International Journal of Management (IJM)
Volume 8, Issue 6, Nov–Dec 2017, pp. 109–129, Article ID: IJM_08_06_013
Available online at
http://www.iaeme.com/ijm/issues.asp?JType=IJM&VType=8&IType=6
Journal Impact Factor (2016): 8.1920 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6502 and ISSN Online: 0976-6510

APPLICATION OF DATA MINING WITH R PROGRAMMING TO IMPLEMENT ERP FOR DIGITIZATION OF VALUATION REPORTS ON COMMERCIAL VEHICLES AND PASSENGER CARS

Debarka Banerjee

Student, MBA, Future Business School, Kolkata, India

Biswajit Roy

Assistant Professor, MBA, Future Business School, Kolkata, India

ABSTRACT

This paper proposes an approach towards data mining with R programming implemented in ERP for complete digitization of valuation reports for commercial vehicles and passenger cars. This paper entails entry of basis data, customary for the operating system of ERP. This data is stored on cloud through server. The model proposed in this paper could be used with the proper selection of cloud and custom designing for ISO system and making the system ready.

The model proposes a system with java programming which will be used for the front end design. In the back end, Oracle database can be used to store and manage data from both ERP database and the Cloud storage with structured orientation. The data contains details about types of car and the age of car and the month for optimum category. For betterment of the predictive model more number of parameters has been considered. In the Core End, The R programming is used for Multivariate statistical and predictive Analysis which plays the most important role for valuation of a used vehicle. The values of the parameters have been shown using Boxplot and histogram. According to the structure, a model has been proposed named as Java Enabled Data Mining Interface (JEMI) Model.

For the purpose, all the data were collected from both the primary and secondary sources. First data had been collected from commercial database and internet. It was on valuation report for all types of cars. These data were segmented based on types of car and insurance companies. These valuation report formats had been differentiated for each company. Now data had been furnished in MS Excel according to their product name, variant's name, classification and company's activity. After that one new excel sheet was used for individual commercial and passenger car. Then all the parameters which were used for valuation of a car, they were collected from previous report and were furnished in one new sheet according to number of individual car.

The research was further conducted to collect primary data from experts of the concerned area about their opinion on the major factors that decide the reselling price of a car. A number of parameters were considered for valuation of car, like TMFL, CMFL, and RCL. This total procedure is network based so it will require the cloud hosting. And different software related issues are also dealt herewith.

Key words: R programming, ERP, Data mining, Predictive analysis.

Cite this Article: Debarka Banerjee and Biswajit Roy, Application of Data Mining with R Programming to Implement ERP for Digitization of Valuation Reports on Commercial Vehicles and Passenger Cars. *International Journal of Management*, 8 (6), 2017, pp. 109–129.
http://www.iaeme.com/IJM/issues.asp?JType=IJM&VType=8&IType=6

1. INTRODUCTION

Enterprise Resource Planning systems are becoming necessary for almost every firm to improve the competitiveness. Accordingly the success for the implementation of ERP system, can provide a competitive advantage to companies in the global market. ERP or Enterprise Resource Planning can be defined as an integrated management of core business processes, generally, in real-time, and simplified with the help of software and technology. The business activities which can be simplified through ERP consist of finance, shipping and payment, inventory management, marketing and sales, manufacturing/service delivery, purchase, and product planning.

This study dealt with the valuation report for all types of cars. These data were segmented based on types of car and insurance companies. These valuation report formats had been differentiated for each company. Next, data had been furnished in MS Excel according to their product name, variant's name, classification and company's activity. After that one new excel sheet was used for individual commercial and passenger car. Then all the parameters which were used for valuation of a car, they were collected from previous report and were furnished in one new sheet according to number of individual car.

This paper concentrated on developing a software system that can primarily load information about different types of used cars in a database. That in turn can be uploaded in a cloud platform and a data mining approach through R can predict the reselling value of the car that can be run through a front end interface. For the purpose, a web based ERP software should be developed. Now the company business data about used cars should be put on software application. One has to log into this application with username & password. Then there would be an option i.e. "ADD MACHINE". In this section there are three entries i.e. asset description, asset type, client name with valuation type. By this procedure all the products name and variants of products and company's details need to be created. In the next stage, each car sections have to be opened and put all the parameters in. In each parameter, there would be different sections i.e. description, default value, table format, availability under label.

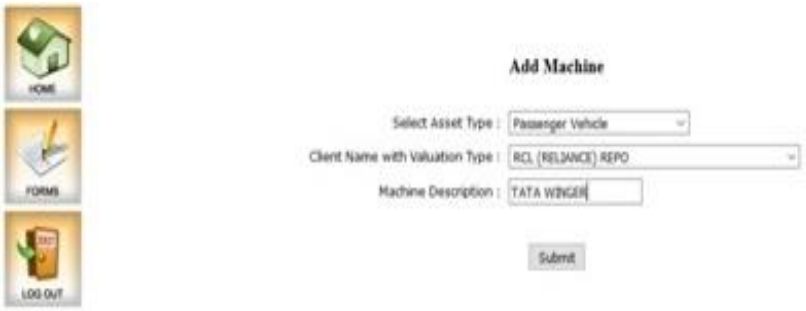


Figure 1 ERP interface

After completion of this work going back to the first window one could see that one column was created and also Number of rows were shown. Here also a number of parameters inserted were shown for valuation of car. Different company has different valuation parameter and two valuation procedures. According to these procedures different parameters of different products were created in companies like TMFL, CMFL, and RCL. Also barge reports had been covered. These were created in ERP software database system for this company.

ERP systems integrate various organizational systems and simplify error-free transactions and production. This, in turn, enhances the efficiency of the organization.

This running application software requires ISO certification for information security management. Documentation had been prepared according to clause. This total procedure is network based. So it will require the cloud hosting. Selection of cloud, hosting id creation, access of software, administrative control and the security management system were placed over the cloud network.

Data mining applications *in various areas including sales/marketing, banking, insurance, health care, transportation and medicine*. A great amount of data are now available in science, business, industry and many other areas due to rapid advances in computerization and digitization techniques. Such data may provide a rich resource for knowledge discover and decision support. Information technology (IT) has become the key enabler of business process expansion if an organization is to survive and continue to prosper in a rapidly changing business environment while facing competition in a global marketplace, they need to spend large IT budgets without accurate performance measurement systems on the business value of IT. This need for a proper data-mining technique, that can examine the impact of IT Data analysis.

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

R is free software. It runs on a wide variety of platforms including UNIX, Windows and MacOS. With the help of R programming to design a proper model selection phase, it is possible to obtain interesting scores for these prediction problems. The main goal of this paper is to describe how to perform some of the most basic data analysis tasks in R for the evaluation of used cars.

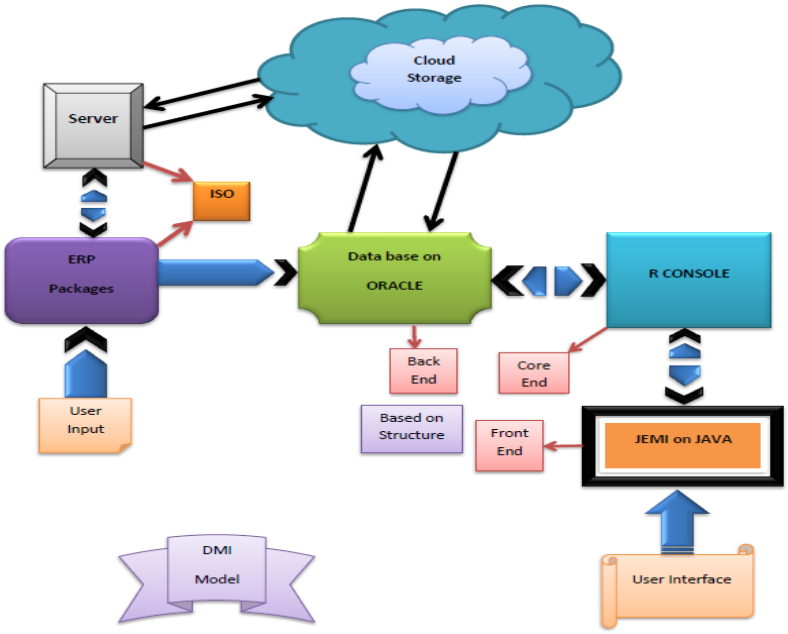


Figure 2 The model

2. LITERATURE REVIEW

Enterprise Resource Planning (ERP) is a software solution that integrates business functions and data into a single system to be shared within a company. While ERP originated from manufacturing and production planning systems used in the manufacturing industry, ERP expanded its scope in the 1990's to other "back-office" functions such as human resources, finance and production planning (Swartz & Orgill, 2001). Moreover, in recent years ERP has incorporated other business extensions such as supply chain management and customer relationship management to become more competitive. Thus ERP system is a set of all application software packages that provide to collect, store, manage and interpret data of operational, Managerial,& Strategic information for many business activities, including: Product planning, cost Manufacturing or service delivery, Marketing and sales, Inventory management, Shipping and payment.

ERP provides an integrated view of core business processes, often in real-time, using common databases maintained by a database management system. ERP systems track business resources—cash, raw materials, production capacity—and the status of business commitments: orders, purchase orders, and payroll. The applications that make up the system share data across the various departments (manufacturing, purchasing, sales, accounting, etc.) that provide the data. ERP facilitates information flow between all business functions, and manages connections to outside stakeholders. Management objectives are: Increased sales, Improved margins, Reduced loss due to rejections, Faster recovery of return on investment.

Enterprise system software is a multi-billion dollar industry that produces components that support a variety of business functions. IT investments have become the largest category of capital expenditure in United States-based businesses over the past decade. Though early ERP systems focused on large enterprises, smaller enterprises increasingly use ERP systems.

The ERP system is considered a vital organizational tool because it integrates varied organizational systems and facilitates error-free transactions and production. However, ERP system development is different from traditional systems development. ERP systems run on a

variety of computer hardware and network configurations, typically using a database as an information repository.

Mohammad J Zaki (2014) mentioned that data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data. We begin this chapter by looking at basic properties of data modelled as a data matrix. We emphasize the geometric and algebraic views, as well as the probabilistic interpretation of data. We then discuss the main data mining tasks, which span exploratory data analysis, frequent pattern mining, clustering, and classification, laying out the roadmap for the book.

Data mining comprises the core algorithms that enable one to gain fundamental insights and knowledge from massive data. It is an interdisciplinary field merging concepts from allied areas such as database systems, statistics, machine learning, and pattern recognition. In fact, data mining is part of a larger knowledge discovery process, which includes pre-processing tasks such as data extraction, data cleaning, data fusion, data reduction and feature construction, as well as post-processing steps such as pattern and model interpretation, hypothesis confirmation and generation, and so on. This knowledge discovery and data mining process tends to be highly iterative and interactive.

The algebraic, geometric, and probabilistic viewpoints of data play a key role in data mining. Given a dataset of n points in a d dimensional space, the fundamental analysis and mining tasks covered in this book include exploratory data analysis, frequent pattern discovery, data clustering, and classification models, which are described next.

Bontempiet al.(1999), mentioned that R programmingis an Open Source scripting language and environment for statistical computing and graphics. If you are interested in knowing more about the international data analysis competition that was behind the data used in this chapter. To compare the data analysis strategies followed by these authors. In terms of data mining, this study has provided information on • Data visualization • Descriptive statistics • Strategies to handle unknown variable values • Regression tasks • Evaluation metrics for regression tasks • Multiple linear regression • Regression trees • Model selection/comparison through k-fold cross-validation • Model ensembles and random forests We hope that by now you are more acquainted with the interaction with R, and also familiarized with some of its features. Namely, you should have learned some techniques for • Loading data from text files and also simultaneously to Oracle database • How to obtain descriptive statistics of datasets • Basic visualization of data • Handling datasets with unknown values • How to obtain some regression models • How to use the obtained models to obtain predictions for a test set Further cases studies will give you more details on these and other data mining techniques.

Siddharth Arora et al mentioned that for a good prediction or classification the learning algorithms must be provided with a good training set from which rules or patterns are extracted to help classify the testing dataset. Main goal of this work is that at the time of purchasing new vehicles when there are a number of choices are available and customer doesn’t have much of the prior knowledge regarding the performance of these vehicles in actual practice then it becomes much typical for the customer to choose the correct one in terms of quality purchase. Decision for purchase of new vehicle, analysis of performance of vehicles, evaluation and comparing the past record of vehicle depend upon for a comprehensive and coherent theoretical and practical understanding of such problems.

Sameer Chand Pudaruth (2014) mentioned that the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bays and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances .In the future, we intend to use more sophisticated algorithms to make the predictions.

In this paper, four different machine learning techniques have been used to forecast the price of used cars in Mauritius. They use different parameters for evaluate use cars, they are: They use linear regression with kNN. The accuracy found to be between 60-70% for different combinations of parameters according to NaiveBayes. They record that, the price attribute had to be classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies. The main limitation of this study is the low number of records that have been used.

3. OBJECTIVE

The valuation of used vehicles constitutes to be a problem in automobile industry that needs the help of managerial decision making. One of those major problems is to evaluate a used car. Being able to monitor and perform an early forecast on a used vehicle's valuation is essential. And also it is important to improve the quality of valuation. With the goal of addressing this prediction problem, several samples were collected during a period of approximately 3 months time. For each sample, deferent properties were measured as well as the frequency of occurrence of seven properties of vehicles needs to be calculated. For, the purpose several data mining tools were employed in the reserch. Also, a database is generated to store all related data. And in addition to this an interface is also generated to make the data mining process a bit more user friendly. So, the main objective of the project is to create a system oriented model that can purpose the job mentioned above.

Limitations

- The model has been developed based on secondary data.
- Time was a constraint to implement the system properly.
- The system is yet not being tested to other systems.
- The database connection with the server could not be tested due to lack of availability of the same.

4. RESEARCH METHODOLOGY

Research process is a series of systematic steps that are followed to solve a business process. It is a frame work of entire plan-of action. It clearly describes that crucial issues like the studies purpose and objective the type of the data needed. This technique to be used for finding the sample, searching, analysing it and other aspects that are essential for guiding business research. The research methodology, which follows, is the backbone of the study.

Research Design

During the research, this paper introduced some basic tasks of data mining, i.e. data pre-processing, exploratory data analysis, and predictive model construction. For this initial case study a small problem by data mining standards is selected. Namely, the problem of predicting the frequency occurrence of several used vehicles.

Types of Data

Secondary data

Secondary data is the data that have been already collected by and readily available from other sources. Such data are cheaper and more quickly obtainable than the primary data and also may be available when primary data cannot be obtained at all. Accuracy of secondary data is not known. Data may be out-dated. This research addresses secondary data for analysis collected from a company database.

System design tools

In the project a data analytics part has been designed by R Programming, data base part has been done by Oracle and user interfacing with R console and oracle data base is done through Java. For the purpose the required coding has been done, that can help to find those parameters which are more important for the valuation of vehicle.
RESEARCH TOOLS used for the study are:

- R Programming
- Oracle data base Edition 9
- JAVA (jdk8)

Data Analysis & findings

Data description

The data available for this problem was collected from a company database which is in the business of evaluating a used car. The dataset consists of data for 200 samples. To be more precise, each observation in the available datasets are an aggregation of several samples collected from the same over a period of 3 months, during the same year (i.e, on 2015). Each observation contains information on 11 variables. Three of these variables are nominal and describe the purpose of the analysis when the samples to be aggregated were collected, as well as the Type and Age of the in question. The eight remaining variables are values of deferent automobile parameters measured in the samples forming the aggregation,
viz:

• Maximum BS value • Minimum value of BD • Mean value of ET • Mean value of CS • Mean value of WHS • Mean of RHDMR • Mean of total EAES• Mean of DA etc.
Where,
Assessors Summary Report = ASR
Engine Type= ET
Chassis =CHS
Vehicle General Appearance=VGA
Operator Cabin / Monitor, Control Levers=OMCL
Engine / Auto Electric System=EAES
Power Train Systems =PTS
Brake & Steering=BS
Body= BD
Chassis and Suspension=CS
Wheels =WHS
Repair History Details & Maintenance Records=RHDMR
Documents Available =DA
Assessor’s Description=AD
Overall Condition of the Equipment=OCE

Associated with each of these parameters are seven frequency numbers of deferent vehicles of the company are found in the respective samples. No information is given

regarding the names of the vehicles that are identified. The second dataset contains information on 140 extra observations. It uses the same basic structure but it does not contain information concerning the frequencies of seven vehicles of the associated evaluating company. These extra observations can be regarded as a kind of test set. The main goal of our study is to predict the frequencies of the seven parameters for these 140 samples. This means that we are facing a predictive data mining task. This is one among the diverse set of problems tackled in data mining. In this type of task, our main goal is to obtain a model that allows predicting the value of a certain target variable (for this case it is value of used vehicles and certain other parameters) given the values of a set of predictor variables. This model may also provide indications on which predictor variables have a larger impact on the target variable; that is, the model may provide a comprehensive description of the factors that impudence the target variable.

Loading the data into R

For getting the data into R, it was simply the text files with the data are downloaded, and then loading them into R. Though the process of creating excel files is obviously more practical and easy to use. But for predictive purposes, data were loaded into R from text file. This data frame contains the first set of 200 observations mentioned above.

The initial R code to load and check data

```
>library(DMwR)
>setwd("D:/1")
> AICH <- read.table("aich.txt", header = TRUE)
>head (AICH)
```

Note:

1: AICH is the name of the data file.

2: package ‘DMwR’ was built under R version 3.1.3

3: package ‘lattice’ was built under R version 3.1.3

```
  Purpose   Type Age Month  ASR  ET  CHS  VGA OMCL EAES  PTS  BS  BD  CS WHS RHDMR  DA  AD OCE
1       1 commer old  june 8.00 9.8 6.80 6.23 8.00 5.00 7.00 5.0 0.0 0.0 0.0   0.0 4.2 8.3 0.0
2       2 passen old  june 8.35 8.0 7.75 1.28 7.00 8.75 8.75 1.3 1.4 7.6 4.8   1.9 6.7 0.0 2.1
3       3 constr old  june 8.10 7.4 4.02 5.33 6.66 5.66 8.05 5.6 3.3 3.6 1.9   0.0 0.0 0.0 9.7
4       4 passen old  june 8.07 4.8 7.36 2.30 9.18 6.18 8.70 1.4 3.1 4.0 8.9   0.0 1.4 0.0 1.4
5       5 constr old  june 8.06 9.0 5.35 5.41 3.70 8.22 7.58 5.5 9.2 2.9 7.5   0.0 7.5 4.1 1.0
6       6 commer old  july 8.25 5.1 6.75 9.24 4.00 8.25 6.66 8.4 5.1 4.6 1.4   0.0 2.5 2.6 2.9
>
```

A data frame can be seen as a kind of matrix or table with named columns, which is the ideal data structure for holding data tables in R. The head () function shows us the first six lines of any data frame. The “aich” link contains the 200 samples in a file named“aich.txt”, while the “Testdata” link points to the“Eval.txt” file that contains the 140 test samples. There is an additional link that points to a file (“Sols.txt”) that contains the AICH frequencies of the 140 test samples. This last file that was used to check the performance of our predictive models and will be taken as unknown information for now. The files have the values for each observation in a different line. Each line of the training and test files contains the values of the variables separated by spaces. Unknown values are indicated with the string “XXXXXXX”. The first thing to do is to store the data file in some directory on the hard disk (preferably on the current working directory of your running R session, which can be checked issuing the command getwd() at the prompt).

Data Visualization and Summarization Given the lack of further information on the problem domain, it is wise to investigate some of the statistical properties of the data, so as to get a better grasp of the problem. Even if that was not the case, it is always a good idea to start our analysis with some kind of exploratory data analysis similar to the one we will show below. A first idea of the statistical properties of the data can be obtained through a summary of its descriptive statistics:

>summary (AICH)

ET	CHS	VGA	OMCL	EAES
Min. : 1.500	Min. : 0.160	Min. :0.050	Min. : 0.000	Min. : 0.500
1st Qu.: 6.625	1st Qu.: 3.292	1st Qu.:1.395	1st Qu.: 4.000	1st Qu.: 2.910
Median : 8.300	Median : 5.445	Median :2.915	Median : 6.000	Median : 5.290
Mean : 7.822	Mean : 5.557	Mean :3.258	Mean : 5.824	Mean : 5.291
3rd Qu.: 9.375	3rd Qu.: 7.815	3rd Qu.:4.683	3rd Qu.: 8.000	3rd Qu.: 7.500
Max. :10.000	Max. :10.000	Max. :9.770	Max. :10.000	Max. :10.000
NA's :2	NA's :10	NA's :2	NA's :2	NA's :2

WHS	RHDMR	DA	AD	OCE
Min. : 0.000	Min. :0.000	Min. :0.000	Min. :0.000	Min. :0.000
1st Qu.: 0.000	1st Qu.:0.000	1st Qu.:0.000	1st Qu.:0.000	1st Qu.:0.000
Median : 1.450	Median :0.000	Median :1.600	Median :0.000	Median :1.000
Mean : 2.314	Mean :1.488	Mean :2.830	Mean :2.131	Mean :1.586
3rd Qu.: 3.500	3rd Qu.:2.325	3rd Qu.:5.325	3rd Qu.:3.825	3rd Qu.:2.400
Max. :10.000	Max. :9.000	Max. :9.900	Max. :9.900	Max. :9.700

This simple instruction immediately gives us a first overview of the statistical properties of the data.

By observing the difference between medians and means, as well as the inter-quartile range (3rd quartile minus the 1st quartile), we can get an idea of the skewness of the distribution and also its spread. Still, most of the time, this information is better captured graphically.

Let us see an example:

>hist(AICH$ASR, prob = T)

>hist(AICH$ASR, prob=T, xlab='',main='Histogram of maximum ASR value',ylim=0:1)

This instruction shows us the histogram of the variable Accessory Summery Report. The result appears in the following figure. With the parameter prob=T we get probabilities for each interval of values, while omitting this parameter setting would give us frequency counts. The following figure tells us that the values of variable Month apparently follow a distribution very near the normal distribution, with the values nicely clustered around the mean value.

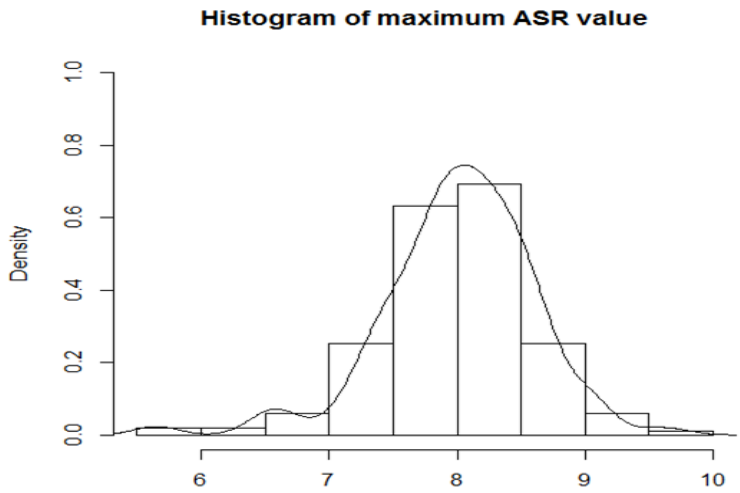


Figure 3 The histogram of variable Month.

The extensive use of function composition in the previous example, with several functions being called with the result of other functions. Every time there will be difficulties in understanding this type of instruction, that can always be called separately, one at a time, to fully understand what they produce. Another example (Figure 2) showing this kind of data inspection can be achieved with the following instructions, this time for variable OMCL:

```
>boxplot(AICH$EAES, ylab = "Engine/Auto Electric System (EAES)")
>rug(jitter(AICH$EAES), side = 2)
>abline(h = mean(AICH$EAES, na.rm = T), lty = 2)
```

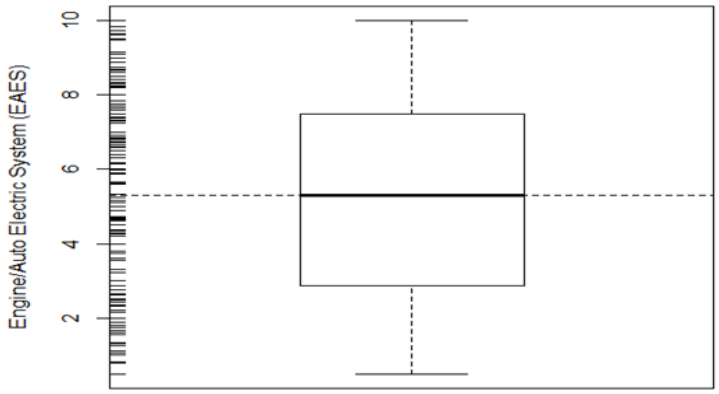


Figure 4 An"enriched" box plot for Engine/Auto Electric system.

The first instruction draws a box plot of variable EAES. Box plots provide a quick summarization of some key properties of the variable distribution. Namely, there is a box whose vertical limits are the 1st and 3rd quartiles of the variable. This box has a horizontal line inside that represents the median value of the variable. Let r be the inter-quartile range. The small horizontal dash above the box is the largest observation that is less than or equal to the 3rd quartile plus 1.5×r. The small horizontal dash below the box is the smallest observation that is greater than or equal to the 1st quartile minus 1.5× r. The circles below or above these small dashes represent observations that are extremely low (high) compared to all others, and are usually considered outliers. This means that box plots give us plenty of information regarding not only the central value and spread of the variable, but also eventual outliers. The second instruction was described before (the only difference being the place where the data is plotted), while the third uses the function abline() to draw a horizontal line at the mean value of the variable, which is obtained using the function mean(). By comparing this line with the line inside the box indicating the median, we can conclude that the presence of several outliers has distorted the value of the mean as a statistic of centrality (i.e., indicating the more common value of the variable). The analysis of Figure 4 shows us that the variable EAES has a distribution of the observed values clearly concentrated on low values, thus with a positive skew.

To study the distribution of the values of, say, BS (Brake & Steering). We could use any of the possibilities discussed before. However, study how this distribution depends on other variables, new tools are required. Conditioned plots are graphical representations that depend on a certain factor. A factor is a nominal variable with a set of finite values. For instance, set of box plots for the variable BS, for each value of the variable Type (figure 5). Each of the box plots was obtained using the subset of samples that have a certain value of the variable Type. These graphs allow us to study how this nominal variable may influence the distribution of the values of BS. The code to obtain the box plots is

```
>library(lattice)
>bwplot(Type ~ BS, data=AICH, ylab='Type',xlab=' BS')
```

The first instruction loads in the lattice package. The second obtains a box plot using the lattice version of these plots. The first argument of this instruction can be read as "plot BS for each value of Type". The remaining arguments have obvious meanings. Figure 3 allows us to observe that higher frequencies of BS are expected in smaller valuation, which can be valuable knowledge. An interesting variant of this type of plot that gives more information on the distribution of the variable being plotted, are box-percentile plots, which are available in package Hmisc. An example of its use with the same BS against the Type of valuation can be found:

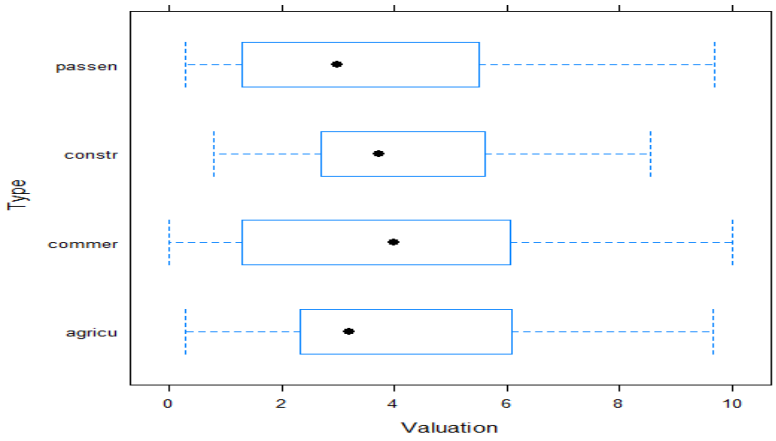
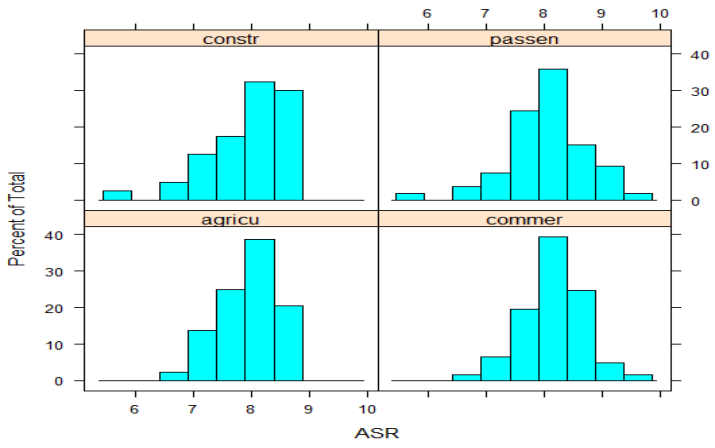


Figure 5 A conditioned box plot of valuation.

Next, it was attempted to explore the correlations between the variables with unknowns and the nominal variables of this problem. Conditioned histograms are used for the "purpose" those are available through the lattice R package with this objective. Cars are classified into four gropus based on the "purpose" of use. They are , cars used for: 1) Construction work (abbreviated as "constr", 2) Passenger cras (passen), 3) cars used for agricultural purposes (agricu) and 4) Commercial cars (comer).

For instance, Figure 4 shows an example of such a graph. This graph was produced as follows:

```
>histogram(~Month | Purpose, data = AICH)
```



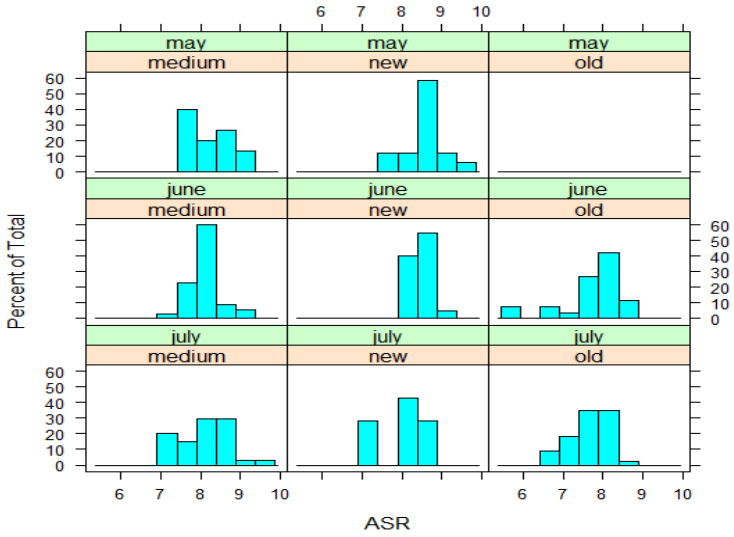


Figure 6 A histogram of variable Month conditioned by "Purpose".

The above instruction obtains a histogram of the values of Month for the different values of "Purpose". Each histogram is built using only the subset of observations with a certain "Purpose" value. It can be noticed that the ordering of the "Purposes" in the graphs is a bit unnatural. For natural temporal ordering of the "Purposes", it is required to change the ordering of the labels that form the factor "Purpose" in the data frame. This could be done by
>AICH$ Purpose<- factor (AICH$ Purpose, levels = c("spring", + "summer", "autumn", +"winter"))

Filling in the Unknown Values by Exploring Similarities between Cases

Instead of exploring the correlation between the columns (variables) of a dataset, it was tried to use the similarities between the rows (observations) to fill in the unknown values. The method to fill in all unknowns with the exception of the two samples with too many NAs are illustrated. The data to override the code of the previous sections are tried again.

>data(AICH)

>AICH<- AICH[-manyNAs(AICH),]

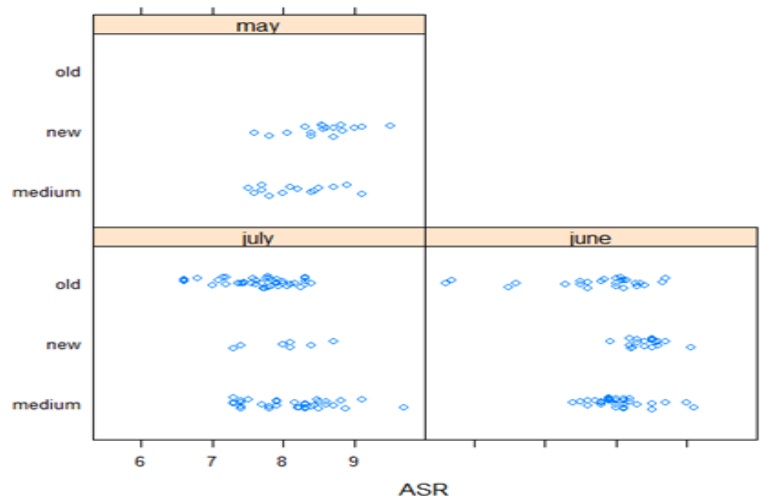


Figure 7 The values of variable Month by Type and Age.

Note: Cars are also classified under three Types, viz, Old, medium and new.

Prediction for different types of cars that are expected during three months for subsequent years

Next, the research goal is to have good forecasts of the of the expected number of different types of cars. This general goal should allow us to easily define what to predict with our models should resort to forecast the future cars time series. However, it is easy to see that even with this simple task we immediately face several questions, namely, for which time in the future? Answering these questions may not be easy and usually depends on how the predictions will be used for generating trading orders.

What to Predict?

The trading strategies we will described based on what we obtain a prediction of the tendency of the market in the next few years. Based on this prediction, we will place orders that will be profitable if the tendency is confirmed in the future.

For the purpose, different prediction models can be considered, viz, Logistic,LDA,QDA, KNN(3), KNN(5), KNN(8), KNN(12), Classification Tree and SVM (Support Vector Machine). The study explored some models that can be used to address the prediction tasks defined. The selection of models was mainly guided by the fact that these techniques are well known by their ability to handle highly nonlinear regression problems. That is the case in this research problem. Still, many other methods could have been applied to this problem. Any thorough approach to this domain would necessarily require a larger comparison of more alternatives. Such exploration does make sense due to its costs in terms of space and computation power required.

Use of Training Data

Complex time series problems frequently exhibit different regimes, such as periods with strong variability followed by more “stable" periods, or periods with some form of systematic tendency. These types of phenomena are often called non-stationeries and can cause serious problems to several modeling techniques due to their underlying assumptions. It is reasonably easy to see, for instance by plotting the number time series, that this is the case for the research data. There are several strategies we can follow to try to overcome the negative impact of these effects. For instance, several transformation techniques can be applied to the original time series to eliminate some of the effects. The use of percentage variations (returns) instead of the original absolute values is such an example. Other approaches include using the available data in a more selective way. The standard approach would use the training data to develop the model that would then be applied to obtain predictions for the testing period. It has a strong reason to believe that there are regimes shifts, using the same model on all testing periods may not be the best idea, particularly if during this period there is some regime change that can seriously damage the performance of the model. In these cases it is often better to change or adapt the model using more recent data that better captures the current regime of the data. These approaches are usually known as incremental learners as they adapt the current model to new evidence instead of starting from scratch. There are not so many modeling techniques that can be used in this way, particularly in R. In this context, the research followed the other approach to the updating problem, which consists of re-learning a new model with the new updated training set. This is obviously more expensive in computational terms and may even be inadequate or applications here the data arrives at a very fast pace and for which models and decisions are required almost in real-time. This is rather frequent in applications addressed in a research area usually known as data streams.

The modeling tools

The research briefly described that the modeling techniques used to address the prediction tasks and also illustrated how to use them in R.

Use of Predictions methods

This research is based on the trading of used cars in future markets. These markets are based on contracts to buy or sell a commodity on a certain date in the future at the price determined by the market at that future time. Still, in objective terms, this means that our trading system will be able to predict three types of used cars, viz, Old, Medium and new.

Model Evaluation and Selection

```
setwd("D:/1")
mydata<- read.table("aich.txt", header=TRUE)
mydata<- na.omit(mydata)
set.seed(1)
training = sample(184,100)
training_data = mydata[training,]
testing_data = mydata[-training,]
Age_testing = testing_data$Age
```

#Logistic Model

==========

Logistic regression deals with the problem of predicting a binary response using multiple predictors. By analogy with the extension from simple to multiple linear regression, the following model is generalized as follows:

==========

```
logistic_model = glm(Age~., data= training_data, family=binomial)
logistic_probs = predict(logistic_model, testing_data, type="response")
logistic_pred = rep("old",84)
logistic_pred[logistic_probs> .33]="medium"
logistic_pred[logistic_probs> .67]="new"
table(logistic_pred, Age_testing)
```

==========

```
logistic_pred = rep("new",84)
logistic_pred[logistic_probs> .33]="old"
logistic_pred[logistic_probs> .67]="medium"
table(logistic_pred, Age_testing)
mean(logistic_pred != Age_testing)
```

==========**Result**

```
>logistic_pred = rep("old",84)
>logistic_pred[logistic_probs> .33]="medium"
>logistic_pred[logistic_probs> .67]="new"
>table(logistic_pred, Age_testing)
Age_testing
```

logistic_pred	**medium**	**new**	**old**
medium	**7**	**5**	**5**
new	**13**	**9**	**21**
old	**17**	**2**	**5**

>

```
>logistic_pred = rep("new",84)
>logistic_pred[logistic_probs> .33]="medium"
>logistic_pred[logistic_probs> .67]="old"
>table(logistic_pred, Age_testing)
```

Age_testing

logistic_pred	**medium**	**new**	**old**
medium	**7**	**5**	**5**
new	**17**	**2**	**5**
old	**13**	**9**	**21**

>

```
>logistic_pred = rep("new",84)
>logistic_pred[logistic_probs> .33]="old"
>logistic_pred[logistic_probs> .67]="medium"
>table(logistic_pred, Age_testing)
```

Age_testing

logistic_pred	**medium**	**new**	**old**
medium	**13**	**9**	**21**
new	**17**	**2**	**5**
old	**7**	**5**	**5**

```
>mean(logistic_pred != Age_testing)
```

[1] 0.7619048

==========Result

Linear Discriminant Analysis (LDA) Model

LDA models the conditional distribution of the response Y ,given the predictor(s) X. the paper considers an alternative and less direct approach to estimating these probabilities. In this alternative approach, the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for Pr($Y = k|X = x$) is modelled. When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression.

```
==========
lda_model = lda(Age~., data=training_data)
lda_pred=predict(lda_model, testing_data)
table(lda_pred$class, Age_testing)
mean(lda_pred$class != Age_testing)
```

==========Result

```
>table(lda_pred$class, Age_testing)
```

Age_testing

medium	**new**	**old**	
medium	**19**	**4**	**11**
new	**6**	**12**	**3**
old	**12**	**0**	**17**

```
>mean(lda_pred$class != Age_testing)
```

[1] 0.4285714

==========**Result**

Quadratic Discriminant Analysis (QDA) model

LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class specific mean vector and a covariance matrix that is common to all *K* classes. Quadratic discriminant analysis (QDA) provides an alternative quadratic discriminant analysis approach. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. That is, it assumes that an observation from the *k*th class is of the form $X \sim N(\mu k, \Sigma k)$, where Σk is a covariance matrix for the *k*th class.

==========

```
library (MASS)
qda_model = qda(Age~ a1+a2+a3+a4+a5+a6+a7, data=training_data) #Applying all predictors giving errors
qda_pred=predict(qda_model, testing_data)
table(qda_pred$class, Age_testing)
mean(qda_pred$class != Age_testing)
```

==========**Result**

```
>table(qda_pred$class, Age_testing)
Age_testing
medium      new       old
medium      19        4         10
new    13         11        18
old     5         1         3
>mean(qda_pred$class != Age_testing)
```

[1] 0.6071429

==========

KNN

The choice of the Number of Neighbors (k) is also an important parameter of predictive methods. Frequent values include the numbers in the set (1; 3; 5; 7; 11), but obviously these are just heuristics. However, we can say that larger values of k should be avoided because there is the risk of using cases that are already far away from the test case. Obviously, this depends on the density of the training data too and sparse datasets incur more of this risk. As with any learning model, the "ideal" parameter settings can be estimated through some experimental methodology. In R, the package class (Venables and Ripley, 2002) includes the function knn() that implements this idea. Below is an how knn is used on the research dataset:

==========

```
library(class)
std_data = scale(mydata[,c(4,5,6,7,8,9,10,11,12,13,14,15,16,17,18)])
training_data = std_data[training,]
testing_data = std_data[-training,]
training_age = mydata$Age[training]
knn_pred = knn(training_data, testing_data, training_age, 3)
```

table(knn_pred, Age_testing)
mean(knn_pred != Age_testing)
==========**Result**
>knn_pred = knn(training_data, testing_data, training_age, 3)
>table(knn_pred, Age_testing)
Age_testing

knn_pred	**medium**	**new**	**old**
medium	**25**	**8**	**17**
new	**3**	**9**	**0**
old	**5**	**5**	**12**

>mean(knn_pred != Age_testing)
[1] 0.452381
>knn_pred = knn(training_data, testing_data, training_age, 5)
>table(knn_pred, Age_testing)
Age_testing

knn_pred	**medium**	**new**	**old**
medium	**23**	**9**	**15**
new	**4**	**7**	**1**
old	**6**	**6**	**13**

>mean(knn_pred != Age_testing)
[1] 0.4880952
>knn_pred = knn(training_data, testing_data, training_age, 8)
>table(knn_pred, Age_testing)
Age_testing

knn_pred	**medium**	**new**	**old**
medium	**26**	**10**	**15**
new	**3**	**6**	**1**
old	**4**	**6**	**13**

>mean(knn_pred != Age_testing)
[1] 0.4642857
>knn_pred = knn(training_data, testing_data, training_age, 12)
>table(knn_pred, Age_testing)
Age_testing

knn_pred	**medium**	**new**	**old**
medium	**26**	**12**	**15**
new	**5**	**4**	**1**
old	**2**	**6**	**13**

>mean(knn_pred != Age_testing)
[1] 0.4880952
==========

Classification Tree

tree_model = tree(Age~., data = training_data)

tree_pred=predict(tree_model, testing_data, type="class")

table(tree_pred, Age_testing)

mean(tree_pred!=Age_testing)

==========**Result**

>tree_pred=predict(tree_model, testing_data, type="class")

>table(tree_pred, Age_testing)

Age_testing

tree_pred	**medium**	**new**	**old**
medium	**26**	**7**	**13**
new	**3**	**7**	**1**
old	**4**	**8**	**15**

>

>mean(tree_pred!=Age_testing)

[1] 0.4285714

########Tree after pruning

The process described here may produce good predictions on the training set, but is likely to over fit the data, leading to poor test set performance. This is because the resulting tree might be too complex. A smaller tree with fewer splits (that is, fewer regions *R*1, . . .,*RJ*) might lead to lower variance and better interpretation at the cost of a little bias. One possible alternative to the process described above is to build the tree only so long as the decrease in the RSS due to each split exceeds some(high) threshold. This strategy will result in smaller trees, but is too short-sighted since a seemingly worthless split early on in the tree might be followed by a very good split—that is, a split that leads to a large reduction in RSS later on.

==========

cv_tree=cv.tree(tree_model, FUN=prune.misclass)

plot(cv_tree$size, cv_tree$dev, type="b") # To determine visually at what level to prune, here =11

pruned_model=prune.misclass(tree_model, best=11)

tree_pred_pruned=predict(pruned_model, testing_data, type="class")

table(tree_pred_pruned, Age_testing)

mean(tree_pred_pruned!=Age_testing)

==========**Result**

>tree_pred_pruned=predict(pruned_model, testing_data, type="class")

>table(tree_pred_pruned, Age_testing)

Age_testing

tree_pred_pruned	**medium**	**new**	**old**
medium	**26**	**7**	**13**
new	**3**	**7**	**1**
old	**4**	**8**	**15**

>mean(tree_pred_pruned!=Age_testing)

[1] 0.4285714

#========== Support Vector Machines (**SVM)==========**

Support vector machines (SMVs) are modeling tools that, as ANNs (Artificial Neural Networks), can be applied to both regression and classification tasks. SVMs have been witnessing increased attention from different research communities based on their successful

application to several domains and also their strong theoretical background. Vapnik (1995, 1998) and Shawe-Taylor and Cristianini (2000) are two of the essential references for SVMs.

#==========

```
tune.out=tune(svm ,Age~., data=training_data, kernel ="linear",ranges =list(cost=c(0.1 ,1 ,10 ,100 ,1000) ))
summary(tune.out)
pred=predict (tune.out$best.model , testing_data)
table(pred, Age_testing)
mean(pred != Age_testing)
```

#==========**Result**

```
>summary(tune.out)
Parameter tuning of ‘svm’:
- sampling method: 10-fold cross validation
- best parameters:
cost
 0.1
- best performance: 0.41
- Detailed performance results:
cost error dispersion
1 1e-01 0.41 0.1197219
2 1e+00 0.44 0.1074968
3 1e+01 0.49 0.1286684
4 1e+02 0.45 0.1080123
5 1e+03 0.48 0.1032796
>pred=predict (tune.out$best.model , testing_data)
>table(pred, Age_testing)
Age_testing
```

pred	**medium**	**new**	**old**
medium	**23**	**13**	**16**
new	**4**	**5**	**1**
old	**6**	**4**	**12**

```
>mean(pred != Age_testing)
```

[1] 0.5238095

==========

Summary (Misclassification Error)

Logistic	**#**	**0.7619048**
LDA	**#**	**0.4285714**
QDA	**#**	**0.6071429**
KNN(3)	**#**	**0.452381**
KNN(5)	**#**	**0.4880952**
KNN(8)	**#**	**0.4642857**
KNN(12)	**#**	**0.4880952**

Classification Tree	**#**	**0.4285714**
SVM	**#**	**0.5238095**

To illustrate the performances of these four classification approaches, data from six different scenarios re generated.

It is observed that LDA and Classification Tree produce the nearest accurate model (minimum misclassification error), hence those considered to be best fit model.
Note: It is advisable in case of response value more than 2, Logistic Regression Model should not be used.

5. CONCLUSIONS

The main goal of this study is to familiarize the reader with java, Oracle and R. For this purpose a small problem is used— at least by data mining standards. Process of performing some of the most basic data analysis tasks in R using Java interface was done. A data base has also been designed to store the data. In terms of data mining, this study has worked on Data visualization, Descriptive statistics, Strategies to handle unknown variable values, Regression tasks, Evaluation metrics for regression tasks, Multiple linear regression, Regression trees, Model selection/comparison through k-fold cross-validation, Model ensembles and random forests. Some techniques used are Loading data from text files, to obtain descriptive statistics of datasets, basic visualization of data, handling datasets with unknown values, some regression models re use to predict the number of cars or best model, using the obtained models to obtain predictions for a test set. It is observed that LDA and Classification Tree produce the nearest accurate model (minimum misclassification error), hence those considered to be best fit model for predicting different types of used cars for evaluation.

Recommendations and Suggestions

- Company can use this technique of finding valuation of vehicles.
- Some other features along with the different types of used can be predicted about vehicles through this technique.
- A user interface can be done through java application that can make things easier for the auctioneers.

REFERENCES

[1] Arora .S, Kaur .P&Arora .P. ”Economical Maintenance and Replacement Decision Making in Fleet Management using Data Mining” ,The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 2.

[2] "Java EE 7 SDK distributions require JDK 7""Java Platform, Enterprise Edition 7 SDK - Installation Instructions". *Installing the Software. Oracle*. Retrieved 10 July 2013.

[3] *Michele. C, Paul. L (2005).* "Process Architecture". *Oracle Database Concepts. Oracle Corporation.* Retrieved 2008-08-13. *A session is a specific connection of a user to an Oracle instance through a user process*

[4] Smith. D (2012); R Tops Data Mining Software Poll, Java Developers Journal, May 31, 2012.

[5] Swartz, D., & Orgill, K. (2001). Higher education ERP: Lessons learned. Educause Quarterly, 24(2), 20-27.

[6] Rexer. K,Allen .H, &Gearan.P. (2011); 2011 Data Miner Survey Summary, presented at Predictive Analytics World, Oct. 2011.

[7] Zaki. M. J., Meira. W.” Data Mining And Analysis”, Library of Congress Cataloging in Publication Data ,ISBN 978-0-521-76633-3.

[8] Pudaruth S**.”** Predicting the Price of Used Cars using Machine Learning Techniques”. International Journal of Information & Computation Technology, Vol. 4.

[9] *Muenchen. R. A. (2012).* "The Popularity of Data Analysis Software".

[10] *Sylvia. T, (29 December 2014).* "Programming tools: Adventures with R". Nature *(517): 109–110.* doi*:*10.1038/517109a.

[11] *Vance, Ashlee (2009-01-06).* "Data Analysts Captivated by R's Power". New York Times. Retrieved 2009-04-28. *R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca.*

[12] Venables, W. N. and Ripley, B. D. (2002). Modern applied statistics with S. fourth edition, Springer.