI. It’s simple and easy to use as an editor. The project in this R studio is saved into the project folder by .Rproj extension.

The background of this project is based on the 5 csv files which are taken from the open source website of KDD. The datasets contains huge amount of the data stored in a CSV file. The CSV (Comma Separated Values) is a simple text file written in text editor of windows and saved with the .csv extension. The files can be used in many applications so as to store the data. The csv files are the simplest form of the dataset. This files are used in the field of marketing, business modelling etc.

In this project we created a logic in the R programming language. This logic is divided into the 4 steps:

1. Importing the Data from the CSV files and packages.
2. Storing the values into the temporary variables.
3. Applying the Random Forest algorithm.
4. Writing the output in the Final.csv file.
5. **Importing the Data**

The packages are imported into the program by library (package\_name) command. But first we have to install the packages from the R project UI to proceed. Then the dataset is imported from the 5 CSV files taken from the open source website of KDD. The csv filenames are testing.csv, training.csv, SampleStores.csv, StoreFeatures.csv and SampleOutput.csv. All these file contains the data about the stores and its departments. The dataset is imported into the program by read.csv command in the R language. The all csv files are stored into the data folder of the project. The logic code is stores in the R folder with the .R extension. The program code imports these files and proceed to the next step. Each of these files are narrowed down to some space so as the limited memory of the system. Each file brought down to the size of maximum of 1.4 MB. The details of each .csv files are as follows:

Holidays in the Holiday column in the table shown are in DD/MM/YY format

Christmas: 31/12/2010, 30/12/2011, 28/12/2012, 27/12/2013.

Super bowl: 12/02/2010, 11/02/2011, 10/02/2012, 08/02/2013.

Thanksgiving: 26/11/2010, 25/11/2011, 23/11/2012, 29/11/2013.

Labor Day: 10/09/2010, 09/09/2011, 07/09/2012, 06/09/2013.

1. SampleStores.csv

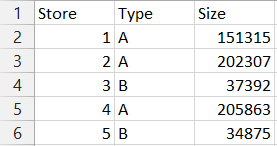


Figure 3 Screen Shot of SampleStores.csv File

This file contains the Information about the store with its type and size i.e. the capacity of the store to hold the commodities. The logic takes the type and size of the store into the consideration while calculating the final output. Here type A and B determines the size of the store whether it’s large or the small. The sample dataset contains the data of the 65 stores but we narrowed down this to the 5 stores due to the limitation of the memory. As stated in the above figure store 4 is the largest store from them with the capacity of 205863.

1. SampleOutput.csv

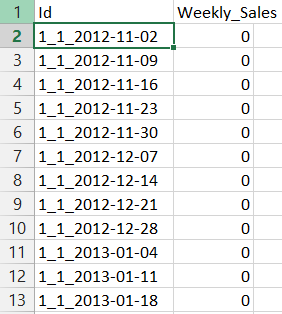


Figure 4 Screen Shot of SampleOutput.csv File

The above file contains how the format of the final output should be. The format to store the data to this file is written as storeid\_departmentid\_date format in the id column of the SampleOutput.csv file whereas the amount of the sales are weekly\_sales will be the final output i.e. amount of the sales. The output of the output Final.csv file will range from date 2nd of November 2012 to 26th July 2013.

1. Testing.csv

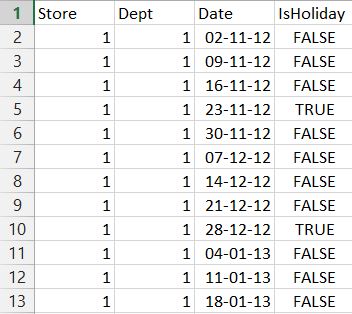


Figure 5 Screen Shot of Testing.csv File

In the above figure 5 testing.csv file the figure contains the store data and future weekly date with the holidays written in true and false. This file is used to test the model from this model our logic retrieves the store id, department id, date and holiday value. The holiday value in the program is written as 1 for true and 0 for false.

1. Store: It contains the store ID of the specific store from the table.
2. Dept: It contains the department id of department of the store.
3. Date: It contains the future weekly dates from which the sales are to be calculated. The date in this column range from 2nd of November 2012 to 26th of July 2013.
4. IsHoliday: This column represents whether that weak contains the holiday or not. This column shows the boolean value of that weekly date in value of TRUE or FALSE.
5. Traning.csv

The training.csv file is one of the most important file from the logic point of view. It will contain the historical sales data from the previous year on which the future sales are to be determined.

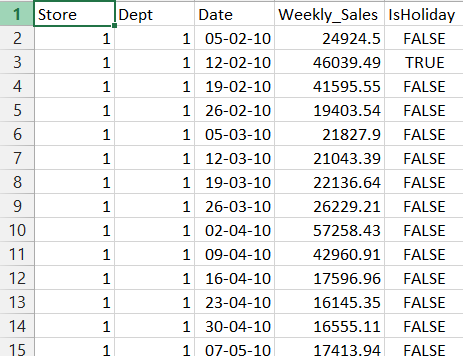


Figure 6 Screen Shot of Training.csv File

In the above screenshot the file contains the column of Store id, Department id, Date, Weekly Sales and Holiday value. This is one of the most important file as our model will trained i.e. predict the amount of sale based upon from the previous year’s value. The column details for this file is as follows:

1. Store: It will contains the store id of the specific store.
2. Dept: This column will contains the department id form the department of the specific store.
3. Date: This column will show the sales date of the previous year. The values of the date in this column will range from date of 5th of February 2010 to 26th of October 2012. The date here is in DD-MM-YY format.
4. Weekly\_Sales: This column will contain the weekly sales of the past years from of 5th of February 2010 to 26th of October 2012. From the above screenshot we can determine the amount of sales on the holiday 12th February 2012 is much greater as compared to non- holiday dates. This data is helpful in determining the future date sales.
5. IsHoliday: This column represents whether that weak contains the holiday or not. This column shows the boolean value of that weekly date in value of TRUE or FALSE.
6. StoreFeatures.csv

Figure 7 Screen Shot of StoreFeatures.csv File

This file will contains all the features of the store such as temperature of the day, fuel prices etc. This file is also important because these constraints can determine the change in value of the weekly sales. The column information is as follows:

1. Store: It will contain the store id of the respective store.
2. Date: It will contain the historical date of the previous year.
3. Temperature: This column will contain the average temperature of that region.
4. Fuel\_prices: This is column of average fuel prices in the region.
5. MarkDown 1-5: This columns has the markdown value of whether the given store has some specific promotional sales running. The markdown data is only available after the November 2011 and will not be available for the certain store. The markdown values which are not available will be marked as NA in the columns.
6. CPI: It is the column for the customer price index for the specific store
7. Unemployment: This column shows the rate of unemployment in the region

1. IsHoliday: This column represents whether that weak contains the holiday or not. This column shows the boolean value of that weekly date in value of TRUE or FALSE.

More elements can be added to this dataset to find the final value of the store. The more the value the more accurate the system will be. Each of these values can alter the final outcome but in this logic we just considered the Holiday values.

1. **Storing the Value into Temporary Variables.**

This is the second part of the logic where the important values which are needed are being stored into the temporary variable to perform the run of logic model. In this logic we have taken the values such as department id, store id, historical sales, date and holiday Boolean value and stored t in temporary variable. In the later part of the program the logic is run on this model to extract the final data.

1. **Applying the Random Forest Algorithm.**

In the Random Forest algorithm each tree is determined from the proximity matrix and when the two trees in the algorithm are related it is said to be the forest. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class (Breiman & Cutler).

### Features of Random Forests:

* Its accuracy is maximum among the other algorithms
* Its efficiency is very high on the other large datasets.
* Its variable handling capacity is very high and it do not delete and leave the variables.
* It provides the estimation of which variables are important in the classification of data.
* It generates an internal unbiased estimate of the generalization error as the forest building progresses.
* It provides the effective method of estimating the missing variables and it does not fail an remain consistent even when large data in the dataset is missing
* This algorithm has the ways of balancing the error in the unbalanced sets
* The final data from the forest tree can be save for the future use.
* Prototypes are computed that give information about the relation between the variables and the classification.
* It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.
* The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
* It offers an experimental method for detecting variable interactions. (Breiman & Cutler)

To understand the working of Random Forest and use the various options, further information about how they are computed is useful. Most of the options depend on two data objects generated by random forests. When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance. After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data. (Breiman & Cutler).

The random forest is included in the program by library(randomForest) at the start of the program if the package is not installed properly the program will come to halt.

C:\Users\Ajinkya\AppData\Local\Temp\x10sctmp7.png

The random forest is donated by the above equation where the year, month, days are the constraints needed to get the future date and the logsales is the logarithmic value of the sale on that specific date. Also the ntree represents number of the trees in the forest which can range from 1 to any value depending upon the amount of proximity you want in the program. This syntax is then stored into the temporary model and with the help of this model the final output is stored into the external .csv file.

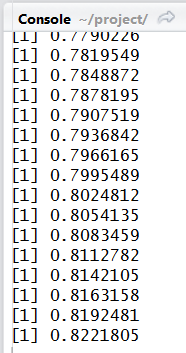


Figure 8 The Trees Proximity Output in the R studio console

In the above figure 8 the console is showing the output for each tree in the forest each tree value is calculated until the value would not reach the maximum proximity i.e 0.99 .After the value reach its maximum proximity the final output phase is carried out. The other package used in this is of Timedate. The timedate package allows the user to include the calendar, date, time and the utilities for the Rmetrices. The timedate package can be installed in R studio to read the time and date from external csv files. Time date package is imported into the program with help of the library(Timedate) command at the start of the program.

1. **Writing the output to the Final.csv file.**

After the execution of the model with the given csv files and after the proximity calculation the program will store the output in the final folder of the project folder by the name of Final.csv file.

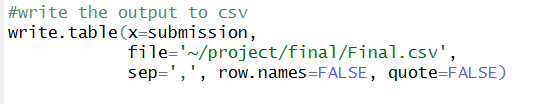


Figure 9 Screen Shot of the output Final.csv File Code

The output is written in the file with the write.table() function. In the arguments of this function the path where the file is to be stored provided and the separator (,) is given to separate the output values.

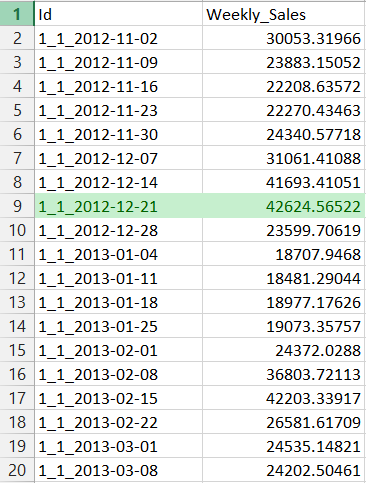


Figure 10 Screen Shot of the output Final.csv File

In the above screenshot the logic writes the value for the each weakly dates with the store id and the department id, as we can see the highlighted green value in the screen shot is the date of the Christmas where the value of the sale is much greater than the previous week’s sales. The values will be high for each holiday sale in the final output. The output date in the final.csv will range from the 02/11/2012 to 27/06/2013 in the YY-MM-DD format for the each store and each department of that store. The output can be calculated for the multiple store and multiple departments depending upon the sample data provided to the logic.

1. **Discussion**

The output file provided by the logic is 88% accurate by comparing it with the actual sales of the store. This accuracy can be increased by providing the more constraints and features in the store.csv file. The

1. **Implications**
2. **Results**

The part of the final result graph shows that the sales are increasing at the end of the year and during the during the holiday season of the Christmas and super bowl the value for the sales are very high as compared to the rest of the year. It’s above 40,000 mark for the Christmas on the 1\_1\_2012-12-21 and on 1\_1\_2013-02-15 for the super bowl. From this figure we can imply that the proposed logic is working fine for the whole system of multiple stores and multiple departments. The same type of output will be finalized for the other departments and stores during the holiday. The proposed logic can work for multiple values of the store with multiple features for the calculation. The proposed sales are 88% accurate according to algorithm execution. The accuracy can be increased further by taking multiple features while calculation the random forest. The proposed logic can also be run on the multiple algorithms for the statistical analysis for the calculation and the comparison of the sales with the system. From this proposed analysis the user can manage the business sales of his store and increase or decrease the quantity of the goods needed according to the sale. The store manager can also manage profit multiple departments by providing the markdown sales on the prices of the goods on that specific department. This logic model helps the store managers and the multiple sales person to model their sales and markdown events which can help them to thrive in their business with regards to profit in the sales.

1. **Conclusion**

In this report I have discussed the sample dataset used and features and values needed to calculate the final data. I have discussed how the KDD final data extraction model work in the real time. I have provided the details of the sample dataset imported into the program for the calculation purpose. I have also discussed how the Random Forest algorithm works by calculation the proximity values and the final values. From comparing to the other model calculation I found out that the Randomforest algorithm is providing the maximum efficiency and accuracy with regards to the calculation of the final values of the store sales. The final outcome of this logic are the future values of the sales whose accuracy can be increased by proposing the better feature values to the logic. Finally I found out that this logic can benefit the store manager to maximize the profit of the sales by giving certain markdown events in the year.

1. **Appendix**

**Software Requirements:** 1. R programming Language 32/64bit.

2. R studio 32/64bit for Windows/Mac OSX/Linux.

3. Windows OS or MAC OSX

4. Microsoft Excel 2013 for reading .csv files.

**Hardware Requirements:** 1. 4 GB of RAM.

2. 512 MB of Hard Disk Space.

3. Internet Connection for Installing R studio Packages.

**User Manual:**

1. First Install the R programming language into the system from the link provided below:

<http://cran.r-project.org/mirrors.html>

1. Install the R Studio for windows, Linux or mac OSX from the website link below:

<http://www.rstudio.com/products/rstudio/download/>

1. After the installation copy the project folder into the system user documents folder example

C: /Users/Administrator/MyDocuments.

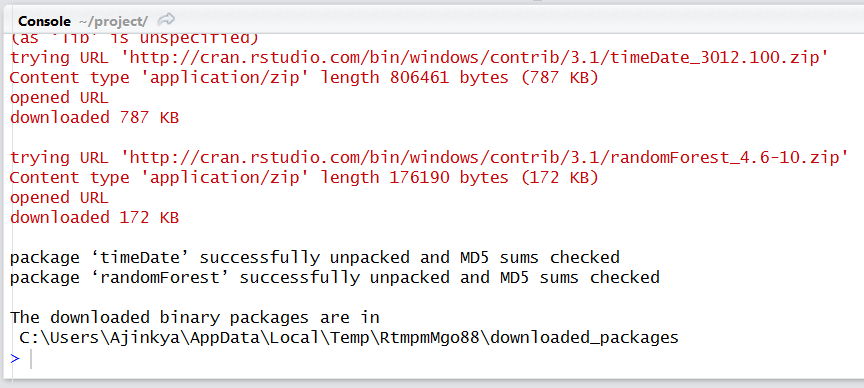
1. Open the R studio software installed into the system.
2. Select the target folder where the project is copied to the hard disk make sure the data folder in the project folder contains all .csv dataset files
3. Open project.Rproj file to load all the variables of the logic.
4. Install the packages form the “Tools” in the toolbar by selecting Tools>>Install Packages make sure the system have the internet connection to install the packages
5. Type in the name of the packages required to install “TimeDate,RandomForest” both the packages must be separated by comma “,”. Repository must be set to CRAN for installation.
6. Check whether the installation is done properly from the output window console in the R studio

Figure 14 Package Installation Console Windows Output

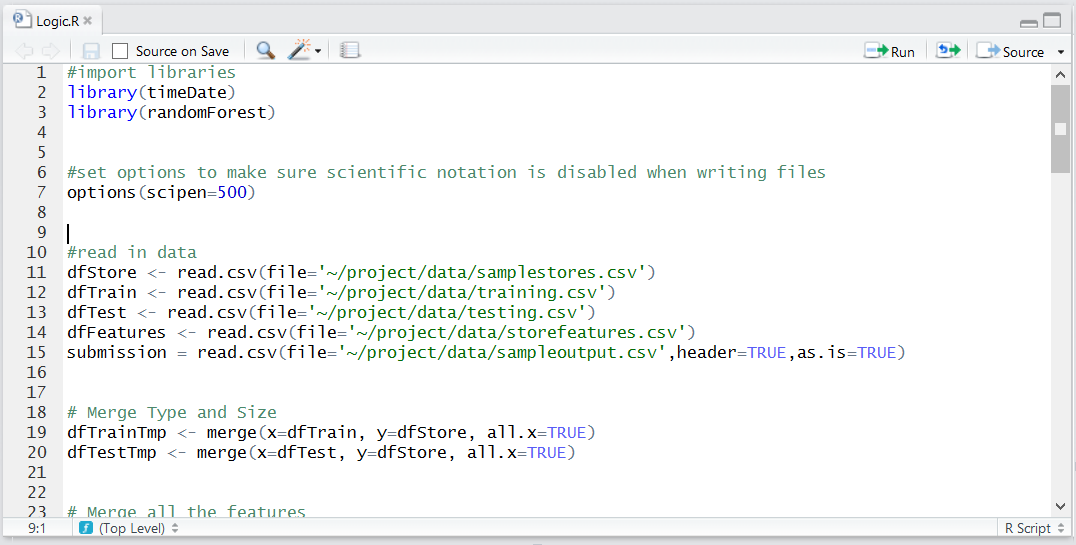
1.  Open the code file from the R folder named Logic.R

Figure 15 Logic.R File

1. Run the code in the file by selecting all code by CTRL+A and then run it by pressing Enter.
2. The output screen will calculate the proximity values into the output windows.

Note: The program may take 5 minutes to execute.

1. The output console window will show that the final output is written into the Final.csv file into the final folder of the project.

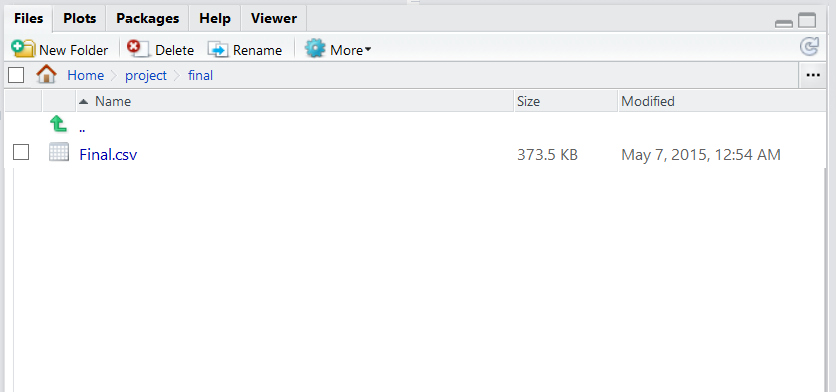


Figure 16 Final Output File Written in the Final Folder of the Project

1. Go to this final file from the explorer and open the Final.csv into the Excel. This file will show the predicted outcome of the sales for the future dates.
2. Possible error may occur when the path of the dataset files are not properly set. Solution is to check the path of the dataset .csv files and change into the input section of the program.

# **References**

Deutsch, G. (2015). *Data Mining in the KDD Environment.* Retrieved from Data Mining blog: http://www.data-mining-blog.com/data-mining/data-mining-kdd-environment-fayyad-semma-five-sas-spss-crisp-dm/

Fayyad. (1996). *An Overview of the Steps That Compose the KDD Process .* Retrieved from Infovis Wiki: http://www.infovis-wiki.net/index.php?title=Knowledge\_Discovery\_in\_Databases\_(KDD)

Leo Breiman, A. C. (n.d.). *Random Forests.* Retrieved from www.stat.berkeley.edu/: https://www.stat.berkeley.edu/~breiman/RandomForests/cc\_home.htm#intro

Marketing Land. (2015, January 14). *Holiday E-Commerce Sales Up 27% For SMB Retailers In 2014*. Retrieved from Marketing Land: http://marketingland.com/holiday-e-commerce-sales-27-small-businesses-114328

*Overview of the KDD Process.* (n.d.). Retrieved from cs.uregina.ca: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\_kdd.html

Technopedia. (2013). *Knowledge Discovery in Databases (KDD)*. Retrieved from Technopedia: http://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd