International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-9 Issue-3, January 2020

Anticipating the Creditworthiness of an Organization using R

Dhanagopal R., Menaka R., Mathana J. M., Eric Clapten J.

Abstract: Numerous start-ups are being created day-by-day. Government also welcomes those by providing funds and loans. As India's economy grows at tremendous pace there is a need for analytical models to help investors track down and predict the performance of industry. Thus, predictive models help us to find and make an informed decision about the financial markets in the future. It allows investors to predict the right shares to obtain profitable investments, banks to invest on repayable customers, mutual funds providers to predict the credit worthiness and shares in order to obtain accuracy about investments and outcomes etc. while there are many models that have been created and perfected by numerous banks and credit rating agencies with their own software tool and data analytics processing there are no such models and systems exists for common retail stock mutual fund investors. This paper mainly focuses on building an open source user friendly model that predict the future performance of concern industry based on the historical records of financial data that is available in BSE/NSE market for various stake holders by focusing on different performance parameters of the concerned company. This prediction is done using R. The Descriptive and predictive models have been created using the financial data collected for more than 3000 companies and tested on accuracy with various statistical methods like ROC.

Keywords: Financial Data, CRA. Bank, R language.

I. INTRODUCTION

The Basel Committee on Banking Supervision (BCBS) defines credit risk as the potential bank borrower a company/organization tends to fail in full filling the repayment constrains accepted with the bank. It comprises of the chances of the borrowers delay as well as failing in repayment of the debt. A credit risk is rise of debt. The risk is for the lender and comprises the unsettled principal and interest, disruption to cash flows, and rise in cost of collection. To maintain the credit risk parameters in acceptable range is to improve the income rate with banks risk-adjusted. The proposed model can be used in following applications.

CREDIT RATING AGENCY(CRA), also called a ratings service is an organization that brands or rates a borrower's capability and ability of repayment from his preceding credit issues (if any) and his timely repayment of debts and also valuing his/her ability by the debtor's livelihood. This branding or rating of a borrower by the organization is known as credit ratings. In accordance with the ratings, one's creditworthiness and loan top up sectors will increase as well as decrease based upon his/her preceding loan briefs as rated by the credit bureaus. Namely credit scores.

The mortgage backed securities and collateralized debt obligations which are rated by debt instruments of CRA's embody government bonds, company bonds, CDs, municipal bonds, preferred shares and collateralized securities. Companies, special purpose entities, state or local governments, non-profit organizations which are the filers of obligations or securities, that affects the price per unit with higher ratings leading to lower interest rates, that a security pays out.

One of the focused trade 'Credit Rating' with the "Big Three" credit rating agencies which controls approximately 95% of the ratings business. Further 15% credit rating agencies in Asian country is controlled by Moody's Investors Service and customary & Poor's (S&P) along management eightieth of the worldwide market and Fitch Ratings. This came into existence only in the later stages of 1980s. As of now, CRISIL, ICRA, CARE, SMERA, Fitch India and Brickwork Ratings are SEBI registered square measure six credit rating agencies below. The character and integrals of the loan provides the ratings of the agencies. If the credit rating is higher it leads to lower rate of interest which is offered to the organization. FORCASTING IN MUTUAL FUNDS & STOCKS: Portfolios treated by world mutual fund consists of two sets of securities, i.e. Domestic and foreign and 2 methodologies won't to measure foretelling ability: domestic differential exposure and assertion rates. Domestic differential exposure is forecasted based on the distinction between each and every fund exposure to the domestic market and also the portfolio exposure to the domestic market where it is outperformed by foreign market. Similar to the differential exposure, Assertion rates measure the ability of fund managers in monthly basis pickup's. The changing economic conditions in both domestic and foreign markets provides huge opportunities to outperform the benchmarks. The output of two empirical tests suggest that exposure to both markets is affected by fund managers. Some evidence is found on a yearly basis based on funds square measurement of excellent foretelling ability.

Revised Manuscript Received on January 05, 2020.

***** Correspondence Author

***Dr. Dhanagopal R.,** Associate Professor, Department of ECE, Chennai Institute of Technology, Kundrathur, Chennai, India.
E-mail: dhanagopal.phd@gmail.com

Dr. Menaka R., Professor, Department of ECE, Chennai Institute of Technology, Kundrathur, Chennai, India.
E-mail: menaka.govindaraj@gmail.com

Dr. Mathana J. M., Professor, Department of ECE, Chennai Institute of Technology, Kundrathur, Chennai, India.
E-mail: jm.mathana@gmail.com

Eric Clapten J., Department of ECE, Chennai Institute of Technology, Kundrathur, Chennai, India. E-mail: eric.shiju@gmail.com

Retrieval Number: B7407129219/2020©BEIESP
DOI: 10.35940/ijitee.B7407.019320

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. LITERATURE SURVEY

Even though many tools are available to support the decision makers, the process of organizing decision making is still at stake to biasing and errors, one of the promising approach for the managers in making well informed and evidence based business decision in Business Analytics (BA). By the means of research papers, two major reasons are stated for the failing of business analytics are Managers are in need for information while performing decision making and Business decisions are often made based on gut feeling and intuition ignoring path or all of the avail data. Author of the research investigate whether the information derived at the time influences the decision makers mind while making business decision, leading to -different decision outcome. The research provides perspective and descriptive decision theory, insights into selective perception and behavior of decision making when giving warning about decision consequences. Based on the research results, the study depicts on how Business analytics can be potentially improved. [1]

For modelling a target of interest on a given set of input variables, modern predictive modeling techniques are commonly used. By this modern techniques input varaiables associated with the target are identified, but not to identify the relation between target and input. because in this scenario observational data is used. Fast and easy collection of data can be done by the latest technology. One of the problem of predictive model method is issuing of data cleansing. This document compares about ten modern predictive modeling techniques over the past 6 years for predicting college graduation rate. The input varaiables used in the modern predictive modeling techniques are ‘pre-college’ performance, ‘first-year’ college performance and variables on various social economic values, variables related to university learning environment. The consequences of quality of data, modeling techniques selection and applying predictive modeling techniques are discussed.[2]

Recently, in bank loans several risks square measure concerned, significantly for the banks therefore on cut their principal loss. The risks analysis and default analysis becomes crucial thenceforth. Banks hold immense information associated with shopper behavior from that they're unable to conclude to a choice. Data processing is employed to analysis the information that aims in extracting valuable information from immense quantity of complicated data blocks. The model is developed supported call tree as base that uses the functions on the market within the R Package. Before developing the model, the dataset is pre-processed, reduced and prepared to produce economical predictions. the ultimate model is used for prediction with the take a glance at dataset and so the experimental results prove the efficiency of the designed model.[3]Accurate financial prediction is of nice interest for investors.

The purpose is to use information analytics in aiding with investors for creating right money prediction in order that right call on investment will be taken. 2 platforms area unit used for operation: Python and R. techniques like Arima, Holt winters, and neural networks (Feed forward and Multi-layer preceptor), regression and statistic area unit enforced to forecast the gap index value performance in R. conjointly in python, Multi-layer instructor and support vector regression area unit enforced for statement dandy fifty stocks. dandy fifty (^NSEI) stock contents area unit thought of as a knowledge input for ways that area unit enforced. Nine years of information is employed. Once examination with the particular value of the stocks, the accuracy was calculated victimization 2-3 years of forecast results of R and a pair of months of forecast results of Python. Mean square error and different error parameters for each prediction system were calculated and it's found that feed forward network solely produces 1.81598342% error once gap value of stock is forecasted victimization constant method [4].

Machine learning may be a growing technique for building analytic models for machines to “learn” from information and be ready to do prognostic analysis. The power of machines to “learn” and do prognostic analysis is extremely vital during this era of massive information and it's a good vary of application areas. i.e. banks and money establishments area unit usually moon-faced with the challenge of what risk factors to think about once motility credit to customers. For many features/attributes of the purchasers area unit usually taken into thought, however most of those options have very little prognostic result on the credit goodness or of the customer's goodness upon his goodwill. Further, a sturdy and effective automatic bank credit risk score which will aid within the prediction of client credit goodness terribly accurately continues to be a serious challenge moon-faced by several banks. We tend to examine real bank credit info and conduct several machine learning algorithms on {the information the knowledge} for comparative analysis and to make your mind up thereon algorithms area unit the foremost effective applicable learning bank credit data. The algorithms gave over eightieth accuracy in prediction. Moreover, the foremost vital options that confirm whether or not a client can default or otherwise in paying his/her credit future month area unit extracted from a complete of twenty three options. We tend to then applied these most vital options on some chosen machine learning algorithms and compare their prognostic accuracy with the opposite algorithms that used all the twenty three options. The results show no important distinction, signifying that these options will accurately confirm the credit goodness of the purchasers. we tend to tend to formulate prophetical model exploitation the foremost important choices to predict the credit goodness of a given shopper.[5] to formulate banks’ risk automatic system.

III. EXISTING MODEL ANALYSIS

Data Analytics though has been there for many years recent advancement in technology like cloud services, new computational language and lower hardware cost has opened many doors. Until now, predictive models were designed and used only by western banks and credit rating agencies like Goldman Sachs, Jp Morgan etc. Several banks and money establishments give a excess of monetary (as well as insurance) merchandise and services, and function one stop buy customers.

Given the one stop searching format, money services corporations ask for to cross sell their merchandise and services to customers. Despite, the supply of knowledge concerning their existing customers, few corporations actively leverages this data to raised style their product and repair offerings. But none of these models are building in an open source manner with a widely accepted technology and packages that can used even by an amateur retail investor, stock brokers traders and mutual fund managers. With BSE and NSE moving so fast into digitalization of financial, market to such extend that even automated algorithmic trading is happening in Mumbai, there is a necessity to build such predictive model with mostly widely user-friendly tool that can serve the purpose of retail investors.

IV. PROPOSED MODEL

Predictive modeling strategies usually deliver a lot of correct predictions of risk than ancient scorecards. With a lot of correct predictions of risk, a lot of credit will be extended to a lot of candidates, whereas maintaining or perhaps reducing the general default rate. Thus, prophetical models will increase profits. Whereas moving from a standard model to prophetical model accuracy will be considerably improved associated additionally will increase profits whereas decreasing risk for an capitalist. The proposed model has been designed to be suitable more for retail investors and financial startups who want to automate and manage their investment decisions over thousands of companies and options available in financial market. With stock brokers like Zerodha providing API and infrastructure facilities for retail investors to do algorithm based automatic trading, small time investors can start using open language like R to build predictive models as proposed here and come up with informed decision before investing especially when one needs to make a decision in a split of a second based on movement in financial markets.

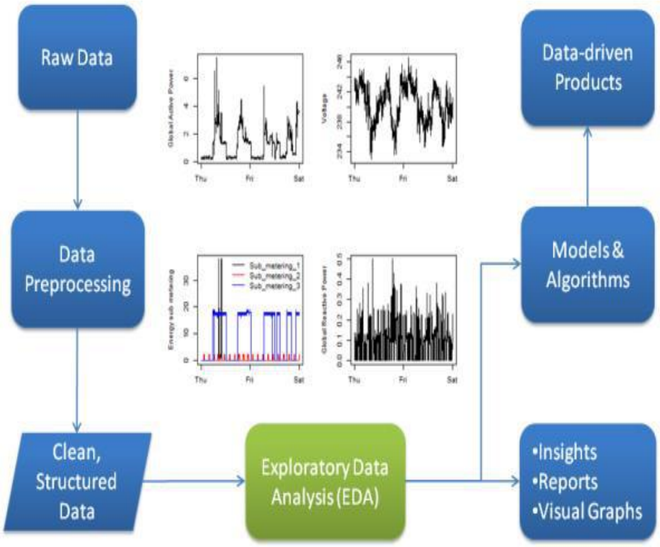


Fig. 1.1.Data Analytics process flowchart

V. MODEL DESIGN

Analytics usually involves:

➢ Identify the problem or opportunity for value creation.

➢ Identify sources of data (primary as well secondary data sources).

➢ Pre-process the data for issues such as missing and incorrect data. Generate derived variables and transform the data if necessary. Prepare the data for analytics model building.

➢ Divide the data sets into subsets training and validation data sets.

➢ Build analytical models and identify the best model (s) using model performance in validation data.

➢ Implement solution/decision/develop product.

The type of analytics to be followed is a supervised learning algorithm i.e: when the coaching knowledge set has each predictors (input) and outcome (output) variables, we have a tendency to use supervised learning algorithms wherever the training is supervised by the actual fact that predictors (x) and therefore the outcome (y) square measure accessible for the model to use. Techniques like regression, logistical regression, call learning, random forest so square measure supervised learning algorithms.

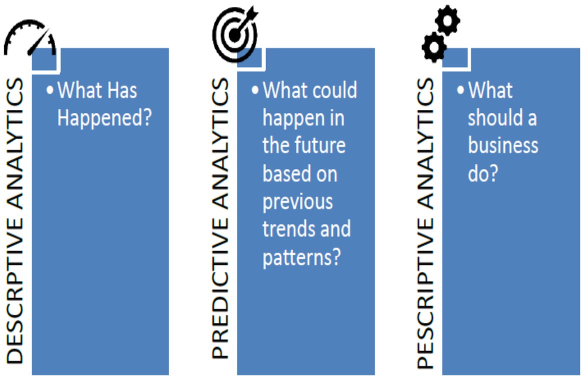

DESCRPTIVE ANALYTICS
• What Has Happened?
PREDICTIVE ANALYTICS
• What could happen in the future based on previous trends and patterns?
PESCRIPTIVE ANALYTICS
• What should a business do?

Fig. 1.2. Different categories of analytics

Data collection & pre-processing:

The conversion of information to valid data through a process is called data processing.

Data is manipulated to provide results that end in a resolution of a haul or improvement of associate existing state of affairs. almost like a production methodology, it follows a cycle wherever inputs (raw data) a$_r$ fed to a method (computer systems, software, etc.) to supply output (information and insights).The process includes activities like information collection and entry, summary of the collected information, calculation based on the requirement and storage of the information, etc. helpful and informative output is given in varied acceptable forms like diagrams, reports, graphics, etc. The data of more than 3000 companies has been collected from the financial statements available in BSE/NSE website and has been reentered into excel for pre-processing.

Num
Default
Random
Is dev data
Networth Next Year
Total assets

Net worth
Total income
Change in stock
Total expenses
Profit after tax
PBDITA
PBT
Cash profit
PBDITA as % of total income
PBT as % of total income
PAT as % of total income
Cash profit as % of total income
PAT as % of net worth
Sales
Income from finical services
Other income
Total capital
Reserves and funds
Deposits (accepted by commercial banks)
Borrowings
Current liabilities & provisions
Deferred tax liability
Shareholders' funds
Cumulative retained profits
Capital employed
TOL/TNW
Total term liabilities / tangible net worth
Contingent liabilities / Net worth (%)
Contingent liabilities
Net fixed assets
Investments
Current assets
Net working capital
Quick ratio (times)
Current ratio (times)
Debt to equity ratio (times)
Cash to current liabilities (times)
Cash to average cost of sales per day
Creditors turnover
Debtors turnover
Finished goods turnover
WIP turnover
Raw material turnover
Shares outstanding
Equity face value
EPS
Adjusted EPS
Total liabilities
PE on BSE

List of performance parameters taken for every company to be analyzed

Preparation of data involves the manipulation of knowledge into a kind appropriate for additional analysis and process. Data can not be processed and will be checked for accuracy. Preparation is regarding constructing a data set from one or further information sources to be used for any exploration and method. Analyzing data that has not been painstakingly screened for problems can prove extraordinarily dishonorable results that unit heavily obsessed on the quality of knowledge prepared. This method has been dead mistreatment the stand out with the assistance of kurtosis tool.1.4.3 Descriptive Analytics:

90% of organizations these days use descriptive analytics that is that the most simple type of analytics, analyses the data in real-time with its preceding data (historical data) for insights on how to approach the future. The main objective of descriptive analytics is to find out the cause behind success or failure in the past. The 'Past' here, refers to any particular time in which an event had occurred and this could be a month ago or even just a minute ago. The majority of big data analytics used by organizations falls into the category of descriptive analytics. Descriptive analytics is the simplest form of analytics that mainly uses simple descriptive statistics, data visualization techniques and business-related queries to understand past data. one of the primary objectives of descriptive analytics is innovative ways of data summarization. Descriptive analytics is used for understanding the trends in past data which can be useful for generating insights. So, based on intuition of performance parameters, they have been classified into positive and negative in order to understand the effect of such parameter on the company performance as follows. Apart from this, the statistical summary of all parameters has also been calculated using the R package and the code for the same.

Types of data measurement scales:

Structured data can be either numeric or alpha numeric and may follow different scales of measurement(levels).it is important to understand the type of variables within the data with respect to the measurement scale since the model specification while building analytics models such as regression may depend on the scale of measurement.

Nominal Scale (qualitative data):

It refers to variables that are basically names and also known as categorical variables. For example, variables such as marital status and industry type fall under nominal scale. During data collection, it is usual to assign a numerical code to represent a nominal variable. While developing a regression model, categorical variables are converted using dummy variables before building the regression models.

Measurement of central tendency:

It is the measures that are used for describing the data using a single value. Mean, median and mode are the three measures of central tendency and are frequently used to compare data sets. Measures of central tendency help users to summarize and comprehend the data.

Measures of variation:

One of the primary objectives of analytics is to understand the variability in the data. Predictive analytics techniques such as regression attempt to explain variation in the outcome variable(Y) using predictor variables(X). Variability in the data is measured using the following:

1. Range
2. Inter – Quartile Distance (IQD)
3. Variance

4. Standard deviation.

VI. PREDICTIVE ANALYTICS

Predictive analytics helps in predicting the end result of a future by utilizing numerous applied mathematics and machine learning algorithms. However the accuracy of predictions is not 100%, as a result of it's predicated on prospects. to form predictions, algorithms take info and fill among the missing info with very best guesses. This info is pooled with historical info gift among the CRM systems, POS Systems, ERP and hour systems to appear for information patterns and determine relationships among numerous variables within the dataset. Organizations ought to maximize hiring a gaggle of information scientists in 2016 World Health Organization will develop applied mathematics and machine learning algorithms to push prognostic analytics and style an efficient business strategy. Prognostic analytics will be any categorized for Forecasting- What if the present trends continue? Organizations like Walmart, Amazon and alternative retailers leverage prognostic analytics to spot trends in sales supported purchase patterns of consumers, foretelling client behavior, foretelling inventory levels, predicting what product customers' unit of measurement most likely to induce on so as that they're going to provide bespoken recommendations, predicting the quantity of sales at the tip of the quarter or year. The simplest example where prognostic analytics notice a nice application is in producing the credit score. A credit score helps financial institutions to decide the possibility of a consumer paying credit bills on time. Within the analytics capability maturity model (ACMM), prognostic analytics comes when descriptive analytics and is the most vital analytic capabilities. It aims to predict the chance of prevalence of a future event like foretelling demand for products/services.

Introduction to correlation:

One of the challenging tasks in analytics, especially in predictive analytics is identifying the variables or features that may be associated to the response variable or the outcome variable that is of interest to the data scientists. Organizations collect data on several variables, sometimes the number of variables can run into thousands (including derived variables such as ratios and interactions).

Classification problems:

It is an important category of problems in analytics in which the response variable(Y) takes a discrete value.in classification problems, the primary objective is to predict the class of a customer (or class probability) based on the values of explanatory variables or predictors.

Binary logistic regression:

Logistic regression is a statistical model in which the response variable takes a discrete value and the explanatory variables can either be continuous or discrete. Logistic regression is one of the supervised learning algorithms. We will be discussing binary logistic regression in which the response variables take only two values. For example, assume the value of Y is either 1(positive outcome) or 0(negative outcome). When there are more than two values of Y, then multinational logistic regression model is used.

$$Ln(P(Y=1)/1-P(Y=1))=z=\beta 0+\beta 1X1+\ldots+\beta mXm \qquad (1)$$

Decision tree:

It is a collection of predictive analytics techniques that use tree-like graphs for predicting the values of a response variable (or target variable) based on the values of explanatory variables (or predictors).it is one of the supervised learning algorithms used for predicting both the discrete and the continuous dependent variable.in a decision tree learning, when the response variable takes discrete values then the decision trees are called classification trees. These are effective for solving classification problems in which the response variable (target variable) takes discrete values. Decision trees employ divide – and-conquer strategy in which the original data is divided into multiple groups or subsets, and the strategy is to establish groups such that within groups the data is homogeneous. This means that the data in the group is dominated by one class. Decision tree use following criteria:

Splitting criteria: it is used to split a node (set of data) into subsets.

Merging criteria: when the predictor variable is categorical with n categories, it is possible that not all n categories may be statistically significant. Thus, few categories may be emerged to create a compound or aggregate category.

Stopping criteria: stopping criteria is used for pruning the tree to reduce the complexity associated with business rules generated from the tree. Usually levels (depth) from root node, minimum number of observations in a node for splitting are used as stopping criteria.

The following steps are used for generating decision tree:

1. Start with the root node in which all the data is present. Decide on a splitting criterion and stopping criteria: the root node is the split into two or more subsets leading to tree branches (called edges) using the splitting criterion. Nodes thus created are known as internal nodes. Each internal node has exactly one incoming edge.
2. Further divide each internal node until no further splitting is possible or the stopping criterion is met. The terminal nodes (aka leaf nodes) will not have any outgoing edges.
3. Terminal nodes are used for generating business rules.
4. Tree pruning (restricting the size of the tree) is used to avoid large trees and over fitting the data.it is achieved through different stopping criteria.

Random Forest:

It is one of popular ensemble method in which several trees are developed using different sampling statergies.one of the most frequently used sampling strategy is the bootstrap Aggregating (or bagging). Bagging is a random sampling with replacement. A new observation is classified by using all the trees developed in the random forest and majority voting is used for deciding the class. They are developed using:

Assume that the training data has N observations.one need to generate several samples of size M(M<N) with replacement (called bagging).let the number of samples based on sampling of the training data set be S1. If the data has n predictors ,sample m predictors (M<N)

➢ Develop trees for each of the samples generated in step 1using the sample of predictors from step 2 using CART.

➢ Repeat step 3 for all the samples generated in step 1.

➢ Predict the class of a new observation using majority voting based on all trees.

Validation of Predictive model:

Sensitivity, specificity, and precision:

In logistic regression, the model performance is often measured using concepts such as sensitivity, specificity and precision. The ability of the model to correctly classify positivity's and negativities are called sensitivity and specificity, respectively. The terminologies sensitivity and specificity originated in medical diagnostics. In medical diagnostics, sensitivity (true positive rate) measures the ability of a diagnostic test to identify disease if it is present in a patient (test positive). That is,

Sensitivity = P (diagnostic test is positive| patient has disease)

In generic case,

Sensitivity=P (model classifies Yi as positive | Yi is positive)

Sensitivity is calculated using the following equation:

Sensitivity =true Positive (TP)/True positive (TP) + False Negative (FN) Where true positive (TP) is the number of positives correctly classified as positives by the model and False negative (TN) is positives misclassified as negative by the model. sensitivity is also called recall. Specificity is the ability of the diagnostics test to correctly classify the test as negative when the disease is not present. That is,

Specificity=P (diagnostic test is negative | patient has no disease)

In general,

Specificity = P (model classifies Yi as negative | Yi is negative)

Specificity can be calculated using the following equation:

Specificity=True Negative (TN)/True negative (TN) + false positive (FP)

Credit score using logistic regression:

Many credit rating organizations use a range of score (say between two values A and B) called credit score to measure the credit worthiness of a customer. For example, one of the popular credit scores, FICO (Fair, Isaac and Company)has a range between 300 and 850.FICO score is calculated using multiple parameters such as payment history, debt burden, length of credit history, type of credit, etc.

A logistic regression model can be used for generating a credit score between any two values A and B using: Credit score of customer +A+ (B-A) * [1-P(Y=1)]

For example, if A=300 and B=850, then the credit score is given by

Credit score of customer =300+550*[1-P(Y=1)].

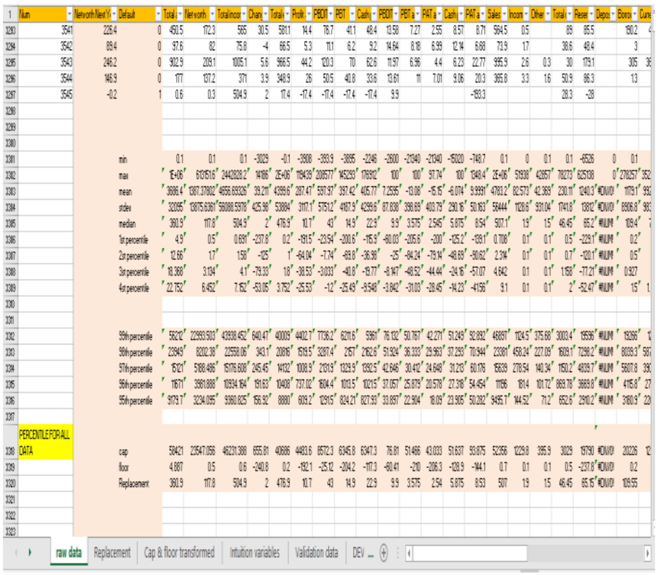
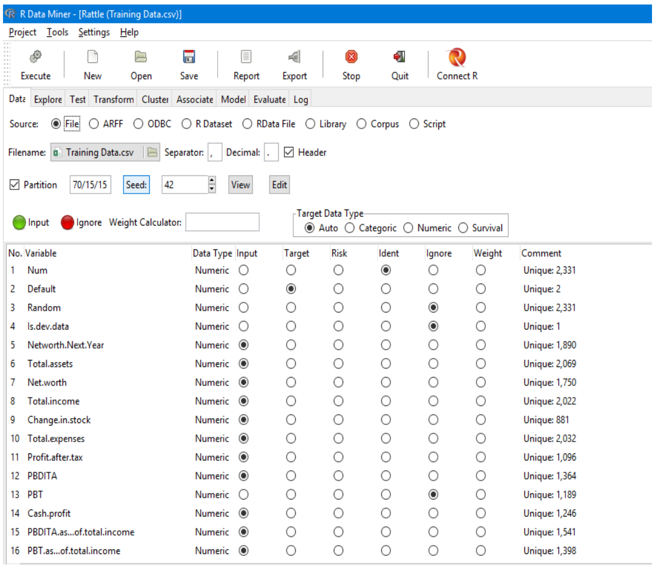


Fig. 1.3. Data Cleaning Process in Microsoft Excel



R Data Miner - [Rattle (Training Data.csv)]

Project Tools Settings Help

Execute New Open Save Report Export Stop Quit Connect R

Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: Training Data.csv Separator: , Decimal: . ☑ Header

☑ Partition 70/15/15 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Num	Numeric	○	○	○	●	○	○	Unique: 2,331
2	Default	Numeric	○	●	○	○	○	○	Unique: 2
3	Random	Numeric	○	○	○	○	●	○	Unique: 2,331
4	Is.dev.data	Numeric	○	○	○	○	●	○	Unique: 1
5	Networth.Next.Year	Numeric	●	○	○	○	○	○	Unique: 1,890
6	Total.assets	Numeric	●	○	○	○	○	○	Unique: 2,069
7	Net.worth	Numeric	●	○	○	○	○	○	Unique: 1,750
8	Total.income	Numeric	●	○	○	○	○	○	Unique: 2,022
9	Change.in.stock	Numeric	●	○	○	○	○	○	Unique: 881
10	Total.expenses	Numeric	●	○	○	○	○	○	Unique: 2,032
11	Profit.after.tax	Numeric	●	○	○	○	○	○	Unique: 1,096
12	PBDITA	Numeric	●	○	○	○	○	○	Unique: 1,364
13	PBT	Numeric	○	○	○	○	●	○	Unique: 1,189
14	Cash.profit	Numeric	●	○	○	○	○	○	Unique: 1,246
15	PBDITA.as...of.total.income	Numeric	●	○	○	○	○	○	Unique: 1,541
16	PBT.as...of.total.income	Numeric	●	○	○	○	○	○	Unique: 1,398

Fig. 1.4.Basic statistics for each numeric variable of the dataset

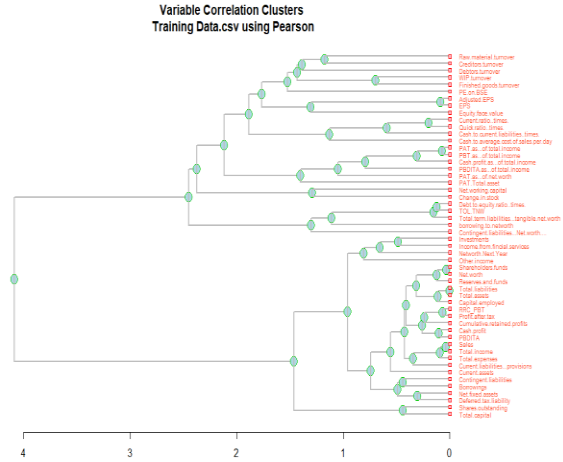


Fig. 1.5.Variable correlation clusters training data using Pearson

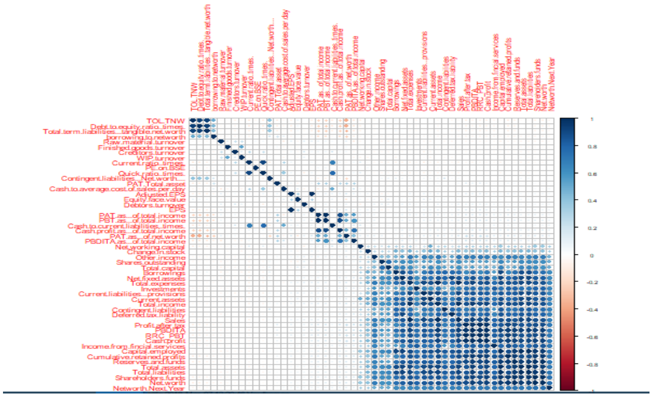


Fig. 1.6. Summary of the Decision Tree model for Classification (built using 'rpart')

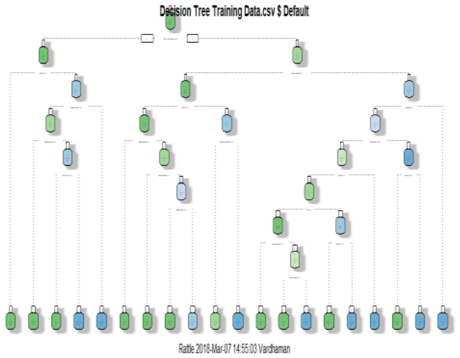

Decision Tree Training Data.csv $ Default
Rattle 2018-Mar-07 14:55:03 Vardhaman

Fig. 1.7. Error matrix for the Decision Tree model on Training Data.csv (counts)

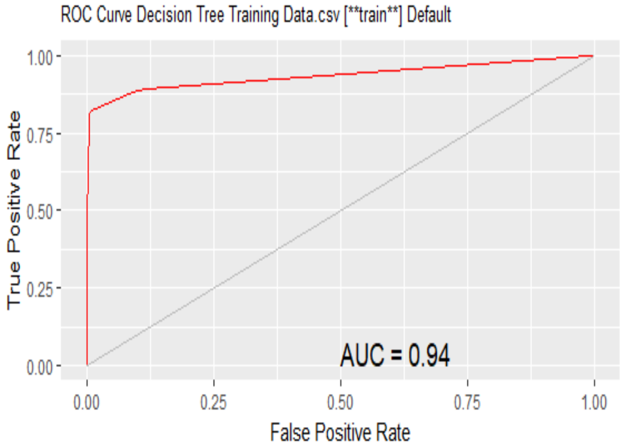


Fig. 1.8. Area under the ROC curve for the rpart model on Training Data.csv is 0.9358

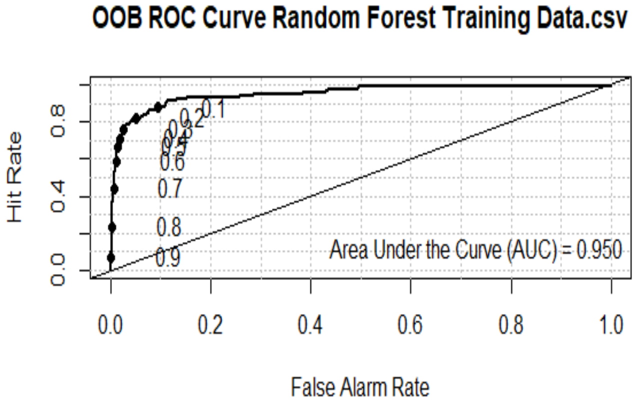


Fig. 1.9. Error matrix for the Decision Tree model on Training Data.csv (counts)

VII. CONCLUSION

Based on the training data set and validation criteria the model created using R has shown an accuracy rate of 94%. Using inferior quality data is one problem that needs to be considered very carefully before using models built using R. Anticipating human behavior is once computers and algorithms fail to contemplate variables from dynamical weather to moods to relationships which may influence selections. This is where Big data integration can come in handy and in future Prescriptive analytics which is the successor of predictive analytics can suggest possible outcomes and leads to actions that area unit seemingly to expand key business metrics .

Prescriptive analytics is a sophisticated analytics conception supported optimization that helps accomplish the most effective outcomes and random optimization that helps perceive a way to accomplish the most effective outcome and determine knowledge uncertainties to create higher selections.

REFERENCES

1. Dursun Delen and Gregory Moscato(2018),” The Impact of Real-Time Business Intelligence and Advanced Analytics on the Behaviour of Business Decision Makers”, 2018 International Conference on Information Management and Processing.
2. Emma A. Gunu, Carl Lee, Wilson K. Gyasi and Robert M. Roe(2017),” Modern Predictive Models for Modeling the College Graduation Rates”, IEEE SERA 2017, June 7-9, 2017, London, UK
3. Kosemani Temitayo Hafiz, Dr. Shaun Aghili, and Dr. Pavol Zavarsky(2016),” The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada”, Department of Information Systems Security and Assurance Management. Concordia University of Edmonton Edmonton, Canada.
4. Sudhamathy G. and Jothi Venkateswaran C(2016),” Analytics Using R for Predicting Credit Defaulters”, 2016 IEEE International Conference on Advances in Computer Applications (ICACA).
5. Shashank Tiwari, Akshay Bharadwaj and Dr. Sudha Gupta(2017),” Stock Price Prediction Using Data Analytics”, Dept. Of Electronics, K J Somaiya College of Engineering Mumbai, India.
6. Regina Esi Turkson, Edward Yeallakuor Baagyere,and Gideon Evans Wenya(2016),” A Machine Learning Approach for Predicting Bank Credit Worthiness”, University of Electronic Science and Technology of China.
7. Mitchell T M(2006),’The Discipline of Machine Learning”,Carnie Mellon University Report.
8. Siegel E(2013),”Predictive Analytics:The Power to predict who will Click, Buy, Lie or Die”,John Wiley and Sons,Hoboken,NJ.
9. Breiman L,Friedman J H, and Olshen R A and Stone C J(1984),”Classification and Regression Trees”,Chapman And Hall,USA.
10. Kass G V(1980),”An Exploratory Technique for Investigating Large Quantities of Categorial Data,” Applied Statistics,20(2),119-127