A PROJECT REPORT
ON
CLASSIFYING CREDIT RISK

Submitted to Osmania University
in partial fulfillment of the requirements for the award of

MASTER OF SCIENCE IN STATISTICS

DEPARTMENT OF STATISTICS UNIVERSITY COLLEGE OF SCIENCE
OSMANIA UNIVERSITY HYDERABAD – INDIA

By

MD. NAZEER	Roll No.: 1007-17-508-019
MD.TAWHEEDJANI	Roll No.: 1007-17-508-014
B.VIDYASAGAR	Roll No.: 1007-17-508-017
V.PRASAD	Roll No.: 1007-17-508-011
G.SAI KUMAR	Roll No.: 1007-17-508-003
R.ASHOK NAIK	Roll No.: 1007-17-508-022
B.NAVEEN	Roll No.: 1007-17-508-028

Under the Supervision of

T. SANDHYA 2018

A PROJECT REPORT
ON
CLASSIFYING CREDIT RISK

Submitted to Osmania University in partial fulfillment of the requirements for the award of Master of Science in Statistics

DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE OF SCIENCE
OSMANIA UNIVERSITY
HYDERABAD – INDIA

By

MD. NAZEER	Roll No.: 1007-17-508-019
MD.TAWHEEDJANI	Roll No.: 1007-17-508-014
B.VIDYASAGAR	Roll No.: 1007-17-508-017
V.PRASAD	Roll No.: 1007-17-508-011
G.SAI KUMAR	Roll No.: 1007-17-508-003
R.ASHOK NAIK	Roll No.: 1007-17-508-022
B.NAVEEN	Roll No.: 1007-17-508-028

Under the Supervision of

T. SANDHYA 2018

CERTIFICATE

This is to certify that

Have submitted the project titled “CLASSIFYING CREDIT RISK” in partial fulfillment for the degree of Master of Science in Statistics.

MD. NAZEER	Roll No.: 1007-17-508-019
MD.TAWHEEDJANI	Roll No.: 1007-17-508-014
B.VIDYASAGAR	Roll No.: 1007-17-508-017
V.PRASAD	Roll No.: 1007-17-508-011
G.SAI KUMAR	Roll No.: 1007-17-508-003
R.ASHOK NAIK	Roll No.: 1007-17-508-022
B.NAVEEN	Roll No.: 1007-17-508-028

Head
Department of Statistics

Internal Examiner   External Examiner

DECLARATION

The research presented in this project has been carried out in the Department of Statistics, University college of Science, Osmania University, Hyderabad. The work is original has not been submitted so far, in part or full, for any other degree of diploma of any university.

Mr. MD.NAZEER

Mr. MD.TAWHEEDJANI

Mr. B.VIDYASAGAR

Mr. V.PRASAD

Mr. R.ASHOK NAIK

Mr. G.SAIKUMAR

Mr. B.NAVEEN

Department of Statistics
Osmania University
Hyderabad – 500007, T.S.
INDIA

ACKNOWLEDGEMENTS

I deem it a great pleasure to express my deep sense of gratitude and indebtedness to my research supervisor T.SANDHYA, Statistics department, University College of Science, Osmania University for his valuable guidance, and enlightening discussions throughout the progress of my project work.

I also express my sincere and heartfelt thanks to PROF.C.JAYALAKSHMI ,Head of Department, Department of Statistics, Osmania University for providing the necessary support and facilities in the department for completion of this work successfully.

It is indeed with great pleasure I record my thanks to Dr . G . JAYASREE , Chairperson, Board of Studies , Department of Statistics , Osmania University for having provided with all the facilities to carry out our work.

I thank Dr.N.Ch.BHATRACHARYULU , Dr.K.VANI, Dr.S.A.JYOTHI RANI, Dr.G.SIRISHA, Mrs.J.L.PADMA SHREE , for their encouragement and constant help during the research.

I would like to express my deepest gratitude to DR. M. VENUGOPALA RAO, BALA KARTHIK for their advice, guidance and involvement at various stages of this work , I would also like to thank them for their understanding and constant encouragement throughout this project.

I thank all Non-Teaching members of the Department of Statistics, who helped me during my Thesis work.

I am thankful to the Osmania University for permitting me to carry out this work.

CONTENTS

CHAPTER-1

INTRODUCTION

INTRODUCTION

1.1 INTRODUCTION

Nowadays, usage of credit card is a way of life. As world going towards digital every individual wants to live is life as digital. Credit card has become an indispensable payment instrument in many countries. Credit card business turned out to be a profitable field for German banks. Worldwide more and more credit card users are becoming wary of using a credit card due to fear of losing their hard earned money to Credit card frauds. For predicting Credit card issue based on some factors like Credit history, present employment, Property, job etc. In this project, Machine learning algorithms are used for predictive analysis.

1.1 SCOPE OF THE PROBLEM

This problem is related to the prediction of Issuing German credit card using data collected from German banks.

Source:https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

Broadly the objectives of the problem are:

- What are the parameters to predict the issue of German credit card in terms of savings account, job etc.
- An algorithm by which we can predict the issuing of new credit card.

1.2 Description of the dataset

- X1:- Status of existing checking account
- X2:- Credit history

- X3:- Savings account/bonds
- X4:- Present employment since
- X5:- Personal status and sex
- X6:- Other debtors / guarantors
- X7:- Property
- X8:- Other installment plans
- X9 :- Housing
- X10:- Job
- X11:- Telephone
- X12:- Foreign worker
- X13:- Purpose Quantitative variables:
- X14:- Duration in month
- X15:- Credit amount
- X16:- Installment rate in percentage of disposable income
- X17:- Present residence since
- X18:- Age in years
- X19:- Number of existing credits at this bank
- X20:- Number of people being liable to provide maintenance for
- Dep:- Credit Risk

Value		Label
x1	1	< 0 DM
	2	0 <= ... < 200 DM
	3	... >= 200 DM
	4	no checking account
x2	1	no credits taken/All credits paid back duly
	2	all credits at this bank paid back duly
	3	existing credits paid back duly till now
	4	delay in paying off in the past
	5	critical account/Other credits existing (not at this bank)
x3	1	... < 100 DM
	2	100 <= ... < 500 DM
	3	500 <= ... < 1000 DM
	4	.. >= 1000 DM
	5	unknown/ no savings account
x4	1.00	unemployed
	2.00	... < 1 year
	3.00	<= ... < 4 years
	4.00	<= ... < 7 years
	5.00	.. >= 7 years
x5	1.00	male : divorced/separated
	2.00	female : divorced/separated/married
	3.00	male : single
	4.00	male : married/widowed
	5.00	female : single
x6	1.00	none
	2.00	co-applicant
	3.00	guarantor
x7	1.00	real estate
	2.00	if not A121 : building society savings agreement/insurance
	3.00	if not A121/A122 : car or other, not in attribute 6
	4.00	unknown / no property
x8	1.00	bank
	2.00	stores
	3.00	None
x9	1.00	rent
	2.00	own
	3.00	for free
x10	1.00	unemployed/ unskilled - non-resident
	2.00	unskilled - resident
	3.00	skilled employee / official
	4.00	management/ self-employed/highly qualified employee/ officer
x11	1.00	none
	2.00	yes, registered under the customer's name
x12	1.00	yes
	2.00	no
x13	1.00	A40 car (new)
	2.00	A41 car (used)
	3.00	A42 furniture/equipment
	4.00	A43 radio/television
	5.00	A44 domestic appliances
	6.00	A45 repairs
	7.00	A46 education
	9.00	A48 retraining
	10.00	A49 business
	11.00	A410
Credit Risk	0.00	Bad
	1.00	Good

CHAPTER-2

REVIEW OF MACHINE LEARNING PROCESS

REVIEW OF MACHINE LEARNING PROCESS

2.0 Need of Machine Learning

In this age of modern technology, there is one resource that we have in abundance: a large amount of structured and unstructured data. In the second half of the twentieth century, machine learning evolved as a subfield of artificial intelligence that involved the development of self-learning algorithms to gain knowledge from that data in order to make predictions. Instead of requiring humans to manually derive rules and build models from analyzing large amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models, and make data-driven decisions. Not only is machine learning becoming increasingly important in computer science research but it also plays an ever greater role in our everyday life.

2.1 Machine Learning Process

The CRISP-DM (Cross-Industry Standard Process for Data Mining) Process was designed specifically for the data mining. However, it is flexible and thorough enough that it can be applied to any analytical project whether it is predictive analytics, data science, or Machine learning. The Process has the following six phases

- Business Understanding
- Data Understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

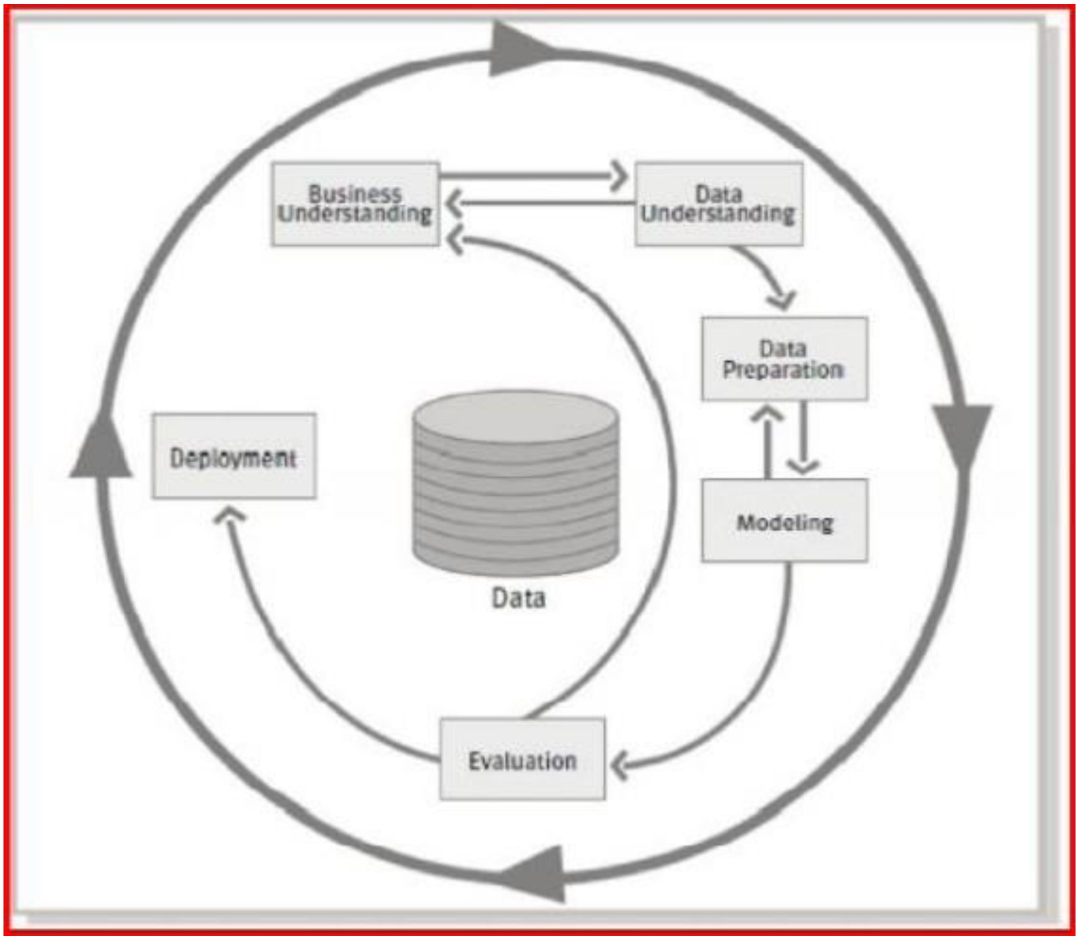


Fig 2.1 crisp- dm diagram

And, each phase has different steps covering important tasks which are mentioned below:

2.1.1 Business Understanding

It is very important step of the process in achieving the success. The purpose of this step is to identif the requirements of the business so that you can translate them into analytical objectives. It has the following tasks:

- Identify the Business objective
- Assess the situation
- Determine the Analytical goals
- Produce a project plan

2.1.2 Data Understanding

After enduring the all-important pain of the first step, you can now get your hands on the data. The task in this process consist the following

- Collect the data
- Describe the data
- Explore the data
- Verify the data Quality

2.1.3 Data Preparation

This step is relatively self-explanatory and in this step the goal is to get the data ready to input in the algorithms. This includes merging, feature engineering, and transformations. If imputation for missing values / outliers is needed then, it happens in this step. The key five tasks under this step are as follows:

- Select the data
- Clean the data
- Construct the data
- Integrate the data
- Format the data

2.1.4 Modelling

Oddly, this process step includes the consideration that you already thought of and prepared for. In this, one will need at least a modicum of an idea about how they will be modeling. Remember, that this is flexible, iterative process and some strict linear flow chart such as an aircrew checklist.

Below are the tasks in this step:

- Select a modeling technique
- Generate a test design
- Build a model
- Assess a Mode

Both cross validation of the model (using train/test or K fold validation) and model assessment which involves comparing the models with the chosen criterion (RMSE, Accuracy, ROC) will be performed under this phase.

2.1.5 Evaluation

In the evaluation process, the main goal is to confirm that the work that has been done and the model selected at this point meets the business objective. Ask yourself and others, have we achieved the definition of success? And, here are the tasks in this step:

- Evaluate the results
- Review the process
- Determine the next steps

2.1.6 Deployment

If everything is done according to the plan up to this point, it might come down to flipping a switch and your model goes live. Here are the tasks in this step:

- Deploying the plan
- Monitoring and maintenance of the plan
- Producing the final report

2.2 Types of Machine Learning

Broadly, the Machine Learning Algorithms are classified into 3 types.

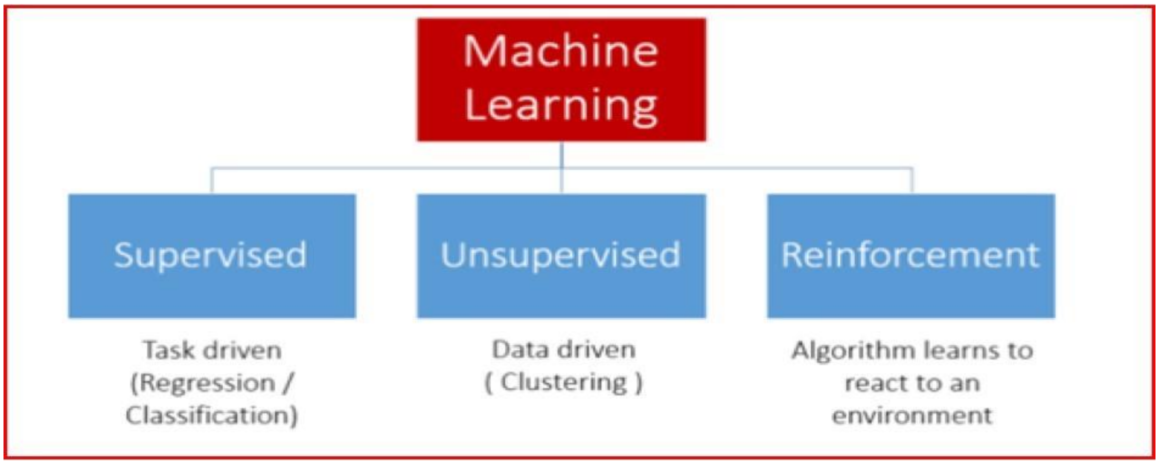


Fig 2.2 Types of Machine learning

2.2.1 Supervised Learning

This algorithm consists of a target / outcome / dependent variable which is to be predicted from a given set of predictors / independent variables. Using these set of variables, we generate a function that maps inputs to desired output. The training process continues until the model achieves a desired level of accuracy on the training data.

The process of Supervised Learning model is illustrated in the below picture:

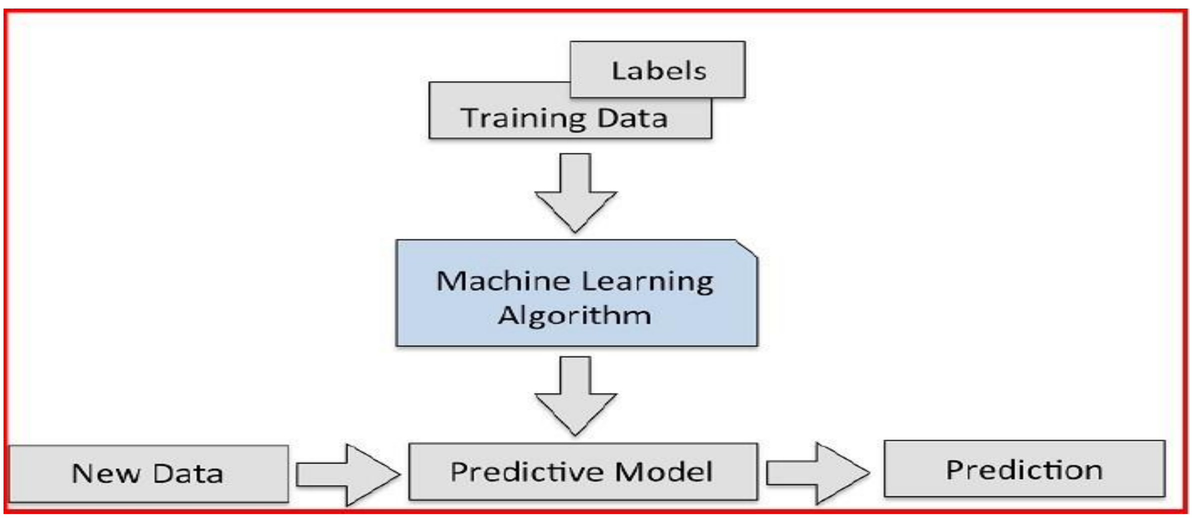


Fig 2.2.1 Supervised Learning

Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression,…etc

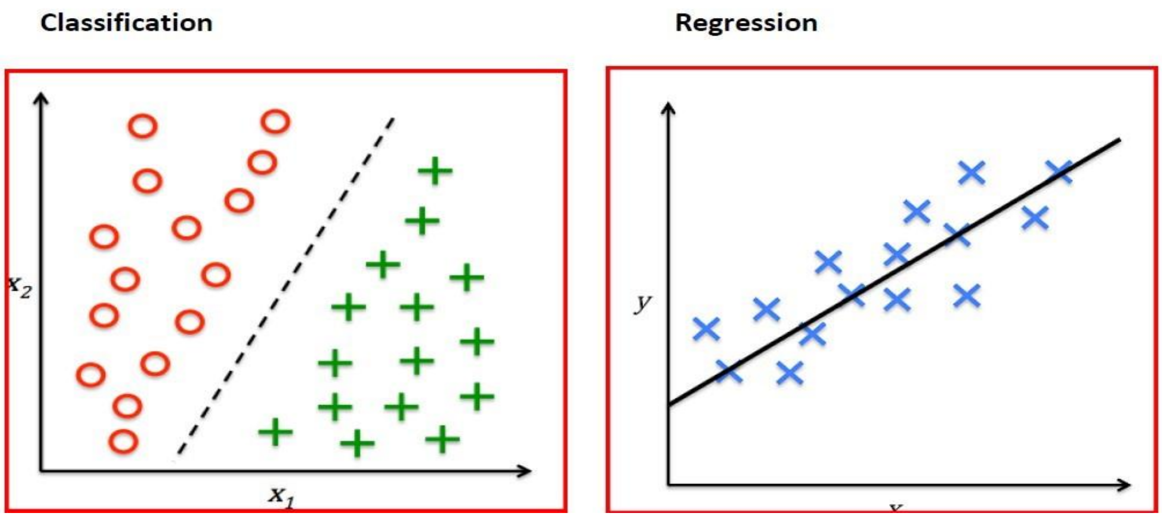


Fig 2.2.1 Classification and regression

2.2.2 Unsupervised Learning

In this algorithm, we will not have any target or outcome variable to predict / estimate. It is used for clustering population into different groups, which is widely used for segmenting customers in different groups for specific intervention. (More of Exploratory Analysis)

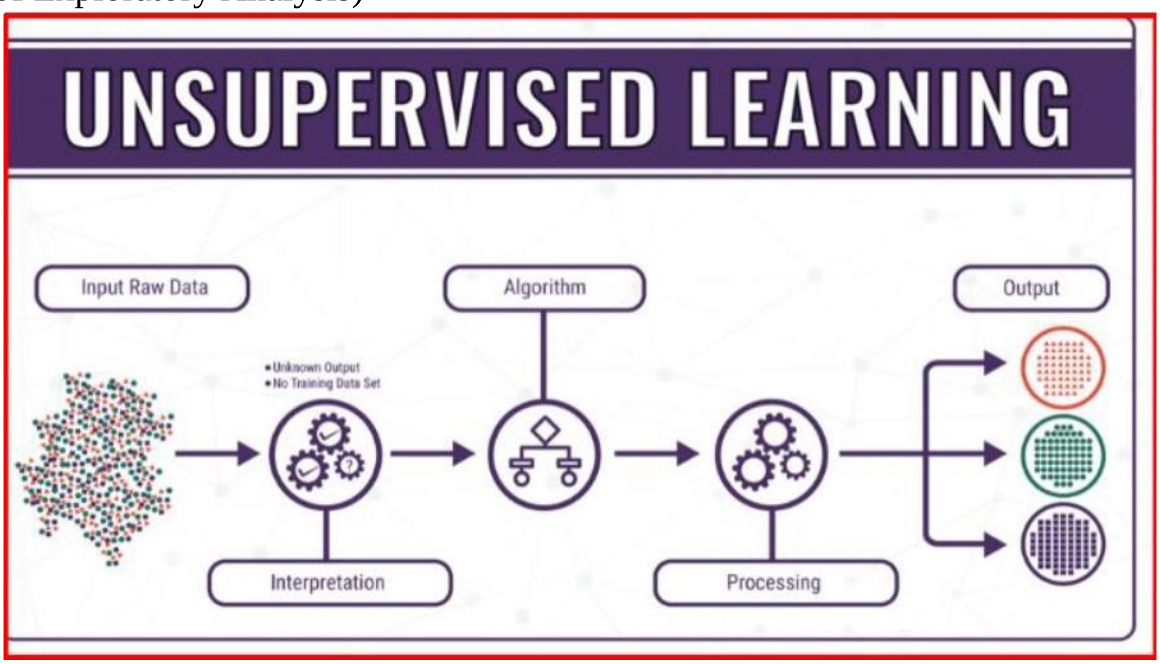


Fig 2.2.2 Unsupervised Learning

Examples of Unsupervised Learning: Data reduction techniques, Cluster Analysis, Market Basket Analysis,…etc

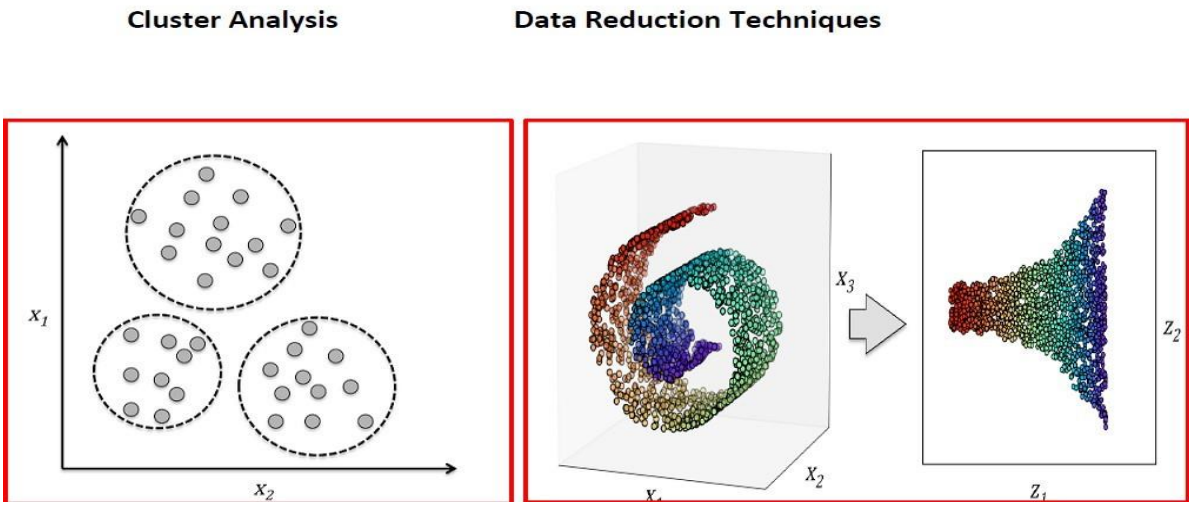


Fig 2.2.2Cluster Analysis Data Reduction Techniques

2.2.3 Reinforcement Learning

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

The process of reinforcement learning is illustrated in the below picture:

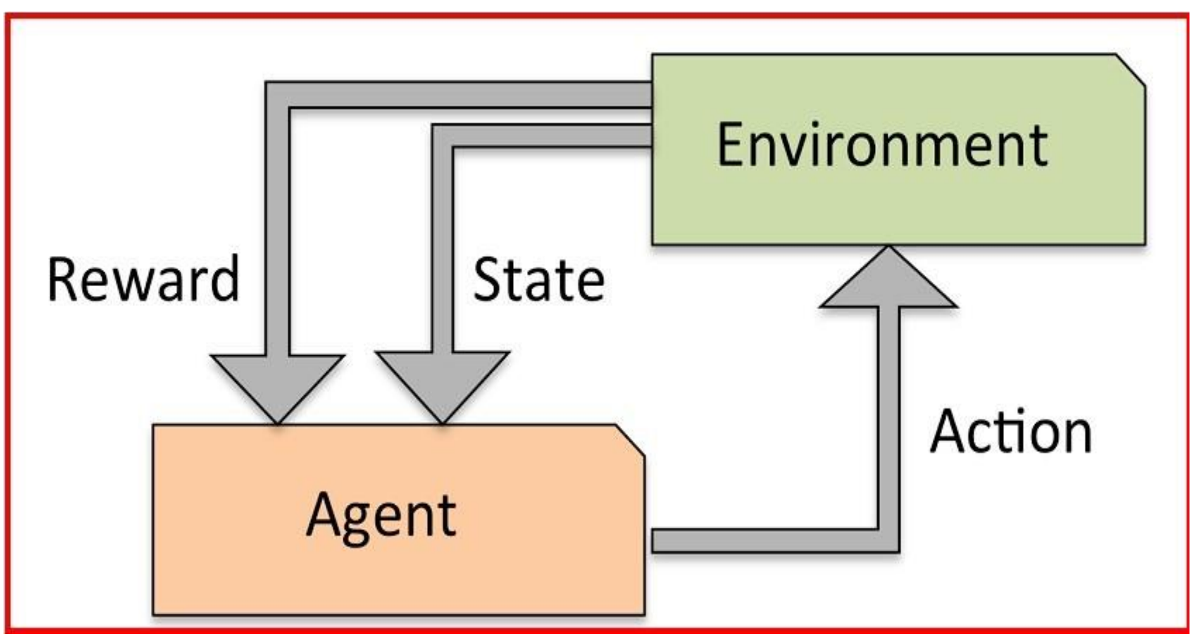


Fig 2.2.3 Reinforcement Learning

Examples of Reinforcement Learning: Markov Decision Process, Self-driving cars,...etc

2.3 Choosing the algorithm

Choosing the right algorithm will depend on the type of the problem we are solving and also depends on the scale of the dependent variable. In case of continuous target variable, we will use regression algorithms and in case of categorical target, we will use classification algorithms and for the model which doesn't have target variable, we will use either cluster analysis / data reduction techniques.

Below picture describes the process of choosing the right algorithm:

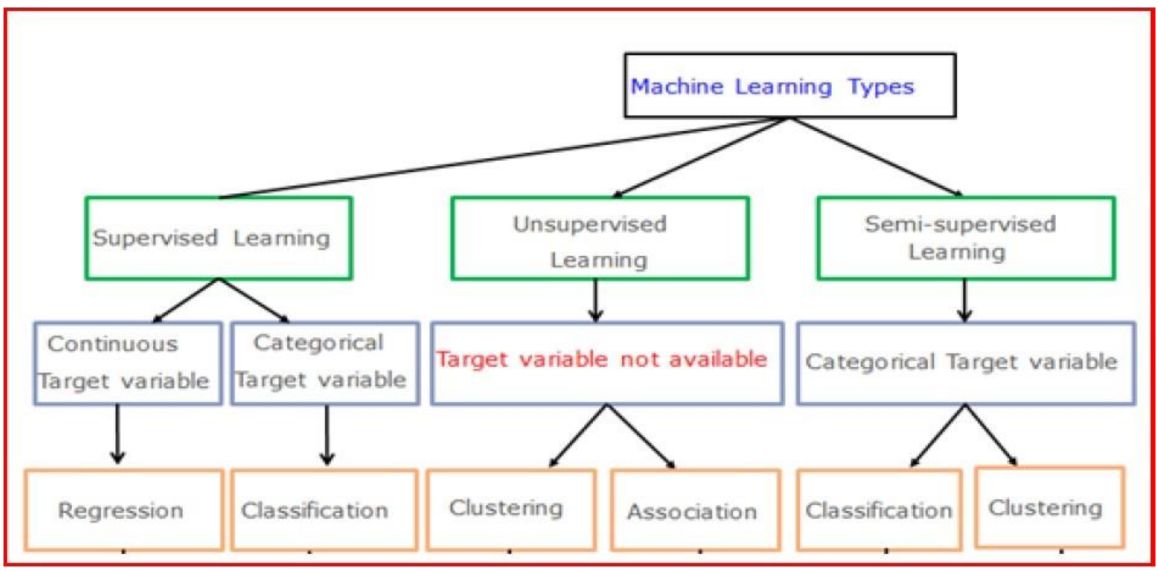


Fig 2.3 Machine learning process

2.3.1 Types of Regression Algorithms

There are many Regression algorithms in machine learning, which will be used in different regression applications. Some of the main regression algorithms are as follows:

- Simple Linear Regression:-

In simple linear regression, we predict scores on one variable from the data of second variable. The variable we are forecasting is called the criterion variable and referred to as Y. The variable we are basing our predictions on is called the predictor variable and denoted as X.

- Multiple Linear Regression:-

Multiple linear regression is one of the algorithms of regression technique, and is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one dependent variable with two or more independent variables. The independent variables can be either continuous or categorical.

- Polynomial Regression:-

Polynomial regression is another form of regression in which the maximum power of the independent variable is more than 1. In this regression technique, the best fit line is not a straight line instead it is in the form of a curve.

- Support Vector Machines:-

Support Vector Machines can be applied to regression problems as well as Classification. It contains all the features that characterizes maximum margin algorithm. Linear learning machine maps a non-linear function into high dimensional kernel-induced feature space. The system capacity will be controlled by parameters that do not depend on the dimensionality of feature space.

- Decision Tree Regression:-

Decision tree builds regression models in the form of a tree structure. It breaks down the data into smaller subsets and while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

- Random Forest Regression:-

Random Forest is also one of the algorithms used in regression technique. It is very a flexible, easy to use machine learning algorithm that produces, even without hyper -parameter tuning, a great result most of the time. It is also one of the most widely used algorithms because of its simplicity and the fact that it can used for both regression and classification tasks. The forest it builds is an ensemble of Decision Trees, most of the time trained with the "bagging" method.

Other than these we have regularized regression models like Ridge, LASSO and Elastic Net regression which are used to select the key parameters and these is also Bayesian regression which works with the Bayes theorem.

2.3.2 Types of Classification Algorithms

There are many Classification algorithms in machine Learning, which can be used for different classification applications. Some of the main classification algorithms are as follows:

- Logistic Regression/Classification:- Logistic regression falls under the category of supervised learning; it measures the relationship between the dependent variable which is categorical with one or more than one independent variables b-y estimating probabilities using a logistic/sigmoid function. Logistic regression can generally be used when the dependent variable is Binary or Dichotomous. It means that the dependent variable can take only two possible values like "Yes or No", "Living or dead".

- K -Nearest Neighbours:- K-NN algorithm is one of the most straightforward algorithms in classification, and it is one of the most used ML algorithms. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. It can also use for regression — output is the value of the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbours.

- Naive Bayes:- Naive Bayes is a type of Classification technique based on Bayes' theorem, with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a Particular feature in a class is unrelated to the presence of any other function. Naive Bayes model is accessible to build and particularly useful for extensive datasets.

- Decision Tree Classification:- Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The first decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.
- Support Vector Machines:- A Support Vector Machine is a type of Classifier, in which a discriminative classifier is formally defined by a separating hyper plane. The algorithm outputs an optimal hyper plane which categorises new examples. In two dimensional space, this hyper plane is a line dividing a plane in two parts where in each class lay in either side.

- Random Forest Classification:- Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. The forest it builds is an ensemble of Decision Trees, most of the times the decision tree algorithm trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. And Random Forest is also very powerful to find the variable importance in classification/ Regression problems.

2.3.3 Types of Unsupervised Learning

Clustering is the type of unsupervised learning in which an unlabelled data is used to draw inferences. It is the process of grouping similar entities together. The goal of this unsupervised machine learning technique is to find similarities in the data points and group similar data points together and also to figure out which cluster should a new data point belong to.

2.3.3(a) Types of Clustering Algorithms:-

There are many Clustering algorithms in machine learning, which can be used for different clustering applications. Some of the main clustering algorithms are as follows:

- Hierarchical Clustering:- Hierarchical clustering is one of the algorithms of clustering technique, in which similar data is grouped in a cluster. It is an algorithm that builds the hierarchy of clusters. This algorithm starts with all the data points assigned to a bunch of their own. Then, two nearest groups are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. It starts by assigning each data point to its bunch. Finds the closest pair using Euclidean distance and merges them into one cluster. This process is continued until all data points are clustered into a single cluster.

- K -Means Clustering:-K-Means clustering is one of the algorithms of clustering technique, in which similar data is grouped into a cluster. K-means is an iterative algorithm that aims to find local maxima in each iteration. It starts with K as the input which is the desired number of clusters. Input k centroids in random locations in your space. Now, with the use of the Euclidean distance method, calculates the distance between data points and centroids, and assign data point to the cluster which is close to its centroid. Re calculate the cluster centroids as a mean of data points attached to it. Repeat until no further changes occur.

2.3.3(b) Types of Dimensionality Reduction Algorithms:-

There are many dimensionality reduction algorithms in machine learning, which are applied for different dimensionality reduction applications. One of the main dimensionality reduction techniques is Principal Component Analysis (PCA) / Factor Analysis.

Principal Component Analysis (Factor Analysis):-

Principal Component Analysis is one of the algorithms of Dimensionality reduction. In this technique, it transforms data into a new set of variables from input variables, which are the linear combination of real variables. These Specific new set of variables are known as principal components. As a result of the transformation, the first primary component will have the most significant possible variance, and each following component in has the highest possible variance under the constraint that it is orthogonal to the above components. Keeping only the best m < n components, reduces the data dimensionality while retaining most of the data information.

2.4 Choosing and comparing models through Pipelines

When you work on machine learning project, you often end up with multiple good models to choose from. Each model will have different performance characteristics. Using re-sampling methods like k-fold cross validation; you can get an estimate of how accurate each model may be on unseen data. You need to be able to use these estimates to choose one or two best models from the suite of models that you have created.

2.4.1 Model Validation

When you are building a predictive model, you need to evaluate the capability or generalization power of the model on unseen data. This is typically done by estimating accuracy using data that was not used to train the model, often referred as cross validation.

A few common methods used for Cross Validation:

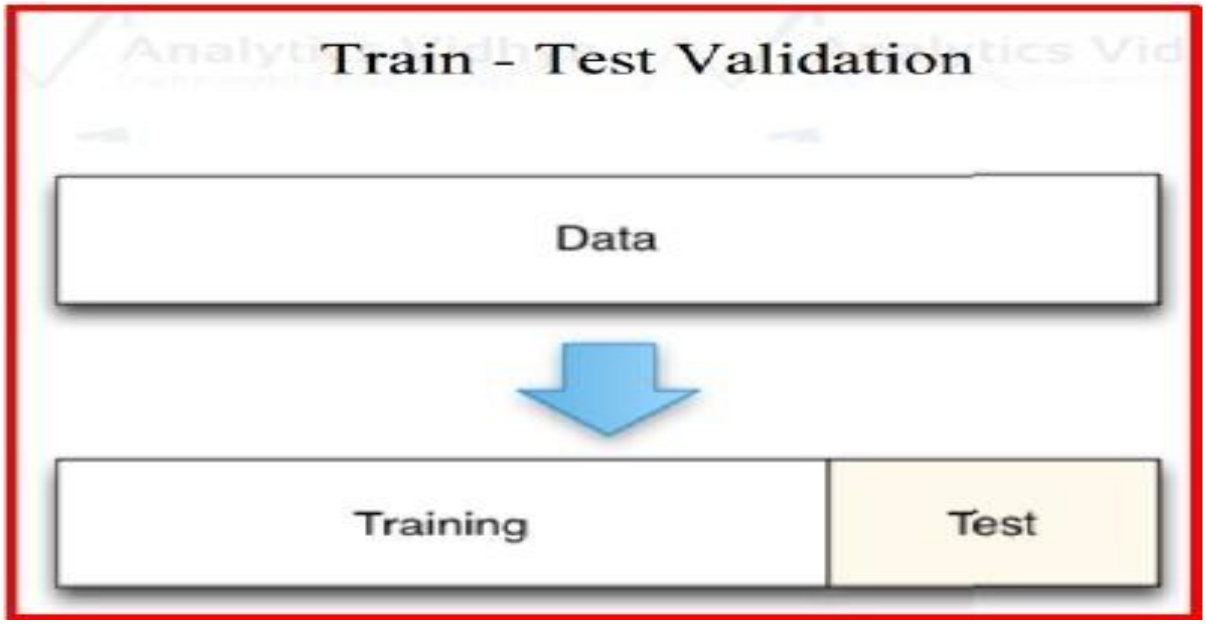


Fig 2.4.1 Model Validation

1) The Validation set Approach (Holdout Cross validation)

In this approach, we reserve large portion of dataset for training and rest remaining portion of the data for model validation. Ideally people will use 70-30 or 80-20 percentages for training and validation purpose respectively.
A major disadvantage of this approach is that, since we are training a model on a randomly chosen portion of the dataset, there is a huge possibility that we might miss-out on some interesting information about the data which, will lead to a higher bias.

2) K-fold cross validation

As there is never enough data to train your model, removing a part of it for validation may lead to a problem of under fitting. By reducing the training data, we risk losing important patterns/ trends in data set, which in turn increases error induced by bias. So, what we require is a method that provides ample data for training the model and also leaves ample data for validation. K Fold cross validation does exactly that.

In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness of our model. As can be seen, every data point gets to be in a validation set exactly once, and gets to be in a training set k-1 times. This significantly reduces the bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set. Interchanging the training and test sets also adds to the effectiveness of this method. As a general rule and empirical evidence, K = 5 or 10 is preferred, but nothing's fixed and it can take any value. Below are the steps for it:

- Randomly split your entire dataset into k "folds".

- For each k-fold in your dataset, build your model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for k-th fold.

- Record the error you see on each of the predictions. Repeat this until each of the k-folds has served as the test set.

- The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model.

Below is the visualization of a k-fold validation when k=5.

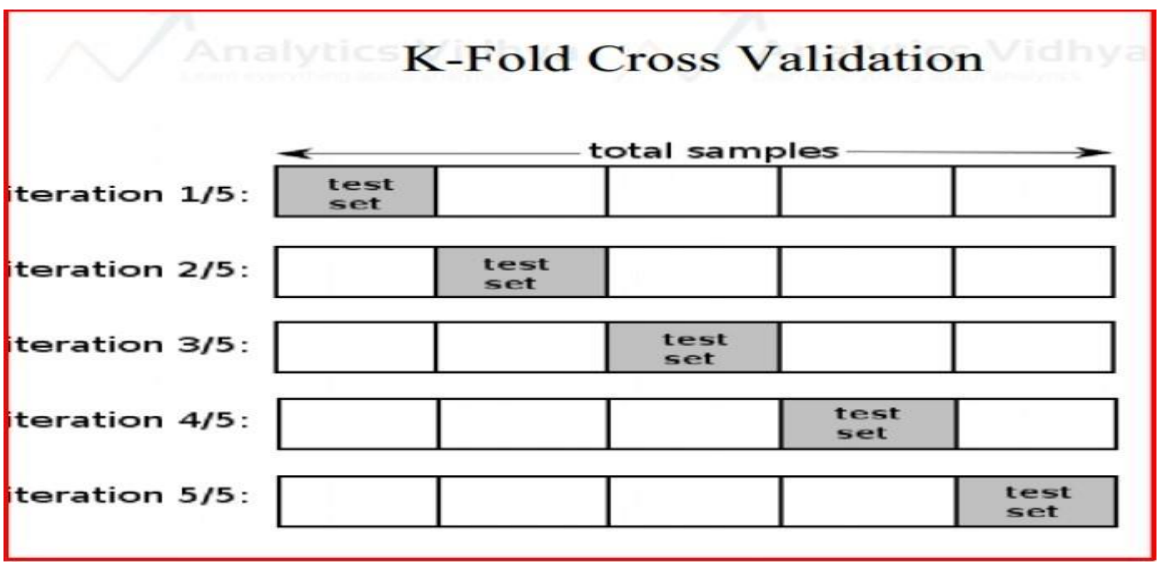


Fig 2.4.1 Model Validation

How to choose K:

- Smaller dataset: 10-fold cross validation is better
- Moderate dataset: 5 or 6 fold cross validation works mostly
- Big dataset: Train – Val split for validation

Other than this, we have Leave one out cross validation (LOOCV), in which each record will be left over from the training and then, the same will be used for testing purpose. This process will be repeated across all the respondents.

2.5 Model Diagnosis with over fitting and under fitting

2.5.1 Bias and Variance

A fundamental problem with supervised learning is the bias variance trade-off. Ideally, a model should have two key characteristics

- Sensitive enough to accurately capture the key patterns in the training dataset.
   Generalized enough to work well on any unseen dataset.

Unfortunately, while trying to achieve the above-mentioned first point, there is an ample chance of over-fitting to noisy or unrepresentative training data points leading to a failure of generalizing the model. On the other hand, trying to generalize a model may result in failing to capture important regularities.

If model accuracy is low on a training dataset as well as test dataset, the model is said to be under-fitting or that the model has high bias. The Bias refers to the simplifying assumptions made by the algorithm to make the problem easier to solve. To solve an under-fitting issue or to reduce bias, try including more meaningful features and try to increase the model complexity by trying higher-order interactions

The Variance refers to sensitivity of a model changes to the training data. A model is giving high accuracy on a training dataset, however on a test dataset the accuracy drops drastically then, the model is said to be over-fitting or a model that has high variance.

To solve the over-fitting issue Try to reduce the number of features, that is, keep only the meaningful features or try regularization methods that will keep all the features. Ideal model will be the trade-off between Under fitting and over fitting like mentioned in the below picture.

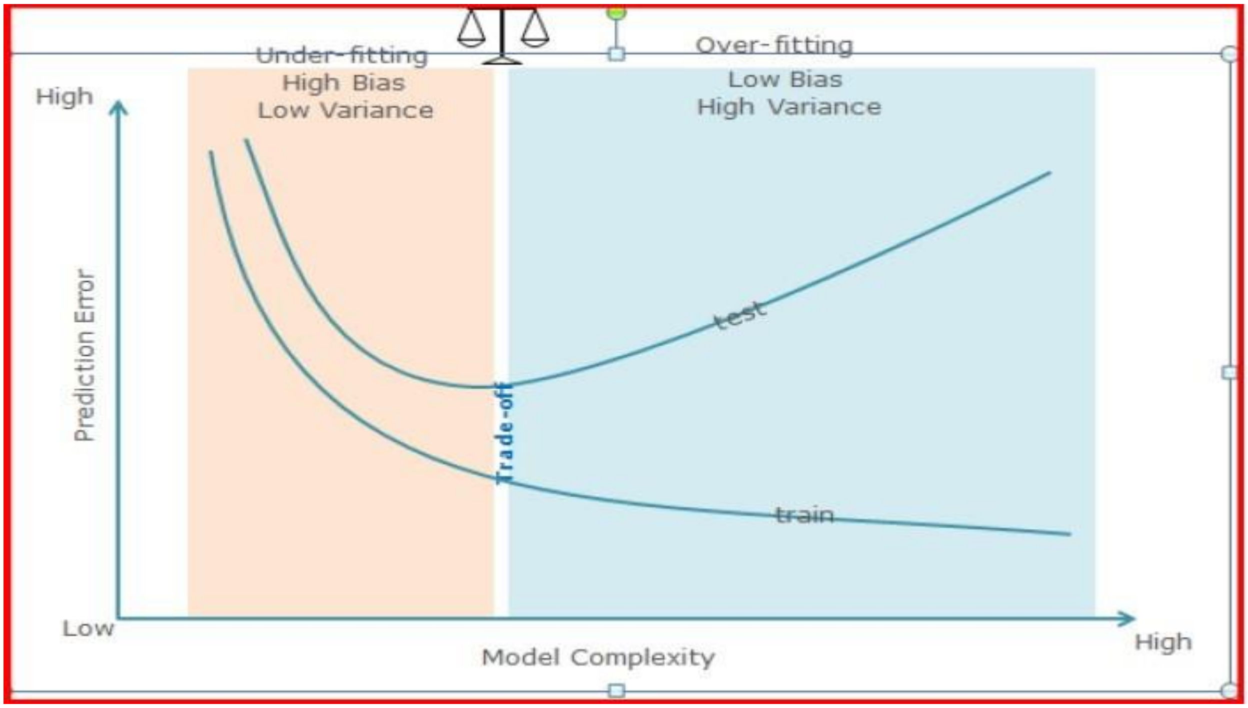


Fig 2.5.1 Bias and Variance

And, the Hyper parameters will be tuned in the below mentioned ways to reach the optimal solution: 1) Grid Search

2) Random Search

3) Manual Tuning

2.5.2 Model Performance Matrix

Model evaluation is an integral part of the model development. Based on model evaluation and subsequent comparisons, we can take a call whether to continue our efforts in model enhancement or cease them and select the final model that should be used / deployed.

 Evaluating Classification Models:

Confusion Matrix:

Confusion matrix is one of the most popular ways to evaluate a classification model. A confusion matrix can be created for a binary classification as well as a multi-class classification model.

A confusion matrix is created by comparing the predicted class label of a data point with its actual class label.

This comparison is repeated for the whole dataset and the results of this comparison are compiled in a matrix or tabular format And, below are the various measures that will be used to assess the performance of the model based on the requirement of the problem and as well as data.

Table 2.5.2 Confusion Matrix

Predicted classed				
		Positive (C_0)	Negative (C_1)	
Actual class	Positive (C_0)	a = number of correctly Classified c0 cases	c = number of c_0 cases Incorrectly classified as c_1	Precision = a/(a + c)
	Negative (C_0)	b = number of c_1 cases Incorrectly classified as c_0	d = number of correctly classified c_1 cases	
		Sensitivity (Recall) = a/(a+b)	Specificity = d/c+d	Accuracy = (a+b)(a+b+c+d)

Specificity : The ratio of actual negative cases that are identified correctly.

shows an example confusion matrix.

. Example of classifications Accuracy measurement

Predicted classed				
		Positive (C_0)	Negative (C_1)	
Actual class	Positive (C_0)	80	30	Precision = 70/110=0.63
	Negative (C_1)	40	90	
		Recall=80/120=0.67	Specificity = 90/240=0.75	Accuracy = 80+90/240=0.71

And, below are the various measures that will be used to assess the performance of the model based on the requirement of the problem and as well as data.

Metric	Description	Formula
Accuracy	What% of predictions were Correct?	(TP + TN)/(TP + TN + EP + FN)
Misclassification rate	What % of prediction is wrong?	(FP + FN)/(TP + TN + FP + FN)
True positive rate OR Sensitivity or recall (completeness)	What % of positive cases did Model catch?	TP/(FN + TP)
False positive Rate	What % 'NO' were predicted as 'Yes'?	FP/FP+TN)
Specificity	What % 'NO' were predicted as 'NO'?	TN/(TN + FP)
Precision(exactness)	What % of positive predictions Were correct?	TP(TP + FP)
FI score	Weighted average of precision And recall	2*((precision*recall)/ (precision + recall))

- Regression Model Evaluation:

A regression line predicts the y values for a given x value. Note that the values are around the average. The prediction error (called as root-mean-square error or RSME) is given by the following formula:

$$RMSE = \sqrt{\frac{\sum_{k=0}^{n}\left(\hat{Y}_k - y_k\right)^2}{n}}$$

And, the regression will also assessed by R square (Co efficient of determination).

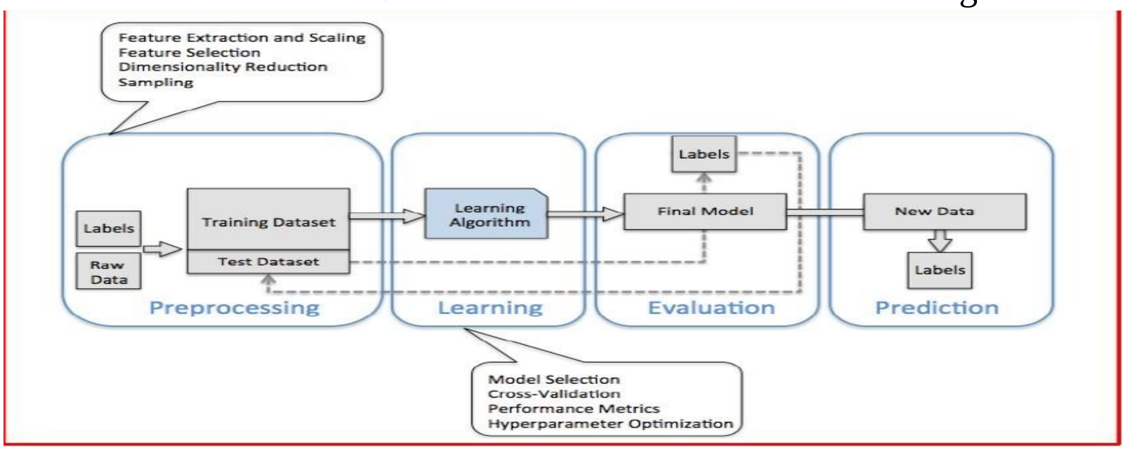
- Evaluating Unsupervised Models:

The Unsupervised algorithms will be assessed by the profile of the factors/ clusters which were derived through the models.

2.6 Overall Process of Machine Learning

To put overall process together, below is the picture that describes the road map for building ML Systems

Table 2.6 Overall Process of Machine Learning



CHAPTER-3

CLASSIFICATION ANALYSIS AT WORK

CLASSIFICATION ANALYSIS AT WORK

3.0 An Approach to the Problem:

In order to carry out the analysis, we have extracted 600 records from the https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29 and the information of the same is mentioned in Chapter 1.

In this Chapter, we are going to discuss about the results of different Machine Learning methods used in order to obtain the solution for the problem mentioned in Chapter 1.

As mentioned in Chapter 2, the first step of a ML Algorithm is Data cleaning and preparing data for the modeling. As a first step, we have to check whether the data was read properly and all the scale types are as per the data. Structure of data:

```
> str(german)
'data.frame':   600 obs. of  21 variables:
 $ x1 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ x2 : int  3 3 5 5 5 5 2 3 5 5 ...
 $ x3 : int  1 5 1 1 1 2 1 1 3 1 ...
 $ x4 : int  5 2 3 5 3 5 5 3 5 5 ...
 $ x5 : int  3 2 3 2 3 3 4 1 3 3 ...
 $ x6 : int  1 1 1 1 3 3 1 1 1 1 ...
 $ x7 : int  1 2 2 4 2 1 3 3 3 1 ...
 $ x8 : int  3 3 3 3 3 3 1 3 3 3 ...
 $ x9 : int  2 2 2 2 2 2 2 2 2 2 ...
 $ x10: int  2 3 3 3 3 3 2 3 3 3 ...
 $ x11: int  1 2 1 2 1 1 1 1 1 1 ...
 $ x12: int  1 2 1 1 1 2 1 1 1 1 ...
 $ x13: int  4 3 1 1 3 6 9 3 4 1 ...
 $ x14: int  10 24 12 6 36 9 12 12 6 21 ...
 $ x15: int  2315 7721 2121 860 5371 1288 339 2577 338 571 ...
 $ x16: int  3 1 4 1 3 3 4 2 4 4 ...
 $ x17: int  4 2 2 4 2 4 1 1 4 4 ...
 $ x18: int  52 30 30 39 28 48 45 42 52 65 ...
 $ x19: int  1 1 2 2 2 2 1 1 2 2 ...
 $ x20: int  1 1 1 1 1 2 1 1 1 1 ...
 $ dep: int  1 1 1 1 1 1 1 1 1 1 ...
```

Output-3.0 Structure of the data

From the above structure of the data we can observe the type of the data, number of observations and number of variables.

As we can see from the table, X1 to X13 are read as Integer variables but, they are actually nominal in nature. Hence, we converted these variables to factors as shown below.

```
'data.frame':   600 obs. of  21 variables:
 $ x1 : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 $ x2 : Factor w/ 5 levels "1","2","3","4",..: 3 3 5 5 5 5 2 3 5 5 ...
 $ x3 : Factor w/ 5 levels "1","2","3","4",..: 1 5 1 1 1 2 1 1 3 1 ...
 $ x4 : Factor w/ 5 levels "1","2","3","4",..: 5 2 3 5 3 5 5 3 5 5 ...
 $ x5 : Factor w/ 4 levels "1","2","3","4": 3 2 3 2 3 3 4 1 3 3 ...
 $ x6 : Factor w/ 3 levels "1","2","3": 1 1 1 1 3 3 1 1 1 1 ...
 $ x7 : Factor w/ 4 levels "1","2","3","4": 1 2 2 4 2 1 3 3 3 1 ...
 $ x8 : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 1 3 3 3 ...
 $ x9 : Factor w/ 3 levels "1","2","3": 2 2 2 2 2 2 2 2 2 2 ...
 $ x10: Factor w/ 4 levels "1","2","3","4": 2 3 3 3 3 3 2 3 3 3 ...
 $ x11: Factor w/ 2 levels "1","2": 1 2 1 2 1 1 1 1 1 1 ...
 $ x12: Factor w/ 2 levels "1","2": 1 2 1 1 1 2 1 1 1 1 ...
 $ x13: Factor w/ 10 levels "1","2","3","4",..: 4 3 1 1 3 6 8 3 4 1 ...
 $ x14: int  10 24 12 6 36 9 12 12 6 21 ...
 $ x15: int  2315 7721 2121 860 5371 1288 339 2577 338 571 ...
 $ x16: int  3 1 4 1 3 3 4 2 4 4 ...
 $ x17: int  4 2 2 4 2 4 1 1 4 4 ...
 $ x18: int  52 30 30 39 28 48 45 42 52 65 ...
 $ x19: int  1 1 2 2 2 2 1 1 2 2 ...
 $ x20: int  1 1 1 1 1 2 1 1 1 1 ...
 $ dep: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Output-3.0 Data frame

3.1 Understanding data using Descriptive Statistics:

The first step in understanding data is to look at the summary.

```
> summary(german)
       x1             x2              x3              x4              x5              x6       
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.00   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000  
 Median :2.00   Median :3.000   Median :1.000   Median :3.000   Median :3.000   Median :1.000  
 Mean   :2.38   Mean   :3.407   Mean   :1.947   Mean   :3.337   Mean   :2.645   Mean   :1.142  
 3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:1.000  
 Max.   :4.00   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.000   Max.   :3.000  
       x7              x8              x9             x10             x11             x12      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00  
 1st Qu.:1.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.00  
 Median :3.000   Median :3.000   Median :2.000   Median :3.000   Median :1.000   Median :1.00  
 Mean   :2.428   Mean   :2.633   Mean   :1.925   Mean   :2.907   Mean   :1.382   Mean   :1.03  
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:1.00  
 Max.   :4.000   Max.   :3.000   Max.   :3.000   Max.   :4.000   Max.   :2.000   Max.   :2.00  
      x13              x14             x15             x16             x17             x18       
 Min.   : 1.000   Min.   : 4.00   Min.   :  250   Min.   :1.000   Min.   :1.00   Min.   :19.00  
 1st Qu.: 2.000   1st Qu.:12.00   1st Qu.: 1370   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:26.00  
 Median : 3.000   Median :18.00   Median : 2347   Median :3.000   Median :3.00   Median :32.00  
 Mean   : 3.868   Mean   :21.75   Mean   : 3378   Mean   :3.008   Mean   :2.83   Mean   :34.97  
 3rd Qu.: 4.000   3rd Qu.:27.00   3rd Qu.: 4156   3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:41.00  
 Max.   :11.000   Max.   :72.00   Max.   :18424   Max.   :4.000   Max.   :4.00   Max.   :75.00  
      x19             x20             dep     
 Min.   :1.000   Min.   :1.000   Min.   :0.0  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0  
 Median :1.000   Median :1.000   Median :0.5  
 Mean   :1.388   Mean   :1.143   Mean   :0.5  
 3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.0  
 Max.   :4.000   Max.   :2.000   Max.   :1.0  
> |
```

Output-3.1 Understanding data using Descriptive Statistics

From the table above, we can look at the minimum, maximum, mean, median and quartiles for each of the Continuous variables and counts for the Nominal variables.

3.2 Checking for missing Values:

Then, check if there are any missing values in the data

```
 x1  x2  x3  x4  x5  x6  x7  x8  x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19 x20 dep 
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 
```

Output-3.2 Checking for missing Values

From above, we can observe that there are no missing values.

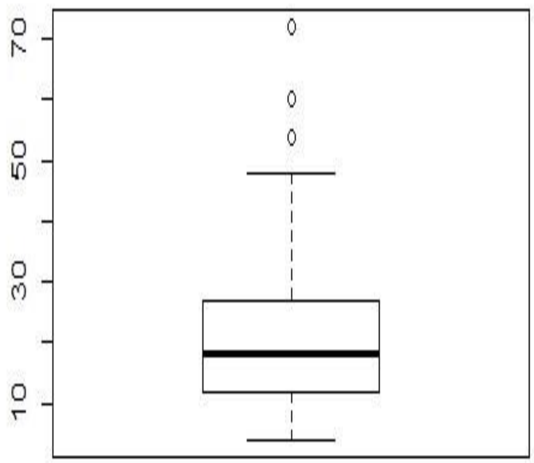
3.3 Checking and Removing of Outliers:

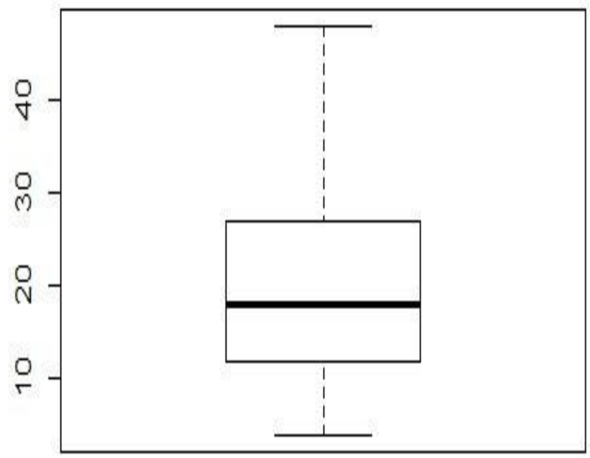
We used Box-plots to check for Outliers in each of the continuous variables.

X14=Duration in month:

Box plot of X14 says that there are outliers (i)

(ii)



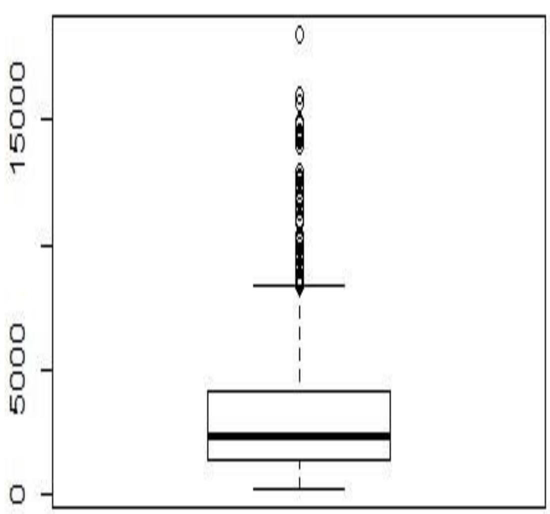


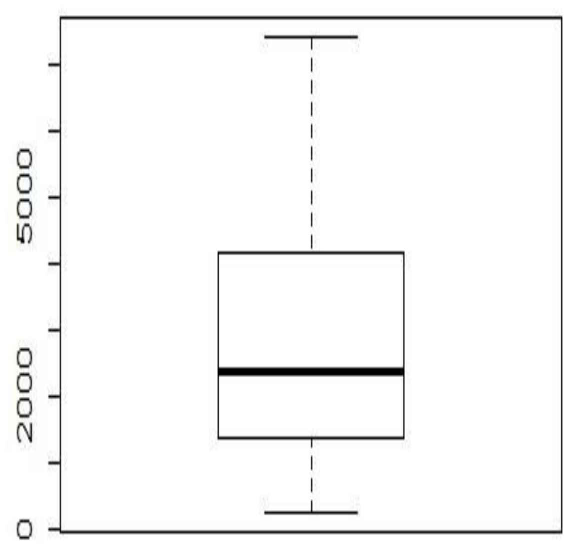
Output 3.3 with outliers output 3.3 without outliers

X15=Credit amount:

Box plot of X15 says that there are no outliers

(i) (ii)



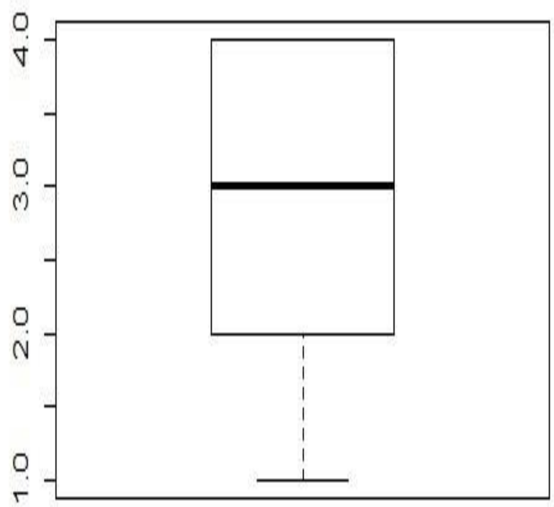


Output 3.3 with outliers output 3.3 without outliers

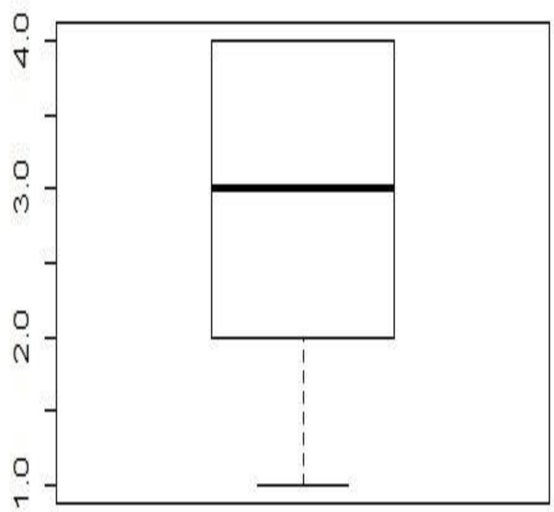
X16=Installment rate in the percentage of disposable income:

Box plot of X16 says that there are no outliers

(i) (ii)



Output 3.3 with outliers

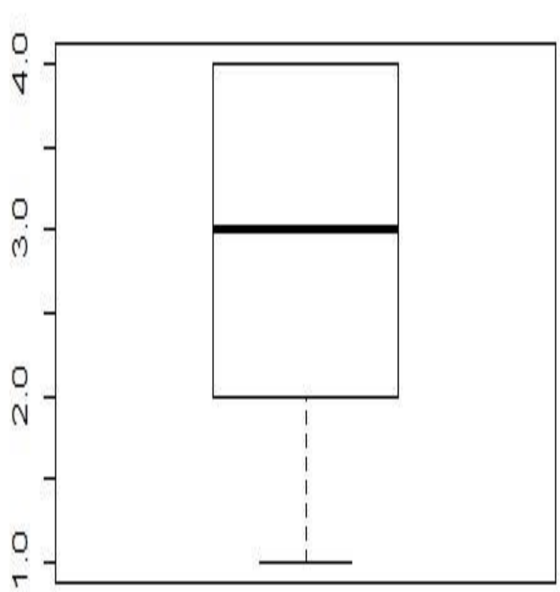


output 3.3 without outliers

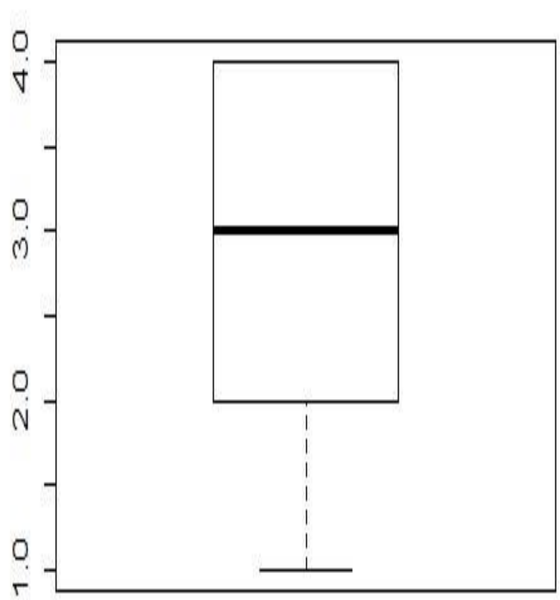
X17=Present residence since

Box plot of X17 says that there are no outliers

(i) (ii)



Output 3.3 with outliers

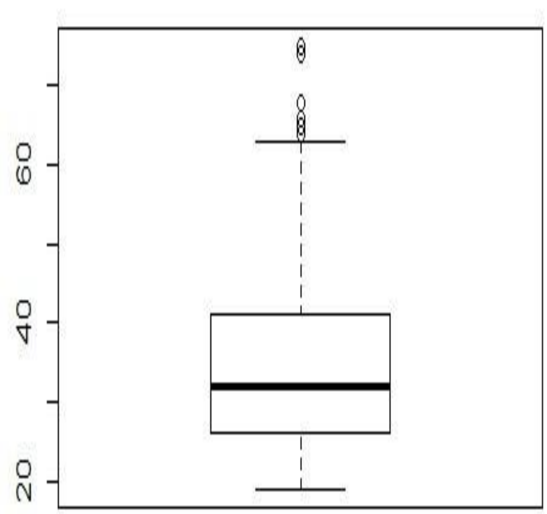


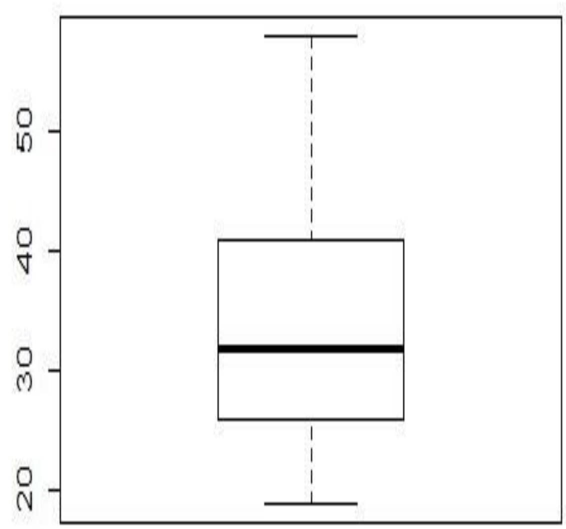
output 3.3 without outliers

X18=Age in year

Box plot of X18 says that their is outliers

(i) (ii)



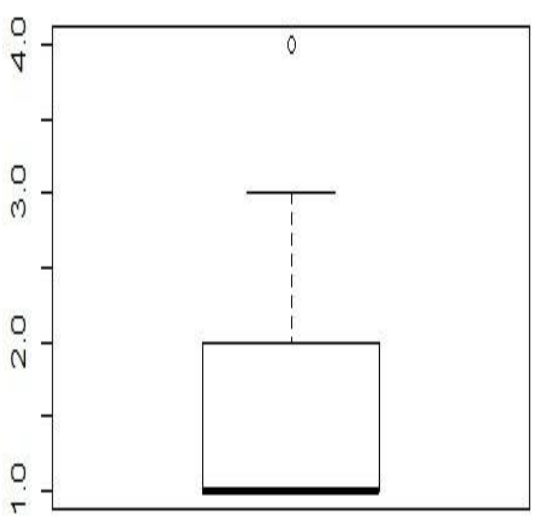


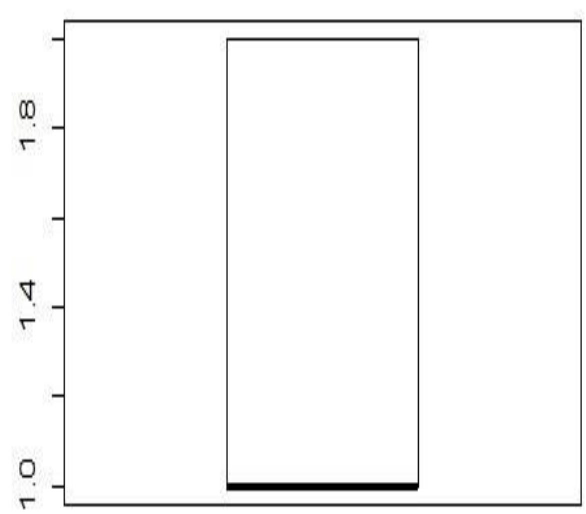
Output 3.3 with outliers output 3.3 without outliers

X19=Number of existing credits at this bank

Box plot of X19 says that there are no outliers (i)

(ii)

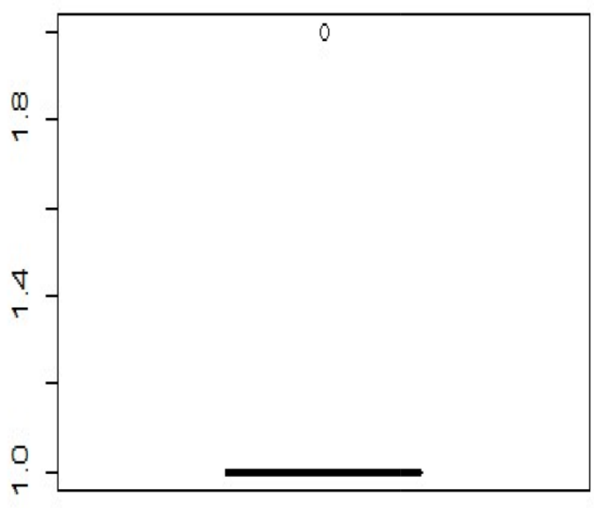


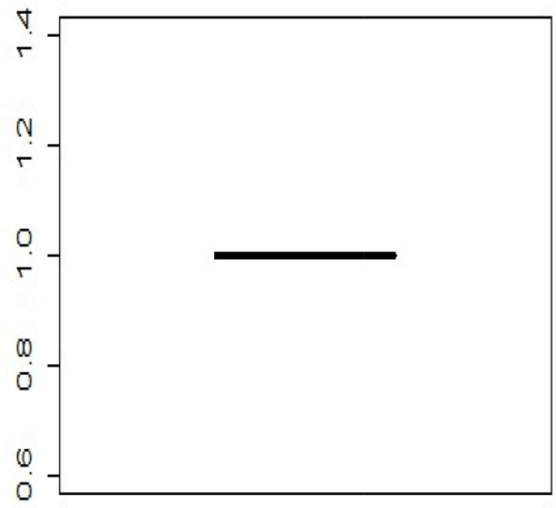


Output 3.3 with outliers output 3.3 without outliers

X20=Number of people being liable to provide maintenance for
Box plot of X20 says that their is outliers

(i) (ii)

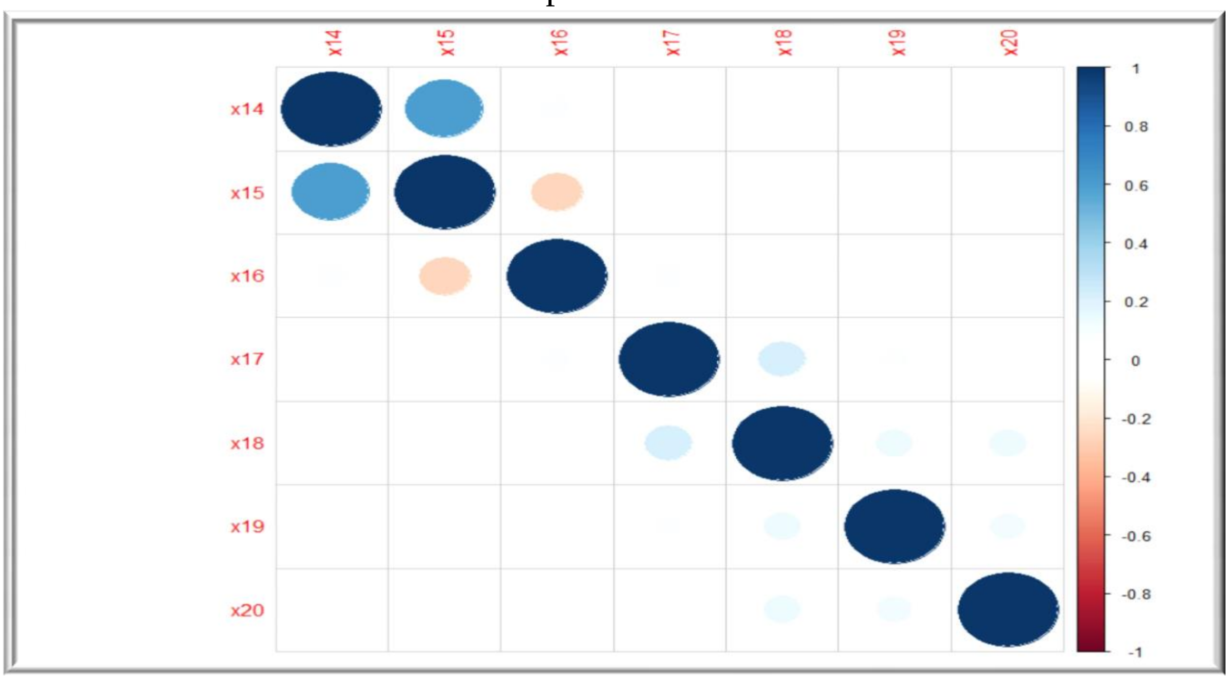




Output 3.3 with outlier   output 3.3 without outliers

3.4 Understanding relationships between variables:

For the continuous variables, we will look at the Correlation plots between variables to understand the relationships between variables.



Output-3.4 Understanding relationships between variables

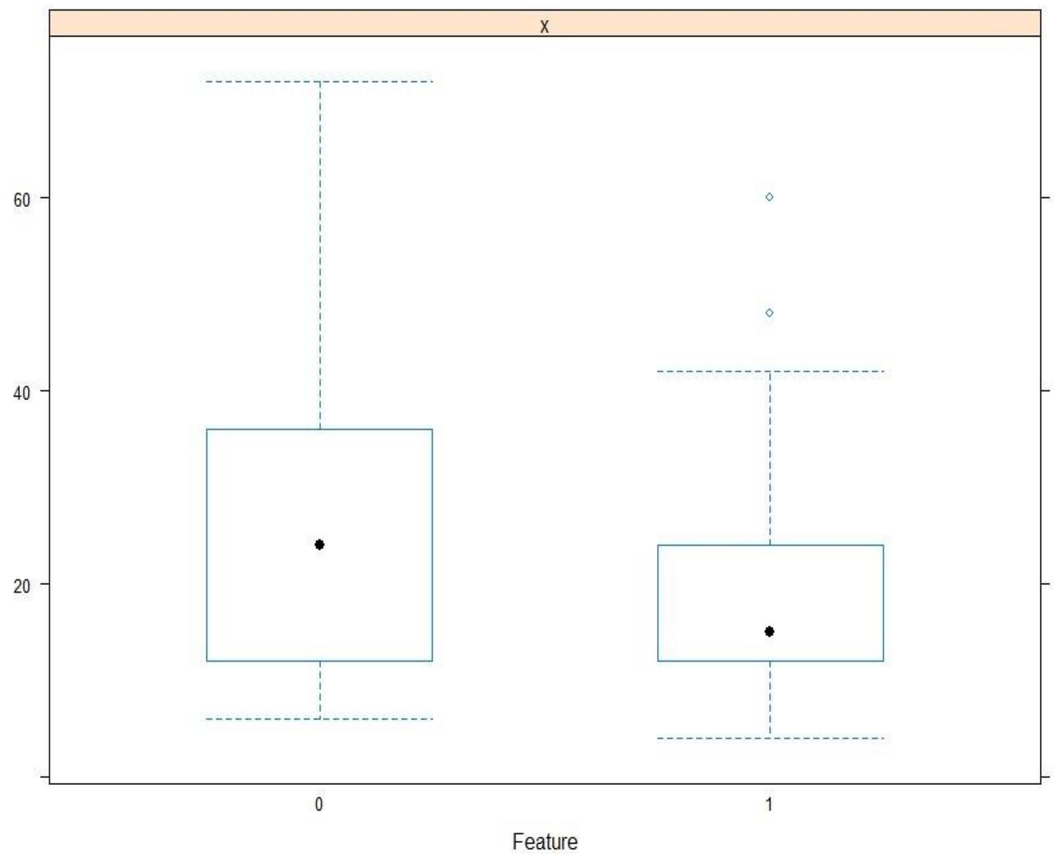
Here, the circle size refers to the strength of the relation and color refers to the direction of the relationship.

From the plot, we can see that X14(Duration in month) and X15(Credit amount) are highly positively correlated.

3.5 Feature plots:

For the continuous vs categorical variable, we will look at Feature plots to understand the relationships.

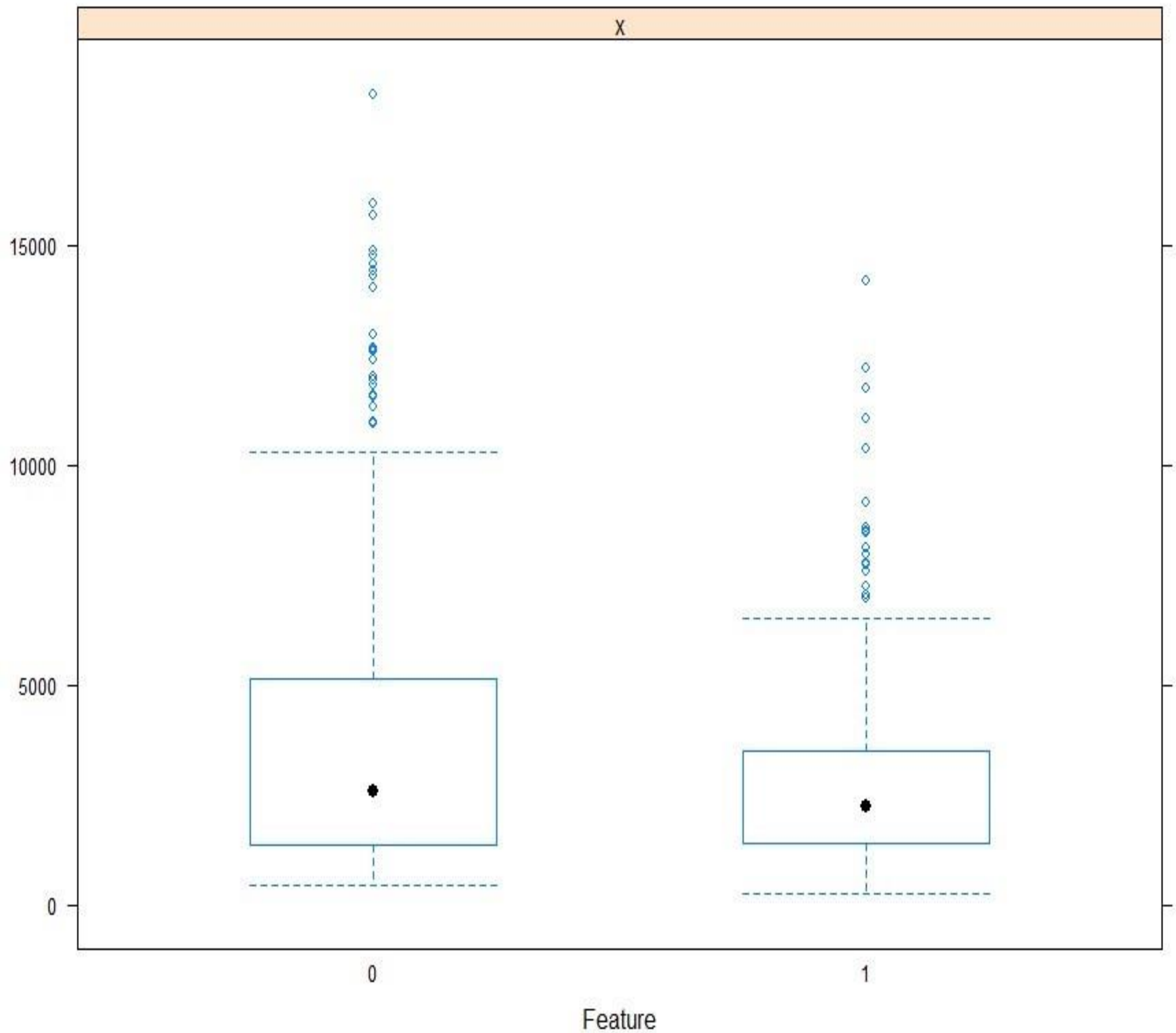
#X14=Duration in month



Output-3.5 Feature plots

From the plot, we can understand that there is small significant difference between Categorical (Dep) and Continuous (X14) variable for the German credit card data.

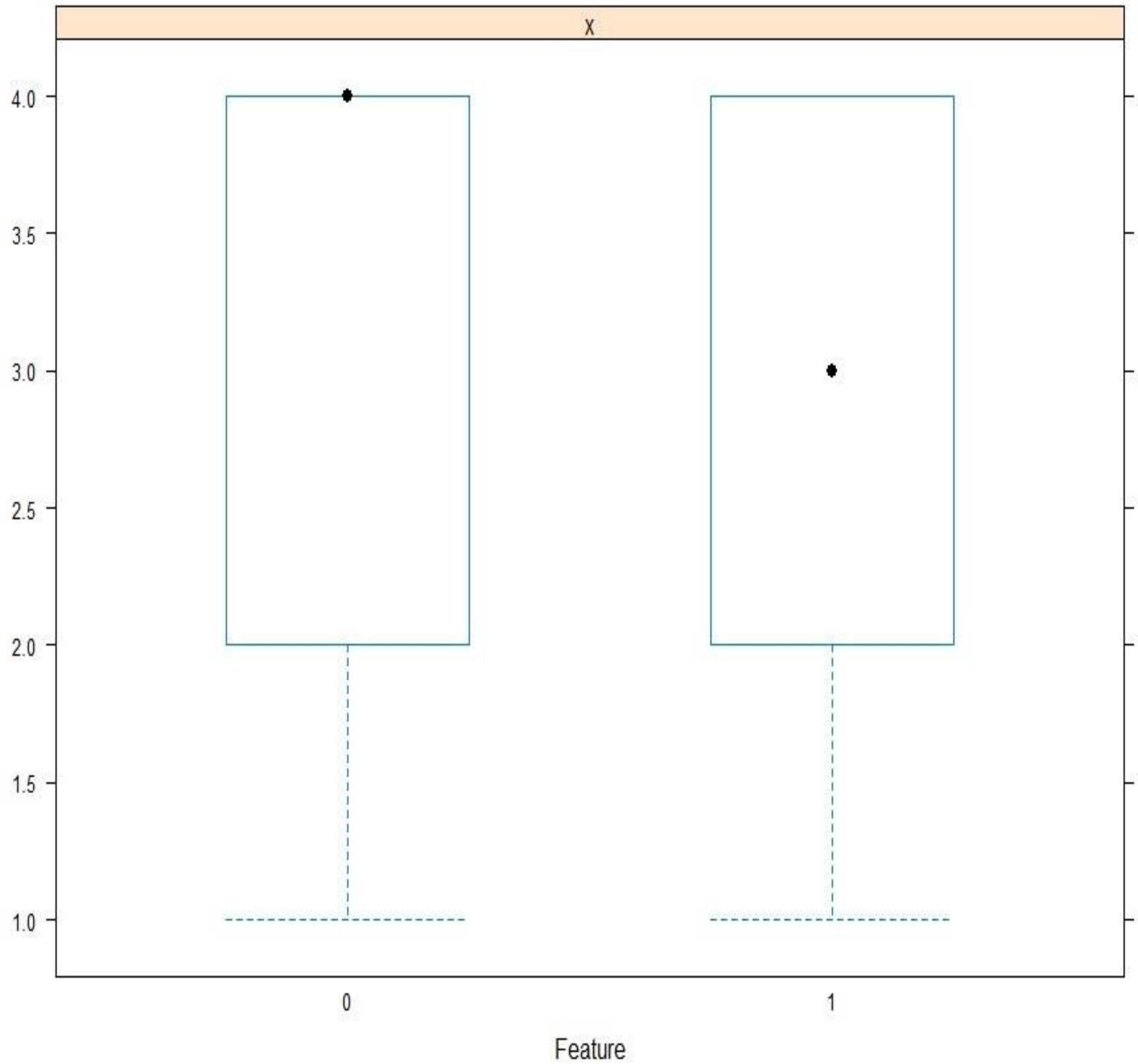
X15=Credit amount



Output-3.5 Feature plots

From the plot, we can understand that there is small significant difference between Categorical (Dep) and Continuous (X15) variable for the German credit card data.

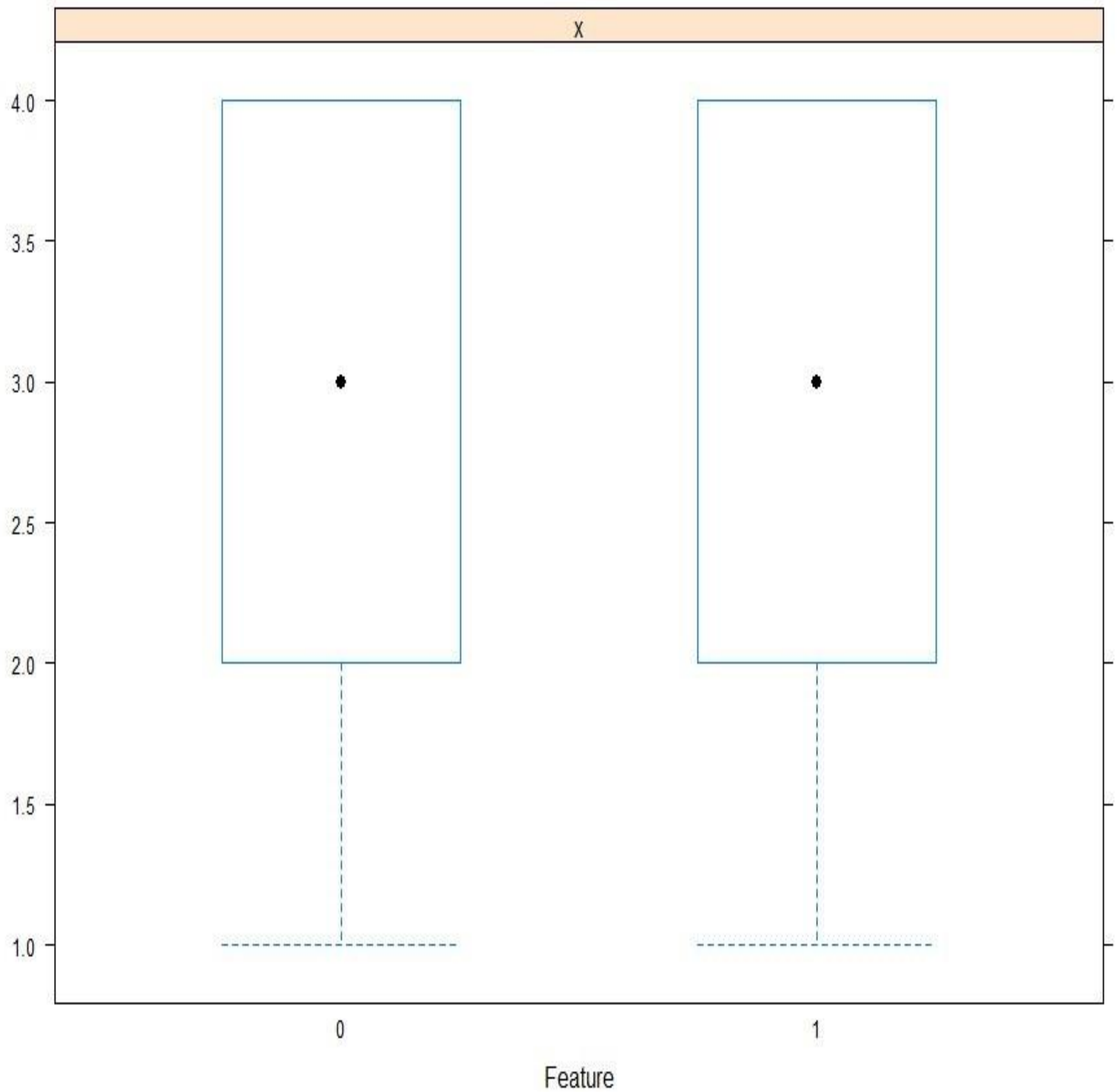
X16=Installment Rate in percentage of disposable income



Output-3.5 Feature plots

From the plot, we can understand that there is large significant difference between Categorical (Dep) and Continuous (X16) variable for the German credit card data.

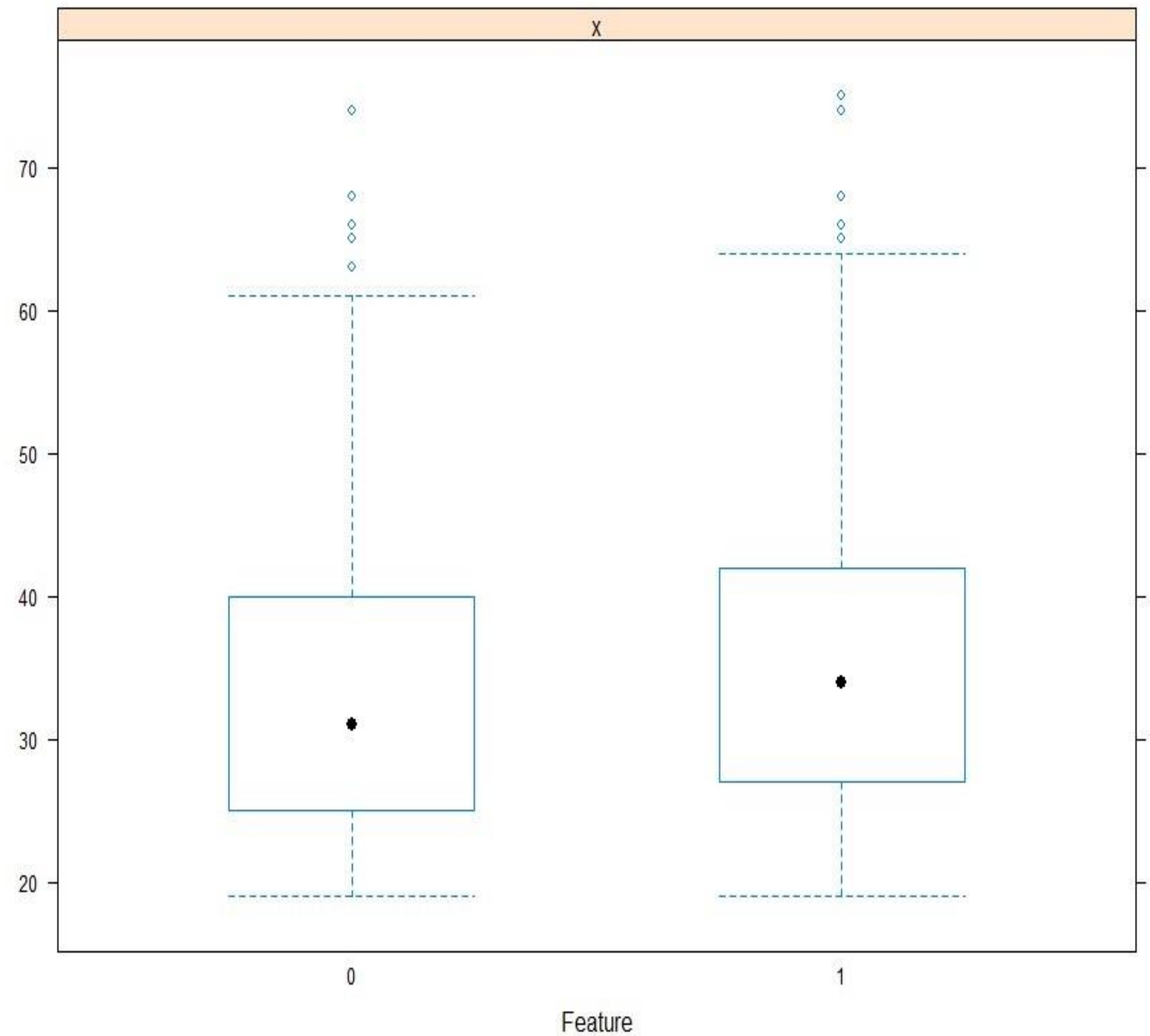
X17=Present Residence since



Output-3.5 Feature plots

From the plot, we can understand that there is small significant difference between Categorical (Dep) and Continuous (X17) variable for the German credit card data.

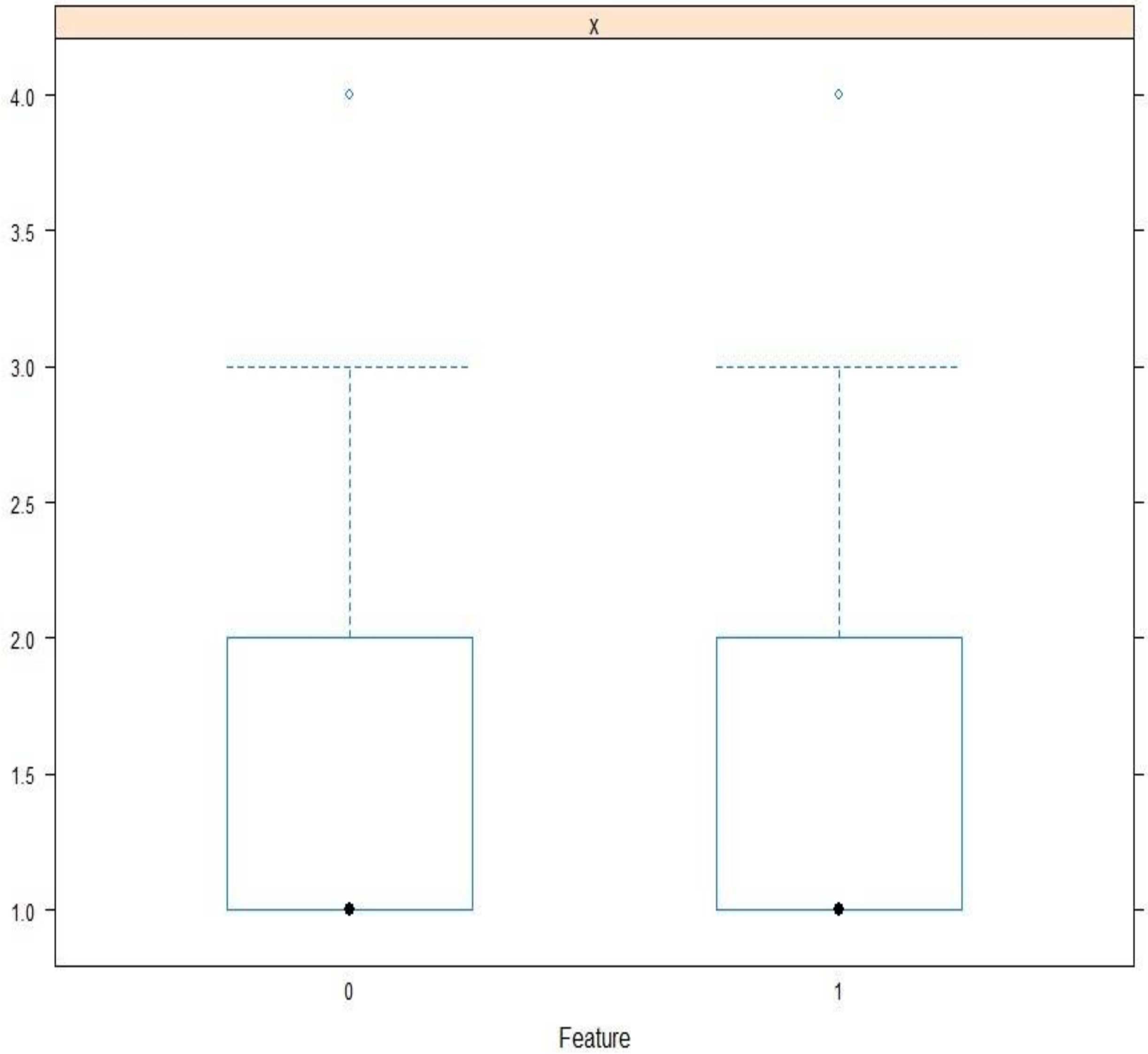
X18=Age in years



Output-3.5 Feature plots

From the plot, we can understand that there is small significant difference between Categorical (Dep) and Continuous (X18) variable for the German credit card data.

X19=Number of existing credits at this bank



Output-3.5 Feature plots

From the plot, we can understand that there is no significant difference between Categorical (Dep) and Continuous (X19) variable for the German credit card data.

3.6 Checking for the significance difference between variables

To test the significance difference between Continuous vs categorical variables, we will look at the t-test value.

X14 vs dep:

```
> a=aov(german$x14 ~ german$dep)
> summary(a)
             Df Sum Sq Mean Sq F value   Pr(>F)    
german$dep    1   5822    5822   39.63 5.94e-10 ***
Residuals   598  87850     147                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Output-3.6 Checking for the significance difference between variables

From the above table, as the p value is < 0.05, we can conclude that there is a significant relationship between X14 and dep.

X15 vs dep:

```
> a=aov(german$x15 ~ german$dep)
> summary(a)
             Df    Sum Sq   Mean Sq F value   Pr(>F)    
german$dep    1 1.885e+08 188469247   22.14 3.16e-06 ***
Residuals   598 5.092e+09   8514492                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Output-3.6 Checking for the significance difference between variables

From the above table, as the p value is < 0.05, we can conclude that there is a significant relationship between X15 and dep.

X16 vs dep:

```
> a=aov(german$x16 ~ german$dep)
> summary(a)
             Df Sum Sq Mean Sq F value Pr(>F)  
german$dep    1    4.7   4.682   3.762 0.0529 .
Residuals   598  744.3   1.245                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Output-3.6 Checking for the significance difference between variables

From the above table, as the p value is > 0.05, we can conclude that there is no significant relationship between X16 and dep.

X17 vs dep:

```
> a=aov(german$x17 ~ german$dep)
> summary(a)
             Df Sum Sq Mean Sq F value Pr(>F)
german$dep    1    0.2   0.240   0.196  0.658
Residuals   598  730.4   1.221
```

Output-3.6 Checking for the significance difference between variables From the above table, as the p value is > 0.05, we can conclude that there is no significant relationship between X17 and dep.

X18 vs dep:

```
> a=aov(german$x18 ~ german$dep)
> summary(a)
             Df Sum Sq Mean Sq F value Pr(>F)
german$dep    1    606   606.0   4.688 0.0308 *
Residuals   598  77296   129.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Output-3.6 Checking for the significance difference between variables

From the above table, as the p value is < 0.05, we can conclude that there is a significant relationship between X18 and dep.

X19 vs dep:

```
> a=aov(german$x19 ~ german$dep)
> summary(a)
             Df Sum Sq Mean Sq F value Pr(>F)
german$dep    1   0.28  0.2817   0.876   0.35
Residuals   598 192.24  0.3215
```

Output-3.6 Checking for the significance difference between variables From the above table, as the p value is > 0.05, we can conclude that there is no significant relationship between X19 and dep.

X20 vs dep:

```
> a=aov(german$x20 ~ german$dep)
> summary(a)
             Df Sum Sq Mean Sq F value Pr(>F)
german$dep    1   0.06  0.0600   0.487  0.485
Residuals   598  73.61  0.1231
```

Output-3.6 Checking for the significance difference between variables

From the above table, as the p value is > 0.05, we can conclude that there is no significant relationship between X20 and dep

In order to validate model, we spilt data into Train and Test with 70 and 30 percentage respectively and we have 416 for train and 214 for test data.

3.7 Split Data into Train & Test:

```
> ind <- sample(2, nrow(german), replace = T, prob = c(0.7, 0.3))
> train <- german[ind==1,]
> test <- german[ind==2,]
```

Output-3.7 Split Data into Train & Test

```
> dim(train)
[1] 416  21
>
> table(train$dep)

  0   1
211 205
> dim(test)
[1] 184  21
> table(test$dep)

 0  1
89 95
```

Output-3.7 Split Data into Train & Test

3.8 Logistic model

The below is the output obtained from logistic model Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.3001	-0.7362	-0.1276	0.7649	2.0890

Coefficients:

Estimate Std. Error z value Pr(>|z|) (Intercept) -1.455e+00 1.611e+00 -0.903 0.36641 x12 1.840e-01

3.497e-01   0.526 0.59888   x13   6.926e-01 5.124e-01
1.352 0.17648   x14   1.741e+00 3.522e-01   4.943
7.71e-07 *** x22   7.363e-01 9.308e-01   0.791 0.42893
x23   1.391e+00 7.686e-01   1.810 0.07032 .   x24
1.497e+00 8.182e-01   1.830 0.06731 .   x25   1.524e+00
7.869e-01   1.937 0.05271 .   x32   7.013e-01 4.534e-01
1.547 0.12193   x33   3.915e-01 5.748e-01   0.681
0.49578   x34   1.928e+00 9.002e-01   2.141 0.03224 *
x35   1.164e+00 4.239e-01   2.745 0.00604 ** x42
6.829e-01 7.725e-01   0.884 0.37670   x43   1.028e+00
7.533e-01   1.365 0.17231   x44   1.572e+00 7.917e-01
1.985 0.04710 *   x45   1.156e+00 7.499e-01   1.542
0.12314   x52   8.519e-02 5.867e-01   0.145 0.88456
x53   6.323e-01 5.698e-01   1.110 0.26708   x54   -
5.579e-01 7.451e-01 -0.749 0.45402   x62   -1.209e-01
5.714e-01 -0.212 0.83244   x63   1.156e+00 6.540e-01
1.767 0.07721 .   x72   -6.059e-01 3.873e-01 -1.564
0.11773   x73   -3.685e-01 3.500e-01 -1.053 0.29236
x74   -3.765e-01 6.973e-01 -0.540 0.58925   x82   -
1.884e-01 6.794e-01 -0.277 0.78156   x83   6.355e-01
3.796e-01   1.674 0.09412 .   x92   5.443e-01 3.533e-01
1.541 0.12340   x93   1.317e-01 7.913e-01   0.166
0.86780   x102   -1.412e+00 1.022e+00 -1.381 0.16715
x103   -1.363e+00 9.837e-01 -1.385 0.16591   x104
-1.877e+00 1.011e+00 -1.857 0.06324 .   x112   6.901e-
01 3.221e-01   2.143 0.03214 *   x122   1.153e+00
8.838e-01   1.305 0.19200   x132   2.632e+00 6.171e-01
4.265 2.00e-05 *** x133   8.572e-01 3.949e-01   2.171
0.02997 *   x134   1.149e+00 3.951e-01   2.907 0.00365
** x135   1.258e+00 1.130e+00   1.113 0.26583   x136
1.124e+00 9.229e-01   1.218 0.22328   x137   -3.773e-01
6.837e-01   -0.552   0.58106   x139   3.011e+00
1.437e+00   2.095 0.03620 *   x1310   3.559e-01 5.317e-
01   0.669 0.50333   x1311   8.656e-01 1.382e+00
0.626 0.53110   x14   -3.054e-02 1.449e-02 -2.107
0.03507 *   x15   -2.100e-04 7.598e-05 -2.763 0.00572
** x16   -3.572e-01 1.349e-01 -2.648 0.00810 ** x17
-1.388e-01 1.399e-01 -0.992 0.32096   x18   2.643e-02
1.485e-02   1.780 0.07500 .   x19   -6.714e-02 2.863e-01

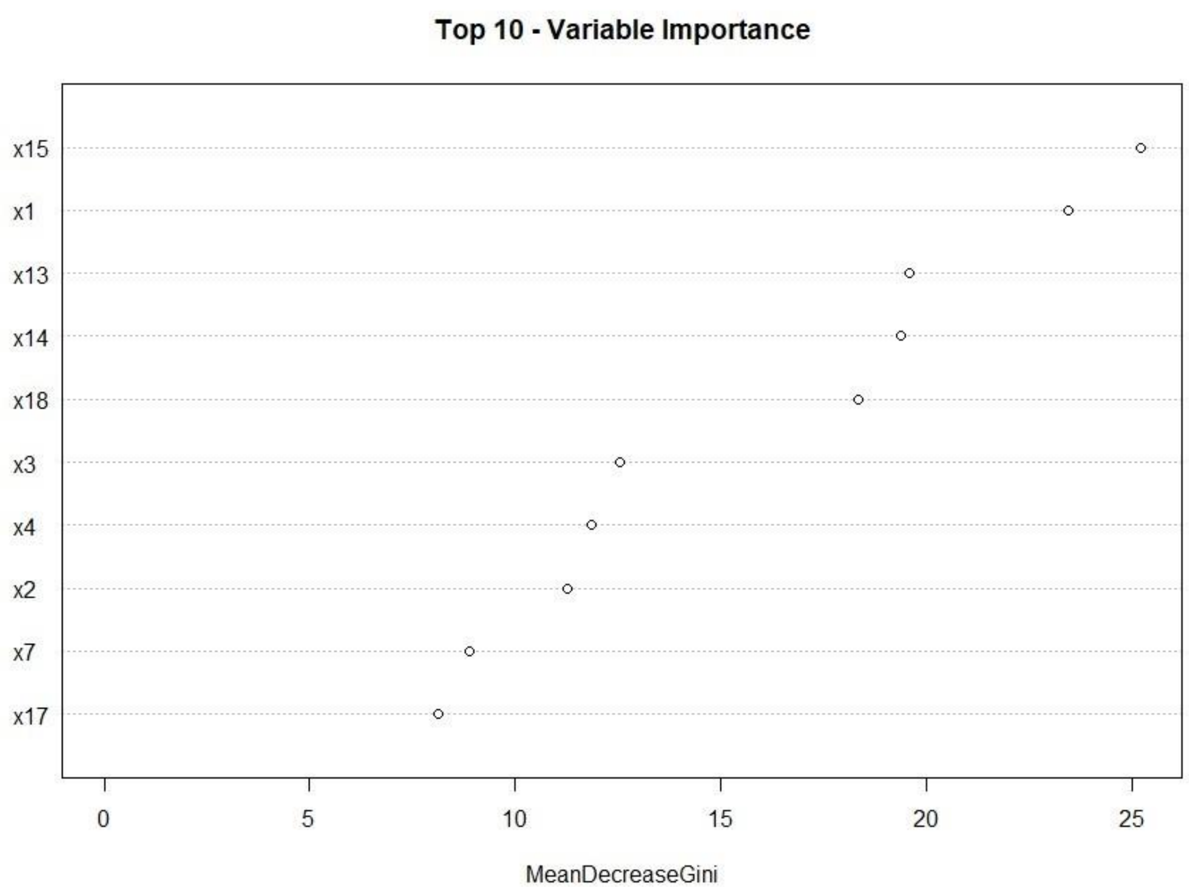
-0.234 0.81460   x20   -4.254e-01 3.898e-01 -1.091
0.27517

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Output-3.8 Logistic model

From the above the table X14=duration in month is more significant

3.9 Finding the Key variables using Random Forest:



Output-3.9 Finding the Key variables using Random Forest

In order to find the key variables we run the Random Forest using grid search and obtained the key hyper parameters. From above graph, we can observe that X15, X1, X13, X14 are the variables which have more importance than other variables.

3.9 (a)Generalised Linear Model using Key variables

```
> summary(ite1)

Call:
glm(formula = dep ~ x1 + x2 + x3 + x4 + x5 + x7 + x13 + x14 +
    x15 + x18, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -0.8780  -0.2174   0.8740   2.2674

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.671e+00  1.228e+00  -2.174  0.02968 *
x12          2.771e-01  3.229e-01   0.858  0.39082
x13          7.719e-01  4.735e-01   1.630  0.10304
x14          1.604e+00  3.302e-01   4.858 1.19e-06 ***
x22          7.029e-01  8.646e-01   0.813  0.41624
x23          1.366e+00  7.250e-01   1.884  0.05957 .
x24          1.352e+00  7.900e-01   1.711  0.08705 .
x25          1.457e+00  7.559e-01   1.928  0.05384 .
x32          5.642e-01  4.056e-01   1.391  0.16425
x33          3.459e-01  5.510e-01   0.628  0.53010
x34          1.547e+00  8.440e-01   1.833  0.06687 .
x35          1.081e+00  3.944e-01   2.740  0.00614 **
x42          2.075e-01  6.321e-01   0.328  0.74265
x43          5.823e-01  5.930e-01   0.982  0.32614
x44          1.017e+00  6.327e-01   1.607  0.10807
x45          5.755e-01  5.930e-01   0.970  0.33181
x52         -2.428e-03  5.588e-01  -0.004  0.99653
x53          3.851e-01  5.363e-01   0.718  0.47273
x54         -5.866e-01  7.028e-01  -0.835  0.40391
x72         -5.094e-01  3.547e-01  -1.436  0.15104
x73         -4.141e-01  3.263e-01  -1.269  0.20440
x74         -6.681e-01  4.591e-01  -1.455  0.14563
```

```
x132        2.304e+00  5.656e-01   4.074 4.61e-05 ***
x133        6.925e-01  3.675e-01   1.884  0.05956 .
x134        9.990e-01  3.617e-01   2.762  0.00575 **
x135        1.214e+00  1.185e+00   1.025  0.30548
x136        1.078e+00  8.789e-01   1.227  0.22001
x137       -3.725e-01  6.450e-01  -0.578  0.56353
x139        2.046e+00  1.384e+00   1.479  0.13924
x1310       3.283e-01  5.050e-01   0.650  0.51555
x1311       4.447e-01  1.275e+00   0.349  0.72722
x14        -4.231e-02  1.313e-02  -3.222  0.00127 **
x15        -1.098e-04  6.254e-05  -1.756  0.07906 .
x18         2.118e-02  1.332e-02   1.590  0.11176
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 576.61  on 415  degrees of freedom
Residual deviance: 428.01  on 382  degrees of freedom
AIC: 496.01

Number of Fisher Scoring iterations: 5
```

Output-3.9 (a) Generalized Linear Model using Key variables

Test Accuracy:

To validate the model we have used the confusion matrix and got the below accuracy

Accuracy For Train Data:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 164  53
         1  47 152

               Accuracy : 0.7596
                 95% CI : (0.7156, 0.7999)
    No Information Rate : 0.5072
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5189
```

Output-The accuracy for train data is 75.96%

Accuracy For Test Data:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 66 30
         1 23 65

               Accuracy : 0.712
                 95% CI : (0.6407, 0.7762)
    No Information Rate : 0.5163
    P-Value [Acc > NIR] : 5.003e-08

                  Kappa : 0.4247
```

Output-The accuracy for test data is 71.2%

CHAPTER-4

SUMMARY

4.0 SUMMARY:

In order to solve the above problem we have applied logistic regression technique and also we apply random forest technique to find the key variable for predicting German credit card. Hence we have applied logistic regression for this key variable and obtained the train and test data and computed the accuracy of confusion matrix for train and test as below.

ACCURACY FOR TRAIN DATA IS :- 75.96 ~76%

ACCURACY FOR TEST DATA IS :- 71.2 ~72%

Since, the accuracy of train and test data are more or less similar accurate. Hence, the model is a Generalised model, so we can use this model to predict the future data.

CHAPTER-5

APPENDIX

APPENDIX

R-CODE:

```
getwd()
setwd("E:/")

#Reading the data

german=read.csv("german credit card.csv")

#Viewing the head, tail of the data

head(german)
tail(german)

#Viewing the data

View(german)

#To get descriptive statistics

summary(german)

#Checking for missing values

is.na(german$x1)
is.na(german$lc)

#Checking the percentage of missing values in each column (variable)

sum(is.na(german))
sapply(german, function(df) {
```

```
(sum(is.na(df)==TRUE)/ length(df))*100; })
```

```
#Checking for outliers

boxplot(german$x1)
boxplot(german$x2)
boxplot(german$x3)
boxplot(german$x4)
boxplot(german$x5)
boxplot(german$x6)
boxplot(german$x7)
boxplot(german$x8)
boxplot(german$x9)
boxplot(german$x10)
boxplot(german$x11)
boxplot(german$x12)
boxplot(german$x13)

#Removing the outliers

german$x1[german$x1>quantile(german$x1,0.95)] <-quantile(german$x1,0.95)
german$x2[german$x2<quantile(german$x2,0.05)] <-quantile(german$x2,0.05)
german$x3[german$x3>quantile(german$x3,0.80)] <-quantile(german$x3,0.80)
german$x4[german$x4>quantile(german$x4,0.80)] <-quantile(german$x4,0.80)
german$x5[german$x5>quantile(german$x5,0.95)] <-quantile(german$x5,0.95)
german$x6[german$x6>quantile(german$x6,0.90)] <-quantile(german$x6,0.90)
german$x7[german$x7>quantile(german$x7,0.95)] <-quantile(german$x7,0.95)
german$x8[german$x8>quantile(german$x8,0.05)] <-quantile(german$x8,0.05)
german$x9[german$x9>quantile(german$x9,0.10)] <-quantile(german$x9,0.10)
german$x10[german$x10>quantile(german$x10,0.01)]<-quantile(german$x10,0.01)
german$x11[german$x11>quantile(german$x11,0.95)]<-quantile(german$x11,0.95)
german$x12[german$x12>quantile(german$x12,0.95)]<-quantile(german$x12,0.95)
german$x13[german$x13>quantile(german$x13,0.80)]<-quantile(german$x13,0.80)
```

#Converting categorical variable to factor

```
german$x1=as.factor(german$x1)
german$x2=as.factor(german$x2)
german$x3=as.factor(german$x3)
german$x4=as.factor(german$x4)
german$x5=as.factor(german$x5)
german$x6=as.factor(german$x6)
german$x7=as.factor(german$x7)
german$x8=as.factor(german$x8)
german$x9=as.factor(german$x9)
german$x10=as.factor(german$x10)
german$x11=as.factor(german$x11)
german$x12=as.factor(german$x12)
german$x13=as.factor(german$x13)
german$dep=as.factor(german$dep) str(german)

mymodel <- glm(dep ~ ., data = train, family = 'binomial') summary(mymodel)
```

#Changing to numeric all the imputed variables

```
pre1=german[c(14:20)]
library(corrplot)
str(pre1) pre1.cor  =
cor (pre1)
corrplot(pre1.cor, method="circle")
```

#Continuous vs categories

```
library(caret) x
<- german[,14]
str(german) y <-
german[,21]
featurePlot(x=x, y=y, plot="box")
```

```
library(caret) x
<- german[,15]
str(german)
y <- german[,21]
featurePlot(x=x, y=y, plot="box")

library(caret) x <- german[,16]
str(german) y <- german[,"dep"]
featurePlot(x=x, y=y, plot="box")

library(caret) x <- german[,17]
str(german) y <- german[,"dep"]
featurePlot(x=x, y=y, plot="box")

library(caret) x <- german[,18]
str(german) y <- german[,"dep"]
featurePlot(x=x, y=y, plot="box")

library(caret) x <- german[,19]
str(german) y <- german[,"dep"]
featurePlot(x=x, y=y, plot="box")

library(caret) x <- german[,20]
str(german) y <- german[,"dep"]
featurePlot(x=x, y=y, plot="box")

str(german)

#Testing the significance diffrence using anova

a=aov(german$x14 ~ german$dep)
summary(a) a=aov(german$x15 ~
german$dep) summary(a)
a=aov(german$x16 ~ german$dep)
summary(a) a=aov(german$x17 ~
german$dep) summary(a)
a=aov(german$x18 ~ german$dep)
summary(a) a=aov(german$x19 ~
german$dep) summary(a)
```

a=aov(german$x20 ~ german$dep)
summary(a)

Preparing data for training and testing - train (70%) & test (30%)

set.seed(1234) ind <- sample(2, nrow(german), replace = T,
prob = c(0.7, 0.3)) train <- german[ind==1,] test <-
german[ind==2,]
View(train)
dim(train)

table(train$dep)

dim(test)
table(test$dep)

#Evaluate algorithms:baseline #Fitting Logistic Regression

mymodel <- glm(dep ~ ., data = train, family = 'binomial')
summary(mymodel)

Prediction

p1 <- predict(mymodel, train, type = 'response')
pred1 <- ifelse(p1>0.5, 1, 0) pred1<-
as.factor(pred1) table(train$dep) str(pred1)
library(caret) confusionMatrix(
pred1,train$dep)

#Testing

p2 <- predict(mymodel, test, type = 'response')
predtest <- ifelse(p2>0.5, 1, 0) predtest<-
as.factor(predtest) library(caret)
confusionMatrix(predtest,test$dep)

Identifying key variables

Final model

```
library(randomForest) rf_model <-
randomForest(dep ~ ., data =train)
print(rf_model)


varImpPlot(rf_model,
sort = T,
        n.var = 10,          main = "Top 10 - Variable
Importance")
```

#Final model with Shortlisted variables

```
ite1 <- glm(dep~x1+x2+x3+x4+x5+x7+x13+x14+x15+x18, data =  train , family =
'binomial') summary(ite1) p2 <-
predict(ite1, train, type = 'response')
pred2 <- ifelse(p2>0.5, 1, 0) str(pred2)
pred2<-as.factor(pred2)
library(caret)      confusionMatrix(
pred2,train$dep)
```

Testing on test data with final Model

```
pred=predict(ite1,newdata=test,type = 'response')
p3 <- ifelse(pred>0.5, 1, 0) p3<-as.factor(p3)
library(caret)
confusionMatrix(p3,test$dep)
```

DATASET

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	dep
1	3	1	5	3	1	1	3	2	2	1	1	4	10	2315	3	4	52	1	1	1
1	3	5	2	2	1	2	3	2	3	2	2	3	24	7721	1	2	30	1	1	1
1	5	1	3	3	1	2	3	2	3	1	1	1	12	2121	4	2	30	2	1	1
1	5	1	5	2	1	4	3	2	3	2	1	1	6	860	1	4	39	2	1	1
1	5	1	3	3	3	2	3	2	3	1	1	3	36	5371	3	2	28	2	1	1
1	5	2	5	3	3	1	3	2	3	1	2	6	9	1288	3	4	48	2	2	1
1	2	1	5	4	1	3	1	2	2	1	1	9	12	339	4	1	45	1	1	1
1	3	1	3	1	1	3	3	2	3	1	1	3	12	2577	2	1	42	1	1	1
1	5	3	5	3	1	3	3	2	3	1	1	4	6	338	4	4	52	2	1	1
1	5	1	5	3	1	1	3	2	3	1	1	1	21	571	4	4	65	2	1	1
1	3	1	3	4	1	1	3	2	2	1	1	1	12	1168	4	3	27	1	1	1
1	1	1	4	3	1	2	1	2	3	2	1	10	18	3104	3	1	31	1	1	1
1	3	1	3	1	1	1	3	1	2	1	1	3	24	3021	2	2	24	1	1	1
1	3	5	3	2	1	2	1	1	3	2	1	7	12	1200	4	4	23	1	1	1
1	3	1	5	3	3	2	3	2	3	1	1	4	30	2522	1	3	39	1	2	1
1	5	4	4	2	1	1	3	2	2	2	1	1	6	666	3	4	39	2	1	1
1	3	1	3	3	1	1	3	2	3	1	1	3	12	1657	2	2	27	1	1	1
1	3	1	5	2	1	3	3	2	3	1	1	4	24	1603	4	4	55	1	1	1
1	5	1	5	2	1	4	3	3	4	2	1	2	24	6419	2	4	44	2	2	1
1	5	1	4	3	2	2	3	2	3	2	1	1	10	1038	4	3	49	2	1	1
1	3	1	3	3	3	2	3	2	2	1	1	3	12	708	2	3	38	1	2	1
1	4	1	5	2	1	2	3	2	2	1	1	3	15	3643	1	4	27	2	1	1
1	3	1	4	3	1	4	3	3	4	2	1	2	24	2910	2	1	34	1	1	1
1	3	4	3	3	1	3	3	2	3	1	1	3	18	2659	4	2	28	1	1	1
1	5	1	4	3	1	1	3	2	2	1	2	1	8	3398	1	4	39	2	1	1
1	3	1	2	2	1	3	3	1	3	1	1	3	18	3650	1	4	22	1	1	1
1	3	5	4	3	1	3	3	2	3	2	1	3	20	2212	4	4	39	1	1	1
1	3	1	4	3	1	4	1	3	4	2	1	1	36	3249	2	4	39	1	2	1
1	3	1	1	2	1	4	3	3	4	1	1	3	12	2578	3	4	55	1	1	1
1	5	1	3	4	1	2	3	2	3	2	1	3	36	2348	3	2	46	2	1	1
1	4	5	3	2	1	3	3	1	3	2	1	7	18	8471	1	2	23	2	1	1
1	3	1	3	3	1	1	3	2	3	1	1	3	9	2136	3	2	25	1	1	1

1	5	1	1	3	2	3	3	2	1	1	1	6	42	3394	4	4	65	2	1	1
1	3	1	3	2	2	2	3	2	3	1	1	3	12	1620	2	3	30	1	1	1
1	5	1	3	3	1	1	3	1	3	1	1	1	11	3905	2	2	36	2	2	1
1	3	1	2	2	1	1	3	2	3	1	1	5	6	343	4	1	27	1	1	1
1	5	1	5	3	1	3	3	2	3	2	1	3	15	1478	4	4	44	2	2	1
1	3	4	5	2	1	2	3	1	3	2	1	3	30	3622	4	4	57	2	1	1
1	5	1	3	3	1	1	3	1	2	1	2	1	12	2122	3	2	39	2	2	1
1	5	1	4	2	1	2	3	2	3	1	2	1	6	609	4	3	37	2	1	1
1	3	2	5	3	1	2	3	2	3	2	1	1	6	1203	3	2	43	1	1	1
1	3	1	2	3	1	1	3	2	2	2	1	1	6	662	3	4	41	1	2	1
1	3	5	1	2	1	2	3	2	4	2	1	1	6	1374	4	3	75	1	1	1
1	3	5	5	3	1	4	1	3	3	2	1	2	24	2964	4	4	49	1	2	1
1	3	1	3	3	3	2	3	2	2	1	1	3	12	1289	4	1	21	1	1	1
1	2	3	3	3	1	1	2	2	3	2	1	3	24	2828	4	4	22	1	1	1
1	2	1	3	2	3	3	1	1	3	1	2	2	24	3632	1	4	22	1	1	1
1	5	1	3	3	1	1	3	1	3	1	1	1	6	3676	1	3	37	3	2	1
1	3	1	2	2	1	3	3	1	3	1	1	3	12	1858	4	1	22	1	1	1
1	3	1	3	2	1	3	3	2	3	1	1	4	24	3660	2	4	28	1	1	1
1	5	1	4	4	1	2	3	2	4	2	1	4	18	1880	4	1	32	2	1	1
1	1	1	3	1	3	1	3	2	3	1	1	3	30	4583	2	2	32	2	1	1
1	3	4	3	3	1	2	3	2	3	1	1	1	12	3651	1	3	31	1	2	1
1	5	1	3	3	1	1	3	2	2	1	1	4	9	1138	4	4	25	2	1	1
1	3	2	3	3	1	3	3	2	2	2	1	1	18	4380	3	4	35	1	2	1
1	2	2	4	3	1	3	1	2	3	1	1	1	24	2325	2	3	32	1	1	1
1	2	3	3	3	1	1	2	2	3	2	1	3	24	2483	4	4	22	1	1	1
1	5	5	4	3	1	2	3	2	4	2	1	3	39	14179	4	4	30	2	1	1
1	5	1	2	3	1	2	1	2	2	1	1	10	13	1797	3	1	28	2	1	1
1	4	1	3	3	1	4	3	3	3	1	1	1	24	4870	3	4	53	2	2	0
1	3	1	2	2	1	2	3	1	3	1	1	10	48	4308	3	4	24	1	1	0
1	5	1	5	3	1	3	3	2	2	1	1	1	24	1199	4	4	60	2	1	0
1	3	2	3	2	1	3	3	2	2	1	1	4	24	1282	4	2	32	1	1	0
1	4	1	5	3	1	4	3	2	3	2	1	10	60	6836	3	4	63	2	1	0
1	5	1	5	2	1	4	2	3	2	1	1	2	48	6143	4	4	58	2	1	0
1	5	1	2	2	2	4	3	1	2	2	1	3	36	6229	4	4	23	2	1	0

1	3	5	5	3	1	4	3	2	4	2	1	7	36	1977	4	4	40	1	1	0
1	3	1	2	3	1	3	3	2	3	1	1	4	42	3965	4	3	34	1	1	0
1	1	1	4	3	1	1	3	2	3	1	1	6	12	1108	4	3	28	2	1	0
1	3	5	4	2	1	3	3	2	4	2	1	4	42	7174	4	3	30	1	1	0
1	5	3	3	2	1	3	3	2	3	1	1	3	33	4281	1	4	23	2	1	0
1	3	1	3	2	1	1	3	2	3	2	1	4	21	1835	3	2	25	2	1	0
1	5	1	3	2	2	1	3	2	3	1	1	1	12	3499	3	2	29	2	1	0
1	4	1	3	3	1	2	2	2	3	2	1	7	36	6887	4	3	29	1	1	0
1	3	1	3	3	1	3	3	2	3	1	1	3	18	2462	2	2	22	1	1	0
1	3	1	3	2	1	3	3	1	3	1	1	3	12	1282	2	4	20	1	1	0
1	3	1	1	2	1	3	3	2	3	1	1	3	18	1131	4	2	33	1	1	0
1	1	1	5	3	1	3	3	1	3	1	1	1	15	950	4	3	33	2	2	0
1	4	1	2	3	1	2	3	2	3	1	1	7	21	3414	2	1	26	2	1	0
1	2	5	3	3	1	2	3	2	2	1	1	1	21	1647	4	2	40	2	2	0
1	3	2	4	4	1	2	3	2	3	1	1	4	12	674	4	1	20	1	1	0
1	3	1	4	4	1	2	3	1	3	1	1	9	12	902	4	4	21	1	1	0
1	1	1	1	3	1	2	2	2	3	2	1	10	27	5293	2	4	50	2	1	0
1	3	1	5	3	1	4	3	3	4	2	1	3	12	7865	4	4	53	1	1	0
1	3	1	1	3	1	3	2	2	4	1	1	4	24	1823	4	2	30	1	2	0
1	3	5	5	2	1	3	1	2	3	1	1	1	24	915	4	2	29	1	1	0
1	1	1	5	3	1	4	3	3	3	1	1	2	48	4605	3	4	24	2	2	0
1	2	1	3	1	1	4	3	3	3	1	1	4	12	2149	4	1	29	1	1	0
1	3	1	5	1	1	2	3	2	4	2	2	1	14	8978	1	4	45	1	1	0
1	3	1	2	1	1	4	3	2	2	2	1	6	30	11998	1	1	34	1	1	0
1	4	1	2	4	1	1	2	2	3	1	1	4	24	1024	4	4	48	1	1	0
1	3	2	2	4	1	4	3	2	2	2	1	4	12	727	4	3	33	1	1	0
1	2	1	5	3	1	3	1	2	3	1	1	3	36	2746	4	4	31	1	1	0
1	3	1	2	1	1	2	3	2	3	1	1	4	24	1938	4	3	32	1	1	0
1	1	1	5	3	1	4	1	1	3	1	1	3	24	4110	3	4	23	2	2	0
1	5	1	3	1	1	1	3	1	4	2	1	3	6	3384	1	4	44	1	1	0
1	3	5	3	2	1	3	3	1	3	1	1	5	15	1275	4	2	24	1	1	0
1	1	1	3	3	1	2	3	1	3	2	1	4	12	6199	4	2	28	2	1	0
1	3	1	4	2	1	1	3	1	3	1	1	3	30	2406	4	4	23	1	1	0

1	3	1	1	3	1	3	3	2	1	1	1	3	18	2473	4	1	25	1	1	0
1	3	1	5	3	1	3	3	2	3	1	1	3	36	3446	4	2	42	1	2	0
1	2	1	4	2	3	3	3	1	3	1	1	10	48	7685	2	4	37	1	1	0
1	3	5	3	2	1	2	3	2	3	1	1	1	24	1381	4	2	35	1	1	0
1	3	1	2	3	1	1	3	2	2	1	1	1	12	2579	4	1	33	1	2	0
1	5	1	1	2	1	4	3	3	1	1	1	6	18	1190	2	4	55	3	2	0
1	3	1	2	2	1	3	3	2	3	2	1	1	36	1842	4	4	34	1	1	0
1	3	1	4	3	2	2	3	2	3	2	1	1	24	4817	2	3	31	1	1	0
1	3	1	5	3	1	4	3	3	3	2	1	2	12	3386	3	4	35	1	1	0

1	4	1	2	2	1	3	3	1	2	2	1	4	24	1659	4	2	29	1	1	0
1	3	1	2	2	1	2	3	2	3	1	1	7	6	448	4	4	23	1	1	0
1	3	1	3	3	1	3	3	2	2	1	1	1	9	654	4	3	28	1	1	0
1	1	1	2	2	1	2	3	1	3	1	1	3	18	3114	1	4	26	1	1	0
1	3	5	3	4	1	3	3	2	3	1	1	3	24	2996	2	4	20	1	1	0
1	3	1	2	2	1	2	3	2	3	1	1	1	24	3123	4	1	27	1	1	0
1	3	1	2	2	1	2	3	1	3	1	1	1	24	1207	4	4	24	1	1	0
1	3	1	4	3	1	1	3	2	3	1	1	1	12	759	4	2	26	1	1	0
1	3	1	3	2	1	1	3	2	3	1	1	4	18	3190	2	2	24	1	1	0
1	1	1	3	3	1	4	3	3	3	1	1	3	48	7119	3	4	53	2	2	0
1	3	1	3	1	1	3	3	1	3	1	1	4	36	2302	4	4	31	1	1	0
1	3	1	5	3	1	4	1	3	4	1	1	1	48	7763	4	4	42	1	1	0
1	3	5	5	3	1	2	3	2	3	1	1	3	30	6350	4	4	31	1	1	0
1	4	1	1	3	1	1	3	3	3	1	1	1	24	1333	4	2	43	2	2	0
1	2	1	3	2	1	1	1	2	2	1	1	4	12	626	4	4	24	1	1	0
1	3	1	4	4	3	1	3	2	3	2	1	4	48	6999	1	1	34	2	1	0
1	2	1	4	3	1	3	1	2	3	1	1	3	24	3552	3	4	27	1	1	0
1	2	1	4	3	3	3	1	1	2	1	1	4	24	1546	4	4	24	1	1	0
1	3	1	4	1	1	3	3	2	3	1	1	1	12	1372	2	3	36	1	1	0
1	3	1	3	2	1	3	3	2	3	2	1	4	48	6758	3	2	31	1	1	0
1	3	1	2	2	1	1	3	1	2	2	1	3	24	3234	4	4	23	1	1	0
1	3	1	3	2	1	1	1	1	3	1	1	3	18	2039	1	4	20	1	1	0
1	3	1	3	4	1	1	3	2	2	2	1	5	18	1217	4	3	47	1	1	0
1	5	1	5	3	1	2	3	2	3	1	1	3	12	2246	3	3	60	2	1	0

1	2	5	5	3	1	3	2	2	4	2	1	11	24	1358	4	3	40	1	1	0
1	3	1	4	2	1	3	3	1	3	1	1	1	24	1442	4	4	23	2	1	0
1	3	2	1	2	1	2	3	2	3	1	1	5	12	741	4	3	22	1	1	0
1	3	1	3	3	1	4	3	3	3	1	1	4	18	2600	4	4	65	2	1	0
1	2	1	3	3	1	3	1	2	3	1	1	3	18	1553	4	3	44	1	1	0
1	4	1	4	3	1	2	1	2	3	2	1	4	42	4370	3	2	26	2	2	0
1	3	1	1	2	1	1	3	2	1	1	1	7	18	750	4	1	27	1	1	0
1	3	1	3	3	1	3	3	1	2	1	1	7	12	684	4	4	40	1	2	0
1	3	1	5	3	2	1	3	2	3	1	1	1	24	2303	4	1	45	1	1	0
1	3	5	3	4	1	3	3	2	3	1	1	1	12	900	4	2	23	1	1	0
1	5	1	5	3	2	2	3	1	3	2	1	1	12	4843	3	4	43	2	1	0
1	5	1	5	3	1	4	3	3	3	2	1	2	48	6331	4	4	46	2	1	0
1	2	1	5	2	1	4	3	3	3	1	1	7	6	1198	4	4	35	1	1	0
1	3	1	3	2	1	1	3	2	2	1	1	1	12	1228	4	2	24	1	1	0
1	3	1	3	2	1	3	1	1	3	1	1	4	18	1882	4	4	25	2	1	0
1	3	1	3	3	1	1	3	1	2	1	1	4	24	1987	2	4	21	1	2	0
1	3	1	2	2	1	1	3	2	2	1	1	1	12	1274	3	1	37	1	1	0

1	3	1	2	2	1	3	3	2	2	1	1	1	18	976	1	2	23	1	1	0
1	3	5	4	2	1	4	3	1	3	1	1	1	24	1285	4	4	32	1	1	0
1	5	1	5	3	1	2	3	2	3	1	1	1	12	691	4	3	35	2	1	0
1	5	1	3	2	1	1	3	1	3	1	1	3	18	2124	4	4	24	2	1	0
1	3	1	2	1	1	2	3	2	2	1	1	3	30	3108	2	4	31	1	1	0
1	5	1	3	2	1	4	3	2	4	2	1	7	36	8065	3	2	25	2	1	0
1	3	5	3	2	1	1	3	1	3	1	1	1	24	1371	4	4	25	1	1	0
1	5	1	3	3	1	4	1	2	3	2	1	7	40	5998	4	3	27	1	1	0
1	3	5	5	3	1	2	3	3	3	2	1	2	18	7511	1	4	51	1	2	0
1	3	1	4	3	1	3	3	2	3	2	1	1	36	9271	2	1	24	1	1	0
1	5	1	4	3	1	3	3	2	3	2	1	2	36	9629	4	4	24	2	1	0
1	3	1	3	3	1	3	3	2	3	1	1	5	48	3051	3	4	54	1	1	0
1	3	1	4	3	1	4	3	3	3	1	1	1	48	3931	4	4	46	1	2	0
1	3	1	3	4	1	1	1	2	3	1	1	4	18	1345	4	3	26	1	1	0
1	3	1	5	3	1	2	1	2	3	1	1	3	36	2712	2	2	41	1	2	0
1	5	1	5	2	1	1	1	1	3	2	1	1	18	3966	1	4	33	3	1	0

1	3	5	5	3	1	4	3	3	3	1	1	2	36	8335	3	4	47	1	1	0
1	3	1	2	2	1	3	3	1	3	2	1	1	18	1216	4	3	23	1	1	0
1	1	1	5	3	1	3	3	1	3	1	1	10	45	11816	2	4	29	2	1	0
1	1	1	3	3	1	3	1	2	3	1	1	1	12	1082	4	4	48	2	1	0
1	3	1	4	3	1	2	3	2	3	1	1	3	36	5179	4	2	29	1	1	0
1	3	1	5	3	1	1	2	2	2	1	1	4	12	709	4	4	57	1	1	0
1	5	1	3	4	3	2	1	1	3	1	2	1	20	2235	4	2	33	2	1	0
1	2	1	4	3	1	4	3	3	2	1	1	1	18	1442	4	4	32	2	2	0
1	3	1	3	2	1	2	3	2	3	2	1	1	15	3959	3	2	29	1	1	0
1	3	1	2	2	1	1	3	2	3	2	1	4	24	2439	4	4	35	1	1	0
1	3	1	2	3	1	4	3	3	4	2	1	1	9	1422	3	2	27	1	1	0
1	3	1	2	2	1	2	3	2	3	1	1	7	12	795	4	4	53	1	1	0
1	3	1	3	3	2	3	3	2	3	1	1	3	18	4153	2	3	42	1	1	0
1	5	1	5	3	3	2	1	1	3	2	1	1	16	2625	2	4	43	1	1	0
1	2	1	3	3	1	2	3	1	3	2	1	10	24	3161	4	2	31	1	1	0
1	3	1	5	3	1	4	1	2	4	2	1	1	6	14896	1	4	68	1	1	0
1	3	2	1	1	1	2	3	2	3	1	1	3	24	2359	1	1	33	1	1	0
1	3	1	5	3	1	2	3	1	4	2	1	3	24	3345	4	2	39	1	1	0
1	3	1	2	2	1	2	3	1	3	1	1	4	9	1366	3	4	22	1	1	0
1	2	1	2	1	1	2	1	2	3	2	1	3	24	6872	2	1	55	1	1	0
1	2	1	2	3	1	3	1	2	3	2	1	1	12	697	4	2	46	2	1	0
1	3	1	4	3	1	4	2	3	3	2	1	2	48	10297	4	4	39	3	2	0
1	2	3	2	3	1	4	3	3	3	2	1	3	24	3349	4	4	30	1	2	0

1	4	1	4	3	1	3	3	2	3	2	1	10	36	2145	2	1	24	2	1	0
1	3	1	3	3	1	3	3	2	2	1	1	1	28	4006	3	2	45	1	1	0
1	2	1	1	2	2	4	3	1	1	1	1	1	24	1193	1	4	29	2	1	0
1	3	1	5	3	2	4	3	1	3	1	1	10	60	7297	4	4	36	1	1	0
1	3	1	3	3	1	2	3	2	3	1	1	2	36	8229	2	2	26	1	2	0
1	3	1	3	3	1	4	3	3	3	2	1	4	45	1845	4	4	23	1	1	0
2	5	5	5	3	3	1	3	2	3	2	1	4	15	1537	4	4	50	2	1	1
2	5	1	4	2	1	1	1	2	2	1	1	10	42	5954	2	1	41	2	1	1

2	4	1	4	3	1	2	3	2	3	1	1	4	11	4771	2	4	51	1	1	1
2	3	1	2	4	1	2	3	2	3	1	1	3	30	3832	2	1	22	1	1	1
2	5	1	2	2	1	2	3	2	3	1	1	3	18	1295	4	1	27	2	1	1
2	3	4	5	3	1	1	3	2	3	2	1	1	12	1318	4	4	54	1	1	1
2	3	2	3	3	1	3	3	2	3	1	1	10	18	2622	4	4	34	1	1	1
2	4	5	4	3	1	2	1	1	4	2	1	2	20	7057	3	4	36	2	2	1
2	3	1	4	3	1	3	3	1	3	1	1	4	36	2323	4	4	24	1	1	1
2	3	1	4	3	1	1	3	1	2	1	1	1	18	6260	3	3	28	1	1	1
2	5	3	4	1	1	3	3	2	4	2	1	1	12	2366	3	3	36	1	1	1
2	5	1	1	3	1	3	3	2	4	2	1	2	12	1860	4	2	34	2	1	1
2	3	5	4	2	1	3	3	2	3	1	1	1	48	8487	1	2	24	1	1	1
2	3	1	2	2	1	4	3	2	4	2	1	4	36	4795	4	1	30	1	1	1
2	5	4	5	2	1	4	3	1	3	1	1	2	24	7758	2	4	29	1	1	1
2	4	2	4	3	1	3	3	1	4	2	1	10	24	6967	4	4	36	1	1	1
2	3	2	4	3	1	3	1	2	3	2	1	1	24	3512	2	3	38	2	1	1
2	3	1	3	3	3	1	3	2	3	1	2	4	8	1414	4	2	33	1	1	1
2	5	1	1	3	1	3	3	3	1	2	1	1	16	1175	2	3	68	3	1	1
2	3	5	5	3	1	4	1	3	3	2	1	2	24	2760	4	4	36	1	1	1
2	5	1	4	3	1	2	3	2	3	1	1	5	12	1424	4	3	26	1	1	1
2	3	1	2	4	3	1	1	2	2	1	1	4	12	625	4	1	26	1	1	1
2	3	5	4	3	1	3	3	2	3	2	1	3	9	2030	2	1	24	1	1	1
2	4	2	1	3	1	4	3	3	4	2	1	11	48	7582	2	4	31	1	1	1
2	3	1	3	2	3	1	3	2	3	2	1	4	12	1092	4	4	49	2	1	1
2	5	4	4	3	1	3	3	2	3	2	1	3	21	2745	3	2	32	2	1	1
2	3	5	1	3	1	4	1	3	1	2	1	11	48	5381	3	4	40	1	1	1
2	3	1	2	2	3	2	3	2	2	1	1	4	13	2101	2	4	23	1	1	1
2	3	1	3	1	1	3	3	2	2	1	1	3	10	1521	4	2	31	1	1	1
2	3	3	2	4	1	1	3	1	3	1	1	4	18	3213	1	3	25	1	1	1
2	3	1	5	3	2	1	1	2	4	2	1	10	18	4439	1	1	33	1	1	1
2	3	4	2	2	1	1	3	1	2	1	1	3	12	983	1	4	19	1	1	1
2	4	2	4	3	1	4	3	1	3	2	1	7	39	11760	2	3	32	1	1	1
2	3	1	5	2	1	3	3	2	3	2	1	4	24	1967	4	4	20	1	1	1
2	3	5	3	3	1	3	3	2	3	1	1	2	12	2028	4	2	30	1	1	1

2	3	1	3	4	1	3	3	1	3	2	1	2	18	2779	1	3	21	1	1	1
2	4	1	3	3	1	3	1	2	3	2	1	11	20	2629	2	3	29	2	1	1
2	3	1	2	2	1	1	3	1	4	1	1	3	12	3017	3	1	34	1	1	1
2	3	5	5	3	1	2	3	2	3	1	1	9	12	754	4	4	38	2	1	1
2	3	2	4	3	1	1	3	2	2	1	1	10	12	1037	3	4	39	1	1	1
2	5	1	3	2	1	1	3	2	2	1	1	4	12	3573	1	1	23	1	1	1
2	3	5	5	3	1	2	3	2	3	1	1	4	6	368	4	4	38	1	1	1
2	3	1	5	4	3	1	3	2	2	1	1	4	18	1301	4	2	32	1	1	1
2	3	1	3	2	3	1	3	2	2	1	1	4	18	1113	4	4	26	1	2	1
2	3	1	2	3	1	3	3	2	3	1	1	2	26	7966	2	3	30	2	1	1
2	3	1	4	3	1	3	3	2	3	1	1	7	36	2273	3	1	32	2	2	1
2	4	5	3	3	1	2	1	2	4	2	1	10	24	4712	4	2	37	2	1	1
2	4	2	4	2	1	2	3	1	3	2	1	4	24	1553	3	2	23	2	1	1
2	3	1	5	3	1	3	3	2	2	1	1	4	9	1082	4	4	27	2	1	1
2	3	1	4	2	1	1	3	2	3	1	1	4	12	2930	2	1	27	1	1	1
2	5	1	2	3	1	1	1	2	2	1	1	1	12	3124	1	3	49	2	2	1
2	3	1	2	2	3	1	3	2	3	1	1	4	7	2329	1	1	45	1	1	1
2	1	5	3	3	1	3	1	2	4	2	1	10	48	12204	2	2	48	1	1	1
2	4	5	3	3	1	4	3	3	4	1	1	4	60	9157	2	2	27	1	1	1
2	5	1	4	3	1	2	3	2	3	1	1	10	21	3652	2	3	27	2	1	1
2	3	5	2	3	1	2	3	2	3	1	1	4	15	1444	4	1	23	1	1	1
2	3	1	5	3	1	3	3	2	2	1	1	6	15	1308	4	4	38	2	1	1
2	4	2	5	3	1	4	3	3	3	1	1	1	36	2862	4	3	30	1	1	1
2	3	2	3	2	1	1	3	2	2	1	1	6	15	2631	3	2	25	1	1	1
2	4	2	5	3	1	2	2	2	3	1	1	10	30	2503	4	2	41	2	1	1
2	3	1	5	2	1	1	3	2	3	1	1	4	9	1206	4	4	25	1	1	1
2	3	1	3	3	1	1	3	2	2	1	1	4	9	2118	2	2	37	1	2	1
2	3	3	3	2	1	1	3	2	2	1	1	4	9	790	4	3	66	1	1	1
2	4	1	2	3	1	3	3	2	3	1	1	4	24	6403	1	2	33	1	1	1
2	3	1	4	2	3	1	3	2	2	1	1	4	8	760	4	2	44	1	1	1
2	2	5	2	2	1	4	1	2	1	1	1	5	36	3990	3	2	29	1	1	1
2	2	1	3	3	1	1	2	2	2	1	1	4	10	1048	4	4	23	1	1	1
2	3	2	4	2	1	1	3	1	2	1	1	10	12	841	2	4	23	1	1	1
2	5	1	5	3	1	3	3	1	3	2	1	3	12	3617	1	4	28	3	1	1

2	3	1	3	2	1	1	3	2	3	1	1	4	48	5951	2	2	22	1	1	0
2	5	1	1	4	1	3	3	2	4	1	1	1	30	5234	4	2	28	2	1	0
2	3	1	2	2	1	3	3	1	3	1	1	1	12	1295	3	1	25	1	1	0
2	3	1	5	2	1	4	3	3	4	2	1	2	24	12579	4	2	44	1	1	0
2	5	1	2	3	1	2	3	2	2	1	1	4	45	4746	4	2	25	2	1	0
2	4	1	5	3	1	4	1	3	3	2	1	1	36	2225	4	4	57	2	1	0

2	3	5	1	3	1	4	3	2	4	2	1	4	12	6468	2	1	52	1	1	0
2	1	1	5	3	1	4	3	3	4	2	1	10	36	1953	4	4	61	1	1	0
2	1	1	3	3	1	3	3	2	3	2	1	10	48	14421	2	2	25	1	1	0
2	3	2	3	3	1	4	3	3	3	2	1	7	36	12612	1	4	47	1	2	0
2	1	1	2	3	1	4	3	1	3	2	1	10	54	15945	3	4	58	1	1	0
2	5	1	3	3	2	3	3	2	4	2	1	11	24	11938	2	3	39	2	2	0
2	3	5	2	2	1	1	3	1	3	1	1	3	18	1924	4	3	27	1	1	0
2	3	1	3	3	1	3	3	2	3	1	1	6	12	639	4	2	30	1	1	0
2	3	3	3	3	1	1	3	2	2	1	1	4	12	766	4	3	66	1	1	0
2	5	1	5	1	1	1	3	2	3	2	1	10	24	1935	4	4	31	2	1	0
2	4	1	1	2	1	2	3	2	4	2	1	3	24	2064	3	2	34	1	1	0
2	4	1	3	1	1	1	2	2	4	2	1	10	36	4455	2	2	30	2	1	0
2	5	1	5	3	1	3	1	2	3	2	1	1	18	884	4	4	36	1	2	0
2	2	1	5	2	1	4	1	3	4	2	1	2	9	5129	2	4	74	1	2	0
2	1	2	4	3	1	4	3	3	2	1	1	10	48	3844	4	4	34	1	2	0
2	3	1	3	3	1	3	3	2	3	2	1	10	27	3915	4	2	36	1	2	0
2	3	2	3	3	3	2	3	1	3	1	1	4	45	3031	4	4	21	1	1	0
2	5	1	5	2	1	3	3	2	4	2	1	7	9	1501	2	3	34	2	1	0
2	3	2	2	2	1	3	1	1	3	1	1	3	12	951	4	4	27	4	1	0
2	4	1	5	1	1	4	3	2	4	2	1	3	18	4297	4	3	40	1	1	0
2	3	4	4	3	2	4	1	2	3	2	1	4	48	10961	1	2	27	2	1	0
2	3	5	1	3	1	2	3	2	1	2	1	1	6	14555	1	2	23	1	1	0
2	3	2	5	1	1	3	1	1	2	1	1	10	21	2767	4	2	61	2	1	0

2	3	1	3	2	3	4	1	2	3	1	1	1	30	2150	4	2	24	1	1	0
2	3	1	5	3	1	3	3	2	3	1	1	4	15	802	4	3	37	1	2	0
2	3	1	4	3	1	1	3	2	3	1	1	4	48	3060	4	4	28	2	1	0
2	3	1	2	4	1	3	3	1	4	2	1	2	36	9398	1	4	28	1	1	0
2	3	1	2	2	1	3	3	2	3	2	1	3	48	9960	1	2	26	1	1	0
2	1	1	3	2	1	3	3	2	3	2	1	4	36	3804	4	1	42	1	1	0
2	3	1	3	2	1	1	3	2	3	1	1	3	8	1237	3	4	24	1	1	0
2	3	1	4	4	1	3	1	2	2	1	1	1	12	685	2	3	25	1	1	0
2	5	1	1	4	1	3	3	2	4	1	1	1	30	4249	4	2	28	2	1	0
2	3	2	2	2	1	2	3	2	4	1	1	1	60	7408	4	2	24	1	1	0
2	5	4	5	3	1	4	3	3	3	1	1	7	9	1136	4	3	32	2	2	0
2	3	1	3	2	1	3	3	2	3	1	2	3	9	959	1	2	29	1	1	0
2	3	1	2	3	1	1	2	2	2	1	1	1	24	1246	4	2	23	1	1	0
2	2	2	5	2	1	4	1	3	4	2	1	11	60	14782	3	4	60	2	1	0

2	3	1	5	3	1	4	3	3	4	2	1	1	36	14318	4	2	57	1	1	0
2	3	1	1	2	1	4	3	3	4	2	1	2	18	12976	3	4	38	1	1	0
2	3	1	5	1	1	1	3	1	3	1	1	1	12	1223	1	1	46	2	1	0
2	3	1	2	4	1	2	3	2	3	2	1	4	24	2039	1	1	22	1	1	0
2	3	5	3	2	1	2	3	2	3	1	1	1	18	1042	4	2	33	1	1	0
2	3	5	5	2	1	2	1	2	3	2	1	3	12	2762	1	2	25	1	1	0
2	3	1	3	3	2	3	1	2	4	2	1	11	24	11328	2	3	29	2	1	0
2	3	5	3	4	1	1	3	2	3	2	1	4	12	1484	2	1	25	1	1	0
2	5	1	3	2	1	3	3	2	4	2	1	3	48	5096	2	3	30	1	1	0
2	4	4	3	4	1	2	2	2	3	1	1	6	15	1512	3	3	61	2	1	0
2	3	2	2	4	1	3	3	1	3	2	1	4	24	3092	3	2	22	1	1	0
2	5	1	3	4	1	3	3	2	3	1	1	4	18	1245	4	2	33	1	1	0
2	4	1	1	3	1	2	3	2	4	2	1	1	6	1209	4	4	47	1	1	0
2	1	1	5	2	1	4	3	3	4	2	1	10	27	8318	2	4	42	2	1	0
2	3	2	2	3	2	4	3	1	4	2	1	3	36	9034	4	1	29	1	1	0
2	5	2	2	3	1	4	1	2	3	1	1	4	24	1216	4	4	38	2	2	0
2	2	2	1	3	1	2	3	2	4	2	1	1	15	6850	1	2	34	1	2	0

2	3	2	3	2	1	3	3	1	3	2	1	1	15	2631	2	4	28	2	1	0
2	3	1	2	4	1	1	3	1	3	1	1	4	12	1534	1	1	23	1	1	0
2	3	1	2	3	1	3	2	2	3	1	1	4	12	1331	2	1	22	1	1	0
2	1	2	2	2	1	3	3	2	3	1	1	1	18	2278	3	3	28	2	1	0
2	5	1	2	3	1	1	3	2	2	1	1	3	18	1928	2	2	31	2	1	0
2	3	5	3	3	1	4	3	3	3	2	1	1	36	12389	1	4	37	1	1	0
2	5	1	2	1	1	3	3	2	3	1	1	1	36	2820	4	4	27	2	1	0
2	5	1	5	3	3	1	1	2	3	1	1	1	18	1056	3	3	30	2	1	0
2	3	1	2	3	1	4	3	1	2	1	1	6	36	2384	4	1	33	1	1	0
2	3	1	3	2	1	2	3	1	2	2	1	1	24	2718	3	4	20	1	1	0
2	2	2	2	2	1	2	2	2	2	1	1	1	6	931	1	1	32	1	1	0
2	1	1	3	3	1	1	3	2	2	2	1	10	24	4241	1	4	36	3	1	0
2	3	1	3	2	1	2	3	2	3	1	1	3	9	918	4	1	30	1	1	0
2	2	1	4	2	1	4	1	3	2	1	1	7	24	1837	4	4	34	1	1	0
2	3	2	3	2	2	4	3	3	3	1	1	4	36	2671	4	4	50	1	1	0
2	3	2	3	2	2	3	3	1	3	1	1	3	30	3441	2	4	21	1	1	0
2	2	2	4	3	1	4	3	2	3	1	1	1	9	1437	2	3	29	1	1	0
2	3	1	2	2	1	3	3	2	2	2	1	1	24	1355	3	4	25	1	1	0
2	3	1	3	3	1	2	3	2	2	1	1	3	12	1922	4	2	37	1	1	0
2	4	2	5	3	1	3	3	2	4	2	1	1	36	8086	2	4	42	4	1	0
2	3	1	5	3	1	3	1	2	2	1	1	1	12	888	4	4	41	1	2	0
2	3	2	3	4	1	3	3	2	3	1	1	4	72	5595	2	2	24	1	1	0
2	3	2	4	3	1	2	3	2	3	1	1	1	48	6560	3	2	24	1	1	0
2	1	1	2	2	1	2	3	1	3	1	1	3	12	2969	4	3	25	2	1	0
2	3	1	4	3	1	4	3	2	4	2	1	1	60	14027	4	2	27	1	1	0
2	2	4	2	2	1	2	1	1	3	1	1	7	6	433	4	2	24	1	2	0
2	2	1	5	2	1	4	3	1	3	1	1	10	48	6416	4	3	59	1	1	0
2	3	1	3	2	1	3	3	1	4	1	1	2	24	11560	1	4	23	2	1	0
2	1	2	3	2	1	3	3	1	2	1	1	10	30	4280	4	4	26	2	1	0
2	5	1	4	3	3	1	3	2	3	1	1	4	39	4933	2	2	25	2	1	0
2	3	2	2	2	1	2	1	2	2	1	1	1	12	836	4	2	23	1	1	0

2	4	1	5	3	1	4	3	3	3	1	1	7	48	6224	4	4	50	1	1	0
2	3	1	5	2	1	2	3	2	3	1	1	10	21	1188	2	4	39	1	2	0
2	1	1	2	2	1	1	3	1	1	1	1	1	15	1778	2	1	26	2	1	0
2	3	1	4	1	1	3	3	2	3	2	1	3	24	4057	3	3	43	1	1	0
2	3	1	3	3	1	3	3	2	3	2	1	10	48	15672	2	2	23	1	1	0
2	5	1	2	2	1	3	1	2	2	1	1	3	24	4736	2	4	25	1	1	0
2	1	1	3	2	1	2	1	2	4	2	2	11	48	18424	1	2	32	1	1	0
2	3	1	2	2	1	3	3	2	3	2	1	4	9	1670	4	2	22	1	1	0
2	4	2	2	3	1	4	2	2	4	1	1	4	30	1919	4	3	30	2	1	0
2	3	1	3	3	1	4	3	3	3	1	1	7	60	6288	4	4	42	1	1	0
2	3	3	2	2	1	3	3	1	3	1	1	2	24	4113	3	4	28	1	1	0
2	5	3	3	3	1	2	3	2	2	1	1	4	27	2520	4	2	23	2	1	0
2	2	2	3	4	1	2	3	1	3	1	1	1	15	1264	2	2	25	1	1	0
2	5	1	4	3	1	2	3	2	3	1	1	3	30	8386	2	2	49	1	1	0
3	5	5	5	3	2	4	3	1	3	1	1	3	24	3617	4	4	20	2	1	1
3	3	1	2	2	1	2	3	1	3	2	1	7	15	392	4	4	23	1	1	1
3	5	1	5	3	1	4	3	3	3	2	1	1	10	781	4	4	63	2	1	1
3	3	1	3	2	1	3	3	1	2	2	1	4	12	1881	2	2	44	1	1	1
3	3	4	2	4	1	1	3	2	1	1	2	1	6	709	2	2	27	1	1	1
3	3	1	3	3	1	3	3	2	3	1	1	4	36	5848	4	1	24	1	1	1
3	3	1	3	3	1	1	3	2	3	2	1	4	36	3913	2	2	23	1	1	1
3	3	1	3	3	1	1	3	2	3	1	1	3	24	1925	2	2	26	1	1	1
3	3	1	3	2	1	3	3	2	2	1	1	3	12	2251	1	2	46	1	1	1
3	2	5	4	3	1	3	1	2	2	1	1	4	18	1445	4	4	49	1	1	1
3	5	1	5	3	1	4	3	3	3	1	1	2	42	4796	4	4	56	1	1	1
3	3	5	2	3	1	1	3	2	2	1	2	1	4	1494	1	2	29	1	2	1
3	3	1	5	3	1	3	3	2	3	1	1	4	36	4473	4	2	31	1	1	1
3	4	1	3	2	1	3	2	2	3	2	1	1	12	2247	2	2	36	2	1	1
3	3	1	2	3	1	1	3	2	3	1	1	1	12	1330	4	1	26	1	1	1
3	3	1	4	3	1	3	1	2	3	1	1	4	24	5152	4	2	25	1	1	1
3	5	2	5	1	1	3	3	2	3	2	1	1	6	1323	2	4	28	2	2	1
3	3	2	5	1	1	1	3	2	3	1	1	4	9	1126	2	4	49	1	1	1

3	3	1	2	2	1	2	2	2	2	1	1	3	18	3049	1	1	45	1	1	1
3	4	1	2	2	1	2	1	2	3	1	1	4	6	683	2	1	29	1	1	1
3	3	1	3	3	1	1	3	2	3	2	1	3	6	2116	2	2	41	1	1	1
3	5	1	3	2	1	2	3	2	2	1	1	7	6	1047	2	4	50	1	1	1
3	5	1	5	3	1	2	3	2	3	1	1	4	30	3017	4	4	47	1	1	1
3	5	5	4	3	1	2	3	2	3	2	1	4	10	1347	4	2	27	2	1	1
3	3	1	5	3	1	3	3	1	4	2	1	7	15	1905	4	4	40	1	1	1
3	3	2	5	2	1	4	3	3	3	2	1	4	24	1377	4	2	47	1	1	1
3	5	5	3	3	1	3	3	2	3	2	1	4	24	3148	3	2	31	2	1	1
3	5	1	3	3	1	1	3	2	1	1	2	1	6	1299	1	1	74	3	2	1
3	3	1	3	3	2	1	2	2	3	1	1	4	18	2100	4	2	37	1	1	0
3	4	1	5	3	1	1	3	2	4	2	1	10	30	1908	4	4	66	1	1	0
3	3	1	3	3	1	3	3	2	3	1	1	4	36	4210	4	2	26	1	1	0
3	5	5	4	3	1	1	1	2	2	1	1	1	24	1344	4	2	37	2	2	0
3	5	3	4	4	1	1	3	2	3	2	1	1	12	939	4	2	28	3	1	0
3	2	1	2	2	1	1	3	2	1	1	1	10	12	609	4	1	26	1	1	0
3	5	1	2	1	1	3	3	1	3	1	1	7	21	2319	2	1	33	1	1	0
3	3	1	3	3	1	1	3	2	2	1	1	3	18	2864	2	1	34	1	2	0
3	3	2	5	2	1	4	3	3	2	1	1	1	10	1240	1	4	48	1	2	0
3	3	1	4	3	1	4	1	3	3	1	1	1	24	947	4	3	38	1	2	0
3	1	1	2	3	1	3	3	2	4	2	1	4	9	1337	4	2	34	2	1	0
3	5	5	3	3	1	4	3	3	3	2	1	4	15	1271	3	4	39	2	1	0
3	3	1	2	2	1	1	3	2	2	1	1	4	15	2327	2	3	25	1	1	0
3	3	1	3	2	1	1	3	2	2	1	1	4	9	745	3	2	28	1	1	0
4	3	5	3	3	1	2	3	2	3	2	1	2	18	3378	2	1	31	1	1	1
4	2	5	3	3	3	1	2	2	2	1	1	1	6	783	1	2	26	1	2	1
4	3	1	2	2	1	2	3	2	3	2	1	4	24	3181	4	4	26	1	1	1
4	3	5	5	3	1	2	3	2	3	2	1	6	27	5190	4	4	48	4	2	1
4	3	1	2	2	1	3	1	2	3	1	1	4	12	2171	2	2	29	1	1	1
4	3	5	3	2	1	3	3	2	3	1	1	4	36	2394	4	4	25	1	1	1
4	3	1	3	2	1	2	1	2	3	1	1	2	36	8133	1	2	30	1	1	1
4	5	1	5	3	1	1	3	2	3	1	1	4	12	618	4	4	56	1	1	1
4	3	5	2	4	1	3	3	1	3	2	1	2	12	2445	2	4	26	1	1	1
4	3	1	2	3	1	2	1	2	2	1	1	3	15	2708	2	3	27	2	1	1

4	3	1	3	3	1	4	1	3	3	1	1	3	18	1984	4	4	47	2	1	1
4	5	5	5	2	1	1	3	2	3	2	1	4	12	1240	4	2	38	2	1	1
4	5	1	2	2	1	2	3	2	3	2	1	3	21	2288	4	4	23	1	1	1
4	5	4	5	3	1	3	3	2	3	1	1	4	18	1582	4	4	46	2	1	1
4	3	2	5	3	1	4	1	3	3	2	1	4	6	1346	2	4	42	1	2	1
4	5	5	3	3	1	1	3	2	2	1	1	4	6	1898	1	2	34	2	2	1
4	3	3	5	3	1	2	3	2	3	1	1	1	36	909	4	4	36	1	1	1
4	3	3	3	3	1	3	3	2	3	2	1	3	6	2978	1	2	32	1	1	1
4	3	1	4	2	1	2	3	1	3	2	1	3	24	3972	2	4	25	1	1	1
4	3	1	4	3	1	3	3	1	4	2	1	4	12	1963	4	2	31	2	2	1
4	5	4	3	3	1	1	3	2	3	2	1	3	24	3777	4	4	50	1	1	1
4	5	1	5	3	1	1	3	2	3	2	1	3	12	1935	4	4	43	3	1	1
4	5	5	3	3	1	1	3	2	3	1	1	4	9	3074	1	2	33	2	2	1
4	5	5	5	3	1	3	3	2	3	2	1	4	36	3342	4	2	51	1	1	1
4	3	1	3	3	1	1	1	2	3	1	1	4	12	707	4	2	30	2	1	1
4	4	5	3	3	1	1	3	2	3	2	1	10	24	2978	4	4	32	2	2	1
4	3	1	3	3	1	3	3	2	3	2	1	2	15	4657	3	2	30	1	1	1
4	1	1	3	3	1	3	3	2	3	1	1	6	36	2613	4	2	27	2	1	1
4	3	3	5	3	1	4	3	3	2	1	1	1	9	2507	2	4	51	1	1	1
4	3	2	4	3	1	4	3	2	3	1	1	4	12	2141	3	1	35	1	1	1
4	5	1	4	3	1	1	3	2	2	1	1	4	4	1544	2	1	42	3	2	1
4	3	1	4	3	1	2	3	2	3	1	1	4	6	1595	3	2	51	1	2	1
4	5	1	3	2	1	1	3	2	3	1	1	10	12	1185	3	2	27	2	1	1
4	2	3	3	2	1	1	3	2	2	1	1	9	12	3447	4	3	35	1	2	1
4	5	1	5	3	1	1	3	2	3	1	1	4	12	717	4	4	52	3	1	1
4	1	2	3	3	1	4	1	1	3	1	2	1	6	1204	4	1	35	1	1	1
4	5	1	3	3	1	1	3	2	2	1	1	4	24	2684	4	2	35	2	1	1
4	3	2	2	2	1	3	3	2	3	2	1	2	15	3812	1	4	23	1	1	1
4	5	2	1	2	1	1	3	2	2	1	1	4	11	1154	4	4	57	3	1	1
4	5	1	5	2	1	4	1	3	2	1	1	7	12	2748	2	4	57	3	1	1
4	5	5	5	3	1	1	3	2	3	1	1	10	36	6304	4	4	36	2	1	1
4	3	1	2	2	1	3	2	2	3	2	1	4	24	1533	4	3	38	1	1	1
4	5	3	4	2	1	3	3	1	3	2	1	3	12	1402	3	4	37	1	1	1

4	3	2	3	3	1	3	3	1	4	2	1	4	24	1901	4	4	29	1	1	1
4	3	1	3	3	1	1	3	2	3	1	1	3	12	1574	4	2	50	1	1	1
4	5	5	5	3	1	2	3	2	3	1	1	3	15	1520	4	4	63	1	1	1
4	3	5	2	2	1	3	3	1	3	1	1	2	12	4675	1	4	20	1	1	1
4	3	5	5	3	2	2	3	2	2	1	2	1	10	1287	4	2	45	1	1	1
4	3	1	3	3	1	1	3	2	3	2	1	3	18	2515	3	4	43	1	1	1
4	3	1	1	2	1	1	3	2	1	2	1	1	6	672	1	4	54	1	1	1
4	3	4	3	1	1	1	3	2	3	1	1	3	6	1543	4	2	33	1	1	1
4	3	2	3	3	1	3	2	2	3	1	1	3	24	5511	4	1	25	1	1	1
4	5	1	3	3	1	3	3	2	3	1	1	7	12	701	4	2	32	2	1	1
4	4	2	3	3	1	1	3	2	2	1	1	10	6	1743	1	2	34	2	1	1
4	5	5	1	3	1	4	3	2	4	2	1	2	18	3229	2	4	38	1	1	1
4	5	5	5	3	2	1	3	2	3	2	1	4	12	2331	1	4	49	1	1	1

4	3	1	3	4	1	1	3	2	3	1	1	4	12	776	4	2	28	1	1	1
4	3	1	3	4	1	2	3	1	3	1	1	3	6	1766	1	2	21	1	1	1
4	3	1	4	3	1	3	3	2	3	2	1	10	12	1542	2	4	36	1	1	1
4	5	1	4	3	1	3	3	2	3	1	1	2	18	3850	3	1	27	2	1	1
4	3	5	3	3	1	1	3	2	3	1	1	1	36	3079	4	4	36	1	1	1
4	5	1	2	2	1	1	3	1	3	1	1	3	10	2146	1	3	23	2	1	1
4	4	1	1	4	1	1	3	2	3	1	1	4	18	2320	2	3	34	2	1	1
4	4	5	5	3	1	4	3	3	3	1	1	4	7	846	3	4	36	1	1	1
4	3	5	4	2	1	2	3	1	3	1	1	1	22	1283	4	4	25	1	1	1
4	5	1	3	2	1	3	3	2	3	1	1	4	18	2238	2	1	25	2	1	1
4	3	5	2	2	1	1	3	1	3	2	1	4	18	1126	4	2	21	1	1	1
4	3	1	4	2	1	3	3	2	3	1	1	10	9	1449	3	2	27	2	1	1
4	5	1	1	3	1	3	3	2	4	1	1	3	9	2406	2	3	31	1	1	1
4	3	1	2	4	1	1	3	2	3	2	1	4	18	1473	3	4	39	1	1	1
4	4	1	3	3	1	3	3	2	2	1	1	1	14	802	4	2	27	2	1	1
4	3	3	5	1	1	3	3	2	4	2	1	4	24	3235	3	2	26	1	1	1
4	5	1	4	3	1	3	3	2	4	2	1	2	33	7253	3	2	35	2	1	1
4	3	3	3	3	1	4	3	1	2	1	2	1	10	1597	3	2	40	1	2	1
4	5	1	5	3	1	2	3	2	3	2	1	2	36	5842	2	2	35	2	2	1
4	3	4	4	2	1	3	3	1	3	1	1	1	15	3186	2	3	20	1	1	1

4	5	1	5	3	1	3	3	2	3	1	1	4	28	2743	4	2	29	2	1	1
4	3	1	4	3	1	1	3	2	2	1	1	10	5	3448	1	4	74	1	1	1
4	5	1	3	1	1	1	3	2	3	2	1	6	24	2058	4	2	33	2	1	1
4	5	4	5	2	1	1	3	2	3	2	1	1	24	1287	4	4	37	2	1	1
4	3	1	5	3	1	3	3	2	4	2	1	1	12	1884	4	4	39	1	1	1
4	5	5	3	3	1	3	3	2	4	2	1	2	36	11054	4	2	30	1	1	1
4	1	1	5	3	1	2	3	2	3	1	1	3	12	2759	2	4	34	2	1	1
4	3	1	5	3	1	3	3	2	4	2	1	2	24	2670	4	4	35	1	1	1
4	3	1	4	3	1	3	2	2	3	2	1	10	18	1950	4	1	34	2	1	1
4	4	1	3	2	1	3	3	2	3	2	1	10	33	2764	2	2	26	2	1	1
4	3	3	3	3	1	3	3	2	3	1	1	2	39	2569	4	4	24	1	1	1
4	3	1	3	3	1	3	3	2	3	2	1	4	12	3077	2	4	52	1	1	1
4	3	1	5	2	2	4	3	3	3	2	1	1	24	3757	4	4	62	1	1	1
4	3	2	3	3	1	1	3	1	2	1	2	1	10	1418	3	2	35	1	1	1
4	3	1	3	3	3	2	3	1	3	1	1	1	6	3518	2	3	26	1	1	1
4	5	1	5	3	1	4	3	2	3	1	1	4	12	1934	2	2	26	2	1	1
4	5	2	3	2	1	2	3	2	3	1	1	4	6	1237	1	1	27	2	1	1
4	3	1	3	4	1	2	3	2	3	1	1	4	24	1413	4	2	28	1	1	1
4	3	2	5	3	1	3	3	2	4	2	1	2	39	8588	4	2	45	1	1	1
4	3	5	2	3	1	2	3	2	3	1	1	1	24	2255	4	1	54	1	1	1
4	5	1	5	3	1	4	3	3	3	1	1	6	24	5507	3	4	44	2	1	1

4	5	1	4	3	1	1	3	2	2	1	1	4	4	1503	2	1	42	2	2	1
4	3	2	4	4	1	2	3	2	3	2	1	4	24	1311	4	3	26	1	1	1
4	3	5	2	3	1	3	3	2	3	1	1	4	24	3105	4	2	25	2	1	1
4	4	1	5	3	1	4	3	3	3	2	1	1	24	2032	4	4	60	2	1	1
4	5	1	5	3	1	1	3	2	2	1	1	1	12	1255	4	4	61	2	1	1
4	5	1	3	2	1	3	3	2	3	2	1	1	24	2022	4	4	37	1	1	1
4	3	1	4	3	1	3	1	2	3	1	1	4	24	1552	3	1	32	1	2	1
4	5	5	4	2	1	1	3	2	3	1	1	6	12	996	4	4	23	2	1	1
4	5	4	5	3	1	3	3	2	3	1	1	4	24	2578	2	2	34	1	1	1
4	5	1	4	3	2	3	3	2	3	1	1	4	30	5954	3	2	38	1	1	1
4	3	5	1	2	1	2	3	1	4	2	1	2	24	5433	2	4	26	1	1	1
4	5	1	4	2	2	3	1	2	3	1	1	3	15	2788	2	3	24	2	1	1

4	5	5	3	2	1	3	3	2	3	2	1	7	24	1927	3	2	33	2	1	1
4	3	1	3	2	1	1	3	1	3	1	1	3	9	1388	4	2	26	1	1	1
4	3	5	5	2	1	4	3	3	4	2	1	1	12	2133	4	4	52	1	1	1
4	4	1	3	3	1	4	3	3	3	1	1	10	24	3863	1	2	32	1	1	1
4	5	1	3	3	1	1	3	2	4	2	1	10	24	4526	3	2	74	1	1	1
4	3	1	2	3	1	1	3	2	3	1	1	4	18	2051	4	1	33	1	1	1
4	3	5	5	3	1	4	1	3	3	1	1	2	15	1300	4	4	45	1	2	1
4	3	5	5	3	1	3	3	2	3	1	1	4	15	1979	4	2	35	1	1	1
4	5	3	3	3	1	1	3	2	3	2	1	3	42	4042	4	4	36	2	1	1
4	3	5	5	3	1	1	3	2	2	1	1	7	9	3832	1	4	64	1	1	1
4	3	1	5	3	1	2	1	2	3	2	1	7	12	1393	4	4	47	3	2	1
4	3	1	2	2	1	3	3	2	3	1	1	4	12	1493	4	3	34	1	2	1
4	5	1	4	3	1	3	3	2	3	2	1	2	24	2346	4	3	35	2	1	1
4	3	1	5	3	1	1	3	2	3	1	1	3	21	2241	4	2	50	2	1	1
4	2	1	3	2	1	1	2	2	3	1	1	10	48	3609	1	1	27	1	1	1
4	3	2	4	3	1	3	3	2	3	2	1	10	36	5742	2	2	31	2	1	1
4	3	1	5	3	1	2	3	2	4	2	1	1	60	10366	2	4	42	1	1	1
4	5	1	4	2	1	3	3	1	4	2	1	4	30	4530	4	4	26	1	1	1
4	3	1	2	2	1	4	3	2	3	1	1	7	6	1538	1	2	56	1	1	1
4	4	1	3	4	1	1	1	2	3	1	1	4	15	1478	4	3	33	2	1	1
4	3	5	3	3	1	4	3	3	3	2	1	1	60	6527	4	4	34	1	2	1
4	3	2	5	3	2	3	3	2	3	2	1	4	9	2753	3	4	35	1	1	1
4	3	3	3	3	1	2	3	1	3	1	1	2	6	1236	2	4	50	1	1	1
4	5	1	3	3	1	3	3	2	3	1	1	4	18	1800	4	2	24	2	1	1
4	3	3	5	2	1	4	3	3	3	1	1	1	15	5324	1	4	35	1	1	1
4	5	3	5	3	1	2	1	2	4	2	1	4	18	629	4	3	32	2	1	1
4	5	5	5	3	1	3	3	2	3	1	1	2	30	7596	1	4	63	2	1	1
4	5	5	5	3	1	3	3	2	3	2	1	4	30	3077	3	2	40	2	2	1
4	3	1	3	3	1	4	3	3	4	2	1	4	18	1505	4	2	32	1	1	1
4	5	4	3	2	1	1	1	2	2	1	1	1	6	250	2	2	41	2	1	1
4	5	3	3	4	1	2	3	2	2	1	1	4	15	1316	2	2	47	2	1	1
4	3	5	5	3	1	2	3	2	3	2	1	4	21	3160	4	3	41	1	1	1
4	3	5	3	3	1	3	3	2	3	1	1	2	21	5248	1	3	26	1	1	1

4	5	1	4	2	1	3	3	2	3	1	1	4	30	5771	4	2	25	2	1	1
4	3	5	2	2	1	1	3	1	3	1	1	2	10	2901	1	4	31	1	1	1
4	3	1	3	3	1	4	2	3	3	2	1	7	36	1819	4	4	37	1	1	0
4	3	5	2	2	1	3	3	2	3	2	1	4	24	5943	1	1	44	2	1	0
4	5	5	5	2	1	2	1	2	2	1	1	4	12	797	4	3	33	1	2	0
4	2	1	5	3	1	4	1	2	4	2	1	1	18	6458	2	4	39	2	2	0
4	5	1	3	2	1	1	2	2	3	2	1	1	36	7855	4	2	25	2	1	0
4	2	5	1	2	1	1	1	2	4	2	1	2	30	7485	4	1	53	1	1	0
4	4	1	2	1	1	3	3	2	3	1	1	10	36	9572	1	1	28	2	1	0
4	3	4	3	3	1	2	3	2	4	2	1	10	24	4591	2	3	54	3	1	0
4	3	1	4	3	1	3	3	2	3	1	1	4	9	1478	4	2	22	1	1	0
4	3	1	1	2	2	1	3	1	3	1	1	4	18	433	3	4	22	1	1	0
4	3	1	2	2	1	2	3	2	3	1	1	3	6	4611	1	4	32	1	1	0
4	5	3	3	3	1	4	1	3	3	1	1	1	48	10127	2	2	44	1	1	0
4	5	2	3	2	1	1	3	2	3	1	1	7	18	1864	4	2	30	2	1	0
4	5	2	3	2	1	3	1	1	2	1	1	2	48	11590	2	4	24	2	1	0
4	4	1	3	3	1	3	3	2	4	2	1	4	36	4463	4	2	26	2	1	0
4	4	1	2	3	1	3	1	2	3	2	1	10	24	8648	2	2	27	2	1	0
4	5	1	1	3	1	3	2	2	4	2	1	10	12	2292	4	2	42	2	1	0
4	3	1	5	3	1	3	1	2	2	1	1	7	12	719	4	4	41	1	2	0
4	3	5	3	3	3	2	3	2	2	1	1	1	10	1309	4	4	27	1	1	0
4	5	1	2	2	1	2	3	1	3	2	1	3	36	7127	2	4	23	2	1	0
4	1	5	3	2	1	2	1	2	3	2	1	1	21	5003	1	4	29	2	1	0
4	5	1	4	3	1	2	1	2	3	1	1	1	18	2775	2	2	31	2	1	0
4	3	1	3	2	1	3	3	2	4	2	1	3	36	3349	4	2	28	1	1	0
4	4	1	4	2	1	1	3	2	3	1	1	3	18	1808	4	1	22	1	1	0
4	5	1	3	3	1	2	1	2	3	1	1	1	18	1530	3	2	32	2	1	0
4	3	5	3	1	1	1	1	2	3	1	1	10	48	3914	4	2	38	1	1	0
4	3	1	2	2	2	3	3	1	3	1	1	3	9	1980	2	2	19	2	1	0
4	3	2	3	3	1	2	3	2	4	2	1	7	15	4623	3	2	40	1	1	0
4	4	5	2	3	1	3	3	1	3	2	1	10	36	7980	4	4	27	2	1	0
4	3	3	3	2	1	2	3	2	3	1	1	1	12	1386	2	2	26	1	1	0
4	4	3	2	3	1	1	1	2	2	1	1	10	21	2580	4	2	41	1	2	0

4	3	3	3	2	1	3	3	1	2	1	1	3	12	1123	4	4	29	1	1	0
4	3	1	3	2	1	1	3	1	3	1	1	1	27	2570	3	3	21	1	1	0
4	4	4	5	3	1	4	3	3	3	1	1	6	12	1555	4	4	55	2	2	0
4	5	5	5	3	1	4	3	3	4	2	1	1	21	12680	4	4	30	1	1	0
4	1	1	3	3	1	3	2	2	3	1	1	10	18	4165	2	2	36	2	2	0
4	3	1	2	2	1	1	3	2	3	1	1	6	18	1943	4	4	23	1	1	0
4	3	5	3	3	1	3	3	1	3	1	1	1	18	6761	2	4	68	2	1	0
4	4	1	3	4	1	3	3	2	3	2	1	10	18	2169	4	2	28	1	1	0
4	3	1	3	3	1	1	1	1	2	1	1	3	10	2210	2	2	25	1	1	0
4	3	1	2	4	2	2	3	2	2	1	1	3	18	1533	4	1	43	1	2	0
4	3	2	5	3	1	3	3	2	3	1	1	4	24	3621	2	4	31	2	1	0
4	3	1	1	2	1	3	3	2	4	2	1	3	36	10974	4	2	26	2	1	0
4	3	3	5	3	1	3	1	2	3	2	1	4	24	2397	3	2	35	2	1	0
4	4	1	5	3	1	3	3	2	2	1	1	1	24	2538	4	4	47	2	2	0
4	3	1	1	3	1	3	1	1	4	2	1	10	48	4844	3	2	33	1	1	0

CHAPTER-6

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Multivariate data analysis (Fifth Edition) --- Joseph F.Hair, RolphE.Anderson, Ronald l Tatham and William C.Black

2. Data Mining- Theories, Algorithms, and Examples – NoNG YE

3. A Practical Guide to Data Mining for Business and Industry -- Andrea Ahlemeyer-Stubbe, Shirley Coleman

4. Data Mining and Predictive Analytics – Daniel T. Larose, Chantal D.Lorse

5. Machine_learning_mastery_with_r. – Jason Brownlee

6. Master_machine_learning_algorithms -- Jason Brownlee

7. Statistical_methods_for_machine_learning - Jason Brownlee

8. Machine Learning Using R -- KarthikRamasubramanian ,Abhishek Singh

9. Data Science for Business - Forster Provost & Tom Fawcett

10. Deep learning with Deep learning R by François Chollet