A Data Mining Approach: Classification and Regression Trees (CART) for the Determinants of Earnings for Pakistan

By

NEELAM YOUNAS

Department of Econometrics and Statistics

Pakistan Institute of Development Economics, Islamabad

2015

A Data Mining Approach: Classification and Regression Trees (CART) for the Determinants of Earnings for Pakistan

A Research Dissertation submitted to the Pakistan Institute of Development Economics (PIDE), Islamabad, in partial fulfillment of the requirements for the award of the degree of Masters of Philosophy in Econometrics.

Submitted by

NEELAM YOUNAS

16/M.Phil-ETS/PIDE/2013

Supervised by

DR. ZAHID ASGHAR

DEDICATION

All dedications go to my mother and sisters

My inspiration, my world!

ACKNOWLEDGEMENT

First and foremost to my best friend, my Allah, the most loving and merciful, who created the universe with beauty and perfection and gave me the ability to ponder and refine my thinking. I offer my humblest feelings and thanks to the prophet Muhammad (peace be upon him), who is forever the source of guidance and knowledge for the humanity as a whole.

There is a debt of gratitude I can't repay to my adviser, Dr. Zahid Asghar for his superb guidance and encouragement during the whole course of research. His wise counseling has always stimulated my mind and makes me truly understand what this manuscript is all about.

It is hard to express my feelings for my family, all my accomplishments whether large or small, is possible because of my parent's love, support and constant prayers

Neelam Younas

Table of Contents

CHAPTER 1

"Big data is crude oil…

But you need to refine the crude oil,

Enter Data science"

(Carlos somohano)

INTRODUCTION

1.1 The Background

Every day, we create 2.5 quintillion bytes of data, so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere like sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data. "Big Data" refers to a combination of an approach to informing decision making with analytical insight derived from data, and a set of enabling technologies that enable that insight to be economically derived from at times very large, diverse sources of data. Big data has hit the business, government and scientific sectors. Big data applies to information which cannot be handle by traditional techniques. Big data is not only about the volume of data it's about importance (think big deal).Four V's gives the definition of the big data, Volume, velocity, veracity and Variety. Volume is the size of the data, velocity is the speed of the data by which the data generates, veracity is the noise in the data and variety is the different source of the data generation.

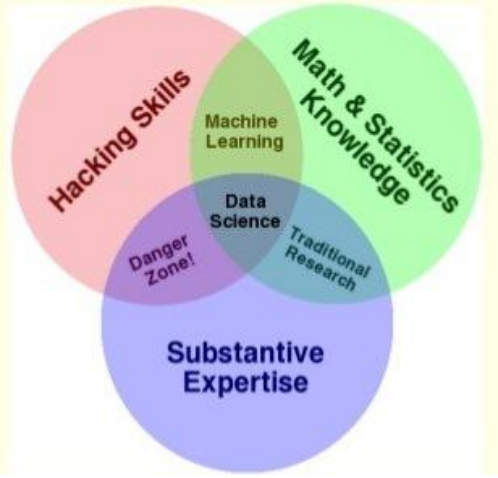
Now-a-days computers are used in many economic activities and these activities create large sets of data which we can manipulate and analysis. Econometrics and statistical techniques such as regression, work well for small data but we face problems while dealing with big data. For big data we need more powerful data manipulating tools. Large data sets may allow complex relationships than simple linear model. For modeling the complex relationship of big data we use

Machine learning techniques. Machine learning is a semi-automated extraction of Knowledge from data; starts with a question that might be answerable using data .Automated extraction mean a computer provides the insight. Semi-automated mean it requires many smart decisions by a human. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning. Machine learning is mainly concerned with prediction, which is one of the fields of data mining. Econometrician, statisticians and data mining specialist are looking for insights that can be extracted from data.

Statistics can be defined as the science of "learning from data" or making sense out of data, where data science is not just a rebranding of statistics, large scale statistics or statistical science.

The term "data science" and "data mining" often use interchangeably. The key words describing the field have changed from "knowledge discovery" to "data mining" to "predictive Analytics" and now to "data science".

Figure 1.1 Venn diagram of data science

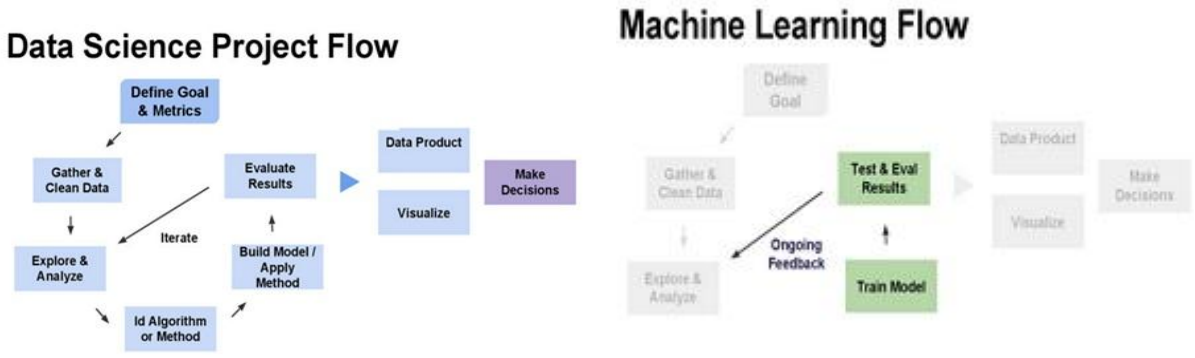


Source: A statistician's view on big data and data science by Diego kuonen.

Data science is the scientific study of the creation, validation and transformation of data to get meaning from the data. In 2001, the statistician S. Cleveland introduced the notion of data

science as an independent discipline. Cleveland extended the field of statistics to incorporate "advance in computing with data". Data science is concerned with prediction, summarization and data manipulation. Econometrians usually use linear regression analysis for detecting and summarizing relationships in data where Machine learning offer tools for summarizing nonlinear relationships in data. The purpose of Machine learning is to find some function which give good prediction for y as a function of x, where in statistical prediction statistician focus on conditional distribution of y given some other variables x. Statistics can be defined as the science of "learning from data" or making sense out of data, where data science is not just a rebranding of statistics, large scale statistics or statistical science.

Figure 1.2: Flow chart of data science and Machine Learning



Source: Melanie Warrick: How to get started with machine learning

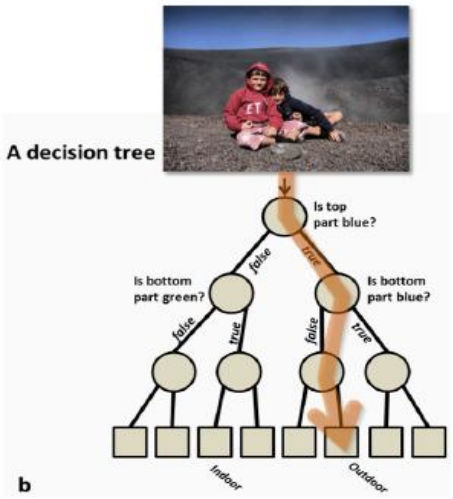
In machine learning as we have dig data set so to get better prediction we divide the data in sets which are used for training, testing and validation. We estimate the model through training data; choose our model through validation data and model is evaluated using testing data. Dealing with problem of good prediction of y for new values of x, which minimize loss function, economist would go for linear and logistic regression while there are better options than these techniques in Machine learning, if we have a lot of data. These nonlinear methods are Classification and Regression tree (CART), Random Forest and penalized regression which gives better out-of-sample forecast.

Here we will discuss one technique of Machine learning in detail, which is decision tree. Decision tree is a flow-chart-like structure which is used for segmenting or stratifying the predictor space in to a number of regions or subsets, to make prediction for a given value, mean and mode of the training data set is used. The set of splitting rules used to segment the predictors space can be summarized in a tree, this approach is referred as Decision tree methods.

1.2 Importance

CART methodology was introduced in 1984 by Leo Breinman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer the following type of the decision trees. CART Algorithm is used for classification and regression tree. It is structured as a sequence of questions the answer to which determines what the next question should be. The results of the sequence of questions make a tree. It ends a terminal node at which there are no more questions**.** (Gordon 2013).

Figure 1.3: Decision making through decision tree



Source: From criminisi et al.MSR-TR-2011-114

Trees perform better than linear model where the response variable and predictors have nonlinear and complex relationship but the relationship of the response variable and predictors is linear then tree based method perform like linear regression it does not exploit it linear relation. The aim of regression analysis is to discover the relationship between the response variable and the predictor variables, and eventually to use the relationship to make predictors based on the

information. After a tentative model is fitted, one can assess how well the model fits and modify it to improve the prediction. In this process, it is very much important to decide which variables are to be included or removed in the model. If there are many irrelevant variables, the variable selection procedure may play a much more important role in the prediction power. Doksum, Tang, and Tsui (2006) point out this problem and proposed a solution called EARTH, which is based on the conditional expectation of the response given all but one of the predictor variables. This study consider regression tree algorithm as an alternative approach.

The performance of tree based method and linear regression can be assess through test error where test error is estimated through cross validation or validation set. If the pictorial presentation of the model is required than we go for tree based methods. CART technique has used a lot in public health and finance but now a days used in economics.

1.3 Purpose of the study

The distribution of the earnings is an important issue for public policy especially when income distribution is skewed. To find the cause of difference in earnings of an individual or to find the determinants of earnings of individual whether personal characteristics play important role in effecting the earning of an individual of labor market characteristics. Once we determine the factors effecting the earning of an individual, it's not only to improve the socio end economic conditions of individual but overall the distribution of income in a country (Nasir 1998)

Different researches estimated earning function for different countries using conventional techniques of estimation i-e OLS and multiple linear regressions(Nasir 1987) investigated determinates of earning in Pakistan using ordinary least square techniques which required a lot of assumptions regarding variables, model and error terms. Functional form of the variables was required. According to our best knowledge, no one has estimated determinants of earning using classification and regression technique.

The main problem in linear regression is that the attributes must be numeric so that the model obtained will also be numeric, simple equation in a dimension. As a solution to this problem, decision trees have been used in data mining for long time as a supervised learning technique (model is learned from the data).Most statistical work starts from the specification of a model .The model say, how we believe the data is generated and contains both a systematic and a

random component. The model is not completely specified and so we use the data to select a particular model by either estimating parameters or perhaps by fitting functions. Clearly this strategy has been effective in a wide range of situation but it is often impractical to develop inferential methods for more complex statistical model.

This study has used CART for finding the determinants of earning because of its interesting features. The purpose of using regression and classification tree (CART), unlike simple regression its fit the model at each splitting node of the tree. Where simple linear regression fit one model for the complete set of data.

The Statistical earning function is given as follows, Lny_i=$f(s_i, x_i, z_i)$ + u_i , lny_i: is the log of earning ,s_i : is schooling ,x_i : is experience ,z_i : Represents other factors affecting earning such as race, gender or geographical region of individual. u_i : is the disturbance term assumed to be normally distributed (Berndt 1991).

1.4 Objective of the study

- *Estimating determinants of earnings through Classification and regression trees (CART).*
- *How to improve the results obtained from CART using bagging, boosting and random forest.*
- *How the results of Classification and regression trees differ from the usual regression or least square technique.*

1.5 Significance of the study

The focus of this research study is to determine the factors affecting earning of an individual. Unsurprisingly, so far there is no other same research carried out with the similar purpose in the context of CART. Many models are good for in-sample prediction but perform poorly for out of sample prediction. Economists say because of small sample it's not possible to divide the data into training and testing. If large data set is available one must separate data in sets of training and testing, which give better out of sample prediction because we can validate and evaluate the chosen model. In this study we will divide the data set into testing and training data which may give better prediction. CART does not give us the on average estimate like usual regression technique. Instead it makes a decision tree depending on sequence of questions that's why it is a white box technique. It is a pictorial representation so one gets a clear idea of the significance of

each variable. In stage wise CART we do not need to choose significant variables in advance like usual regression technique.

1.6 Organization of the study

We provide a brief review of the literature in Chapter 2 which focuses on the relationship between earning and various predictors of earning and literature of classification and regression tree. Then, Chapter 3 and 4 represent methodology and data to get analysis by using this methodology. Chapter 5 shows results and finally, we conclude in Chapter 6.

Chapter 2

Literature review

2.1 Introduction

Before we proceed with our study, it is important to have a broad idea of the current developments in the theoretical and empirical literature on determinants of earning and Classification and regression tree. The aim of this chapter is thus to review major studies that have been done so far in this field. A large body of literature is available on this subject that theoretically and empirically analyzes the determinants of earning and classification and regression trees. The literature reviewed in this chapter has been chosen for its relevance to the proposed research in this thesis.

The rest of the chapter is divided into two sections. In section 2.2, we highlighted the Application of CART in different fields. A detailed discussion on the advantages and disadvantages of CART is also presented in this section. In section 2.3, we reviewed the empirical studies investigating the determinants of earning for Pakistan and other different countries. Section 2.4 concludes the whole discussion.

2.2 Literature review for the classification and regression tree (CART)

Regression trees are similar to additive models in that, the compromise between the linear model and the completely nonparametric approach is represented. Tree methodology has roots in both statistics and computer literature. A precursor to current methodology was "CHAID" developed by Morgan and Songuist (1963) although the book by Breiman, Friedman, Olshen, and Stone (1984) introduced the main idea to statistics. Tree methodology was developed in machine learning in the starting of 1970s.Tukey (1977) advocated explanatory data analysis (EDA) in his book. Graphical and descriptive statistics can sometimes make the message of data very clear or at least suggest a suitable form for a model, but EDA is not the solution. Regression trees are an example of statistical methods that is best describes by the algorithms used in their construction.one can uncover the explicit model underlying regression trees.

CART is a method of classification to construct decision tree using historical data. Then decision tree is used to classify new data for making prediction. To build a decision tree, CART uses learning sample - a set of historical data (Timofee 2004).CART is non-parametric approach which does not rely on the distribution of data set. It is not affected by outliers (Sutton 2005).CART uses testing with testing data and cross validation to assess the goodness of fit of model. It uses same variables more than once in different part of the tree, to uncover the complex interdependencies among variables. Tree based methods are simple and useful. CART is model free. No distribution assumptions are required and treat data generating process as unknown. No functional form of predictors is required. Tree methods are very easy to interpret. It's give simple and powerful analysis.

CART does not give average effect of predictors like traditional approach of regression. Its divide the data into sub groups which allow the identification of groups of interest. It is highly efficient in dealing with high dimension data and its flexibility to deal with missing value. It deals with missing values in two ways either its make it a new category of missing values or make a surrogate variables and at the time of split if the original variable is missing that variable is used (Gordon 2013).

Different algorithms for classification tree are, THAID was the first published algorithm for classification tree, which uses a measure of node impurity based on the distribution of Y values in the node.it search X and S for split$\{X \in S\}$which minimizes, which minimizes the total impurity in two child node. The process is applied recursively to each child node .stop splitting if there is a relative decrease in impurity is below a prespecified threshold .C4.5 and CART are other methods of classification tree .C4.5 uses Entropy as a measure of impurity and CART uses Gini index. Gini index is the generalization of binomial variance.C4.5 is used when a categorical variable has more than two categories. CRUISE, GUIDE and QUEST algorithms use two steps approach, based of test of significance of split. First most significance variable is selected through test for split and then search for S is performed. GUIDE and CRUISE use chi squared test while QUEST uses chi squared for unordered variables and ANOVA tests for ordered variables CHAID uses significance and Bonferroni correction to try to merge iteratively pairs of child nodes (Wei-Yin Loh 2011).

Algorithms for Regression tree, AID was the first regression tree algorithm .M5 is a regression tree algorithm by Quinlan .it first make a piecewise constant tree then fits a linear regression model to the data at each node. GUIDE uses classification tree techniques to solve regression problem.at each node it fits the regression model and find the residuals (Wei-Yin Loh 2011).

Comparison was made between logistic regression and Trees using titanic data set. Age and class were used as predictors. Age was barely important in logistic regression where in trees showed it's an important variable for survival. Logistic regression performs well only for small data set where trees are better for big data (Varian 2014).

Many researchers used CART for public health data to develop clinical decision rule. This can be used for new patients to classify in to categories .Traditional statistical techniques are hard to apply because there are many potential variables so variable selection becomes difficult. Many variables are not normally distributed and clinical decision requires large data (Roger et.al 2000). Growing a big tree mean that split the predictor space till last observation. Over grow trees perform poorly, then two approaches are used in cutting the insignificant nodes, one is optimization by number of points in each and second approach is cross validation. Among other important features of CART, trees do not change if any of variables is change by its square root or logarithm (Timofee 2004). CHAID (Chi-Square automated interaction detection) and AID (automated interaction detection) algorithms were used for trees and its loss function expressed in terms of a goodness of fit statistics- the proportion of reduction in error. For regression multiple R^2 is used as a loss function. Gini rule is used in classification and squared residuals minimization rule is used in Regression trees. Classification trees are parallel to linear discriminant analysis where Regression trees are parallel to regression/ANOVA. For regression trees least square, trimmed mean and least absolute deviations are used as a loss function (Wilkinson 2004).

AID method is used for fitting tree, which deal interaction naturally. AID (automated interaction detection was the first algorithm used for decision tree which deals with interaction automatically. It first make a piecewise constant tree then fits a linear regression model to the data at each node. Then CHAID (Chi-Square automated interaction detection) as was introduce which was based on Chi-Square. Different algorithms were discussed by Loh 2011) in his article

and the different stopping rule in these algorithms. CHAID (Chi square automatic interaction detection) uses significance and Bonferroni correction to try to merge iteratively pairs of child nodes.C4.5 and CART are other methods of classification tree. C4.5 uses Entropy as a measure of impurity and CART uses Gini index. Gini index is the generalization of binomial variance. C4.5 is used when a categorical variable has more than two categories (Loh 2011).

Bagging and boosting are used for Classification and Regression Tree. Bagging improves the performance of CART by averaging many trees where boosting assign weights to trees and then average them (Sutton 2005). CART and Logistic regression were used for predicting severity of road crashes for Iran which is major health issue. They found driving license and safety belt are significant variables where driver age were negatively associated with injury in road crash (pakgohar et al 2010).

2.3 Literature review for determinants of earning

Separate analysis was made for male and female using human capital and non- human capital variables by ordinary least square technique to estimate the determinants of earnings in Pakistan. He used the data of Labor Force Survey (1993-94).His result suggested that market structured is different for male and female .compensation in the formal sector is higher than in the non- formal sector. Size of establishment in the formal sector was important for female (Nasir 1987**).**The coefficient of schooling used in Mincer earning function for Sweden and different cases when it yields misleading information and its assumptions about length of working life .He found the decline in rate to schooling from 1068 to 1981 in college education where return to high school is stable. There estimate suggest that impact of education on length of working life is an important topic for future research. Education has a causal effect on earnings (Bjorklund and Kjellstrom 2000**)**.

The factors affecting the earnings of an individual and returns to education for Lahore district Pakistan for teaching and non-teaching staff in university, college and school using multiple linear regression. The factors that significantly contributed to earning of all employees ,university employees, college employees and school employees were age experience, occupation, gender, worked hour, computer literacy, family status, family background, spouse education. Those who have passed SSC from private institute earn 8.7 more than those who have

passed SSC from Government institute. Family background has positive and significant effect on earnings. Teaching staff earn more than non -teaching staff (Afzal 2011)

Earnings functions for industrial works in Punjab, to analyze the difference in earnings of individuals due to gender ,marital status, regional location and other socio economic variables using linear single equation least squares regression analysis(Kapoor and Puri 1971).Parents effect the earning of a child potentially through genes and family environment by using variance component model to find the contribution of genetics, family and environments to the variance of the log earnings of white males around 50. The model is estimated through linear additive equation. The contribution of non-common environment is 46 percent for the log of earnings and 24 percent for the years of schooling. After making a lot of assumptions, they partition the remaining variance. Using more plausible estimates, the partitioning of the variance of the ln of earning suggests 18 to 41 percent was due to genetics and 8 to 15 percent to common environment (Taubman 1976).

2.4 Conclusion

Different researches estimated earning function for different countries using conventional techniques of estimation i-e OLS and multiple linear regressions (Nasir 1987) investigated determinates of earning in Pakistan using ordinary least square techniques. No one has estimated determinants of earning using classification and regression technique. CART has been used in financial econometrics and in health economics but rarely used in Economics. The contradictory outcomes of empirical literature provide a room for further analysis.

CHAPTER 3

METHODOLOGY

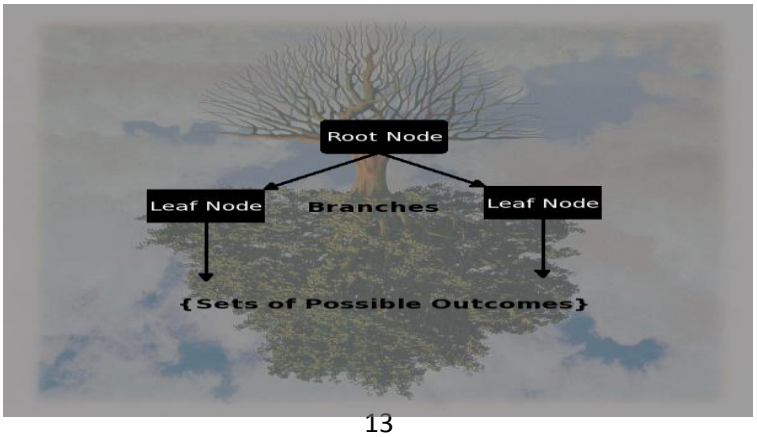
3.1 Introduction

The methodology used in this study is classification and regression tree. This chapter is the detail study of decision tree and its advance techniques which improve the accuracy. In section 3.2 CART algorithms is explained. Decision trees are explained in detail in section 3.3. In section 3.4 is about Regression tree and the procedure of regression tree. Section 3.5 explained pruning of regression tree. Section 3.6 is about algorithm of regression tree. Section 3.7 is about Model assessment. Section 3.8 is about cross validation .In section 3.9 we explained classification tree. In section 3.10 is the detail study of different measures use for the node purity in classification tree. The advance technique ensemble trees are explained in section 3.11.

3.2 Decision Trees

Decision tree is a flow- chart-like structure .which is used for segmenting or stratifying the predictor space in to a number of region or subsets .To make prediction for a given value mean and mode of the training data set is used. The set of splitting rules used to segment the predictor's space can be summarized in a tree; this approach is referred as Decision tree methods. Decision trees are based on predictive modeling approach which predicts the value of target variable based on several input /independent variable .Its widely used in data mining, as one of successful techniques of supervised learning.

Figure 3.1: components of Decision tree

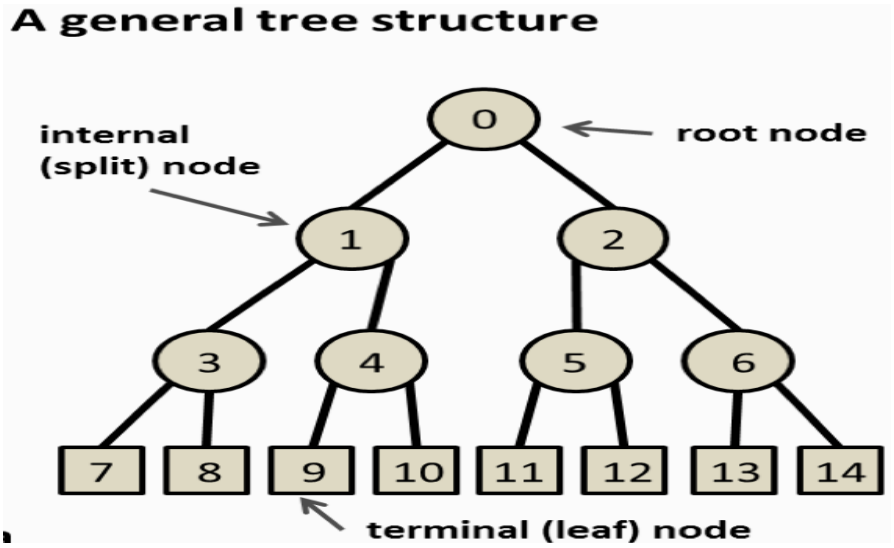


Source: classification and Regression tree: A practical guide for describing a dataset.

Components/Elements of a Decision Tree are, root node is the parent node for root node there is no incoming edge but it has outgoing edges. At root node we have all the predictors space X. Internal node/Non-Leaf node, that point of tree where the predictor space splits is referred as internal node. Leaf/Terminal node represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. We make prediction at leaf node and average all the training data points which belong to that leaf.

Decision trees are made by segmenting the source set into subsets based on an attribute value test. In a recursive manner we repeat the process of splitting for each subset, this process is called recursive partitioning**.** This process completes where the subsets at node has all the same values of the target variable. Which is called node purity.so we stop the process we a node gets pure or there is relative reduction in impurity. This is the top down induction of the decision trees. Decision trees drawn upside down the leaves are at the bottom of the tree and the root is at the top of the tree.

Figure 3.2: Decision tree structure



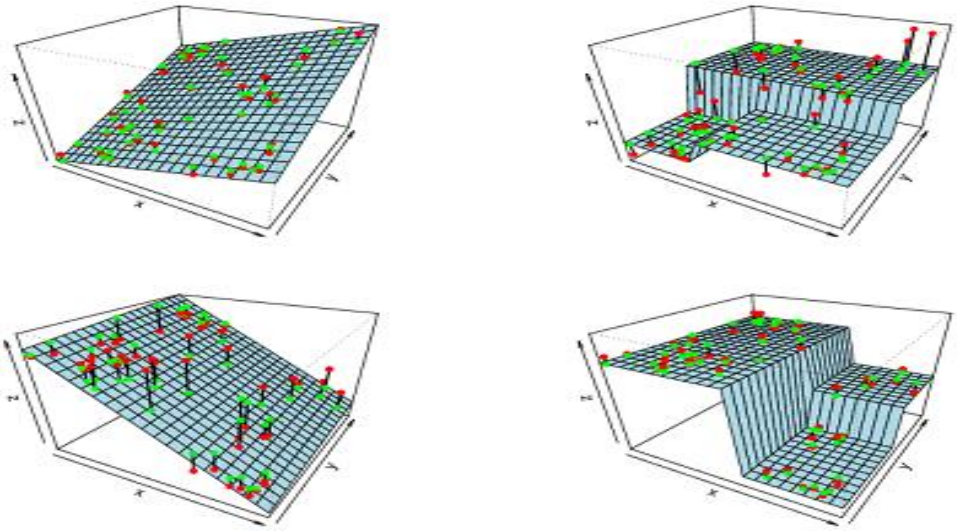
Source: From criminisi et al.MSR-TR-2011-114

Decision tree is a supervised technique, where supervised learning is, where we predict the target variable Y using predictor variables X_1, X_2, $X_3 \ldots X_p$. Supervised learning is a predictive modeling. Where in unsupervised methods, there is no target variable. There are just predictors. It's discovering the pattern of the data.

Decision tree provides a solution to the problem of high non linearity and interaction in the data by first splitting the data into small part and the fit the model which is easy to interpret. Linear, polynomial and other regression fit a global model which holds the complete predictor space, but there are features in the data which interact in a nonlinear way so fitting a global model is very hard and not easy to interpret. There are some non-parametric methods which fit the model locally and then combine them together but again it hard to interpret.

An alternate approach is to nonlinear regression is to partition the predictor space into sub sets or groups which can handle the interaction. Then again we sub divide the subsets in a recursive manner. And then fits the model.

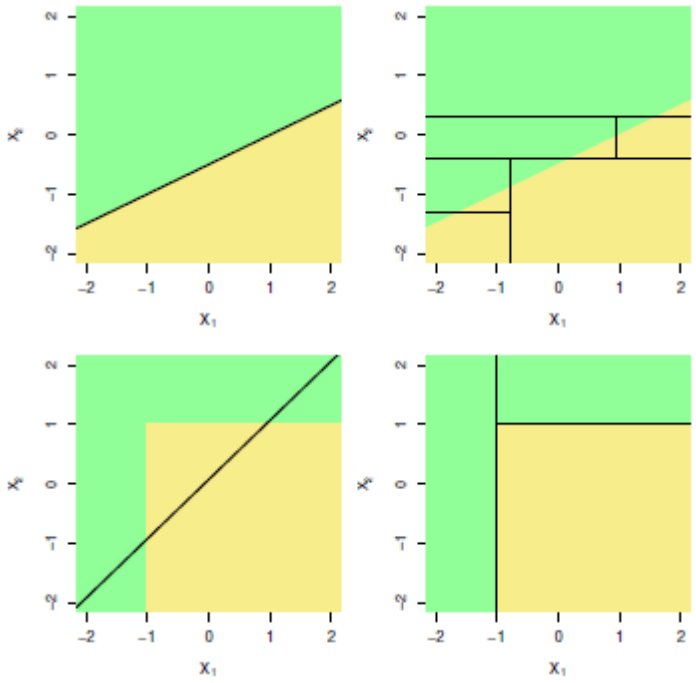
Figure 3.3: Comparison of Linear model and Trees Graphically



Source: From An introduction to statistical learning with Application in R

The left hand side of figure 3.1 shows the surface plot of simple linear regression, where we fit one model to the complete data set and get an average estimate for each predictor ,where the right hand side of figure 3.1 shows that it's not a plane surface like simple regression. Here we fit model at each splitting node of the regression and classification tree, as model is fitted to the each subset of the data.

Figure 3.4Trees Verses Linear Models



Source: From An introduction to statistical learning with Application in R

Trees have to work harder to capture the linear relationship.

3.3 CART Algorithm

CART Algorithm is used for classification and regression tree. It is structured as a sequence of questions, the answer to which determines what the next question, if any should be. The results of the sequence of questions make a tree. It ends a terminal node at which there are no more questions (Leonard Gordon 2013).

Main Elements of CART

1. Rules of splitting data at a node based on the values of one variable.
2. Stopping rule for deciding when a branch is terminal and can be split no more.
3. Finally a prediction for the target variable at each terminal node.

3.4 Regression Tree

Regression in supervised learning is different than the regression in statistics. Regression in statistics can have a numerical or categorical target variable while regression in supervised learning can have only continues target variable, categorical variable is handle in classification.

The functional form of regression tree is $\mathbf{f(X)} = \sum_{m=1}^{M} \boldsymbol{c_m} \cdot \mathbf{1}_{(X \in R_m)}$

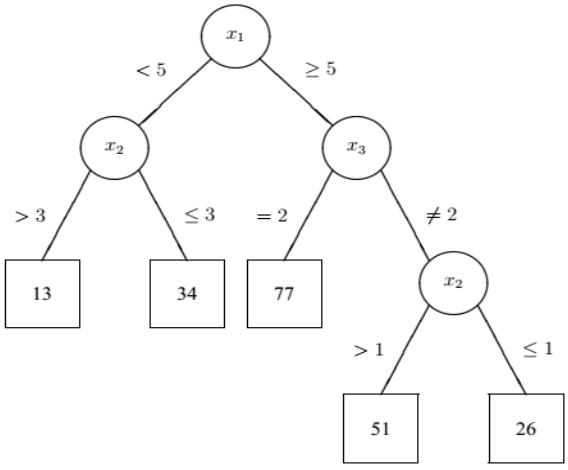
$R_1, R_2, R_3 \ldots \ldots R_m$ Represents the partition of feature space .How to choose $c_1, c_2, c_3 \ldots c_m$?

For loss function l ($\hat{y}$, y) = ($\hat{y} - y$)2, best is $\hat{c}_m$= ave ($Y_i / X_i \in R_m$)

Where the functional form of the linear regression is $\mathbf{f(X)} = \boldsymbol{\beta_0} + \sum_{j=1}^{p} \boldsymbol{X_j}\ \boldsymbol{\beta_j}$

Suppose we have quantitative response variable and p predictors where Y$\in R$.

Figure 3.5 Example of Regression Tree:



Source :Classification algorithms and regression trees by Breiman et al.

When regression tree partitions the predictors into $A_j$disjoint sets, then assigns the fitted value E(Y/X$\in$ A_j) in each region**.**

3.4.1 Process of building a regression tree

It consists of two steps.

Step 1:

Dividing the predictor space that is the set of possible values of$X_1, X_2, X_3 \ldots, X_p$into J distinct and non-overlapping regions$R_1, R_2, \ldots, R_j$.

Step 2:

We make same prediction for each observation fall in the same regionR_j, which is the mean response value for the training observation in region/groupR_j. For example, if we divide the training observation into two regions/subsets R_1 andR_2, if the mean response value of R_1 is 10 and $R_2$is 20 for a given observation X=x, if X belongs to R1, we predict a value of 10 and if x belongs toR_2, we predict a value of 20.

3.4.2 How to construct the Region

We divide the data set/predictor space into J distinct groups or regions R1 to Rj, the regions/boxes should be such that minimizes the RSS.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_{i-} \hat{y}_{R_j})2$$

$\hat{y}_{R_j}$ is the mean response of the jth box. It is computationally hard to consider every possible partition of feature space into Jth box. For this reason we take a top down, greedy approach that is known as recursive binary split.

It begins from the top of the tree, which is called the root node, at which all the observations belongs to one region. Then we successively split the predictor's space. Each split is shown by two new branches further down on the tree, that's why it is called a greedy approach.

3.4.3 Performing recursive binary splitting

We choose predictor Xj and cut point S, such that the subset or region produces through splits lead to reduction in RSS. Through splits we get regions:

$\{X|X_j < S\}$ & $\{X|X_j > S\}$

We consider all predictors and cut point S for each predictor that result a tree which has minimum RSS.

$R_{1(j,s)}$=$\{X|X_j < S\}$ and $R_{2(j,s)}$=$\{X|X_j > S\}$

We choose the value of J and S such that minimize the following equation.

$\sum_{i:x_1 \in R_1(j,s)} (y_i - \hat{y}_{R1}) + \sum_{i:x_1 \in R_2(j,s)} (y_i - \hat{y}_{R2})$

$\widehat{y_{R1}}$ is the mean response value for the training observations in $R_{1(j,s)}$

$\widehat{y_{R2}}$ is is the mean response value for the training observations in $R_{2(j,s)}$

If p , no of predictors are small then it's easy to find s,j .Now again we look for s and j such that it minimize RSS in each of the region .we just split one of the previously identified region .we split one of these three region regions/boxes which minimize RSS .we repeat the process till the stopping criterion or until no region contains more than 5 observations.

3.5 Tree Pruning

Once we grow the tree it might fit well for the training data but not for the testing data .In that case we face the problem of over fitting. Over fitting is the problem which arises due to certain reasons, due to lack of representative sample, presence of noise etc. A good model must have small Generalization error (Expected error based on unseen records or observations) as well as low training error (based on training data or re substitution error). In other words we can say when a model having small training error or re substitution error or CART methodology was introduced in 1984 by Leo Breinman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer the following type of the decision trees; (Classification and regression

tree).it's a non-parametric approach but a large testing error or generalization error is called model over fitting.

Reason of over fitting is that the method works so hard on training data or memorizes the training data and finds the pattern of unknown function caused by random chance instead of true pattern. Overgrowing the tree faces the problem of over fitting, the smaller the tree the lower its variance and gives better interpretation but at the cost of little bias .So the solution to this problem is to grow a tree as long as the decrease of RSS due to each split exceeds some threshold. This method produces smaller tree but its short sighted because a seemingly useless split might follow by best split, which gives low RSS. Instead of that we should grow a large tree and then prune it back .the pruned tree gives lowest test error rate.

3.5.1 How to prune tree

In pruning we select a sub tree which gives minimum test error rate .through cross validation or validation set approach we find test error rate of sub trees, but considering all possible sub trees is difficult to find its cross validation error. Instead of considering all possible sub trees, we use cross complexity pruning or weakest link pruning. It takes only a sequence of trees indexed by a non-negative tuning parameter α. For each value of α there corresponds a sub tree TC T_o such that

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{Rm})^2 + \alpha |T|$$

Is as small as possible

$|T|$: Is the number of terminal nodes of the T

$\hat{y}_{Rm}$: Is the predicted response/predicted mean of m^{th} terminal node

R_m : Is the m^{th} region

α: Controls the trade of between subtrees complexity and it fits to the training data.

If α=0 then T= T_o then the above equation only gives training error rate and no pruning is required. As α increase the above equation minimizes for a smaller tree and the branches get

pruned. So instead of considering all sub trees obtaining a sequence of sub trees as a function of α is easy. We find the value of α through cross validation and then choose sub tree corresponding to α.

3.6 Algorithm for building a regression tree

1. Grow a large tree using recursive binary split and stop when few observations left in each terminal node.
2. Apply cost complexity pruning to make a sequence of best trees as a function of α.
3. Use K fold cross validation to choose α for k= 1…K:

(a) Repeat step 1 and on ($\frac{k-1}{k}$) thfold of the training set and hold out the kth fold.

(b) Assess the Mean square prediction error in the held out fold as a function of α.

(c) Average the result, and choose that value of α which minimize the error.

4: Now from step 2 choose that sub tree which correspond to the chosenα,which minimizes the average error(Gareth James Daniela Witten Trevor Hastie 2013).

3.7 Model Assessment

The process of evaluating the performance of Model is known as Model Assessment .We select the Model the model on the bases on Testing error .The Model having test MSE as small as possible is the good fitted model (Gareth James Daniela Witten Trevor Hastie 2013).

In Model evaluation we have two types of error testing and training error. Testing error is the average error of model to predict response when we use new data or testing observation for fitted model, which has been fitted or trained on training data set. Test error is estimated through Cross validation.

Mathematically we can express as: Ave (I ($y_0 \neq \hat{y}_0$)) we always want it to be minimum for the good fit model. The error calculate through training data is called training error. Mathematically it can be represented as: 1/n ($\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$)

3.8 Cross validation

Cross validation is the process or method of estimating testing error, on the bases testing error we evaluate the model .In cross validation we divide the samples set in to two parts, training

and testing observations or hold a set of observation from the training data .Training data is that data which is being used for training or fitting a model. We trained the model using training data set and assess that model through testing data .cross validation is used for quantitative and qualitative response variable.

There are number of methods to find Test error.

3.8.1 Validation set approach

Validation set approach is one of the method of cross validation .In validation set approach we randomly divide the training data into training and validation set we fit the model to training data and then apply that train model to the validation set or hold out set of data or testing data set is used to predict response of the these hold out set and evaluate the performance of the model, using test MSE, if the response/target variable is quantitative and in case of qualitative we use RSS as measure of test error. Test MSE is used as a measure of test error in quantitative response variable and Misclassification as a measure of test error in qualitative response variable, both should be as small as possible for the good fitted model. If we repeat the process of splitting the data into training and validation set Test error varies, as the observations differ in training data .validation set approach is easy to use but it has some limitations.

Graphically it can be represented as,

Figure 3.6: dividing data into two sets for training and testing

Source: From An introduction to statistical learning with Application in R

3.8.2 Limitation of Validation set approach

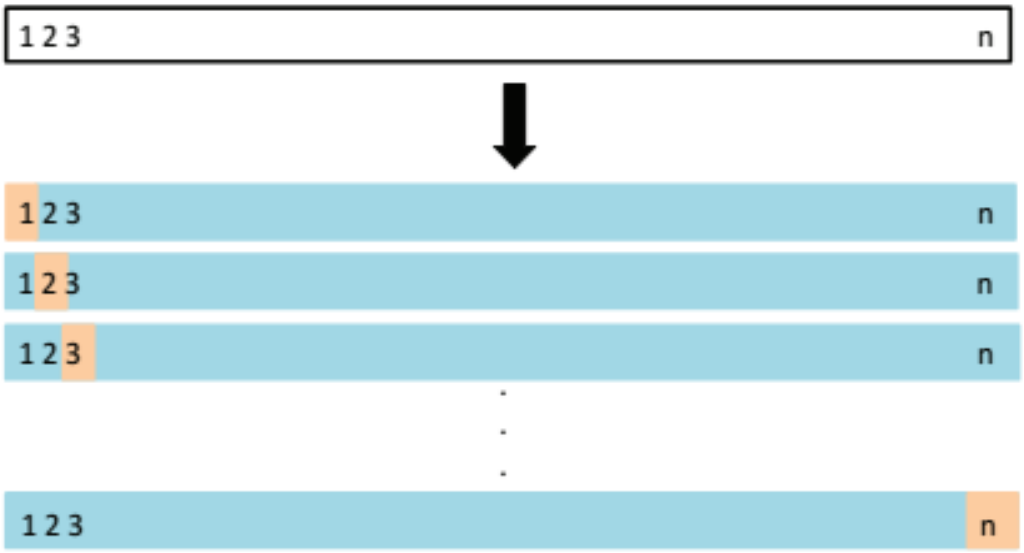
1: validation test error highly varies due to each time different observation includes in the training set and validation set.

2: If model is train on smaller observation validation test error overestimate the test error for the complete data set.

3.8.3 Leave One out Cross validation (LOOCV)

LOOCV is a general approach, can be used in any predictive modeling, like validation set approach. We divide the data in two parts or subsets but not in comparable size like Validation set approach. If we have a data set of n observation (x_1,y_1) (x_2,y_2) (x_3,y_3)…… (x_n,y_n).we make a training set of (n-1) observation and exclude just (x_1,y_1) for validation set. We fit the model to training data and predict $\hat{y}_1$ using x_1 which is not included in training set and we find $MSE_1 = (y_1-\hat{y}_1)^2$. Graphically it can be represented as

Figure 3.7 dividing data into n subsets for training and testing.



Source: From An introduction to statistical learning with Application in R

It's an unbiased test error but a not a good estimate because one observation is used for validation. Now we exclude (x_2,y_2) as validation set and use the rest of (n-1) observation as

training set .predict $\hat{y}_2$using $x_2$and calculate $MSE_2 = (y_2 - \hat{y}_2)^2$,then exclude (x_3,y_3) and find MSE_3 continue this process till $MSE_n$is calculated.(MSE is only used for regression tree)

LOOCV test error: $CV_{(n)} = (\sum_{i=1}^{n} MSE_i)/n$

3.8.4 LOOCV approach overcomes the drawbacks of Validation set approach

1: It's not overestimate the test error, because it uses half of the observation as validation.

2: MSE does not change that much with the change of different observation in training set. Here we just use single observation as a validation. There is not randomness like Validation set approach. Every time it produces the same result.

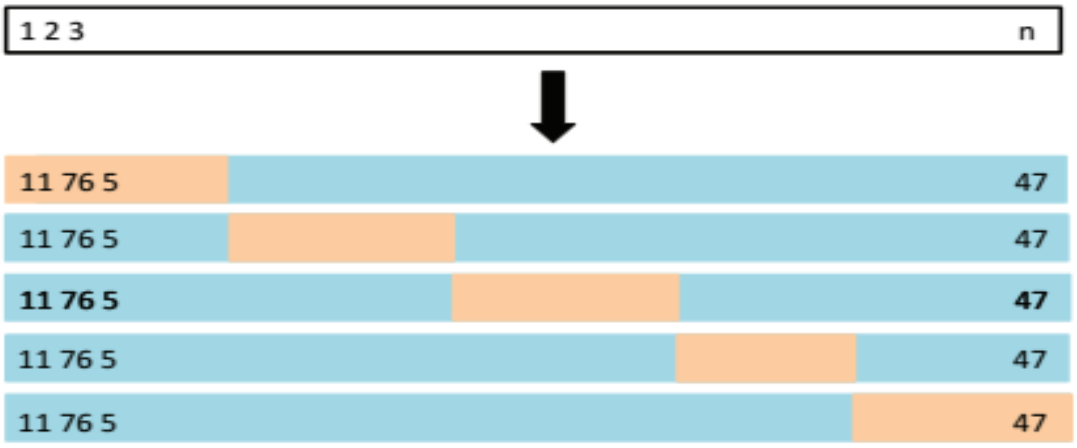
3.8.5: K-fold Cross validation

LOOCV is a special case of k-fold cross validation-fold cross validation divides the data randomly into k non overlapping subsets .one fold is hold out for validation set while remaining k-1 folds are used in training set to fit the method. Test error is through MSE for validation set is known as MSE_1.then second fold is hold out for validation set and remaining k-1 is used to fit the method .MSE_2 is the estimate of test error when second subset is hold out for validation .similarly MSE_3 ,MSE_4.....$MSE_k$is calculated .The estimate of cross validation is give the average of these MSE calculated for k folds.

$$\mathbf{CV_{(k)} = (\sum_{i=1}^{K}(MSE_i))/K}$$

For feasibility and computational convenience k=5 or k=10 is used

Figure 3.8 dividing data into five subsets for training and testing



Source: From An introduction to statistical learning with Application in R

3.8.6 Advantages of k-fold Cross validation over LOOCV

K-fold cross validation approach has less correlated training data sets and less over lapping training data sets.

While in LOOCV training data sets are highly correlated and cause higher variance.

The estimate of test error k=5, k=10 cross validation is neither excessively high bias nor high variance(Gareth James Daniela Witten Trevor Hastie 2013).

3.9 Classification Tree

Classification tree works like Regression tree. Classification tree grows tree by using recursive binary split, but classification tree is used when we have to predict the response of qualitative variable. In classification tree we predict the class corresponding to each terminal node and the class proportion of training observations that fall into that region.

In classification problem we have a training sample of 'n' observation on Y and p predictor variables X_1, X_2, $X_3 \ldots X_p$. When we have to predict Y for new values of X, we partition predictor space X into K disjoint subset A_1, A_2, $A_3 \ldots\ldots\ldots A_j$, for j=1, 2, 3….k. If X belongs to Aj, for j=1,2,3…k then predicted values of Y is j.

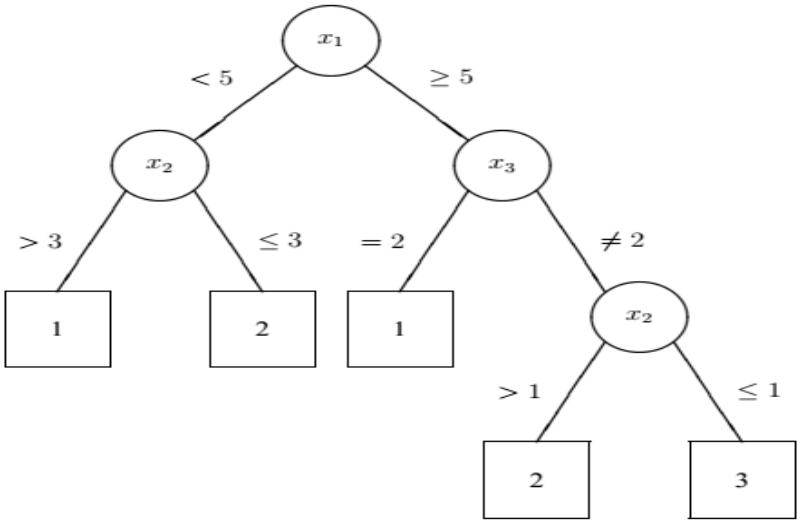
For ordered variables linear discriminant analysis and nearest neighbor classification are used. Linear discriminant analysis yields sets Aj with piecewise linear and nearest neighbor classification nonlinear boundaries, which is not easy to interpret as p (no of predictors) gets large(Wei-Yin Loh 2011).

Classification makes rectangular disjoint sets of Aj by recursively partitioning the predictor space one variable at a time which is easy to interpret.

Suppose we have a categorical response variable Y and p predictors$\in K$ where K=1, 2, 3…j.

The subsets made through splits called Nodes. Those nodes which can't be split further called terminal nodes.

Figure 3.9 Example of Classification Tree



Source: Classification algorithms and regression trees by Breiman et al.

In the given classification tree X_1 represents root node where X_2 and X_3 are internal nodes. Squares represent terminal nodes. Each terminal node belongs to one of the classes of feature space. Classification trees predict the class not predicted value like regression and it gives prediction in the form of probability Pr(Y=s/X$\in A_j$) in each region.

Node purity is used as criteria. A node is said to be pure if all the training observations in that node belong to the same categories of target variable or where the subsets at node has all the same values of the target variable. Node purity is used as a criterion of stopping the split of classification tree. Node impurity is defined as the heterogeneity of the outcome categories within a stratum. Node impurity is measured through classification error rate, cross entropy and Gini index (Leonard Gordon 2013).

3.10 Measures used for recursive binary split /criterion for split on Classification tree or Measure of node impurity in classification tree.

3.10.1 Classification error rate

In classification tree instead of RSS classification error rate is used as a measure/criterion for binary splits .classification error rate is define as a fraction of training observation in that region that don't belong to the most common class. Mathematically it can be express as:

E = 1 - $Max_k(\hat{p}_{mk})$

$\hat{p}_{mk}$: denotes the proportion of training observations in m^{th} region in k^{th} class.

3.10.2 Gini Index

Is another measure used as a criterion for best binary split of classification tree. It shows the total variance across the K classes. Mathematically it can be expressed as:

G =$\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$

G is minimum when $\hat{p}_{mk}$ is close to zero or one.it is used as a measure of node purity.

3.10.3 Entropy/Deviance: is used as another criterion. Mathematically Entropy can be expressed as:

Entropy = -$\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$

$where\ 0 < \hat{p}_{mk} < 1$, entropy is minimum or zero when $\hat{p}_{mk}$either zero or one is. Entropy and Gini index is sensitive to node purity than classification error rate. These given approaches are used as a criterion for pruning the tree as well.

3.10.4 Cross validation for Classification

Cross validation can be used on classification as well. MSE is used as measure of test error when output Y / target variable is quantitative. When target variable is qualitative Misclassification is used as a measure the test error. LOOCV is error rate is equal to

$cv_{(n)} = (\sum_{i=1}^{n} Err_i)/n,$

Err = I (y_i- $\hat{y}_i$).The K-fold cross validation error rate and validation set error rates are defined analogously.

3.10.5 Evaluating the performance of Classification Model

The performance of Classification model is evaluated by using confusion metric and performance metric.

3.10.5.1 Confusion metric:

Confusion metric gives us the number of misclassified and correctly classified test records in a tabular form.

Confusion matric for 2 class problem.

		Predicted class	
		Class=1	Class=0
Actual class	Class=1	f_{11}	f_{10}
	Class=0	f_{01}	f_{00}

$f_{01:}$ Shows number of records from class 0 incorrectly predicted as class 1.

f_{10}: Shows number of records from class 1 incorrectly predicted as class 0.

Where f_{11} and f_{00}represent the correctly classified observation

f_{11}+f_{00}= total number to correct prediction made by the model.

$f_{01}+f_{10}$= total number to incorrect prediction made by the model.

3.10.5.2 Performance metric:

Performance metric summarizes the information of confusion matrix in a single number, which is called Error rate .Error rate makes the comparison of different models easy. The performance of the model is evaluated by error rate or accuracy.

Accuracy = $\frac{\textbf{\textit{Total num of correctly classified records}}}{\textbf{\textit{Total number of predictions}}}$ **or** $\frac{f_{11}+f_{00}}{f_{11}+f_{00}+f_{01}+f_{10}}$

Error rate = $\frac{\textbf{\textit{Total num of incorrectly classified records}}}{\textbf{\textit{Total number of predictions}}}$ **or** $\frac{f_{10}+f_{01}}{f_{11}+f_{00}+f_{01}+f_{10}}$

As the Accuracy of the model increases the Error rate decreases. We choose that model which give lower Error rate and maximum accuracy**.**

3.11 Ensemble trees

The main demerit of CART is does not give same level predictive accuracy for different training sets. To improve the predictive accuracy of the model or to attain same level of predictive accuracy we use the technique of bagging, random forest and boosting where we ensemble the trees. But before explaining these techniques in detail we must know what is bootstrapping. Bootstrapping is a powerful technique of estimating the standard error of any statistical method. It's find the uncertainty attached with estimator and to assess its variability that how much an estimator can vary. Bootstrapping is used to estimate the standard error of the coefficient of the linear regression.

3.11.1 Bootstrap aggregation /Bagging for Regression

In order to improve the prediction accuracy and reduce the variance bootstrap aggregation is used. CART faces the problem of high variance due to random split of data set in to training and testing data. If repeatedly we fit the model on different training set we get different result.

If we have 'n' independent observationX_1,X_2,X_3, ...X_n have variance σ^2 the variance there mean is σ^2/n of the observations. The mean of the observation reduce the variance

Procedure:

Step 1: we take different training set from population and find its prediction model for each training set.

Step 2: then average the result of all those training data set drawn from population.

$$\hat{f}_{ave}(x) = 1/B(\sum_{b=1}^{B} \hat{f}^{*b}{}_{(x)})$$

Where B is the number of training sets

But sometimes it's not possible to take different training sets from the population, we repeatedly take different sample from a single available sample, which is called bootstrapping, then prediction model is fitted to each bootstrap training data set and average the results of all prediction model. Which reduce the variance and increase prediction accuracy of the model.

$$\hat{f}_{bag}(x) = 1/B(\sum_{b=1}^{B} \hat{f}^{*b}{}_{(x)})$$

In this approach we make number of trees using different bootstrapped training data set. Each tree has high variance but when we average all the fitted trees in order to make a single tree which result in low variance. The test error obtain from the single bagged tree is lower that the test error of individual tree.

3.11.2 Bagging for Classification tree

In classification we deal with qualitative response variable while in regression tree response variable is quantitative, so in bagging for classification tree we can't find the average of the trees like regression tree.

Procedure:

Step1: We repeatedly take bootstrapped training sets from a single training data set and make its classification tree.

Step 2: Then we find the class predicted for each tree we choose the most common or most occurring class among the predicted classes.

The number of predicted trees using bootstrapped training data set should be quite large, usually equal to 100.

Although bagging reduces the variance of individual tree but it has some disadvantages too, that it's not easy to represent bagged trees in a single tree and pictorial representation of tree is a beauty of tree based methods. It's hard to interpret bagged trees than a single tree.

3.11.3 Random Forest

Random forest is a cleverer approach than bagging. It's de correlate the trees while in bagging trees are highly correlated and when the average of highly correlated trees are taken there is not that much reduction in variance than taking the average of uncorrelated trees.

In bagging most of the trees look alike because most of the trees use the strongest predictor as a first split because we have random sample of m predictors at each split out of full p predictors. While random forest force each split to consider only a subset of predictors. On average (p-m)/p of the splits are not even allowed to use the strongest predictor like this other predictors will have chance for being split.in random forest if m=p then its same like bagging and there will be not de correlation of trees. Random forest uses m=$\sqrt{p}$ which reduce the variance of the resultant tree.

3.11.4 Boosting

Like bagging boosting also combine a lot of trees but in bagging we take many bootstrapped samples and fit tree to each bootstrapped sample while in boosting we make trees sequentially. Second tree uses the information of original tree and third uses the information of the second tree. All trees are the modified version of original tree

CHAPTER 4

DATA AND VARIABLES DESCRIPTION

4.1 Data and data source

The data used in the study is that of Labor Force Survey 2012-13.It is a regular feature of Federal Bureau of statistics (FBS) since July, 1963. An important feature of LFS is, it separates regular wage employees from irregular wage workers, which helps us to separate formal sector from informal sector. The unique feature of (LFS data provide information about employed and unemployed person. We have used the information of only employed persons that is affecting earning of an individual i-e age, occupation, training, gender, experience, residence, educational level, marital status, and income. R-Programming have used for Classification and Regression Tree (CART) to estimate the determinants of earning function.

4.2 Variable Description

Main Variables used in the study, which are assume are important predictors of earning of an individual, are occupation, training, gender, status, type of industry, location of work, age, education, experience and numbers of hours an individual is working.

Occupation: we have divided occupation in categories white collar job, blue collar and pink collar jobs,

Pink-collar worker performs jobs in the service industry. In contrast, blue-collar workers are working-class people who perform skilled or unskilled manual labor, and white-collar workers typically perform professional, managerial, or administrative work in an office environment. We used the classification of occupations given in Pakistan standard classification of occupation and classified "legislators, senior officials and managers" , "professionals", "technicians and associate professionals", and "clerks" as white collar job. Categories "Service workers and shop and market sales workers" and "skilled agricultural and fishery workers" are classified as blue collar job. "Craft and related trades workers", "plant and machine operators and assemblers" and "elementary occupation" are classified as pink collar job.

Type of industry: Type of industry is a categorical variable and is divided into three categories,

Government sector, public sector, and public sector and other. Categories “federal government”, “provincial government” and “local body government” are classified as government sector. Categories “public enterprise” and “public limited company” are classified as public sector. Categories “co-operative society”, “individual ownership”, “partnership” and “other” is classified as other sectors.

Age: Those individuals are included in the study whose age is greater than 10.Age is continuous variable.

Education: is classified into five categories, no formal education, below middle, below intermediate, 16 year of education Degree (16), M.Phil and Ph.D. In category “No formal education” we have all those individuals who have no formal education or Nursery or K.G. In category “below middle” we have those individuals whose education level is primary or middle. In category “below intermediate” we have those individuals whose education level is matric or below intermediate.in fourth category “Degree (16)” all those individuals who are having 16 year of degree, whether in medical /engineering or M.A/M.Sc. In fifth category “M.Phil/Ph.D.” we have those individuals whose education level is either M.Phil or Ph.D.

Location of work: is having two categories rural and Urban.

Hours of work: is continuous variable and is the total number of hours an individual is working.

Experience: Code for years of education was given in LFS questionnaire; those codes are converted to years of schooling or years of education to make it quantitative variable then experience is calculated using age and years of education. Experience= (Age-years of education -7).where 7 is child school going age in Pakistan.

Training: have two categories whether any individual has given training or not.

Quintiles: The standard Qintiles from Labor Force Survey has used, Q1=16428.Q2=20015, Q3=23273,Q4=29275,Q5=46424.

Sex: has two categories Male and Female.

Marital Status: Has four categories "Never married", “Married", "widow", "Divorced"

Income: is used as a log of monthly income, while variable “income” is monthly income but without log.

Table 4.1: Structure of variables

Data Frame: 14805 obs. of 10 variables:	
Age : integer	21 45 53 24 22 45 23 53 48 45...
Income: integer	6200 7000 27300 20000 6000 38000 7200 7000 21000
Training : integer	20 20 19 20 20 20 18 20 20 20...
Hours worked: integer	48 48 40 40 40 46 49 49 46 48...
Log (income): number	8.73 8.85 10.21 9.9 8.7...
Training: Factor	w/ 2 levels "NO","YES": 1 1 2 1 1 1 2 1 1 1...
Sex : Factor	w/ 2 levels "Female", "Male": 2 2 1 1 1 2 2 2 2 2...
Status: Factor	w/ 5 levels "","Divorced": 4 3 3 4 4 3 4 3 3 3 ...
Location Factor	w/ 3 levels "","Rural", "Urban": 3 3 3 3 3 3 3 3 3 3...
Type of industry Factor	w/ 5 levels "","Gov. Sector", 3 3 2 3 3 5 3 3 2 2 ...
Experience Integer	4 28 34 1 -1 22 8 46 31 28 ...
Quantiles Factor	w/ 6 levels "FALSE","Q1","Q2",2 2 5 3 2 6 2 2 4 4 ...
Occupation Factor	w/ 4 levels "","blue collar"..: 2 2 4 4 4 4 3 3 2 2 ...
Education Factor	w/ 5 levels "","below inter",..: 1 1 1 4 4 4 3 5 1 1 ...

4.3 Statistical Software

MS Excel has been used for coding of variables. R programming is used for construction of classification and regression trees and estimation of the results. Simple multiple Regression is also estimated using R programming.

Chapter 5: Results and Discussion

Table 5.1: Summary statistics of the variable

Experience	Type of industry	Hours worked	Sex	Status	Location	Income
Min. :-3.00		Min. : 1.00	Female: 1671	Divorced : 34	Rural: 3266	Min. : 50
1st Qu.: 8.00	Gov. Sector :5916	1st Qu.:40.00	Male :13134	Married :10185	Urban:11539	1st Qu.: 7500
Median :17.00	others :5520	Median :48.00		Never married: 4387		Median :12000
Mean :18.87	Private Sector: 241	Mean :49.19		widow : 199		Mean :16322
3rd Qu.:28.00	Public Sector :3128	3rd Qu.:56.00				3rd Qu.:20000
Max. :87.00		Max. :99.00				Max. :99000
log (income)	**Age**	**Quantiles**	**Occupation**	**Training**	**Education**	
Min. : 3.912	Min. :10.00	FALSE: 638	blue collar :3184	NO :12155	below inter : 0	
1st Qu.: 8.923	1st Qu.:25.00	Q1 :9673	pink collar :5164	YES: 2650	below middle :3579	
Median : 9.393	Median :34.00	Q2 :1446	white collar:6043		Degree(16) :3571	
Mean : 9.399	Mean :34.85	Q3 : 614			No Formal edu:2871	
3rd Qu.: 9.903	3rd Qu.:44.00	Q4 :1147				
Max. :11.503	Max. :99.00	Q5 :1287				

5.1 Regression tree Results

Table 5.2: summary of regression tree using complete data set

Variables actually used in tree construction Income as dependent variable where sex, status, location of work, hours worked, occupation, Experience, education, and type of industry are used as predictors. **significant variables** 1: "Type of Industry " 2: "sex" 3: "age10" "Education" 4:"Experience"	
Number of terminal nodes:	8
Residual mean deviance:	0.3459 = 5118 / 14800

Distribution of residuals:					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-5.12400	-0.27790	0.01146	0.00000	0.35630	3.08200

Table 5.2 shows that Income is used as dependent variable where sex, status, location of work, hours worked, occupation, Experience, age, training, education and type of industry are used as predictors in the construction of regression tree but the significant variables are five, Type of Industry, sex, age, Education and Experience. The numbers of terminal nodes we obtained from the regression tree, using complete observation of the predictors, are 8.The error rate is 34% .It's the training error rate we need to verify it using unseen data or testing data to avoid the problem of over fitting.

Figure 5.1: Plot of regression tree using complete dataset of LFS

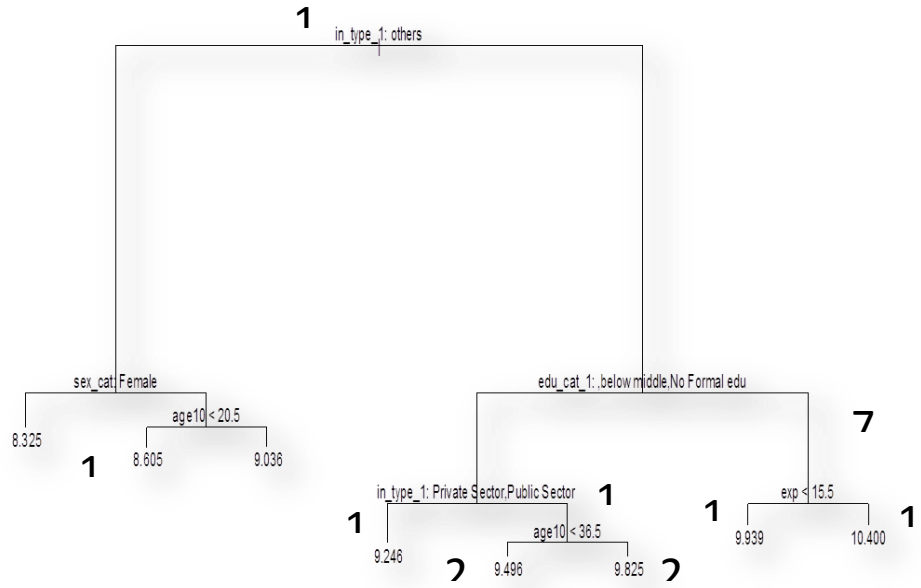


Figure 5.2: Fancy Regression Tree

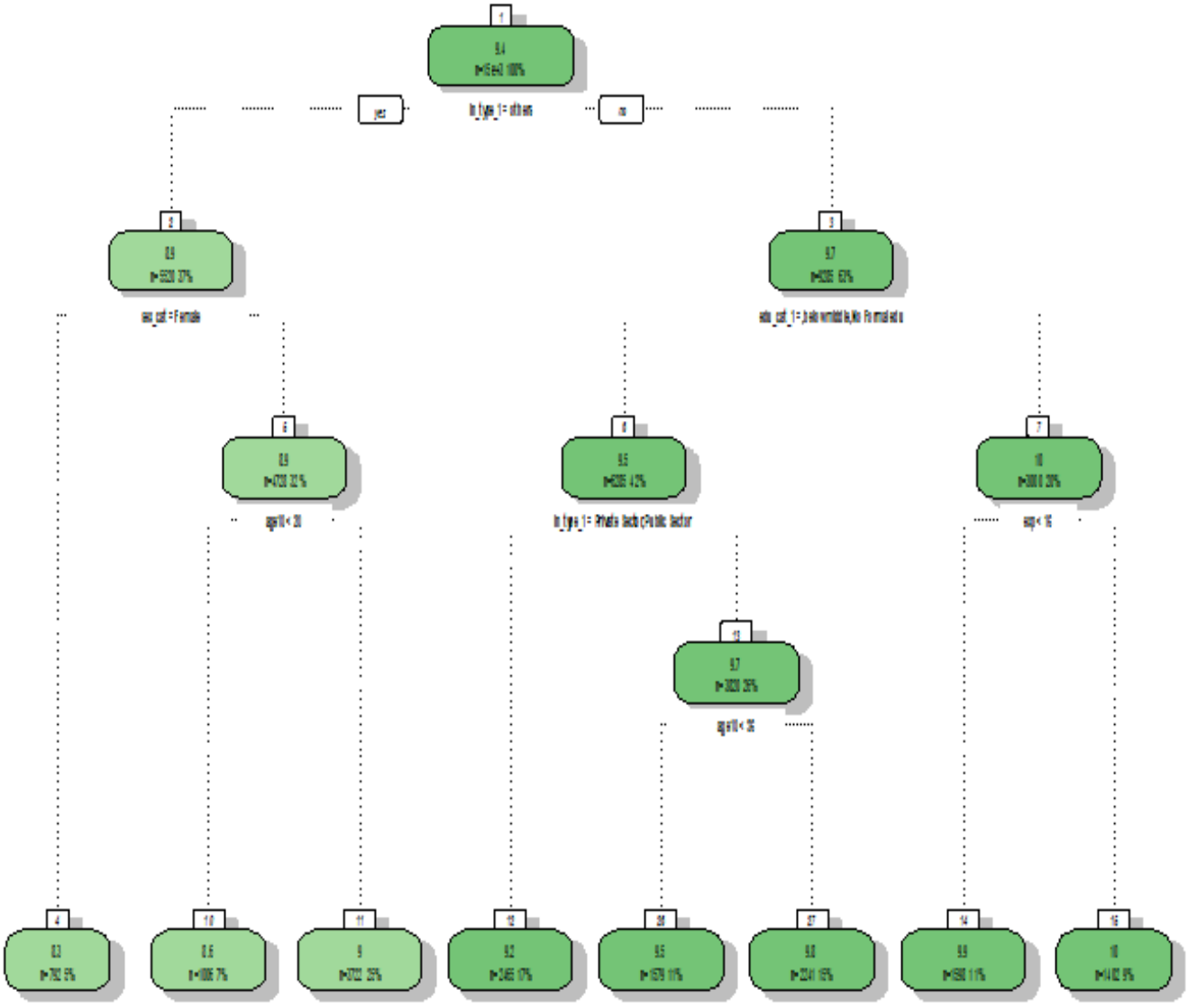


Figure 5.3: Regression Tree with plot. Rpart

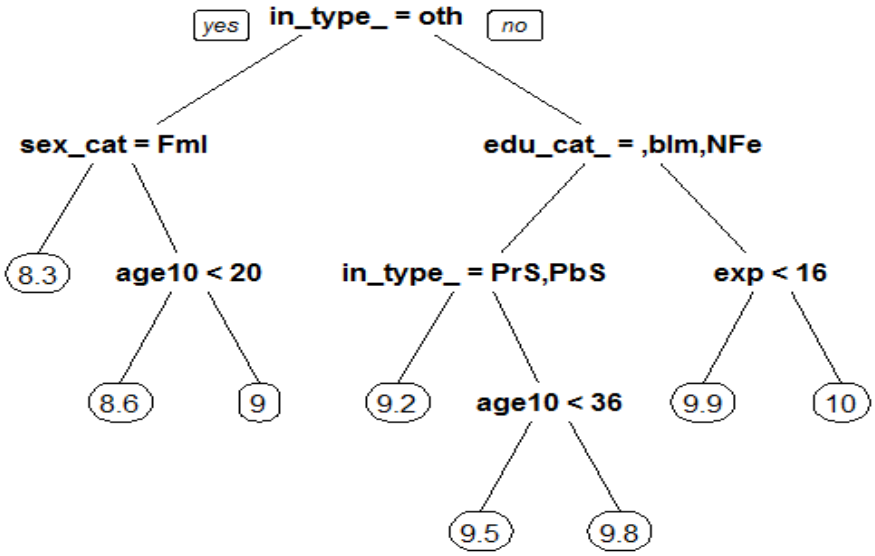
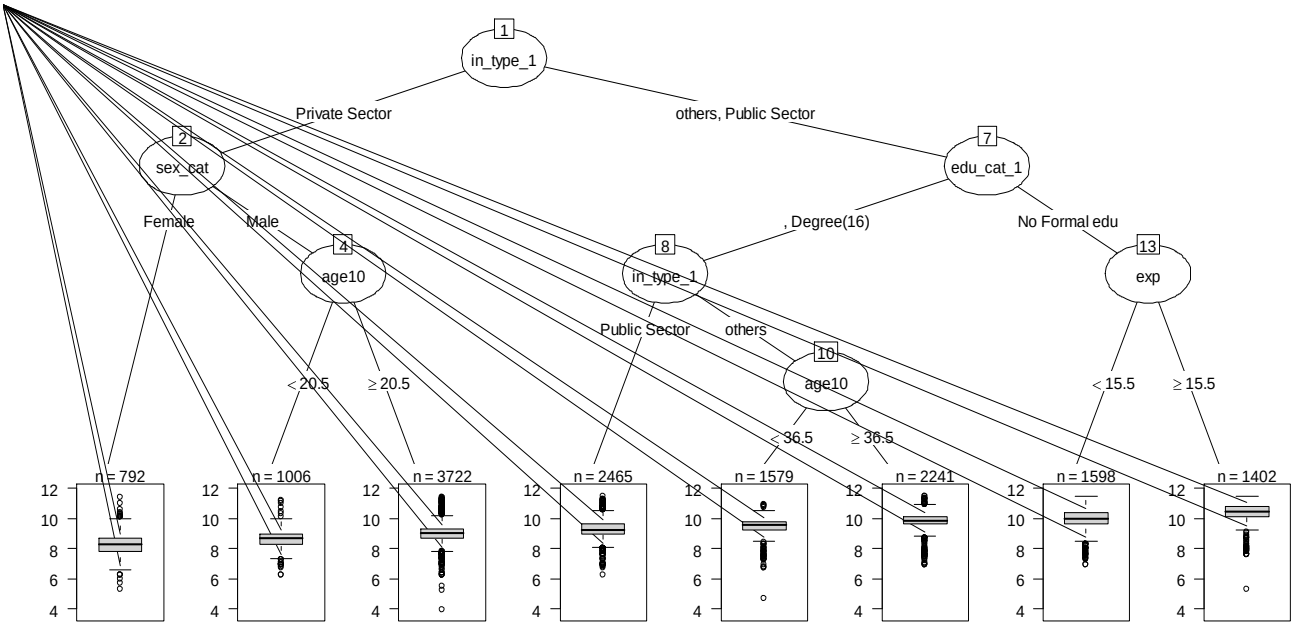


Figure 5.4: Regression tree using R part algorithm



5.1.1 Prediction through Regression Tree

Left branch of the tree**:**

All individuals working in cooperative society, individual ownership, partnership and other, female their average log income is 8.325, so we make the prediction of $e^{8.325}$ i.e. 4125.737.Individuals working in cooperative society, individual ownership, partnership and other sectors but are females and having age less than 20 their average log income is 8.605.so we make prediction of $e^{8.606}$ i.e.5458.885 but those, whose age is greater than 20 their average log income is 9.0306. So the prediction is $e^{9.036}$ i.e. 8400.

Right branch of the tree:

Those who are working in Government, private and public sector and having education below middle and no formal education and specifically working in private and public sector their mean log income is 9.246 so predicted as $e^{9.246}$ i.e. 10363 but working in the sector other than private and public and age is less than 36 their mean log income is 9.496.ie. $e^{9.496}$So prediction is 13306, having age greater than 36 their mean log income is $e^{9.825}. i.e.$ 18490.

Those who are working in government, private and public sector and having 16 years of degree in science, Arts or in medicine, intermediate, matric but experience less than 16 having predicted income as $e^{9.939}$ i.e. 20723.010 and having experience greater than 16 their predicted income is $e^{10.400}$ i.e.32859.So we conclude that the government employees earn more than other sectors employees. Females earn less than male. Empolyees having higher education and experience, earn most. Those females whose age is greater than 20 earn more than those, whose age is less than 20.

Table 5.3: Internal nodes of regression tree

NODE	Split	N	Deviance/Entropy	Y values at terminal node
1	Root	14805	9512.0	9.399
2) Industry type	Other	5520	2465.0	8.856
4) Sex	Female	792	428.1	8.325*
5)Sex	Male	4728	1776.0	8.945
10) Age	Age<20.5	1006	323.5	8.605*
11)Age	Age >20.5	3722	1305.0	9.036*
3:Type of Industry	Gov. sector private sector public Sector	9285	4448.0	9.722
6:Education	Below Middle No formal Education	6285	2235.0	9.515
12:Type of Industry	Private sector Public sector	2465	808.6	9.246*
13:Type of Industry	Gov. sector	3820	1132	9.689
26:Age	Age<36.5	1579	414.6	9.496*
27:Age	Age>36.5	2241	618.0	9.825*
7:Education	16 years degree	3000	1380	10.160
14:Experience	Experience<15.5	1598	686.5	9.939*
15:Experience	Experience >15.5	1402	533.2	10.400*

Table 5.3 shows that type of Industry is used as a root node, which is node number 1, where we have all the predictors. The split point of the node is "other" i-e it divides the data into two parts using a split point "other". If the type of industry is other some predictors will go to the left and if the type of industry is "private sector", "public sector" and "government sector",

some predictors or the rest of the predictors, will go to the right. On the left side of the tree "Sex" is selected as splitting predictor, if Sex=Female the predictors or observation belong to this category will go to the left otherwise right. The right side of the tree, education is selected as a splitting variable, if education is below middle or no formal education the observations will go to the left otherwise right. On the right side of the education we have experience as a splitting variable and on the left again type of Industry is chosen as splitting predictor .Type of Industry is further spited on the bases of age if age is less than 36 observations belong to this category will go to the left otherwise right.

On the left side of the tree internal node Sex is used as splitting variable and splitting point is female, if Sex=female observations will go to the left branch of the tree and if male observation will go the right then further sex is divided in on the bases on age if age is less than 20 it goes to the left otherwise right.

Table 5.4: **Summary of regression tree using training data**

Variables actually used in Regression tree construction:
Income as dependent variable where sex, status, location of work, hours worked, occupation, Experience, education, and type of industry are used as predictors.

5.1.2 Significant Variables

1: Type of Industry 2: Sex 3: Age 4: occupation 5: Education

Number of terminal nodes:	9				
Residual mean deviance:	0.3429 = 2535 / 7393				
Distribution of residuals:					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-5.12000	-0.28960	0.02747	0.00000	0.33280	2.70000

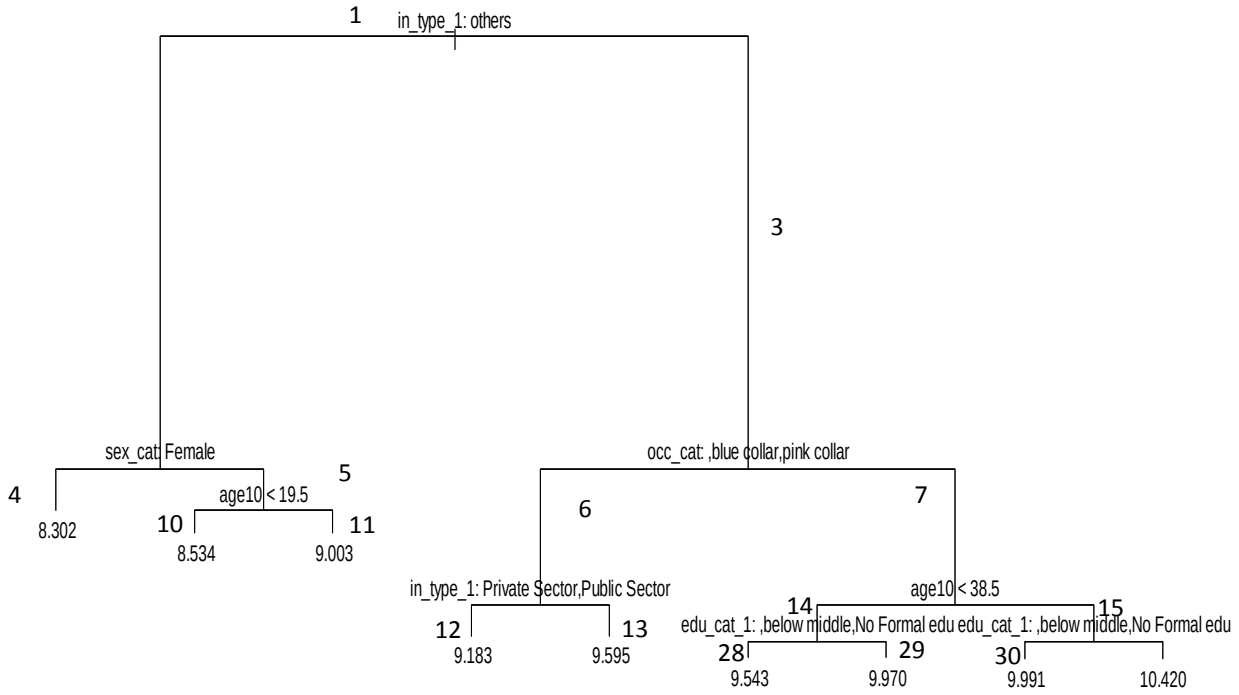
Table 5.5 Internal nodes details using training data only

Node	Split	N	Deviance	Y values at terminal node
1:Root	Type of Industry	77402	4812.0	9.392
2:Type of Industry	Other	2760	1203.0	8.840
4:Sex	Female	390	205.4	8.302*
5:Sex	Male	2370	866.1	8.928
10:Age	Age<19.5	376	116.7	8.534*
11:Age	Age>19.5	1994	679.9	9.003
3:Type of Industry	Gov. Sector Private Sector Public Sector	4642	2269.0	9.720
6:Occupation	Blue collar Pink Collar	2274	730.0	9.412
12: Type of Industry	Private Sector Public Sector	1012	270.2	9.183*
13: Type of Industry	Gov. Sector	1262	364.2	9.595*
7:Occupation	White collar	2368	1116.0	10.020
14:Age	Age<38.5	1160	525.9	9.794
28:Education	Below Middle No Formal Education	478	185.9	9.543*
29:Education	16 years Degree	682	288.8	9.970*
15:Age	Age>38.5	1208	478.3	10.230
30:Education	Below Middle No formal Education	535	150	9.991*
31:Education	16 years of Education	673	273.9	10.420*

Table 5.4 Shows that “type of industry” is used as a root node, using split category other than government, public and private sector. Which divide the data set or predictor space into two parts i.e. node 2 and node 3 which is the left and right branch of the tree respectively. At node 2 predictor sex is used as splitting predictor and Sex=female is use as a splitting point which reduces the deviance. Node 5 has further split at age<19.5 which leads to terminal nodes 11 and 12.

At right branch of tree, node 3 is further divided into two parts using occupation and split point “blue collar and pink”. which leads to node 6 and 7.Node 6 is further slit on variable type of Industry which is further divided into node 12 and 13, using split point type of industry=private and public sector .Node 12 and 13 are terminal nodes. Node 7 uses variable “age” for splitting and produce node 14 and 15. At node 14, education=below middle and no formal education leads to terminal node 28and 29 .Node 15 also use education as splitting variable and splitting point is education=below middle and no formal education and produce terminal node 30 and 31.

Figure 5.5: Plot of regression tree using training data only



5.1.3 Prediction using Regression Tree

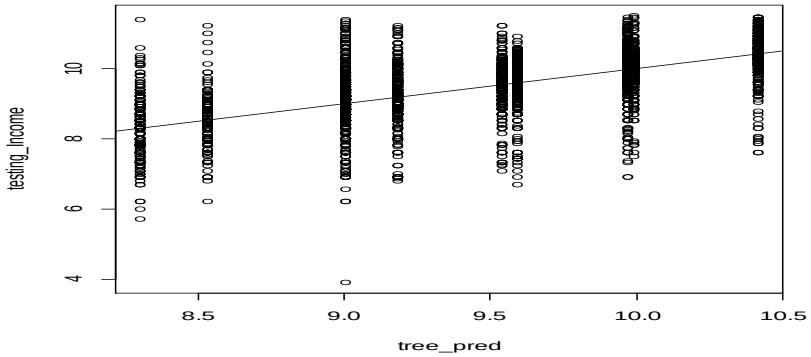
Left branch of the tree shows that, those individuals who are working in industry other than private, public and government and are females their average log income is 8.3 i.e. 4023 and Those who are female and age less than 19.5 their average log income is 8.5 but whose age is greater than 19.5 their average log income is 9.003 i.e.8127.43.

Right branch of the tree show that those individuals who are working in government, private and public sector but having occupation "blue collar" and "pink collar" and working in private or public sector their average log income is 9.18 i.e.9701 and those who are working specifically in government sector and other sector their average log income is 9.595.i.e. 13359.7.Those who are working in government ,private or public sector and occupation is "white collar" but having age less than 38.5 and education is below middle or no formal education their mean log income is 9.54 i.e.13904. and whose are having 16 year of degree, M.A ,M.Phil or PhD, their average log income is 9.97.i.e. 21375

Those individuals who are working in other than government ,private or public sector, occupation is "white collar" and age greater than 38.5 with education below middle or no formal education is 9.99 i.e. 21807 and whose education is 16 year of degree, M.A ,M.Phil or PhD their average log income is 10.42 i.e.3352.So we conclude that workers in government, private and public sectors, white collar workers, age greater than 38 and having higher education get the highest monthly income, where females working in other sectors earn the lowest monthly income.

5.1.4 Making prediction through fitted model using testing data

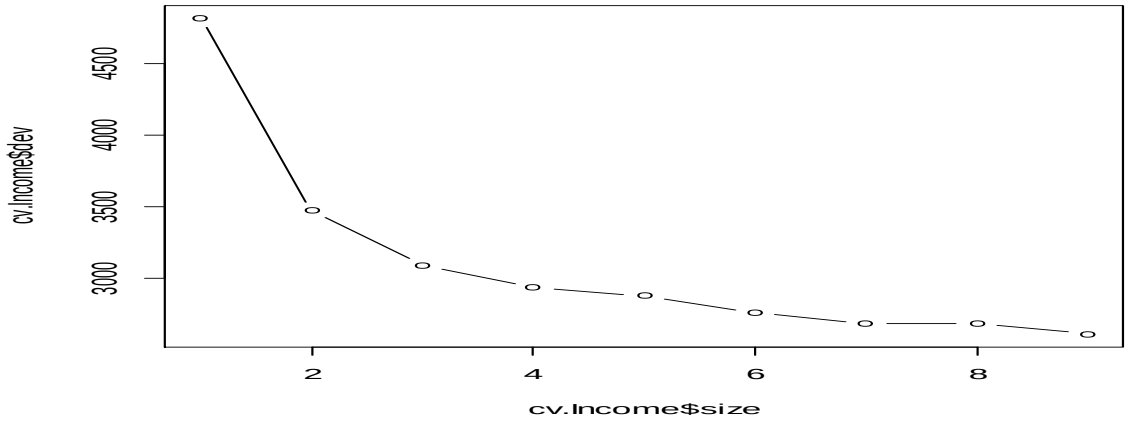
Figure 5.6: Plot of prediction made using testing data



MSE obtain form the training model used for testing data =33.9%

From Figure 7 we can see the prediction is quite reasonable and a numerical measure calculated for Error is MSE which in this case is 33.9%.we want to reduce this error more by pruning, to get the appropriate size of the tree and see the problem of over fitting. To know the right size of the tree and the problem of over fitting we would go for cross validation.

Figure 5.7: Cross validation



Cross validation graph 5.7 shows the plot of size of the tree against the deviance. We choose that point where deviance gets minimum. Here the deviance is minimum at size equal to 9.

Figure5.8: Plot of prune tree

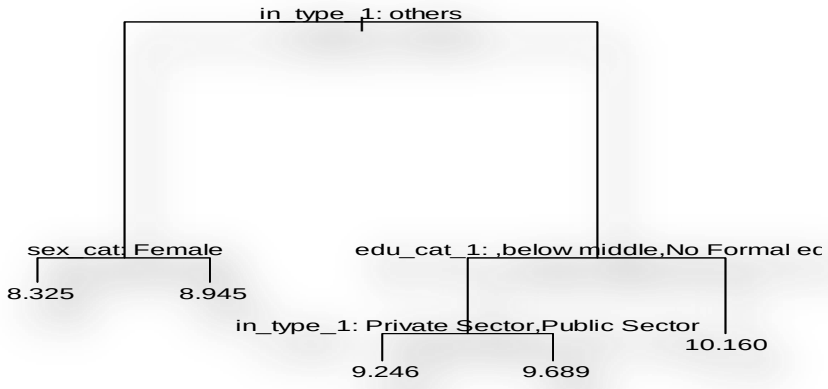
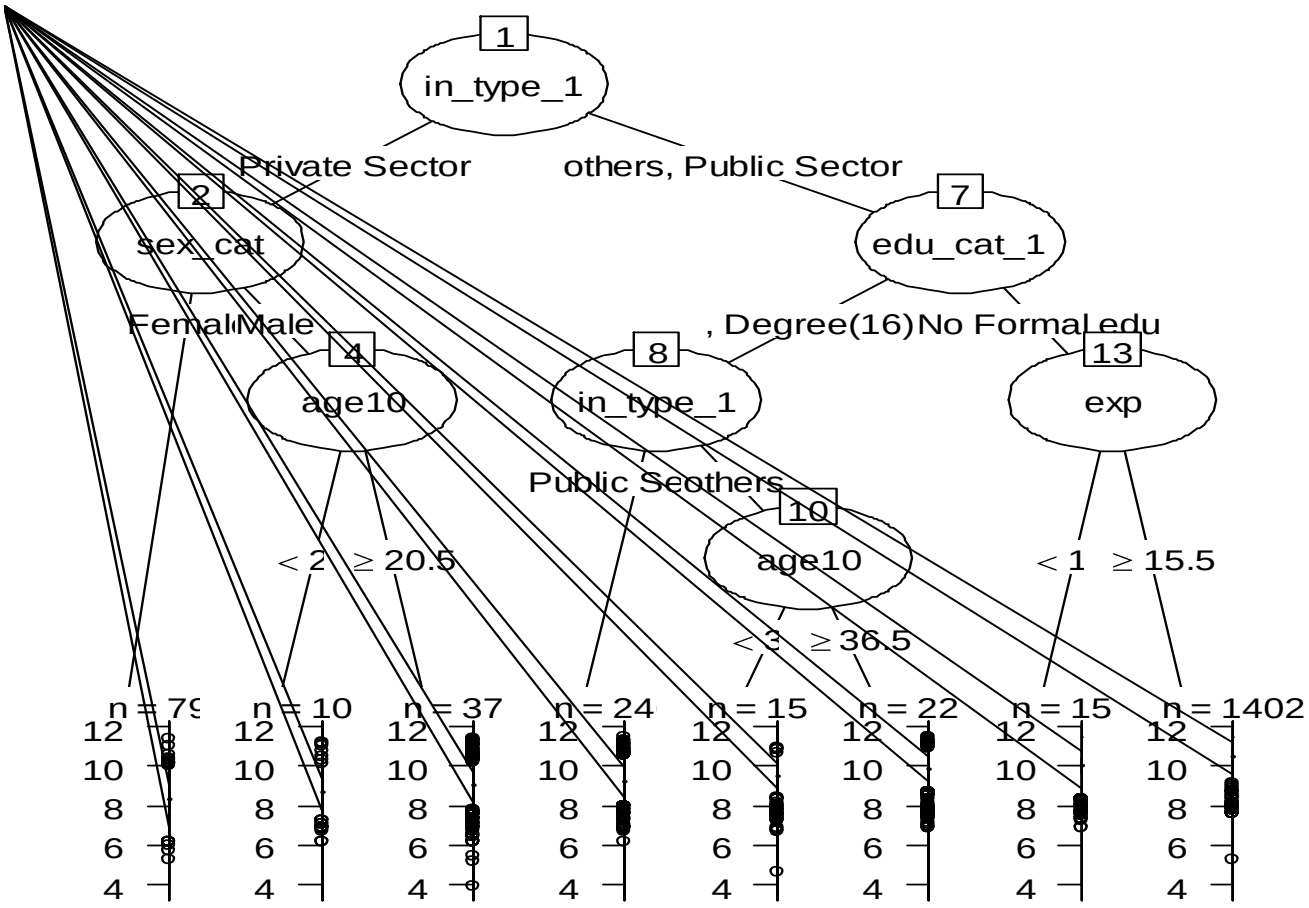


Figure5. 9: Plot of the pruned regression tree using Rpart Algorithm



The pruned tree is shorter than the un pruned tree, the important variables are type of industry of an individual, sex and education. Pruned tree has five terminal nodes and three internal nodes.

5.1.5 Prediction form prune tree: Those individuals who are working in sectors other than government, private and public and are female their average log income is predicted as 8.3 i.e. 4023 and those who are males their average log income is 8.9 i.e. 7331. This shows males earn more than female if they are work in the same type of industry.

Those who are working in government, private and public sector and education below middle or no formal education then specifically working in private and public sector their average log income is 9.2 i.e.9897 and who are working in government or other sector their mean log

income is 9.68 i.e. 15994. It shows that government employees with higher education earn more than employees of other sectors.

Those who are not working in government ,private and public sector in other words working in other sector and education greater than middle their mean log income is 10.160 i.e. 25648.

Figure 5.10: Prediction through pruned tree using testing data

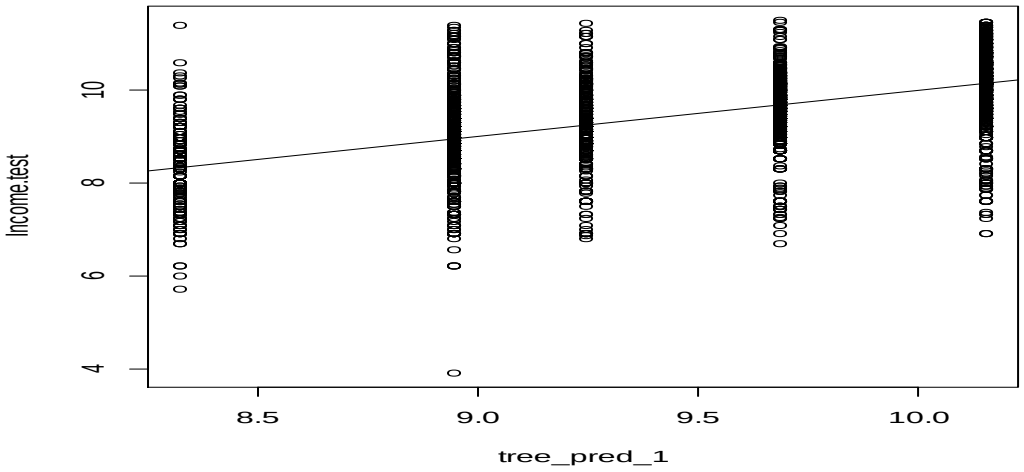


Figure 10 shows that the prediction using testing data through pruned model is quite good and the numerical measure used for calculating the error of the fitted model is MSE, which in this case is 36% and is increased a little for testing data. To reduce the error and improve the accuracy of the model we will go for advance techniques, bagging, random forest and boosting.

5.2 To improve Accuracy of the fitted model using Bagging, Random Forest and boosting

5.2.1 Bagging

Variable used in Random forest :Income , sex , status , training experience, education , location of work , occupation ,Age , hours worked ,Type of Industry

Table 5.6: Summary of the bagged model

Type of random forest	Regression
Number of trees	500
No. of variables tried at each split:	3
Mean of squared residuals	0.3161054
% Variation explained	51.38

Figure 5.11: Plot of number of trees against Error in bagged model

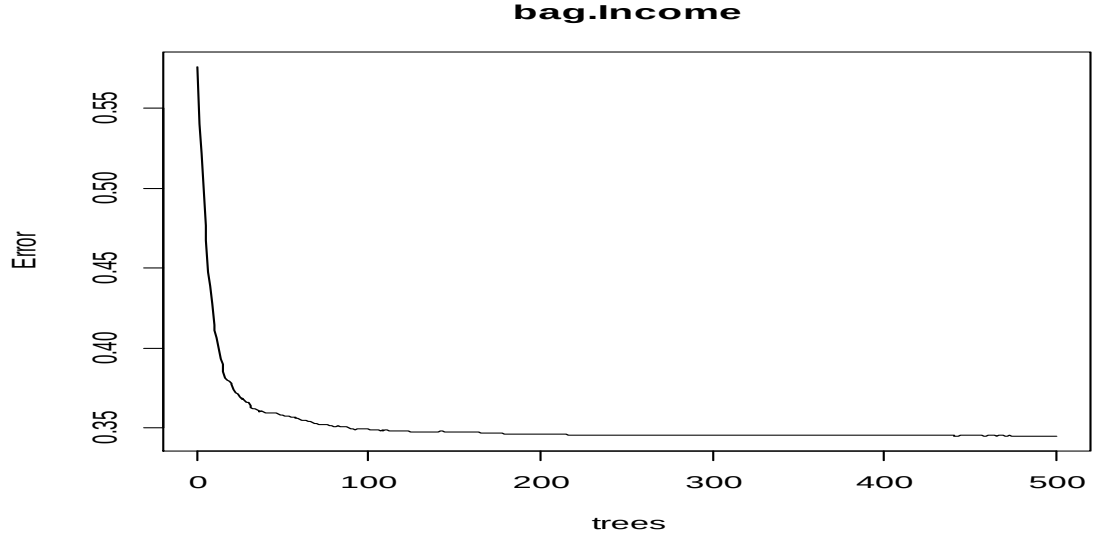
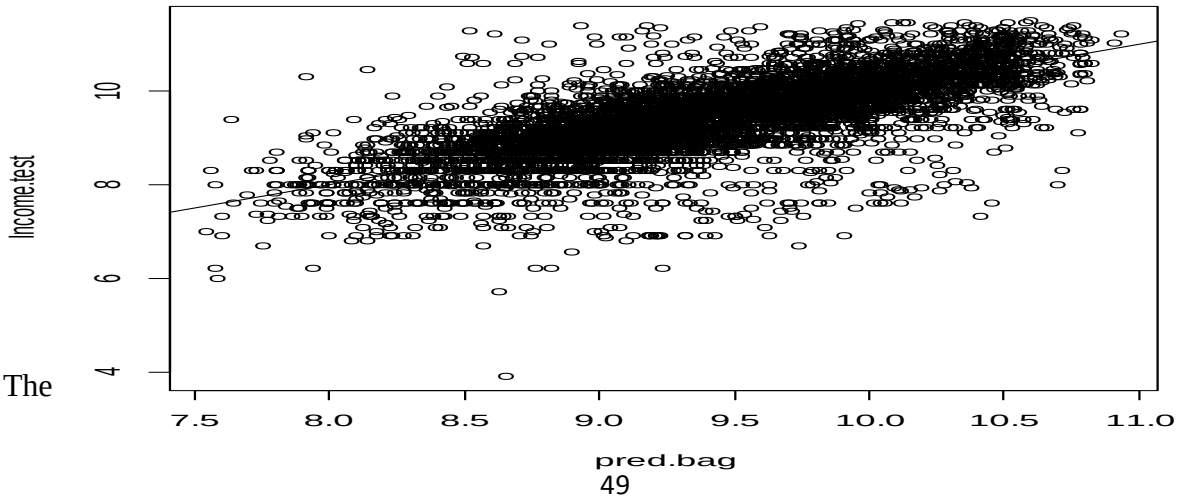


Figure 12: Plot of the Prediction made through bagged model



The plot

shows that the bagged tree gives better prediction than the prune tree, most of the values are close to the line and the MSE calculated for the tree is 30%, which is less than the MSE of the prune tree, so the bagged tree gives better prediction. But it's not possible to plot the bagged tree because bagged tree is average of too many trees so it is not possible to represent it as a single tree.

5.3 Random Forest

Variables used in Random Forest: Income as a predictor and sex, location of work, hours worked, status, age, experience, training, experience and education as a predictors.

Table 5.7: Summary of Random Forest

Type of random forest	Regression
Number of trees	500
No. of variables tried at each split	3
Mean of squared residuals	0.2962897
% Variation explained	51.35

Prediction through random forest:

MSE of the regression tree with random forest =29%.which is less than the MSE calculated in bagging .so random forest reduced the error .when the error reduces it increases the accuracy of the fitted model.

Figure 5.13: Plot of the Prediction using Random forest

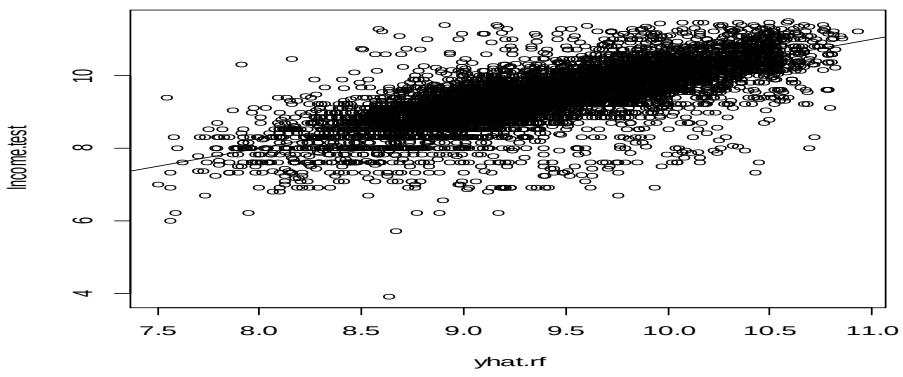


Figure 14: Plot Number of tree against Error in Random forest

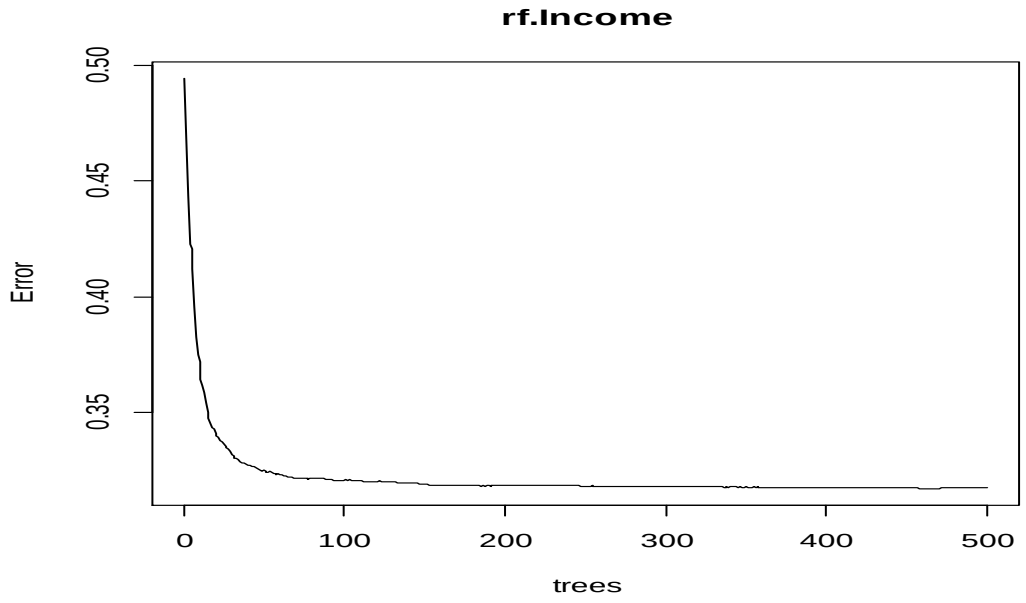


Figure 14 shows that as we increase the number of trees, it reduces the error but like bagging in random forest it's not possible to represent multiple trees in a single tree.

Table 5.8: Variable Importance for regression

Variables	**% Inc. MSE**	**Inc. Node Purity**
Sex	50.56067	147.94174
Status	16.76052	125.85630
Training	18.15291	60.57397
Experience	40.20921	397.61256
Education	63.14786	525.19160
Location of work	23.49164	75.80829
Occupation	43.40243	369.37238
Age	52.52881	598.95869
Hours worked	47.40184	431.55879
Type of Industry	122.96498	1043.26406

Table of variable importance shows the importance of each variable used in the random forest. It shows how much each variable affect the income. It gives the % increases of the MSE

and %increase in the node purity. We use that variable for the split which increase node purity. Type of industry an individual is working in, is the most important factor which affects the income of worker because for Type of industry increase in node purity is with the amount of 1043 and we want the maximum node purity. Secondly age is an important variable, thirdly education then hours of work, experience and education. Training and location of work are least important variables.

Figure 5.15: Plot of variable Importance

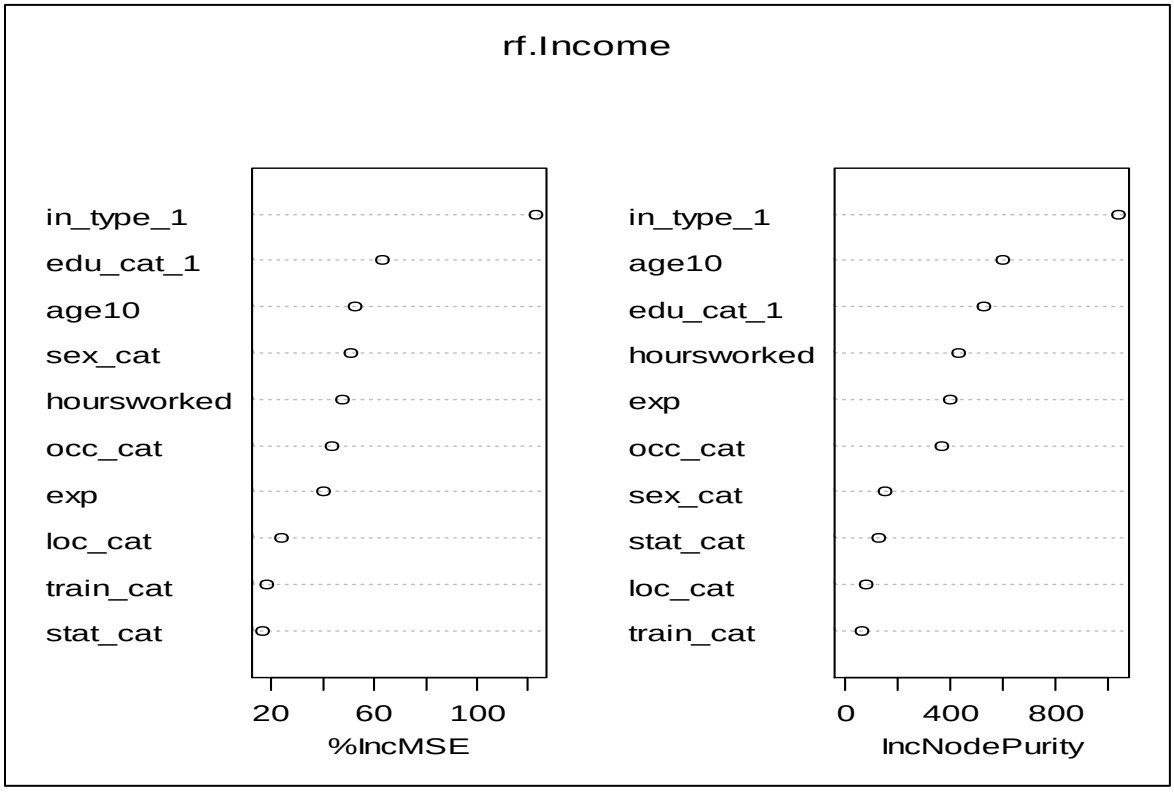


Figure 15 shows variables importance of each variable used in the random forest, that how it's affect the node purity and MSE. This plot represents the variable importance table graphically. Type of industry a worker working in, is the most important variable. Then age, education, hours of work, experience and so on.

5.4 Boosting

5.4.1 Boost model

Variable used in boosting are Income, age, experience, training, status, sex, occupation, hours worked, location of work, education and type of industry. A gradient boosted model, with Gaussian loss function 5000 iterations were performed and there were 10 predictors of which 10 had non-zero influence.

Table 5.9: Summary of Boosted regression Tree

	variables	Relative Influence
Type of industry	Type of industry	48.4324599
Education	Education	16.9777362
Age	Age	15.7406451
Occupation	Occupation	7.5829693
Sex	Sex	5.4811557
Hours worked	Hours worked	2.8268505
Experience	Experience	1.4464898
Location of work	Location of work	1.1468503
Training	Training	0.2430847
Status	Status	0.1217586

Table 5.9 shows that Type of industry, Education, age and occupation are the important variables where location of work, experience and work hours are the least important variables in boosting.

Figure 5.16: Plot of the prediction for testing data using boosted model

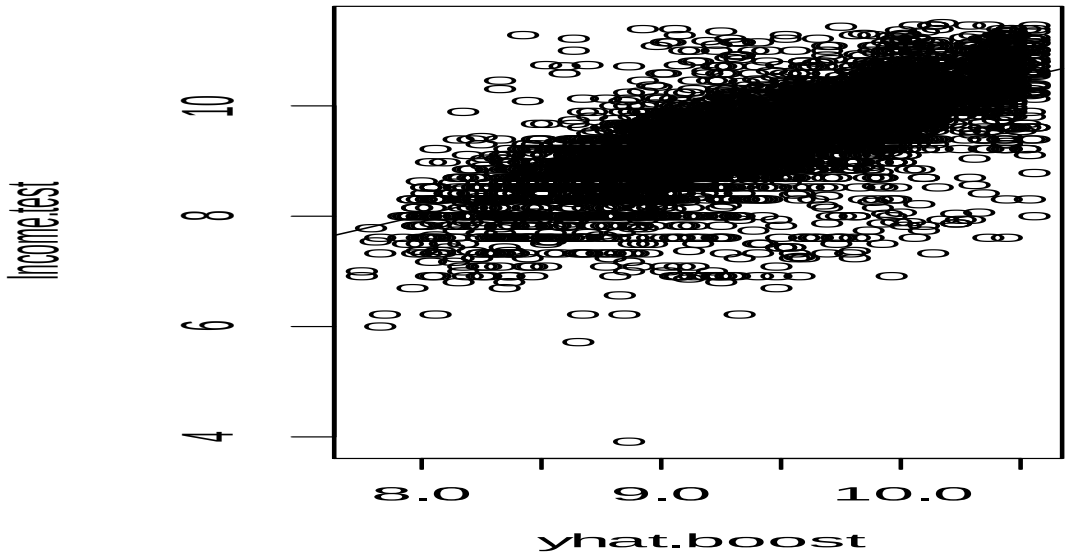
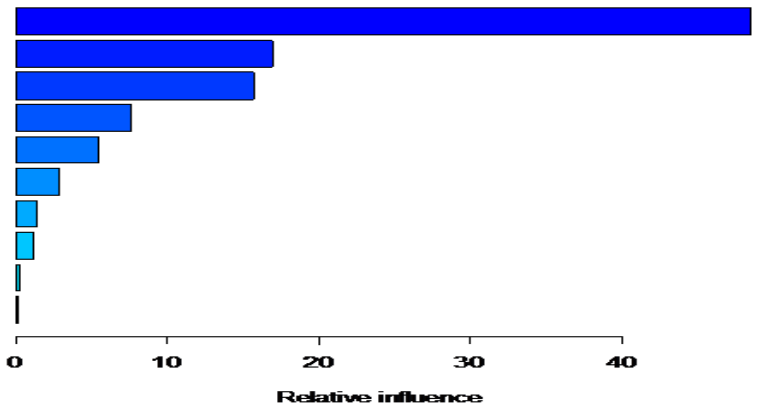


Figure 5.16 shows the prediction made through boosted model which is quite better than the prediction made through prune model and the quantitative measure MSE=29% estimated through boosting same as for random forest.

Figure 5.17: Plot of the relative influence of variables



In Figure 5.17 the height of the bar shows the relative importance of the variables. First or the highest bar shows, Type of industry ,which is the most significant and influential variable affecting the monthly income of an individual, education is the second important variable representing by second highest bar, then age ,occupation and sex. Variables location of work, training and marital status is represented by last three bars respectively and are least important

5.4 Classification tree for Quintiles of income

Figure 5.18 Plot of classification tree using complete set of observation

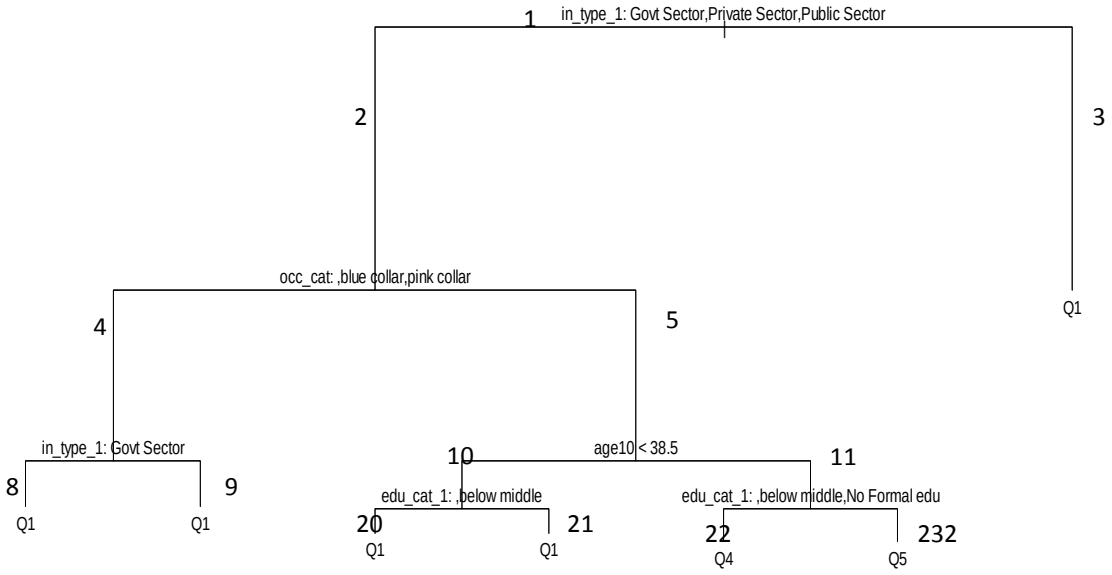


Figure 5.19: Plot of the Classification tree using Rpart algorithm 1:

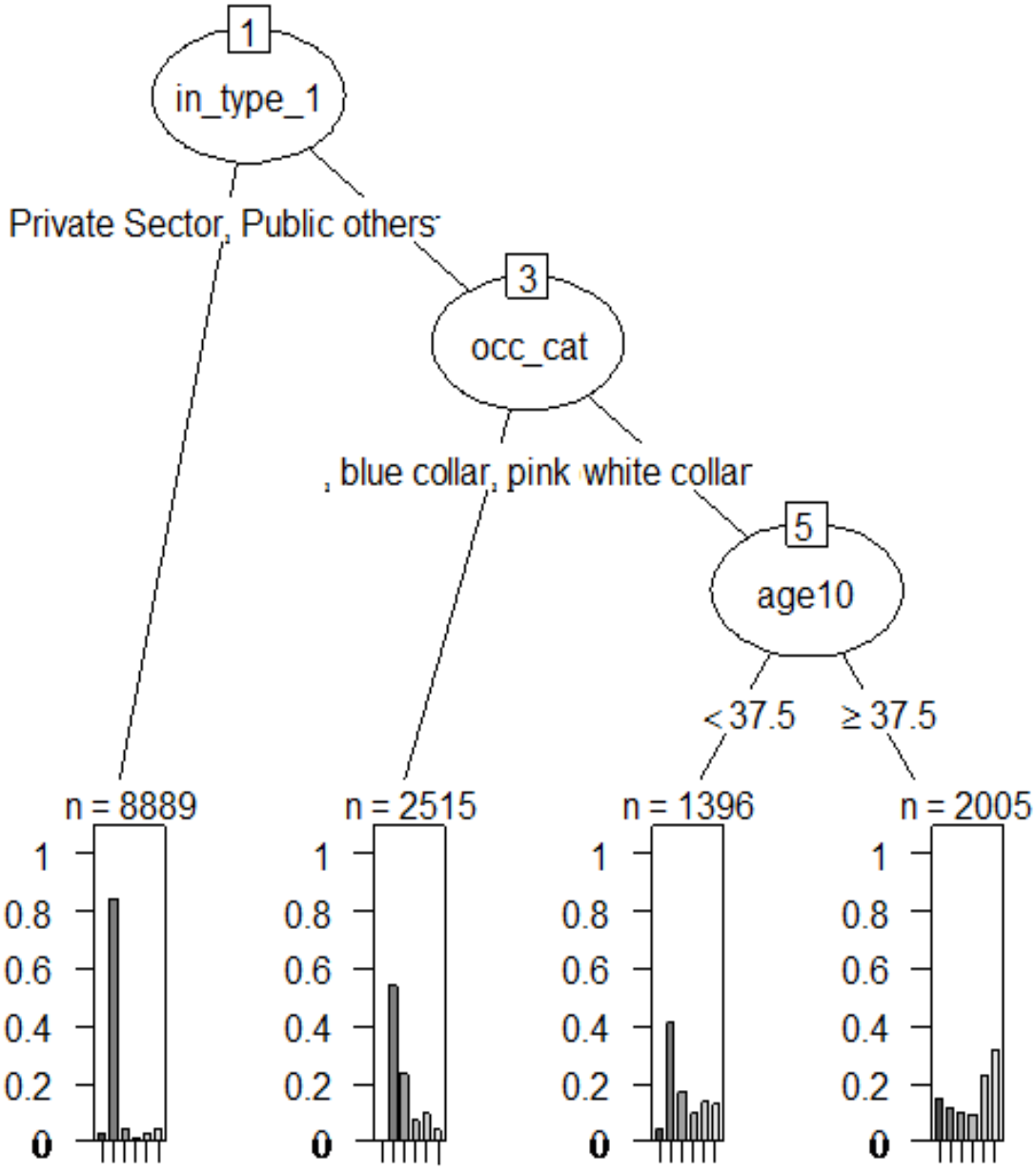


Figure 5.20: Fancy classification tree

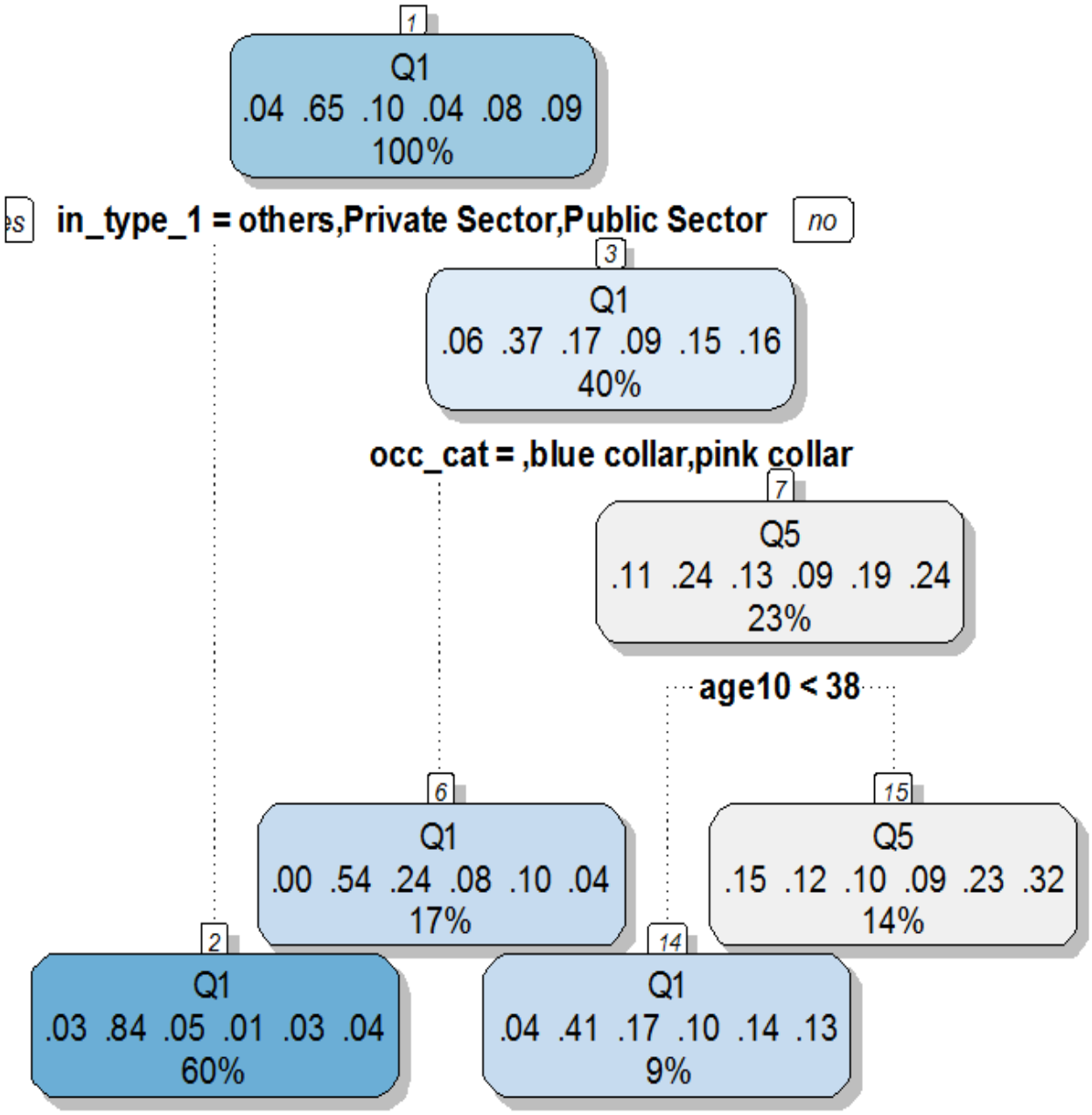
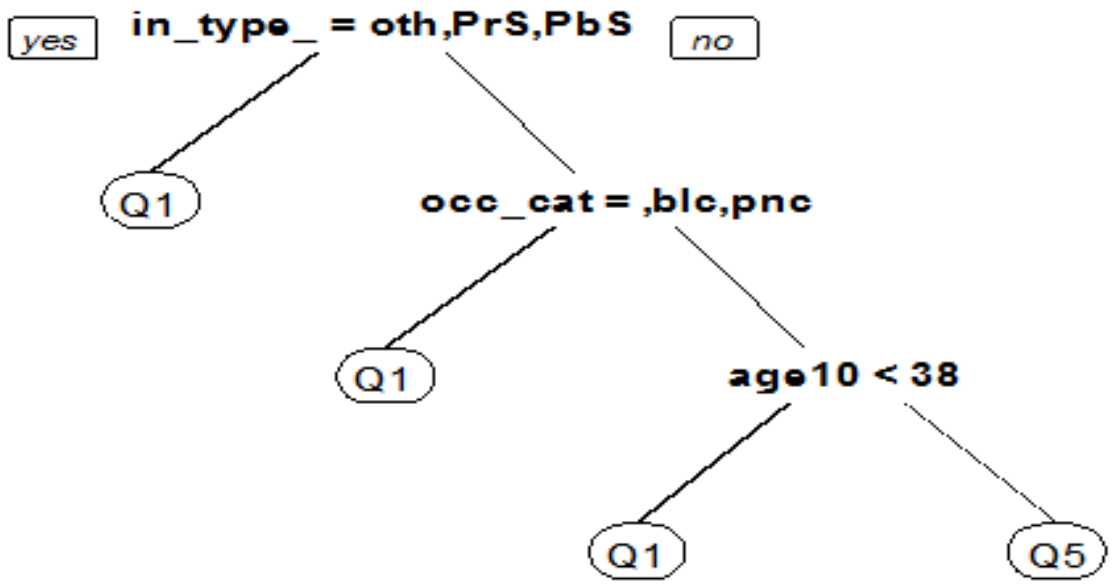


Figure 5.21 Classification tree Using Rpart 2:



5.4.1 Prediction from classification tree:

Those individuals who are working in government, private and public sectors, blue and pink collar workers and then working specifically in government sector are belongs to the Q1 ,the 1st quintile of income group. The range of 1st quintile is (0-16428), and those who are working in public, private or other sector also belong to Q1, the 1st quintile of income.

Those who belongs to occupation category "white collar job" and "other", age is less than 38.5 and education below middle belong to 1ist quintile of income. Those whose age is less than 38.5 but having education above middle and other also fall in 1st quantile of income.

Those whose age is greater than 38.5, education below middle and no formal education fall in 4rth quintile (23273-29275) of income and those whose education category is other than below middle and no formal education fall in 5th quintile (29276-46424) of income. So we conclude that individuals of "other sector" earn less than government, private and public sectors and belong to lower income group. Those who are working in government, private and public sectors, white collar workers, age greater than 38 and education below and above middle belong to the upper income group.

Table 5.10: Summary of classification tree using complete set of observation:

Variables used in Classification tree:			
Quantiles is used as a dependent variable where age , experience, sex, status, training, work location, hours worked, education, occupation and type of industry			
Variables actually used in tree construction or significant variables:			
"Type of Industry"	"Occupation"	"age"	"education"
Number of terminal nodes:	7		
Residual mean deviance:	1.799 = 26630 / 14800		
Misclassification error rate:	0.3183 = 4713 / 14805		

When we fitted the classification tree to complete observation of the predictors, only four variables are significant age, occupation, education and type of industry. The classification error rate is 31% which is quite reasonable but to validate the model for unseen data and avoid the problem of over fitting we divide the data into testing and training and see the model performance on testing data.

Table 5.11: Internal node of classification tree using complete set of observation

Node	Split point	N	Deviance	Yvalues/probability of Y
1)	Root	14805	35040	Q1 (0.043094 0.653360 0.097670 0.041472 0.077474 0.086930)
2) Type of industry	Type of industry: Government Sector, Private Sector, Public Sector	9285	27780	Q1 (0.063651 0.483360 0.138611 0.063328 0.119440 0.131610)
4)occupation	Blue collar ,pink collar	4521	9108	Q1(0.004203 0.691440 0.156381 0.048883 0.067021

				0.032073)
8) Type of industry	Gov. Sector	2515	6254	Q1 (0.001988 0.544334 0.237376 0.075944 0.096620 0.043738) *
9) Type of industry	Private Sector, Public Sector	2006	2200	Q1 (0.006979 0.875872 0.054835 0.014955 0.029910 0.017448) *
5) occupation	white collar	4764	16230	Q1 (0.120067 0.285894 0.121746 0.077036 0.169186 0.226071)
10) age	age < 38.5	2406	7546	Q1 (0.059435 0.426018 0.150873 0.073566 0.136741 0.153367)
20)Education	,below middle	957	2364	Q1 (0.014629 0.594566 0.167189 0.082550 0.102403 0.038662) *
21) Education	Degree(16),No Formal education	1449	4831	Q1 (0.089027 0.314700 0.140097 0.067633 0.159420 0.229124) *
11) age	age10 > 38.5	2358	7994	Q5 (0.181934 0.142918 0.092027 0.080577 0.202290 0.300254)
22) Education	below middle, No Formal education	1025	3474	Q4 (0.047805 0.214634 0.158049 0.119024 0.253659 0.206829) *
23) Education	Degree(16)	1333	4047	: Q5 (0.285071 0.087772 0.041260 0.051013 0.162791 0.372093) *
3) Type of industry	Type of industry= others	5520	3460	Q1 (0.008514 0.939312 0.028804 0.004710 0.006884 0.011775) *

In the above table the detail of all the internal and terminals nodes is given. Type of industry an individual is working is the most important variable then occupation, then age and the education. At root node Type of industry is split at point government, private and public sector. The total number of observation at root node is 14805 and deviance, which is the measure of error, is 35040.The right branch of the tree which is node 3 the Number of observations are 5520 and measure of error is 3450 the probability of Q1 is high at node three, which is .93, so all the individuals working in other than private, public and government sectors belong to the low income group or the first quintile of income group.

On the left side of the tree or at node 2 the number of observations are 9285 and measure of error is 27780 which is considerably reduce at node 2 .Node 2 is further divided into two parts node 4 and 5, node 4 is further divided in node 8 and 9 and node 5 is divided into 10 and 11 then 10 is divided at 20 and 21 where 11 is divided into 22 and 23.

Node 1 is the root node, node 2, 4, 5, 10and 11 are internal nodes and node 8,9,20, 21, 22 and 23 are terminal nodes. The algorithm chooses the split point at each node which minimizes the deviance and chosen that class or category which has high probability.

Table 5.12: Summary of Classification tree using training data

Variables used in the construction of Classification tree: Quantiles is used as a dependent variable where Age, experience, sex, status, training, work location, hours worked, education, occupation and type of industry.	
Variables actually used in tree construction/significant variables : "Type of industry " "Occupation" "Age" "Education"	
Number of terminal nodes:	7
Residual mean deviance:	1.833 = 13560 / 7395
Misclassification error rate:	0.3234= 2394 / 7402

Table 5.12 shows that after using some of the of observations of the predictors for classification tree, the misclassification error is 32 % which is close to the error obtain from classification tree using complete observations of predictors and the same number of terminal nodes.

Table 5.13: Internal nodes of classification tree using training data

Node used for split	Split point	N	Deviance	Y vales /Prob of Y
1) root	Root	7402	17660.0	Q1 (0.043907 0.648608 0.098622 0.040665 0.079573 0.088625)
2) Type of industry	Gov. Sector	3028	9923.0	Q1 (0.065059 0.364267 0.172721 0.084544 0.153567 0.159841)
4) occupation	blue collar, pink collar	1261	3116.	Q1 (0.003172 0.547185 0.242665 0.068200 0.095163 0.043616) *
5) occupation	white collar	1767	6103.0	Q5 (0.109225 0.233729 0.122807 0.096208 0.195246 0.242784)

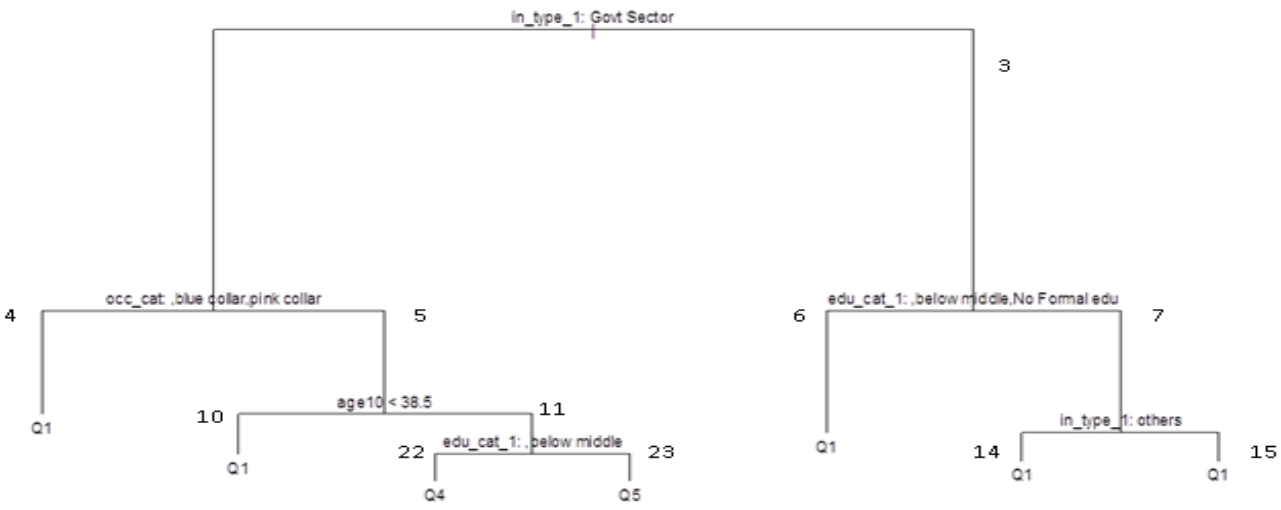
10) Age	age< 38.5	797	2564.0	Q1 (0.046424 0.380176 0.168130 0.089084 0.159348 0.156838) *
11) age	age> 38.5	970	3265.0	Q5 (0.160825 0.113402 0.085567 0.102062 0.224742 0.313402)
22) Education	below middle	437	1475.0	Q4 (0.041190 0.169336 0.146453 0.151030 0.260870 0.231121) *
23) Education	Degree(16),No Formal Education	533	1609.0	Q5 (0.258912 0.067542 0.035647 0.061914 0.195122 0.380863) *
3) Type of industry	others, Private Sector, Public Sector	4374	5817.0	Q1 (0.029264 0.845450 0.047325 0.010288

				0.028349 0.039323)
6) education	below middle, No Formal education	3634	2908.0	Q1 (0.006604 0.916621 0.042102 0.009631 0.015410 0.009631) *
7) Education:	Degree(16)	740	2078.0	Q1 (0.140541 0.495946 0.072973 0.013514 0.091892 0.185135)
14) Type of industry:	Others	294	462.2	Q1 (0.037415 0.799320 0.040816 0.006803 0.034014 0.081633) *
15) Type of industry :	Private Sector, Public Sector	446	1423.0	Q1 (0.208520 0.295964 0.094170 0.017937 0.130045 0.253363) *

Table shows the split of the internal nodes node 1 is the root node, node 2, 3,5,11 and 7 are internal nodes where node 4,10,22,23,6,14 and 15 are terminal nodes. These numbers are just used as label for nodes.

The number of observations at each node, the split point used at each node and the deviance/measure of error is given in the table. At root node there are 7402 and the measure of error is 1766.0, Q1 is having the highest probability. Internal node 2 and 3 are further divided into two parts, node produced node 4 and 5 where is further split into two parts 10 and 11 , and node 11 is divided into 22 and 23 node number. On the right branch which is node 3 which is further divided into 6 and 7. Node 7 produces terminal nodes.

Figure 22: Plot of classification tree using training data only



5.4.2 Prediction made through classification tree:

Left branch of the tree: Those individuals who are working in government sectors, blue and pink collar workers belong to the Q1, the 1st quantile of income group. The range of 1st quantile is (0-16428), and those who are doing white collar job but their age is less than 38.5 also belong to Q1 but those whose age is greater than 38.5 and having education middle or no formal education also belongs to Q4 .those whose education is above middle or higher education belong to Q5 (5th quintile(29276-46424))

Right branch of the tree:

Individuals who are working in public, private or other sectors and education below middle or No formal education belongs to Q1, the 1st quintile of income but those whose education is above middle and working in Industry other than government, private and public also belong to lower income group and those who are working in government, private and public sector belong to lower income group Q1.

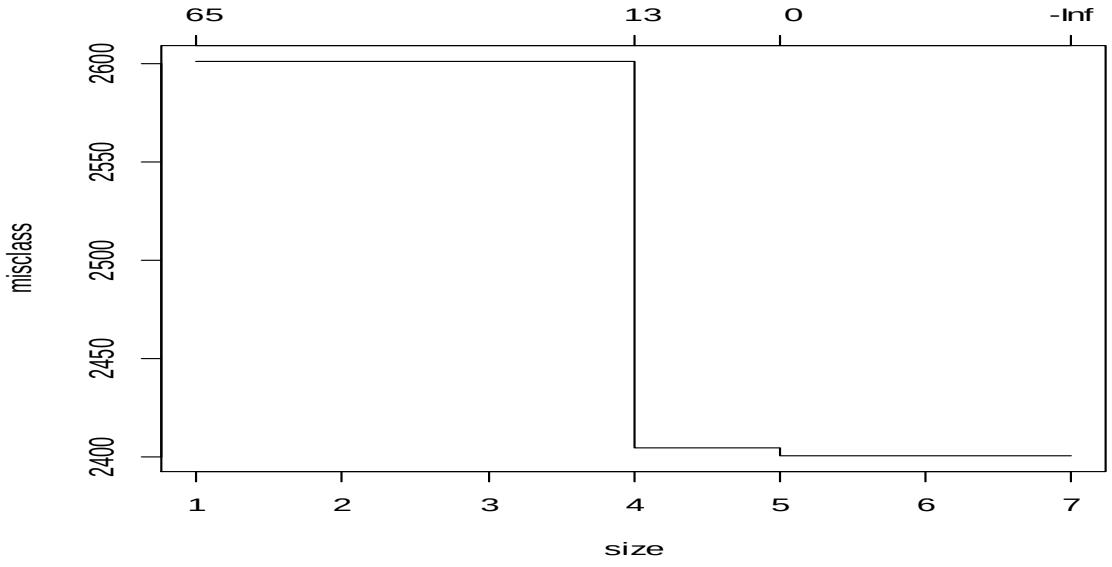
5.4.4 Prediction by using testing data for trained model

We fit the model to training data, to see how it's doing for testing data. we made prediction using trained model for testing data. The classification error rate calculated for testing data is also 31 % which shows that the model is good fit.

Classification Error: for testing data using train model

5.5 Cross validation

Figure 23: plot of the cross validation



Plot of the size of the tree against the misclassification shows the different size of tree against the misclassification but best point of pruning is 5,size of tree mean the number of leaves we have or the level to reach in pruning, when size of the tree is 5 the misclassification error is minimum.

Table 5.14: Summary of prune tree

Variables used in the construction of Classification tree: Quantiles is used as a dependent variable where Age, experience, sex, status, training, work location, hours worked, education, occupation and type of industry.	
Variables actually used in tree construction/significant variables : "Type of industry " "Occupation" "Age" "Education"	
Number of terminal nodes:	5
Residual mean deviance:	1.971 = 14580 / 7397
Misclassification error rate:	0.3234= 2394 / 7402

This table shows that the error rate, which is classification error rate is 32 % which is close to the error rate of unpruned tree or trained model but the number of terminal nodes are fewer than the number of terminal node of trained model. This tree is shorter than the previous one.

Figure 24: Plot of prune classification tree

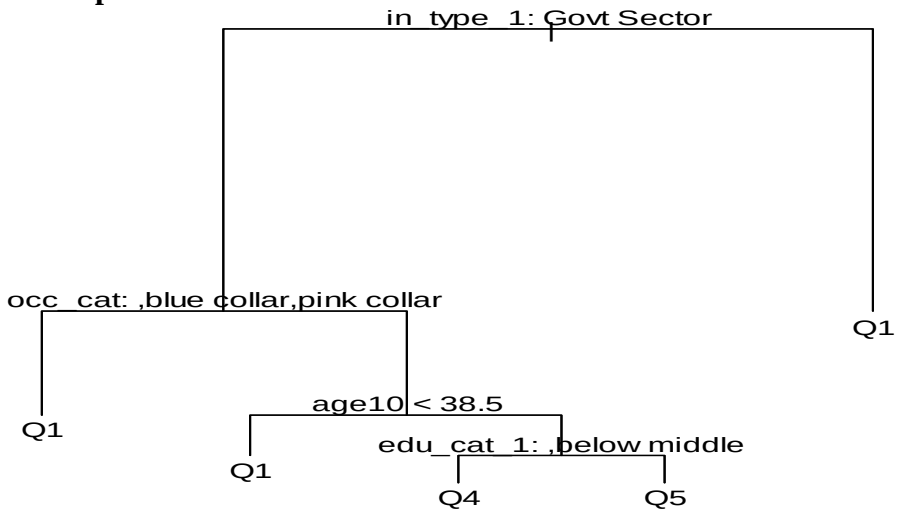
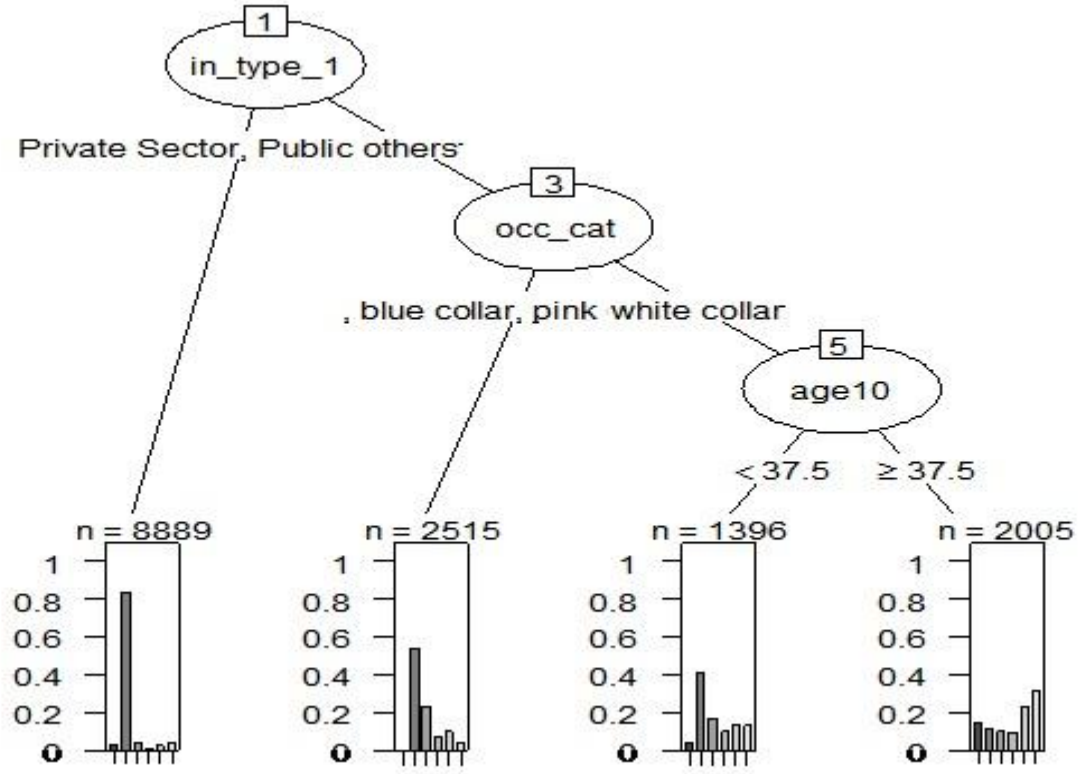


Figure 25: Plot of the Prune Classification tree Using Rpart



5.5.1 Prediction using prune tree:

Left branch of the tree:

Those individuals who are working in government sectors, blue and pink collar workers belong to the Q1, the 1st quantile of income group. The range of 1st quantile is (0-16428),and those who are white collar workers and there age is less than 38.5 also belongs to Q1 but those whose age is greater than 38.5 and having education middle or no formal education also belongs to Q4 but those whose education is above middle belongs to Q5 (5th quantile(29276-46424))

Right branch of the tree:

Individuals who are working in public, private or other sectors belong to Q1, the 1st quintile of income.

5.5.2 Prediction by using pruned model for testing data/unseen data:

Error is still 31% so the model is good fit for training and testing data, there is no problem of over fitting. To improve the accuracy of the CART we go for ensemble tree.

Table 5.15 Summary of Bagging for classification

Type of random forest: classification

Number of trees:	500
No. of variables tried at each split:	3
OOB estimate of error rate:	31.99%

Error obtain through bagging is 31% there is no reduction in Error.

Table 5.16: Summary of Random forest

Type of random forest: classification

Number of trees:	500
No. of variables tried at each split:	3
OOB estimate of error rate:	31.96%

Error obtain from random forest is also 31%.so the train model is good fit for both testing and training data.

Table 5.17: Variable importance for classification tree

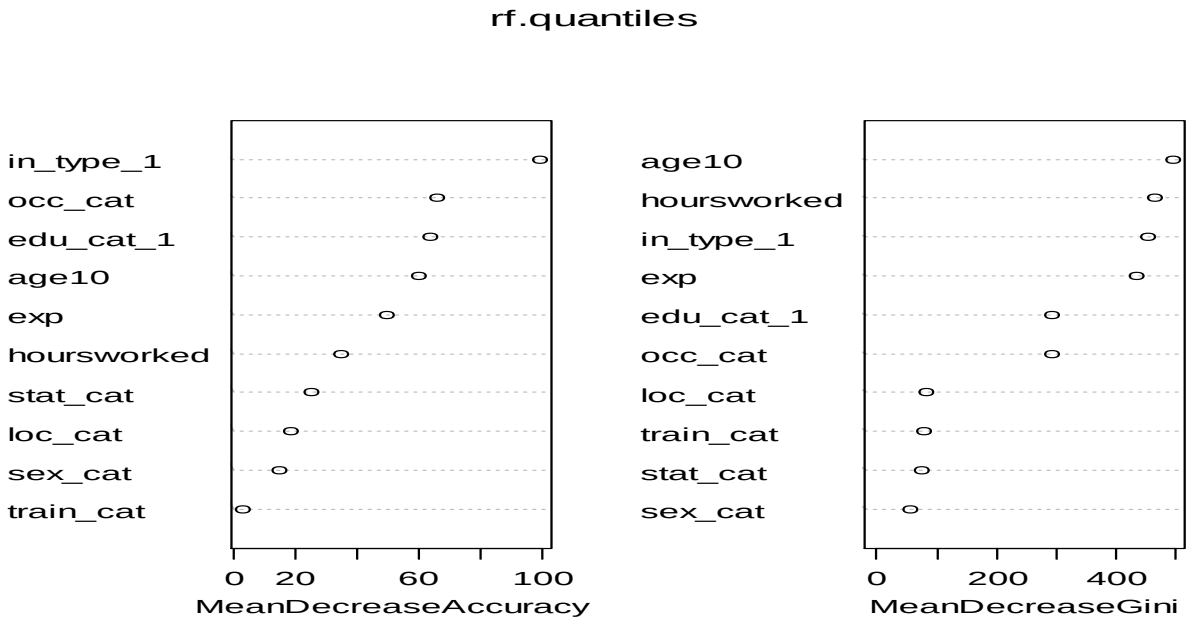
`	Mean Decrease Accuracy	Mean Decrease Gini
Sex	14.272723	56.94599
Status	24.788382	74.92308

Training	3.052402	79.98989
Experience	49.930020	434.75992
Education	64.003479	291.35316
Location	18.368662	83.33083
Occupation	65.732640	291.35015
Age	59.773517	495.43094
Hours worked	34.462996	462.93532
Type of ;industry	99.644075	453.86811

In Table 5.17 two measures of variable importance are given Mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model and Mean Decrease Gini index a measure of the total decrease in th node impurity that results from splits over that variable .As we have used classification so node impurity is measured through Deviance.

Importance of variable summarizes the overall importance of the variables using RSS for regression tree and Gini index for classification tree. Large value of importance of variable shows greater importance of that variable .type of industry is have mean decrease in Accuracy 99 and mean decrease in Gini is 453.which mean type of industry is the most important variable an individual is working then occupation comes and then Age of individual is effecting the income and so on

Figure 5.26 Plot of variable importance in Classification tree



The above figure shows that "Type of industry" is the most important variable then occupation, education, age, experience and so on.

Table 5.18: Summary of boosted model

	Variable	Relative influence
Type of industry	Type of industry	47.24546572
Education	Education	21.44569582
Occupation	Occupation	13.92794086
Age	Age	10.20133396
Experience	Experience	5.42274429
Hours worked	Hours worked	0.84564057
Location	Location	0.70023984
Status	Status	0.14192811
Training	Training	0.03856198
Sex	Sex	0.03044885

Table5.18 shows the influence of each variable , the Large value of the relative influence of a variable shows greater influence .it also shows that type of industry is the most important variable, education is the second influential variable then occupation and so on.

Figure 5.27: Plot of Relative influence for classification

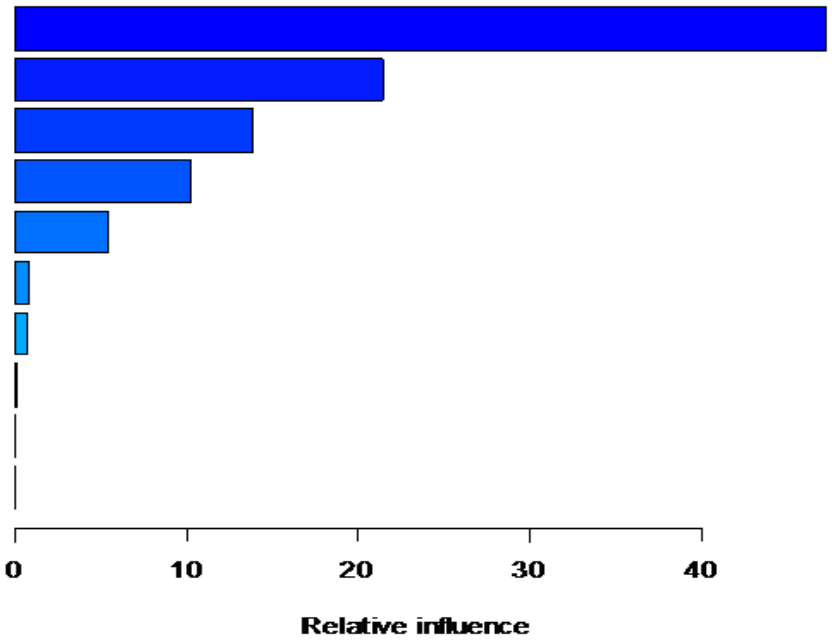


Figure 5.28 shows the influence of the variables graphically. The highest bar shows the influence of the type of industry. Second highest bar shows the influence of Education, third highest bar is for Occupation.

Table 5.19: Usual regression for Mincer Earning Function

Variables actually used in Regression: Income as dependent variable, where sex, status, location of work, hours worked, occupation, Experience, Experience square, education, and type of industry are used as predictors.
Residuals:
Min 1Q Median 3Q Max -4.8840 -0.2653 0.0413 0.3246 2.7869

5.6 Usual Multiple Regressions in R

5.6.1 Estimated Coefficients:

Table 5.18: Usual Multiple Regressions coefficients estimates

(Intercept)	7.7816833	0.1075255	72.371	< 2e-16 ***	
Experience	0.0115929	0.0005168	22.431	< 2e-16 ***	
Sex Male	0.4041182	0.0160086	25.244	< 2e-16 ***	
Status Married	0.1168475	0.0969457	1.205	0.228111	
Status Never married	0.0075589	0.0974581	0.078		0.938179
Status widow	0.0142781	0.1042013	0.137		0.891014
Training YES	0.1012716	0.0127454	7.946	2.06e-15 ***	
Education below middle	-0.1922342	0.0138514	-13.878	< 2e-16 ***	
Education Degree(16)	0.4662119	0.0140396	33.207	< 2e-16 ***	
Education No Formal	-0.3301371	0.0164903	-20.02	< 2e-16 ***	
Location Urban	0.1515391	0.0113754	13.322	< 2e-16 ***	
Occupation blue collar	0.1201026	0.030519	3.935	8.34e-05 ***	
Occupation pink collar	0.1134721	0.0295664	3.838	0.000125 ***	
Occupation white collar	0.3170425	0.0306068	10.359	< 2e-16 ***	
Hours worked	0.005375	0.0004533	11.856	< 2e-16 ***	
Industry type Gov. Sector	0.6881027	0.0289578	23.762	< 2e-16 ***	
Industry type others	-0.0139549	0.0278762	-0.501		0.616658
Industry Private Sector	0.362397	0.0457335	7.924	2.46e-15 ***	
Industry Public Sector	0.3990511	0.0290306	13.746	< 2e-16 ***	

Multiple R-squared: 0.4994, Adjusted R-squared: 0.4988

F-statistic: 848.3 on 18 and 15305 DF, p-value: < 2.2e-16

Significance Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5.18 show that predictors training, sex, experience, type of industry, occupation, location, hours of work are significant variables where only marital status and Type of industry "other" category is insignificant. Males earn more 40 percent more than female. Those individuals working in urban earn more than the individuals in rural .Those individuals who are doing white

collar jobs earn more than those who are doing blue and pink collar job. Individual working in Government sector earns more than people working in private, public and other sector.

5.7 Checking the Assuptions

Figure 5.28 Multiple regression plot

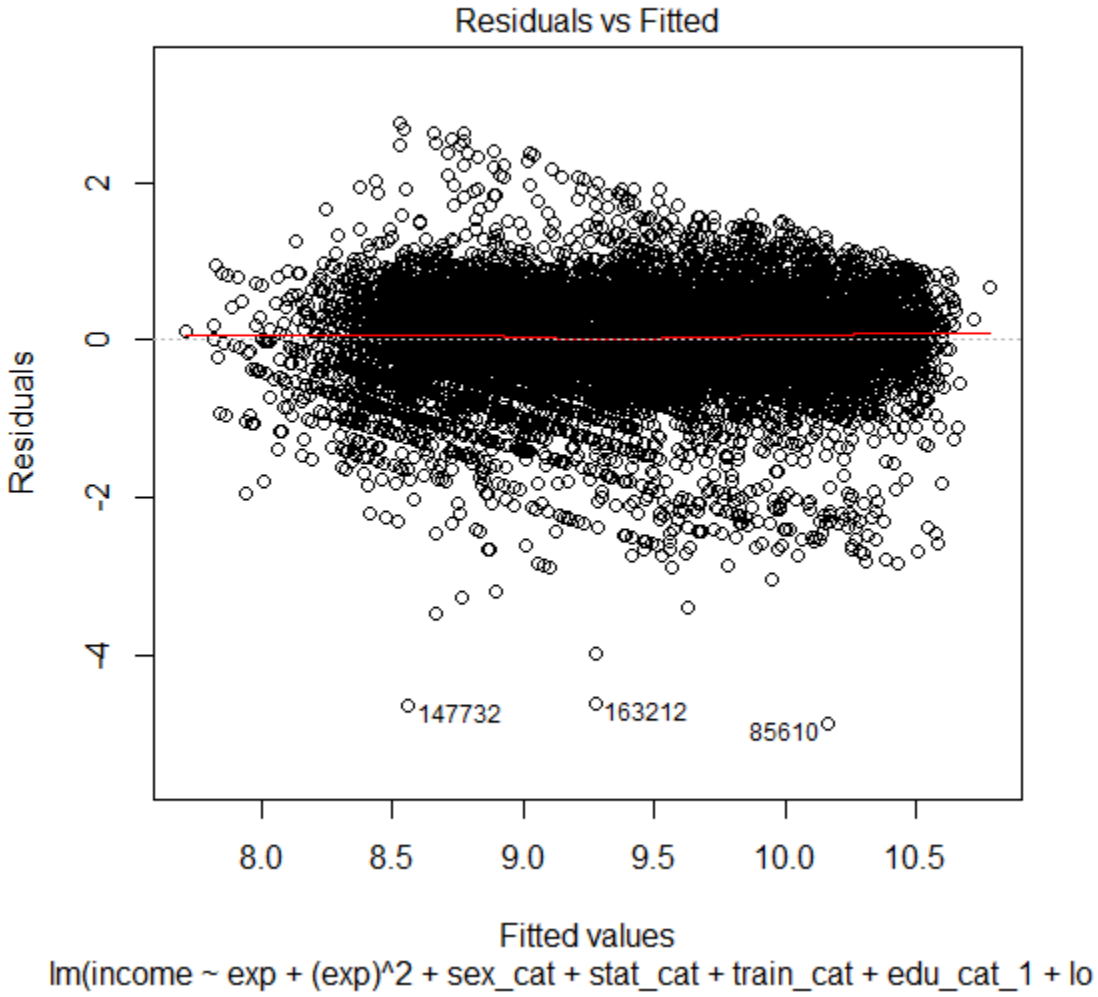


Figure 28 shows that fit of model.in the given plot most of the observations are close to the line so we can say that the model is good fit.

5.29 Normal Q-Q plot for regression

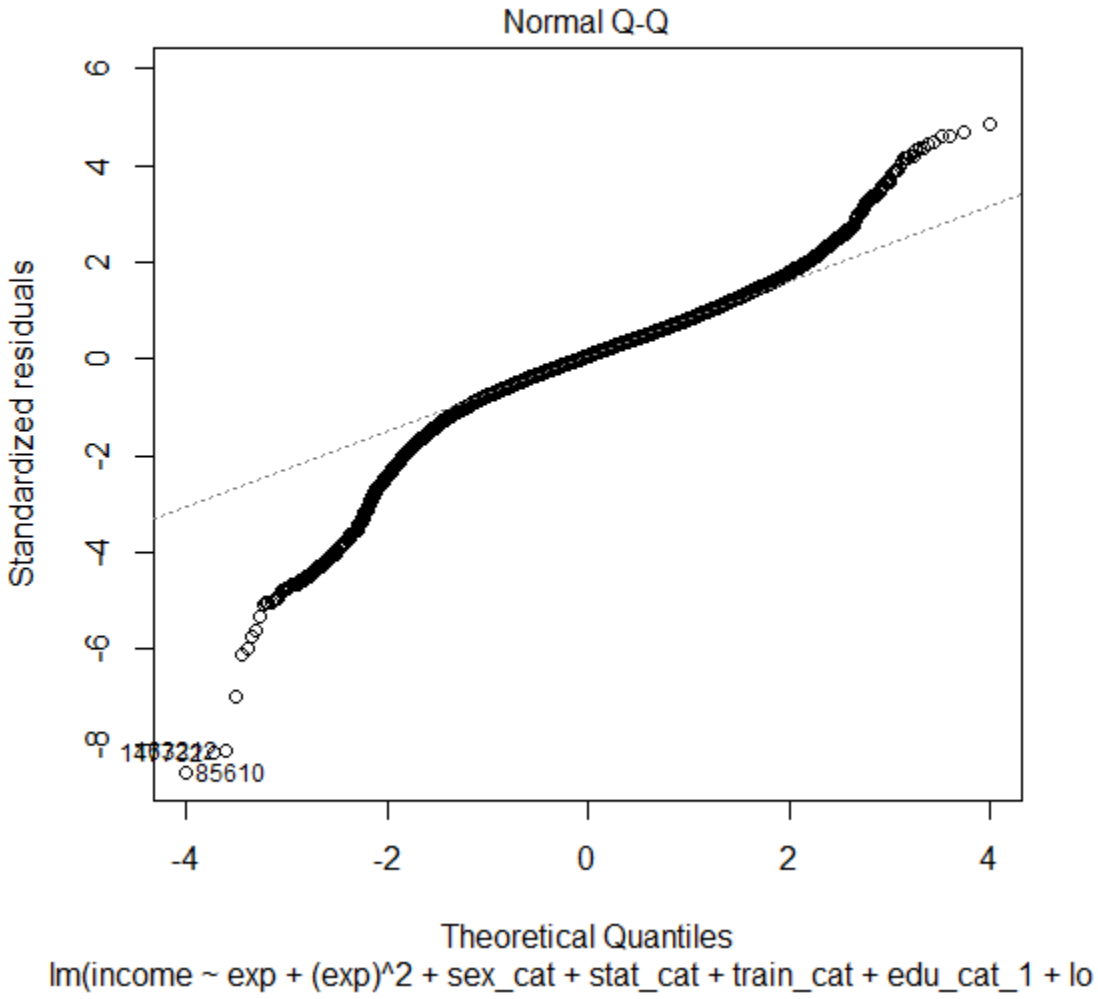


Figure 29 shows the Normality of the model, as Normal Q-Q plot is used to assess the normality of the variables. Here just the tails are away from the line which shows that model is following normal distribution with fat tail.

Chapter 6

Conclusion Summary

The prediction made through regression tree using complete set of observations show that individuals working in cooperative society, individual ownership, partnership and other, females earn less than male irrespective of their age difference. Those individuals who are working in government, private and public sectors and having below middle or no formal education then specifically, working in public and private sector earn less than those who are working in government sector irrespective of their age difference. Those who are having education above middle, working in government, private and public sector and are experienced earn more than those who are less experienced.

Regression tree obtained using half observations shows that type of industry, sex, age, occupation and education are significant variables. Type of industry an individual is working is the most important variable .Those individual who are working in government, private and public sector, white collar workers, their age greater than 38 and have education above middle earn more than those whose education is below middle. whose age is less than 38 earn less than those whose age is greater than 38.Females working in cooperative society, individual ownership, partnership and other sectors ,age less than 19,earn less than those whose age is greater than 19.

In the final prune tree which is free from the problem of over fitting, three variables are significant, type of industry, sex and education. Here age has pruned and show that those individuals who are working in government, private and public sector and having higher education earn more than those whose education is lower than middle and those who are working in government sectors earn more than private and public sector. Female of Individual of cooperative society, individual ownership, partnership and other sectors earn less than male of these sectors. Result of variable importance in case of regression tree show the importance of each variable , type of industry ,education, age and occupation are the most important and influential variables where status, training and location of work are least important.

In case of classification, there are Quintiles of income, which is a qualitative variable. The classification tree made for Quantiles predicted that those individuals who are working in

government ,public and private sectors and are blue and pink collar workers then specifically in government sector, belongs to lower group income and if they are working in private and public sector also belong to lower group income.

Those individual who are working in cooperative society, individual ownership, partnership and other sectors belong to lower income group if they are white collar workers and their age is greater than 38, having education above middle they belong to the highest income group but if individual have below middle education or no formal education and have age greater than 38, white collar worker and working in government ,private or public sector belongs to income group Q4.So we conclude from classification tree that those individuals whose age is greater than 38 ,doing white collar job ,having higher education and working in Government, private and public sector earn more. Type of industry an employee is working, occupation; education and age are important variable in the study. Advance techniques, bagging, random forest used for classification and regression tree show the improvement in the accuracy of the results and show that the model is good fit.

Comparing the results of simple multiple regression with regression tree we found that, in regression tree only type of industry, age, education and occupation is significant variable, where in usual regression all the predictors are significant except marital status and type of industry category “other” . Regression tree show the individual importance of each variable where usual regression does not.

References:

1. Bjorklund , A. and Kjellstrom ,C.,(2000), “Estimating the return to investments in education :how useful is the standard Mincer education?”, Economics of Education Review ,21:195-210
2. Metcalf ,D. ,(1971), “ The determinants of earnings changes: A regional analysis for the U.K.,1960-68 “,International economic Review,12(2).
3. Afzal .M., (2011),”Micro econometric Analysis of Private returns to education and determinants of earnings ”Pakistan economic and social review,49(1):39-68
4. Khan, S .,and Irfan, M.,(1985), “Rates of Returns to Education and the Determinants of Earnings in Pakistan”, The Pakistan Development Review, XXIV(3&4).
5. Tubman , P.,(1976), “The Determinants of Earnings:Genetics, Family, and other Environments; study of White Male Twins”, American Economic Association ,66(5):858-870
6. Nasir, Z., (1998), “Determinants of earnings in Pakistan: Findings from the labor Force Survey 1993-94”, The Pakistan Development Review, 37(3):251-274.
7. Kapoor., and Puri,A.,(1971), “The Determinates of personal Earnings: A study of industrial workers in Punjab”, Economics of Education Review.
8. Sutton,C.D.,(2005), “ classification and Regression Trees, Bagging, and Boosting”, Hand Book ofstatistics,24
9. Pakgohar,A.,Tabrizi,S.,R., khalili,M. and Esmaeili,A.,(2010), “The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining Approach” *procedia computer science*,3:764-769
10. Rogar,J,L.,(2000), “An introduction to classification and regression Tree(CART)Analysis”, *presented at annual meeting of the society for Academic Emergency medicine in san Francisco, California.*
11. Berndt,R.E.,(1991),“The practice of Econometrics :classic and contemporary” .
12. De'ath, G. and K. E. Fabricius (2000). "Classification and regression trees: a powerful yet simple technique for ecological data analysis." *Ecology* **81**(11): 3178-3192.

13. Gordon, L. (2013). "Using Classification and Regression Trees (CART) in SAS® Enterprise Miner TM For Applications in Public Health." *Data Mining and Text Analytics*: 089-2013.

14. Horning, N. (2013). "Introduction to decision trees and random forests." *American Museum of Natural History's.*

15. James, G., witten.D and Hastie,T., (2014). An Introduction to Statistical Learning: With Applications in R, Taylor & Francis.

16. Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest." *R news* **2**(3): 18-22.

17. Loh, W. Y. (2011). "Classification and regression trees." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(1): 14-23.

18. Ohno-Machado, L., et al. (2000). *Decision trees and fuzzy logic: a comparison of models for the selection of measles vaccination strategies in Brazi*l. Proceedings of the AMIA Symposium, American Medical Informatics Association.

19. Patel, H. D., et al. (2013). "Cost-effectiveness of a new rotavirus vaccination program in Pakistan: A decision tree model." *Vaccine***31**(51): 6072-6078.

20. Rokach, L. (2007). *Data mining with decision trees: theory and applications*, World scientific.

21. Thakur, G. S., et al. "UNDERSTANDING THE APPLICABILITY OF LINEAR & NON-LINEAR MODELS USING A CASE-BASED STUDY."

22. Varian, H. R. (2014). "Big data: New tricks for econometrics." *The Journal of Economic Perspectives*: 3-27.

23. Chang, Y. (2008). Robustifying regression and classification trees in the presence of irrelevant variables, ProQuest.

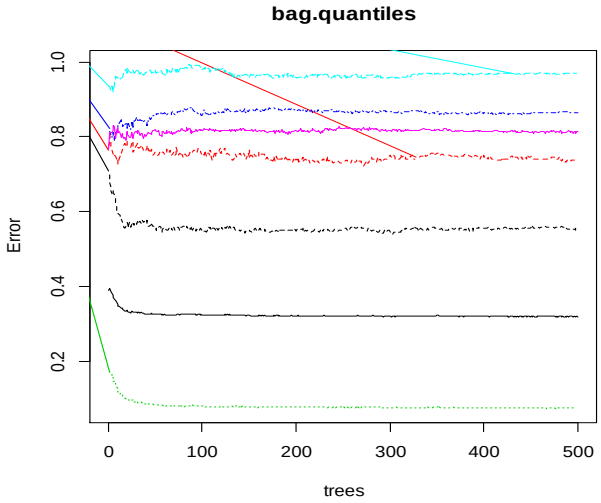
24. Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine." Annals of statistics: 1189-1232.

25. Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest." R news **2**(3): 18-22.

26. Zhang, D. (2006). Advances in machine learning applications in software engineering, IGI Global.

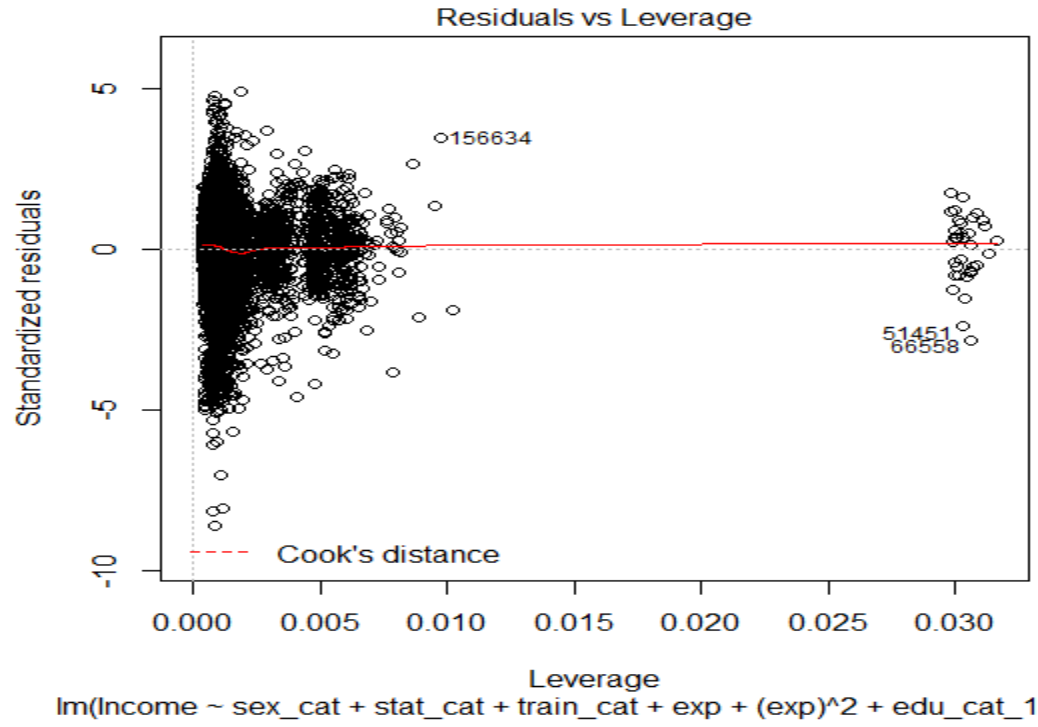
Appendix

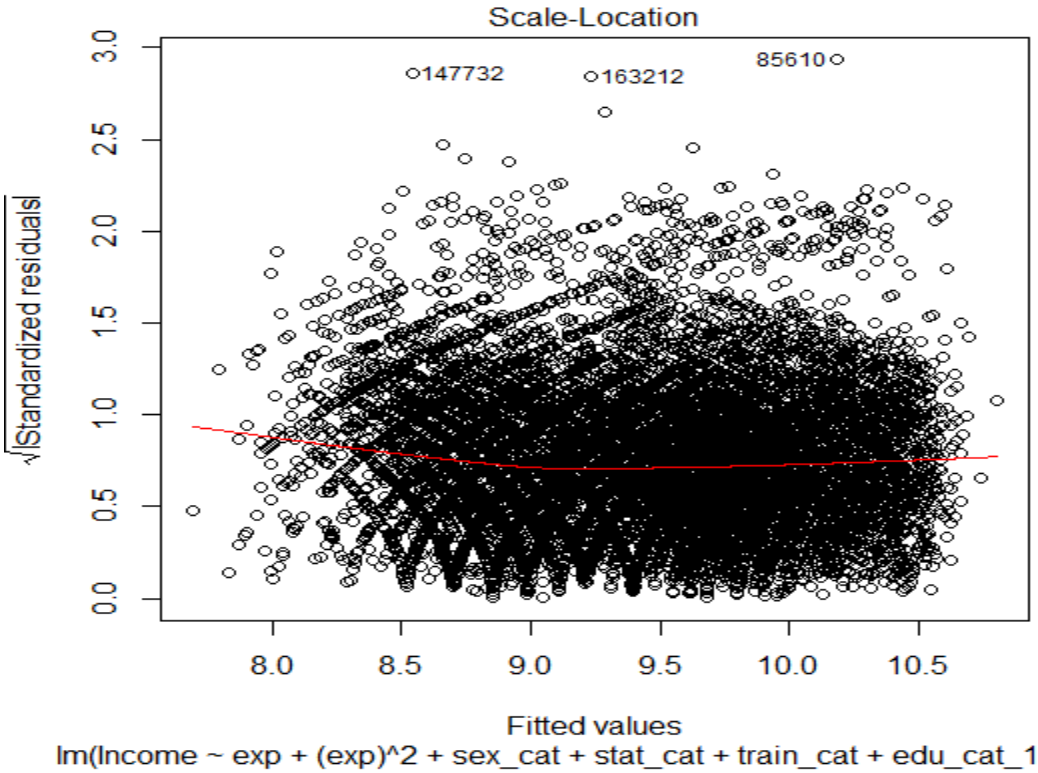
Bag Quintiles



Confusion matrix

	FALSE	Q1	Q2	Q3	Q4	Q5	Class. Error
FALSE	77	57	9	2	18	162	0.763076
Q1	22	4437	123	10	90	119	0.07581754
Q2	16	507	92	10	54	51	0.87397260
Q3	2	143	32	10	47	67	0.966777
Q4	19	224	61	13	117	155	0.801358
Q5	71	182	26	8	66	303	0.53810976


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
156634
51451
66558
5
0
-5
-10
0.000
0.005
0.010
0.015
0.020
0.025
0.030
lm(Income ~ sex_cat + stat_cat + train_cat + exp + (exp)^2 + edu_cat_1 + lo


Scale-Location
√|Standardized residuals|
Fitted values
147732
163212
85610
3.0
2.5
2.0
1.5
1.0
0.5
0.0
8.0
8.5
9.0
9.5
10.0
10.5
lm(Income ~ exp + (exp)^2 + sex_cat + stat_cat + train_cat + edu_cat_1 + lo

Confusion matrix

	FALSE	Q1	Q2	Q3	Q4	Q5	Class. Error
FALSE	85	59	11	3	14	153	0.73846
Q1	21	4440	121	7	93	119	0.07519
Q2	15	496	98	10	54	57	0.86575
Q3	2	145	31	9	49	65	0.97009967
Q4	17	235	58	13	110	156	0.81324278
Q5	71	181	28	10	74	292	0.55487805

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study finds the determinants of earning for Pakistan using data mining technique simple multiple Regression and Classification tree and regression tree (CART) .For improving the accuracy of prediction advance techniques, bagging, random forest and boost, for regression and classification has been used. Labor force survey data (2012-13) is used in the study. Main Variables used as predictors in the study are education, Sex, Marital status, training, and occupation, location of working, training, experience, age etc. Monthly income is used as dependent variable. In case of classification income is divided in Quintiles, which is used as a dependent variable for classification variable. Type of industry, education, age and occupation are found useful predictors in both classification and regression tree. Results of regression shows that female earns less than male even if they are working in the same type of industry and those who are working in government, private or public sectors and have higher education they earn more than individuals working in other sectors. Government employee's monthly average income is greater than private and public sector even they have same level of education. Results of classification tree shows that those individual who are working in government ,doing white collar job and age greater than 38 and education middle or above belong to Q4 and Q5 . While those individual who are working in other than government sector belongs to Q1.In Multiple regression all the predictors are useful except marital status.