Project Report

Prediction Of Consumer Purchase Decision Using Demographic Variables

With Special Reference to Premium SUV Cars

Submitted By

Ishant Kumar

2K18/MBA/910

UNIVERSITY SCHOOL OF MANAGEMENT

& ENTREPRENEURSHIP

DELHI TECHNOLOGICAL UNIVERSITY

DECLARATION

This to certify that I have completed the project titled "**Prediction of Consumer Purchase Decision Using Demographic Variables** (with Reference to Premium SUV Car)" , using Machine Learning Techniques under the guidance of **Dr Deepti Aggrawal** of Delhi Technological University .

The partial fulfillment of the requirement for the award of the degree of "Master in Business Administration" in Business Analytics from Delhi Technological University, Delhi. This is my original work and I have not submitted it earlier elsewhere**.**

Date: May 10th may, 2020

Name – Ishant Kumar

Roll No. - 2k18/MBA/910

Master of Business Administration – Business Analytics

ACKNOWLEDGEMENT

It is privilege to express my sincerest regards to project mentor **Dr Deepti Aggrawal(Assistant Professor)** of **Delhi Technological University** for her valuable inputs, able guidance, encouragement, whole-hearted cooperation and direction throughout the duration of this project.

Name – Ishant Kumar

Roll No. - 2k18/MBA/910

Master of Business Administration – Business Analytics

Table of Contents

Chapter 1

INTRODUCTION

Demography is the investigation of the essential and statistical insights of a populace The segment factors are age, sex, salary and so forth in this undertaking 3 segment factors are thought about . Demographics are utilized to portray a populace , dispersion and structure. Size techniques the amount of individuals in a people while structure depicts the masses the extent that age, pay, and their surveyed remuneration. Every one of these components impacts the conduct of purchaser and adds to the general interest for different items and administrations . Shopper conduct is worried about the investigation of elements that impact people‟s conduct in a purchasing circumstance. can settle on better showcasing choices exactly when they know why and how individuals choose their usage decisions. In the event that publicists can comprehend the conduct of buyers absolutely, they could anticipate how clients are presumably going to react to various instructive and environmental points and could shape their displaying frameworks so they can achieve an uncommon advantage in the market.

Chapter 2

ABSTRACT

The demographic environment is of significant enthusiasm to advertisers since it includes individuals and individuals make up showcase. Discontinuity of the mass market into different scaled down scale markets isolated by age, sex and Salary, etc. Since each gathering has solid inclinations and purchaser qualities that can be handily reached through progressively focused on correspondence and appropriation channels. The vast majority of advertisers taking vital choice intensely rely upon the segment factors of individuals, where they center around showcasing their items. The relationship set up between the segment factors and the various phases of purchaser's buy choice procedure further aides distinguishing the noteworthy factors. This will be certainly useful to the advertisers of vehicles to realize their objective gathering and to develop showcasing systems to make them turning into a suv vehicle car owner with the help of classification techniques for predicting weather consumer is going to buy this premium suv or Not based on this decision of consumers company can take decisions weather they should send the customer the promotional offer or not .

Chapter 3

OBJECTIVE

- Firstly ultimate goal of this project is to classify right users into right categories that is trying to building a classifier that will catch the right users into right categories which are yes they bought the suv and no they haven’t bought the suv.

- Secondly to check weather The demographic variables have much significance in marketing or not .

- Thirdly that is In this project machine learning algorithms are used . Logistic regression ,k nearest neighbours ,naïve Bayes and decision tree a part of model building. In this dataset the dependent variable is a categorical variable. And classifiers logistic regression ,k nearest neighbours ,naïve bayes and decision tree where the dependent variable is a dichotomous variable like ability to buy suv car (yes or no decision). The classification value will be either 0 or 1, showing the reluctance by 0 and the ability by 1. The model fit is tested using the „percent correct prediction‟ and these four models are compared using perfomance parameters which includes accuracy , precision and recall . The more the number of correct predictions, the better the results of performance parameters

- Fourth Which classifier outperforms the others classifiers that should be used in arena of marketing to predict more correct predictions that will improve futher sales as ultimate objective of the company or enterprise is to Reduce the cost and Increase the revenue .

Chapter 4

CONCEPTUAL FRAMEWORK

Like mental and social factors, segment factors are additionally considered as the individual purchasing choice factors. demography as the investigation of populace attributes. These qualities depict individuals – who are there, where they live, and where they are moving. socioeconomics are effectively recognizable and quantifiable insights that are utilized to depict the populace. Patterns in populace size demonstrate future potential and therefore, impact showcase plans. The size of the masses shows the potential market enthusiasm for purchaser things and organizations. demography is a huge premium condition, helping the sponsor to foresee both size and change in target markets. These measurements help to clarify buyer ways of life; the manners in which the individuals live. By getting shoppers, a firm can decide the most proper crowd whom to claim and the mix of an advertising factors that will fulfill this crowd. In this way, it is basic for advertisers to utilize segment information related to and part of social, mental and purchaser decision making examinations. The segment data assists with finding an objective market whose intentions and conduct would then be able to be clarified and anticipated utilizing mental or social examination. Segment data distinguishes potential for deals and utilization of item in spite of the fact that it doesn't recognize why or by whom a particular brand is used. It is additionally moderately available and financially savvy to assemble. markets.

Further, the segment factors uncover patterns pertinent to advertisers, for example, moves in age and pay disseminations, and so forth. They can build up purchaser profiles that may introduce alluring business sector openings. These are the causes why advertisers, in developing numbers, are utilizing segment insights for creating showcasing procedures and projects. Buyer conduct isn't simply settling on a buy choice or the demonstration of buying; it incorporates the full scope of encounters related with utilizing or devouring items and administrations. It likewise incorporates a feeling of joy and fulfilment got from having or gathering things . The yields of utilization are changes in emotions, states of mind, or perspectives; support in ways of life; an upgraded feeling of self; fulfilment of a buyer related need; having a place with gatherings; and communicating and engaging oneself. A consumer's choice to purchase or not to purchase an item is a significant instinct for most advertisers. It can signify whether a promoting system has been shrewd, insightful, and effective, or whether it was misguided and come up short. Consequently, advertisers are especially intrigued by the consumer's dynamic procedure. For a shopper to settle on a choice, more than one option must be accessible. In executing a buy aim, the buyer may settle on up to five buy sub choices: a brand choice, merchant choice, amount choice, timing choice and instalment – technique choice (Kotler, Philip).

Chapter 5

LITERATURE REVIEW

In the wake of looking and evaluating, the chief outcome will be the decision to purchase or not to purchase the choice surveyed as commonly appealing. In case the decision is to buy, a movement of related decisions must be made concerning the features, where and when to make the genuine transaction, how to take proprietorship, the technique for portion and various issues A customer's choice to adjust, defer or stay away from a buy choice is vigorously impacted by seen risk. The proportion of evident hazard contrasts with the level of money in stake, the proportion of uncertainity and the proportion of purchaser's valor. Choosing a source from which to make a buy is one of the purchasing choices. The measure of apparent hazard differs with the degree of cash in question, the measure of characteristic vulnerability and the measure of consumer's fearlessness. Advertisers must comprehend the components that incite a sentiment of hazard in the purchaser and give data and backing to lessen the apparent hazard (Kotler, 2003). Peterson, Balasubramanian and Bronnenberg (1997) conjecture that right off the bat in the twenty first century customers will buy nourishment and other fundamental family unit needs through in-home TV PC frameworks. The customer will settle on decisions in the wake of review brands and costs on the screen. Thus, the buying procedure itself may change significantly in the coming decades. Lilien, Kotler and Moorthy (1999) uncover that a consumer's buy expectation is impacted by changes in foreseen situational factors. The shopper frames a buy expectation based on such factors true to form family salary, the normal all out expense of the item and the normal advantages of the item. Besides, when the customer is going to act, unforeseen situational elements may intercede him from so, (for example, absence of accessibility of a favored item). Subsequently, inclinations and buy goals are not totally dependable indicators of genuine purchasing conduct: while they manage buy conduct, they neglect to incorporate some of extra factors they may mediate. Demographics are wild factors in the outside condition. The explanation behind any market is people. Consequently, contemplating the populace as far as its segment structure is critical for promoting chiefs.

In a genuine situation, when we are given an issue we cannot predict which algorithm perform best. Clearly from the issue, we can advise whether we have to apply Regression or classification. Be that as it may, it is hard to tell which Regression or classification algorithm to apply in advance. It is just through experimentation and checking the performance metrix , we can limit and pick certain algorithms.

In this Undertaking will tell the best way to compare different classifications algorithms and pick the best ones. Instead of actualizing the each of the 7 characterizations algorithms we will utilize 4 most commonly utilized calculations for our Prediction Task and afterward discovering that the presentation of all the model with their particular Calculations , we will play out every one of the 4 algorithms bit by bit and informational index for all the calculations will stay normal first check the exhibition of algorithms and afterward choose which one is best utilizing confusion matrix.

Chapter 6

Research Methodology:

This examination is unmistakable in nature . The explanation behind this task is to predict the purchase decision of a customer to buy a vehicle when his/her precise portion profile is known for this reason, optional information have been gathered of 400 individuals alongside segment factors which incorporates evaluated compensatio, age and sex from information taken from datascience .com. In this project machine learning algorithms are used . Logistic regression ,k nearest neighbours ,naïve bayes and decision tree a part of model building. In this dataset the dependent variable is a categorical variable. In this project logistic regression ,k nearest neighbours ,naïve bayes and decision tree where the reliant variable is a dichotomous variable like capacity to purchase suv vehicle (yes or no choice). The characterization worth will be either 0 or 1, indicating the hesitance by 0 and the capacity by 1. The model fit is tried utilizing the „percent right prediction‟ and these four models are compared using performance parameters which includes accuracy , precision and recall . The more the number of correct predictions, the better the results of performance parameters .and the classifier with highest values of performance parameters will be chosen to choose whether the company should send the promotional offer to the customer or not . R studio has been used for analysing the data and for predicting the purchase decision of consumers.

Chapter 7

INTRODUCTION TO DATA SET

7.1 Data Set

Informational index taken from superdata science .com . The motivation behind this investigation is to foresee the buy choice of a shopper to purchase a vehicle when his/her precise segment profile is known. For this reason, essential information have been gathered from 400 people

In the dataset there are 5 factors in which 4 factors are free and 1 factors is reliant ie userid ,Sex ,Age and Assessed pay of a client are autonomous factors and Bought is a needy variable The worth will be either 0 or 1, showing the reluctance by 0 and the eagerness by 1 . usidid won't help our model so it will be expelled from our informational collection in the underlying stage and in the later phase of this task we may discover the Sexual orientation is likewise not contributing in Anticipating purchasing choice of Shopper with the assistance of Strategic Regression that implies Sex of a buyer is certifiably not a noteworthy variable as Sex has higher P esteem when contrasted with Assessed Compensation AND Age

```
User.ID Gender Age EstimatedSalary Purchased
```

In this Data set User id will be removed in the initial stage and Estimated Salary and Age of the person are two significant variables in this data set in predictin the buying behaviour of the consumer.

7.2 Histogram of Estimated Salary

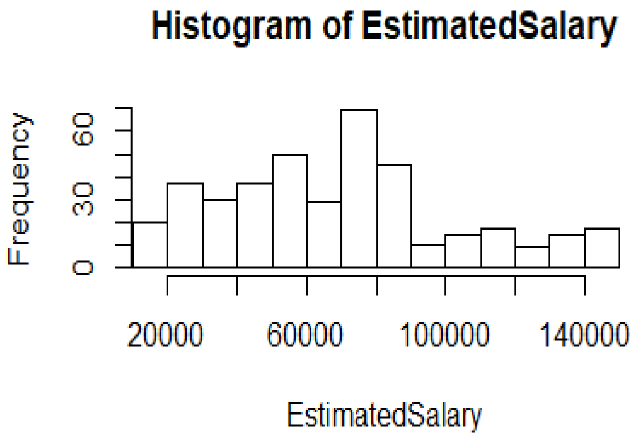


Figure 7.1

The following interpretations are made from the figure 7.1:

- The estimated salary of a customer is in Range of $15,000 to $150,000 with mean of $69,473 and median of $70000.
- Mean of estimated salary of a person is $ 69743.

7.3 Box plot of Estimated Salary

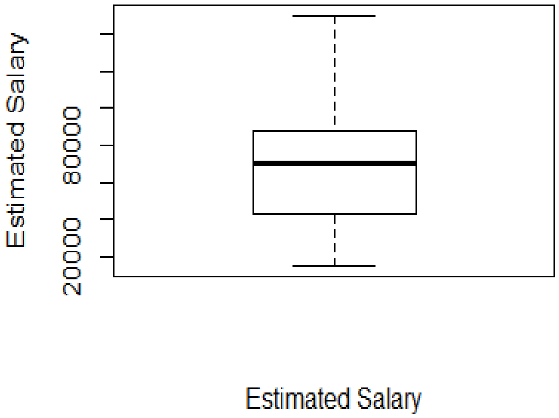


Figure 7.2

Box plot are acclaimed for distinguishing the nearness of outliers in dataset . in figure 7.2 By looking as the Box plot it can be easy suspect that there is no Outlier in In estimated salary of consumer in the data , can leverage this thing in our modelling

Similarly With the Age , In customers variable in the Data are in Range of 18 years to 60 years with mean age of customers is 37.6 years students with 18 years of age are also the owners of this premium Suv car and people with 60 years of age are the owners of this premium Suv it can be infer that people of all age groups the buying this suv then it must be a very cool Sports utility vehicle

Last Dependent variable column is a binary column means customers are divided into Owners and Non owners , proprietors are denoted by 1 means they have bought this Luxurious suv and non proprietors are signified by 0 methods haven't purchased this luxurious vehicle.

7.4 Proportion of Owners and Non owners

The proportion of the Owners and Nonowners are reprented with help of 3D Pie Chart

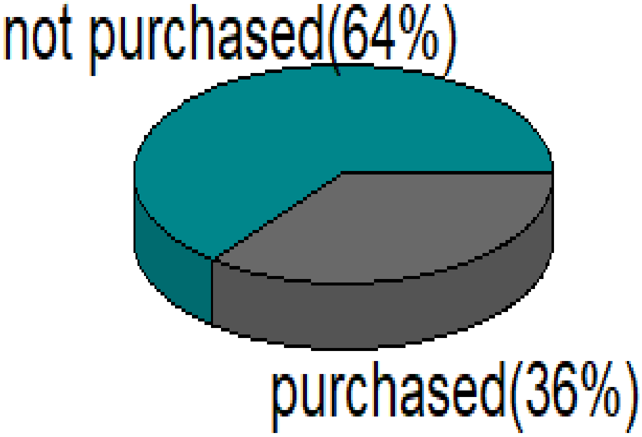
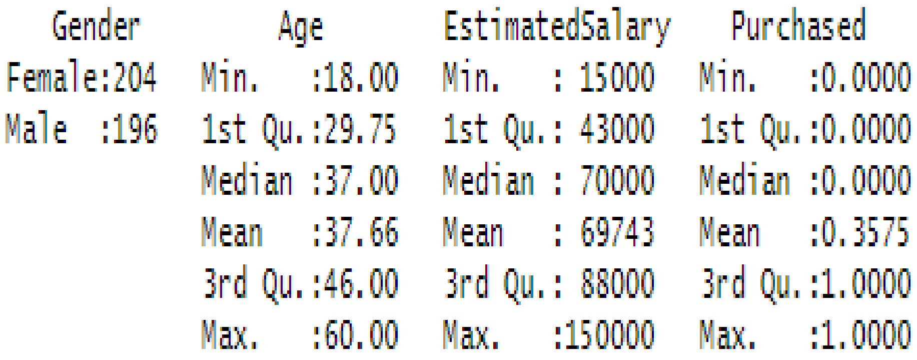


Figure 7.3

As represented in the Pie chart it can be infer people who have bout the SUV are 36% and 64% of people havnt bought the suv below summary of the Data set is given that means out of 400 people as whole only 144 persons are those who can afford this premium suv car and rest of 264 people are those who don't have enough money based on their estimated salary and age to buy this premium suv car .

7.5 Basic statistical measures of the attributes



```
   Gender         Age        EstimatedSalary    Purchased
 Female:204   Min.   :18.00   Min.   : 15000   Min.   :0.0000
 Male  :196   1st Qu.:29.75   1st Qu.: 43000   1st Qu.:0.0000
              Median :37.00   Median : 70000   Median :0.0000
              Mean   :37.66   Mean   : 69743   Mean   :0.3575
              3rd Qu.:46.00   3rd Qu.: 88000   3rd Qu.:1.0000
              Max.   :60.00   Max.   :150000   Max.   :1.0000
```

Figure 7.4

7.5.1 The accompanying understandings are produced using the above table are :

- The is no missing value.
- There are 204 females and 196 males in data set.
- The min age of person under observation is 18 years and Maximum age of person is 60.
- second quarter being the median that is not influenced by the higher and lower values like mean with value – 37 years
- Mean age of person under this experiment is 37.66
- As mean is more than median it can be inferred that It is right skewed in case of Age
- The min salary of person under observation is $15000 and Maximum age of person is $150000.
- 2 st qu - is second quarter being the median that is not influenced by the higher and lower values like mean with value – $70000.
- Mean salary of person under this experiment is $69743.
- As mean is less than median it can be inferred that It is left skewed in case of Estimated salary.

7.5.2 Scatter plot between Estimated salary and Age

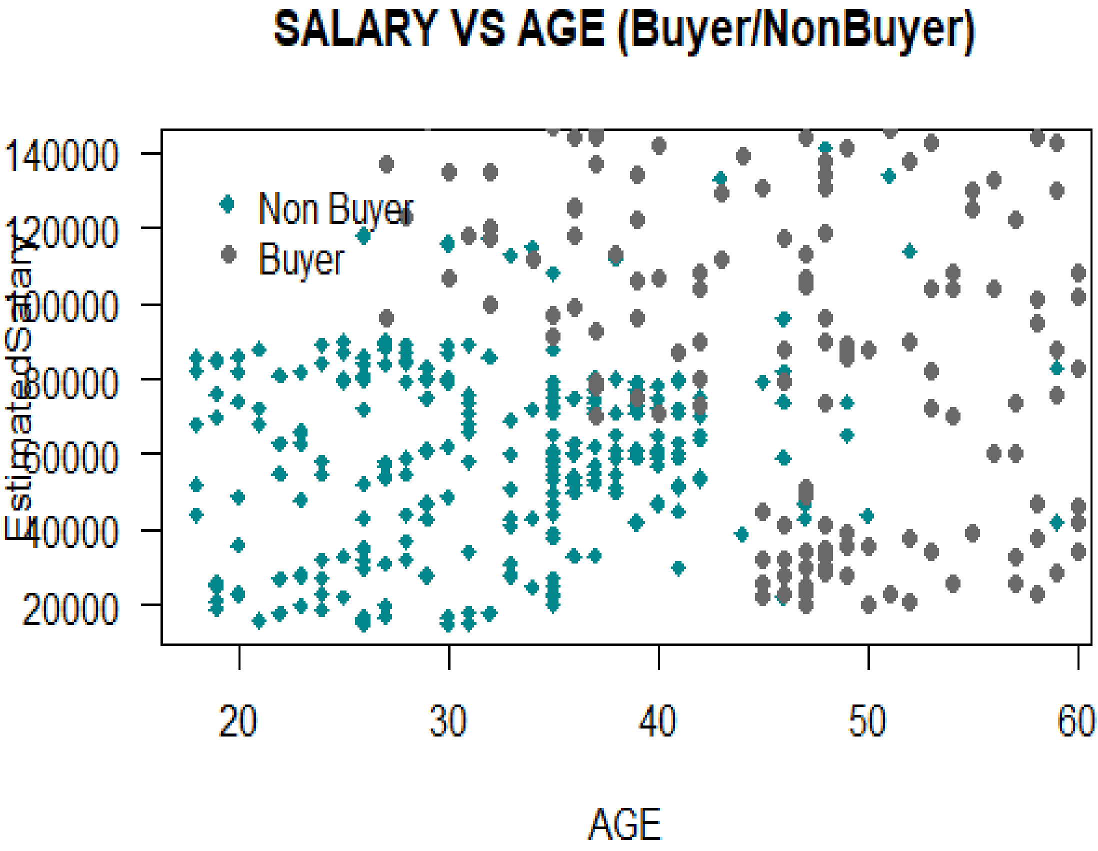


Figure 7.5

A scatter plot, scatter graph, and correlation chart are other names for a scatter diagram.

The accompanying understandings are produced using the above dissipate plot are:

- This plot can be utilized in a manner to discover the relationship between's these two factors evaluated compensation on the Y axis and age of the client is on the X axis
- Dependent variable an individual purchases the suv or not are joined with various different colours

dim grey for the clients who have purchased the suv and blue shading who didn't purchased the suv

- It can be inferred that with the increase in age and estimated salary of an individual they are coming in umbrella of Luxurious suv shopper implies people with high evaluated pay and higher age are more pulled in towards purchasing this superior suv and the individuals with low evaluated pay and age are not buying this suv might be a direct result of the way that this suv is exorbitant and it is out of their spending limit

From the below boxplot Between Age and Purchase decision of person it can be interpreted that the mean age of the persons who have not purchased the suv is lower the the mean age of person who have bought the suv

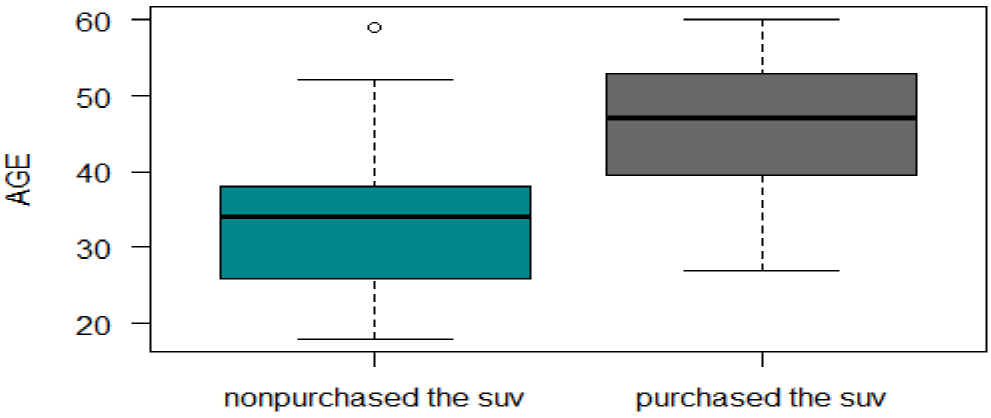


Figure 7.6

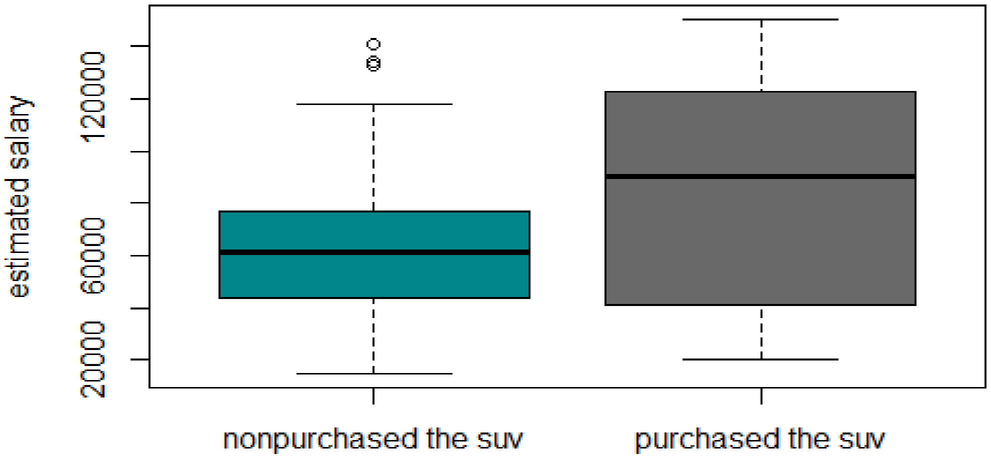


Figure 7.7

In the above box plot of data between the Buying decision of suv and estimated salary of a person it is visible that the person who have bought the suv having mean salary higher than the ones who havnt bought the suv .

The commitments for this work in the coming pages can be summed up as follows:

- By feature the significance of AI classifiers in demonstrating and foreseeing person's buy choice for personalised context .
- By doing modelling and predicting personalized decision with their demographic varibles in multi-dimensional contexts using various machine learning classification techniques .
- What's more, by leading a scope of examinations on the purchasing decison for premium suv vehicles of individual to assess the adequacy of every classifier put together forecast model with respect to concealed logical experiments.

What's more, by leading a scope of examinations on the purchasing decison for premium suv vehicles of individual to assess the adequacy of every classifier put together forecast model with respect to concealed logical experiments.

Chapter 8

Machine Learning Classifiers

Classification is one of the most notable machine learning method to anticipate the class of new examples, utilizing a model gathered from preparing information. At the point when everything is stated, arrangement is characterized as a learning method that maps and requests data events into the relating class names that are predefined in the given dataset.

Information grouping is a two-advance procedure; initial one is the learning step where a characterization model is model is created from a given dataset; the information from which an order capacity or model is found out is known as the preparation set, and second one is an arrangement step where the model is utilized to test or anticipate the class labels or a separate unseen given data; the data set that is used to test the classifying ability of the learned model or function is known as the testing set. In order to predict the buying behaviour of consumers different classification techniques for various context we for the most part centre and sum up various notable and most famous arrangement methods applicable to our objective clients, in the accompanying subsections.

Selecting an Algorithm

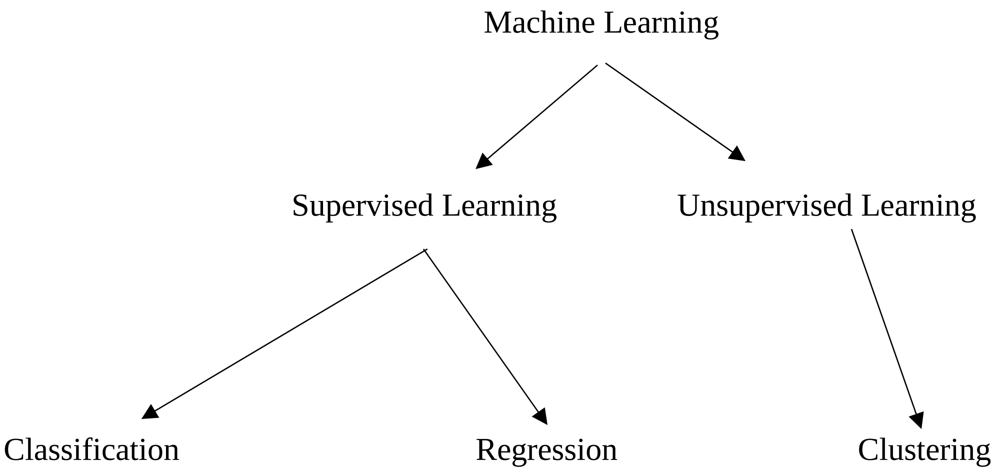


Figure 8.1

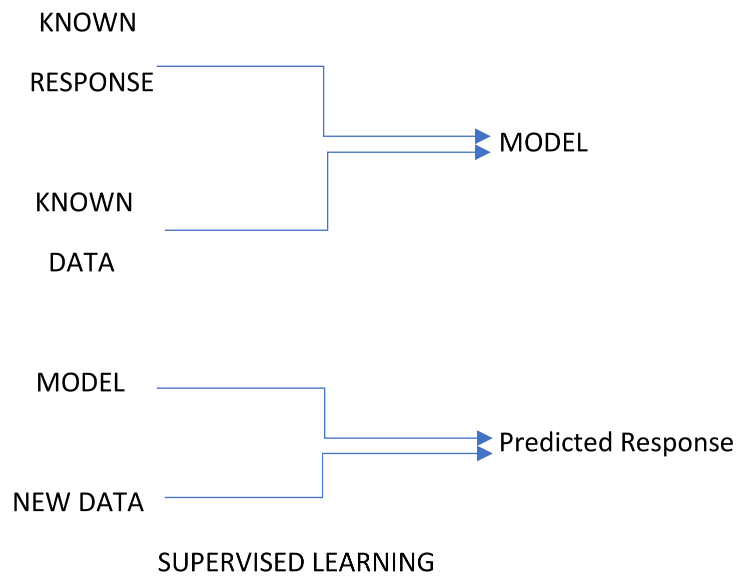


Figure 8.2

7.1 Logistic Regression

Logistic regression is the right calculation to begin with classifications algorithms. Eventhough, the name 'regressor' comes up, it's definitely not a regression model, however a characterization model. It uses a logistic function to frame binary output model. The yield of the logistic regression will be a likelihood (0≤x≤1), and can be utilized to anticipate the double 0 or 1 as the yield (if x<0.5, output= 0, else output=1).

Logistic regression acts to some degree fundamentally the same as linear regression. It additionally accurately ascertains the linear output, followed by a stashing function over the regression output. Sigmoid function is the frequently used logistic function.

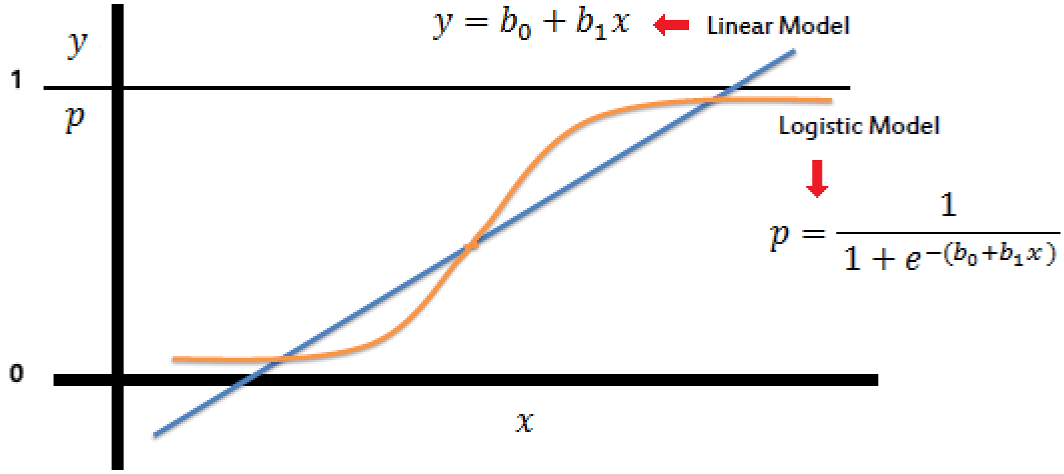


Figure 8.3

The h(θ) value here corresponds to p(y=1|x), ie, probability of output to be binary 1, given input x. P(y=0|x) will be equal to 1-h(θ).

When value of z is 0, g(z) will be 0.5. Whenever z is positive, h(θ) will be more prominent than 0.5 and output will be binary 1. Similarly, at whatever point z is negative, value of y will be 0. As we use a linear equation to find the classifier, the output model also will be a linear one, that means it splits the input dimension into two spaces with all points in one space corresponds to same label.

Advantages **:**

- Easy, quick and very simple classification method.
- Can be utilized for multiclass classifications also.

Disadvantages **:**

- Cannot be applied on non-linear classification issues.
- Proper choice of features is required.
- Good signal to noise ratio is expected.
- Collinearity and outliers tampers the accuracy of Logistic Regression Classifier (LR) model.

7.1.1 Logistic Regression vs Decision Tree :

- Decision tree handles collinearity better than LR.
- Decision trees cannot infer the worthiness of features, but LR can.
- Decision trees are preferable for categorical values than LR.

7.1.2 Logistic Regression vs Naive Bayes :

- Naive bayes is a generative model whereas LR is a discriminative model.
- Naive bayes functions admirably with small datasets, whereas LR+regularization can achieve similar performance.
- LR performs superior as compared to naive bayes upon collinearity, as naive bayes expects all features to be autonomous..

7.1.3 Logistic Regression vs KNN :

- KNN is a non-parametric model, where LR is just a parametric model.
- KNN is comparatively much more slower than Logistic Regression.
- KNN supports non-direct(linear) solutions where LR bolsters just linear solutions.
- LR can infer confidence level (about its prediction), whereas KNN can just output the labels.

7.2 K- Nearest Neighbours KNN

Another characterization calculation, K-nearest neighbors (KNN) is one of the most clear order strategy in machine learning. It is kind of occurance based learning , or lazy learning. In this classification technique, it takes into account neighbourhood guess and all the calculation is deferred until classification. It stores all the accessible cases in the given dataset and characterizes new cases reliant on separation capacities like Euclidean partition as a resemblance measure.

$$D(a,b)=\sqrt{\sum_{i=1}^{n}(b_i-a_i)^2}$$

After that the K most practically identical events, called the neighbours, are controlled by means of looking for through the entire getting ready set for another test data point. Finally the craving result is made by condensing the yield variable for those K cases subject to bigger part throwing a greater part casting a ballot of the neighbours. Various specialists use KNN classifier in foreseeing the shopper purchasing conduct.

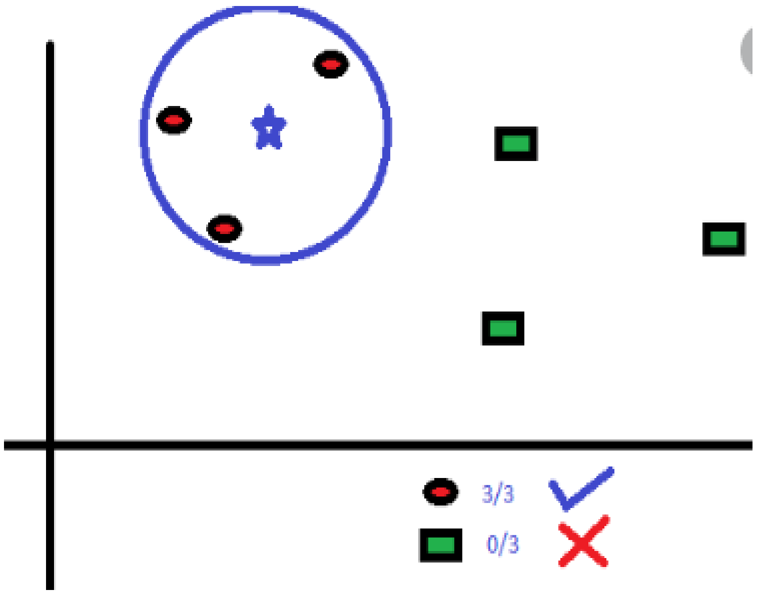


Figure 8.4

Advantages :

- Easy and simple machine learning model.

Disadvantages :

- k should be wisely selected.
- Large calculation cost during runtime if test size is large

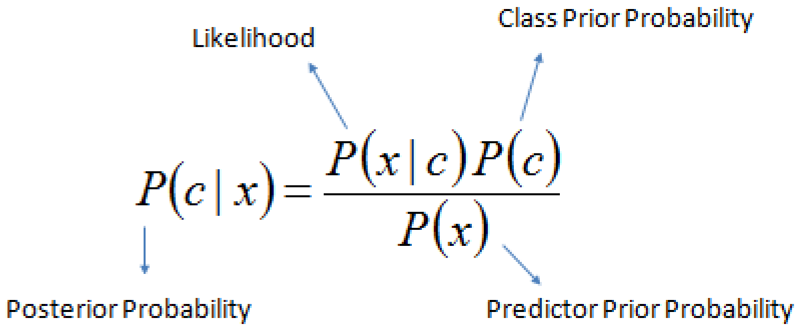
- Proper scaling ought to be accommodated reasonable treatment among highlights..

8.2.1 KNN vs naive bayes :

- Naive bayes is much quicker than KNN due to KNN's real-time execution.
- Naive bayes is parametric though KNN is non-parametric.

8.3 Naïve Bayes

A Naive Bayes classifier is a fundamental probabilistic based method, which can predict the class enrolment probabilities . As it figures the probability of enrolment for each class, it likewise can without much of a stretch handle the missing property estimations by basically precluding the relating probabilities for those characteristics. In a Naive Bayes classifier, the impact of a quality on a given class is additionally free of those of different traits, which is known as class contingent autonomy. This classifier effectively computes the likelihood to group or anticipate the class in a given dataset.. Although, it may cause the zero probability problem, it can play an important role while predicting Premium Suv Car owner classes in demographic contexts . Naive Bayes classifier fulfill the purpose of intelligently handling Owner interruptions. Naive Bayes classifier in their Prediction techniques


Likelihood
Class Prior Probability
$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$.

Naïve Bayesian classifiers assume that there are no dependencies amongst attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, hence is called "naive". This classifier is also called simple Bayes, or independent Bayes.

The Advantages of Naive Bayes are:

- It uses a very intuitive technique. Bayes classifiers, unlike neural networks, do not have several free parameters that must be set. This greatly simplifies the design process.
- Since the classifier returns probabilities, it is more straightforward to apply these outcomes to a wide assortment of undertakings than if a subjective scale was utilized.

- It doesn't require a lot of data before learning can start .
- Naive Bayes classifiers are computationally fast when making decision **.**

8.4 Decision Tree

Decision tree is a tree based calculation regularly used to tackle regression and classification problems. An altered tree is shaped which is fan out from a heterogeneous likelihood dispersed root hub, to profoundly homogeneous leaf hubs, for inferring the yield. Regression trees are utilized for dependent variable with continuous values and classification trees are used for dependent variable with discrete values. Basic Theory :Decision tree is derived from independent variables, with each node having a condition over a feature.The nodes decides which node to navigate next based on the condition. The moment leaf node is reached, an output is predicted. The correct arrangement of conditions makes the tree efficient. entropy/Information gain are used as the criteria to choose the conditions in nodes.

Points of interest :

- No preprocessing required on information.
- No assumptions on distribution of data.
- Handles collinearity proficiently..
- Decision trees can give justifiable clarification over the expectation.

Hindrances : :

- Chances for overfitting the model on the off chance that we continue assembling the tree to accomplish high immaculateness. decision tree pruning will be helpful to unravel this issue.
- Inclined to anomalies(outliers).
- Tree may create to be incredibly while preparing confounded datasets
- Looses important data while taking care of persistent(continuous) factors.

8.4.1 Decision tree versus KNN :

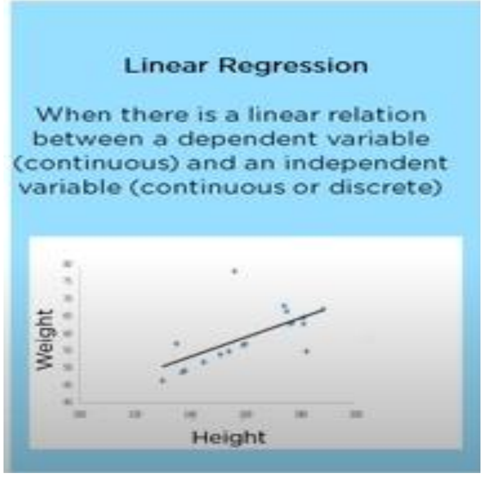
- Both are non-parametric methods.
- Decision tree sup ports automatic feature interaction, whereas KNN cant.
- Decision tree is quicker because of KNN's costly continuous execution**.**

8.4.2 Decision tree vs naive Bayes :

- Decision tree is a discriminative model, whereas Naive bayes is a generative model.
- Decision trees are progressively adaptable and simple.
- Decision tree pruning may disregard some key qualities in preparing information, which can lead the precision for a hurl.

A very notable and most part known for classification and then used for prediction is decision trees . The center calculation for building decision trees called ID3 proposed by Quinlan. ID3 count constructs a decision tree by using a top-down philosophy wherein a greedy looking through the given preparing dataset is utilized to test each quality or setting at each attribute. It computes the entropy and data gain which is a attribute that is utilized to choose which credit to test at every node in the tree . In light of the ID3 calculation, an expansion in particular C4.5 calculation is proposed by Quinlan . C4.5 assembles choice trees from a preparation dataset in the comparative method as ID3 This calculation produces relevant decision rules to anticipate action.

8.5 When to use Logistic Regression

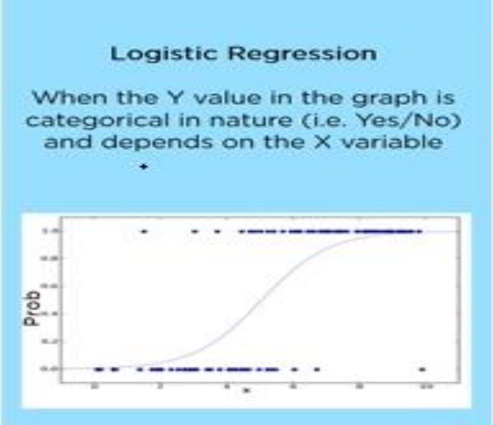


Linear Regression

When there is a linear relation between a dependent variable (continuous) and an independent variable (continuous or discrete)

Figure 8.5

When there is a straight connection among reliant(dependent) and autonomous(independent) variable (continuous) and an independent variable (continuous or discrete)



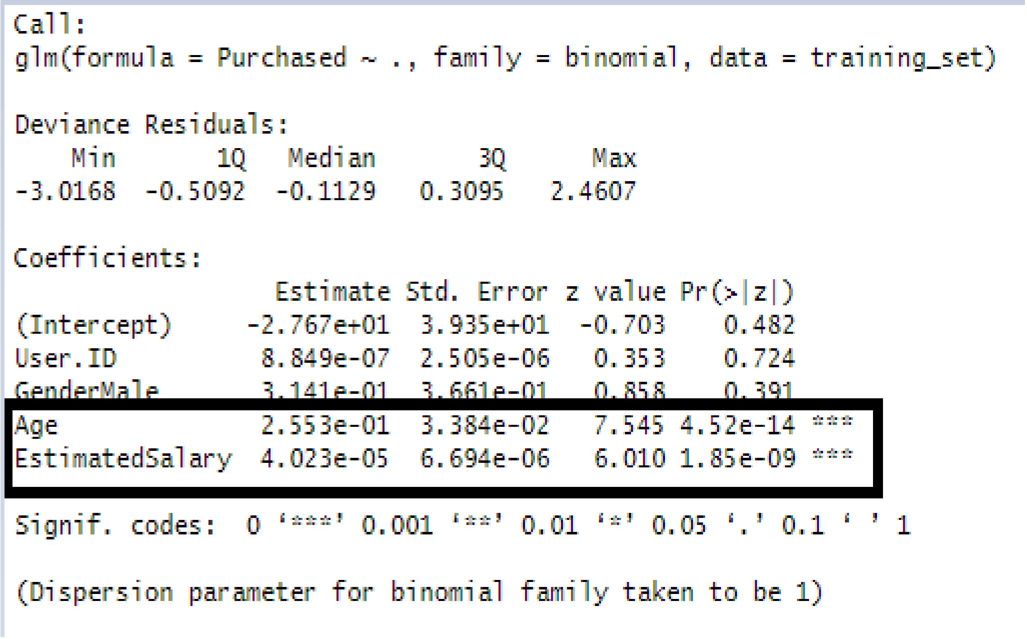
Logistic Regression

When the Y value in the graph is categorical in nature (i.e. Yes/No) and depends on the X variable

Figure 8.6

When there is a no straight connection(linear relation) among reliant and autonomous variable that is when the Y value in the graph is categorical in nature ((ie YES/NO) as in this case Purchased =1 and Not purchased =0)) and depends upon the independent variables(that is age and estimated salary of a person)

Following are the summary results when logistic regression classifier fit into the data set and in the summary function in our R programming language it can be seen that out of three independent variables that is Person's Age ,Gender and Estimated Salary) two variables age and estimated salaries are significant variable's



```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0168  -0.5092  -0.1129   0.3095   2.4607

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.767e+01  3.935e+01  -0.703    0.482
User.ID         8.849e-07  2.505e-06   0.353    0.724
GenderMale      3.141e-01  3.661e-01   0.858    0.391
Age             2.553e-01  3.384e-02   7.545 4.52e-14 ***
EstimatedSalary 4.023e-05  6.694e-06   6.010 1.85e-09 ***

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

(Dispersion parameter for binomial family taken to be 1)
```

Figure 8.7

Chapter 9

Prediction Using Classification Algorithms

9.1 Logistic Regression

The following is the consequences of logistic regression classifier on training set demonstrated our information of 400 observations and separated into 300 perceptions (observations) of training set and 100 perceptions of test set (observations of test set) . 300 observation using function set seed(123) are choosen to learn the model from training set observation and afterward this learning is applied on anticipating the 100 perception of test set

So model will be trained on traing set with 300 observations and then later the models performance will be tested on test set with 100 observations however following is the equation of the classifier applied on training set to learn and further used this model to predict the test set observations utilizing R studio **.**

Classifier = glm(formula = purchased ~.,family = binomial, data=train)

9.1.1 Prediction of Logistic Regression Classifier on Training set of 300 observations:

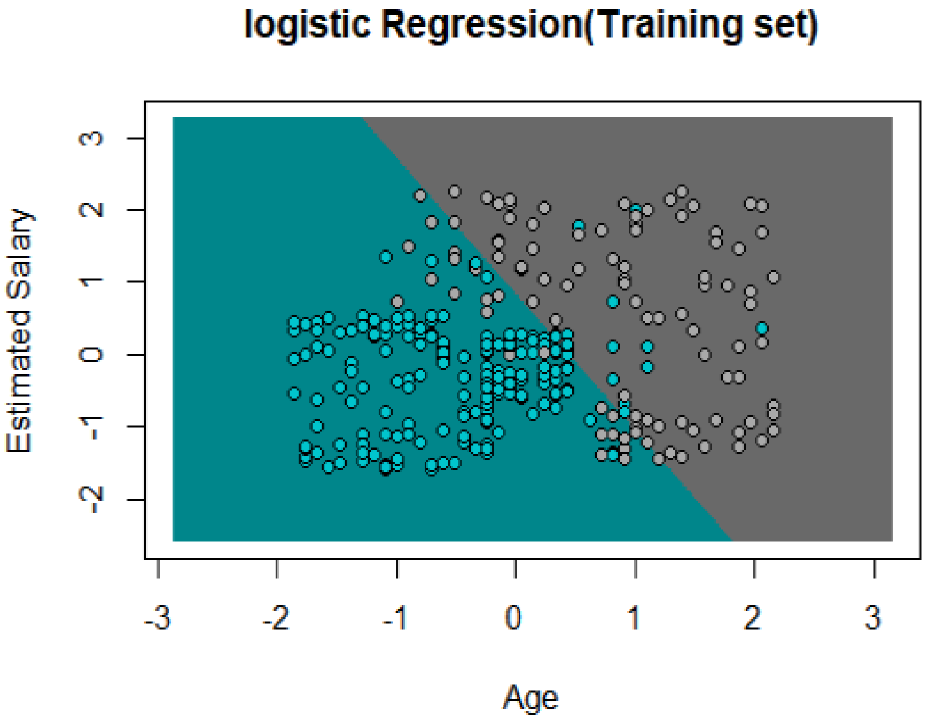


Figure 9.1

9.1.2 Visualize the training set result :

Here we have some blue and grey points so all of these points which are present on the above graph are observation points of training set that is all the users that were selected to go to training set and every one of this client is portrayed by age that is on x hub and evaluated pay that is on y hub .

Presently there are some blue points and some grey points the blue points are training set observations for which dependent variables purchased = 0. The dim grey for which subordinate variable bought = 1. That suggests blue focuses are the clients who didn't bought the SUV and grey points are the customers who bought the indulgence SUV.

Directly the clients who are youthful with low evaluated pay really didn't purchased the suv and the clients who are more seasoned(older) with high assessed pay the majority of them have purchased the sumptuous SUV.

Well that really bode well as SUV is progressively similar to a family vehicle and in this way more interesting for the more seasoned clients with high evaluated pay.

Besides some older people even with lower estimated salary actually bought the SUV those grey points that corresponds to an age above the average age . These older guys although have lower estimated salary bought this luxurious suv probably they were saving up some money or may be they finished paying off their mortgage ,yet thing is they couldn't avoid purchasing the cool lavish suv offered at significant expense.

There are additionally a few youths with higher assessed pay have really purchased the suv might be its cool SUV and presumably they need to dazzle their companion or might be they are pulled in towards the suv.

NOW GOAL OF THIS CLASSIFICATION

Well objective here is to arrange right clients into right classifications that is attempting to building a classifier that will get the correct clients into right classes which are yes they purchased the suv and no they haven't purchased the suv.. And the classifier catches the users by plotting which are often known as prediction regions .

Prediction regions are two regions that are shown on the graph in blue an grey , blue prediction region is the region where classifier catches all the users that don't buy the suv and grey prediction region is a region where classifier catches all the user that buy the suv. But this is according to the classifier that for each user of this blue region the logistic regression classifier predicts that user don't buy the suv and for each user of the grey region classifier predicts that user buy the suv that whats the classifier predicts.

Point = true user , Region = prediction

So for each user logistic regression classifier will tell based on the person's age and estimated salary if the user belongs to blue prediction region therefor doesn't buy the suv and if the user belongs to the grey prediction region therefore buy the suv so in this way business car company can substantially optimize their marketing campaign by targeting the social network ads to the user in the grey region because these are the user that are predicted to buy the suv according to the classifier as there are the two predicted regions separated by a straight line is called prediction boundary ,

because it's the boundary between two predicted regions and the fact it is a straight line is not random it is for a particular reason that is that is the essence of linear regression if the predicted boundary is the straight line that's because logistic regression classifier is a linear classifier.

Logistic regression classifier manages to catch most of the users who didn't bought the suv in blue region and most of the users who bought the suv are in grey region hence we can say the logistic regression classifier did a pretty good job.

However it seems to have problem in catching some of the grey users who in spite of the lower estimated salary bought the luxurious suv as well as the grey user who also bought the suv as these points are in the region where our classifier predicts that user don't nuy the suv and those incorrect predictions are specifically due to the fact that classifier is a linear classifier and the users are not linearly distributed if they all are linearly distributed then we will have all the grey points in the predicted region then linear classifier would perfectly separate them

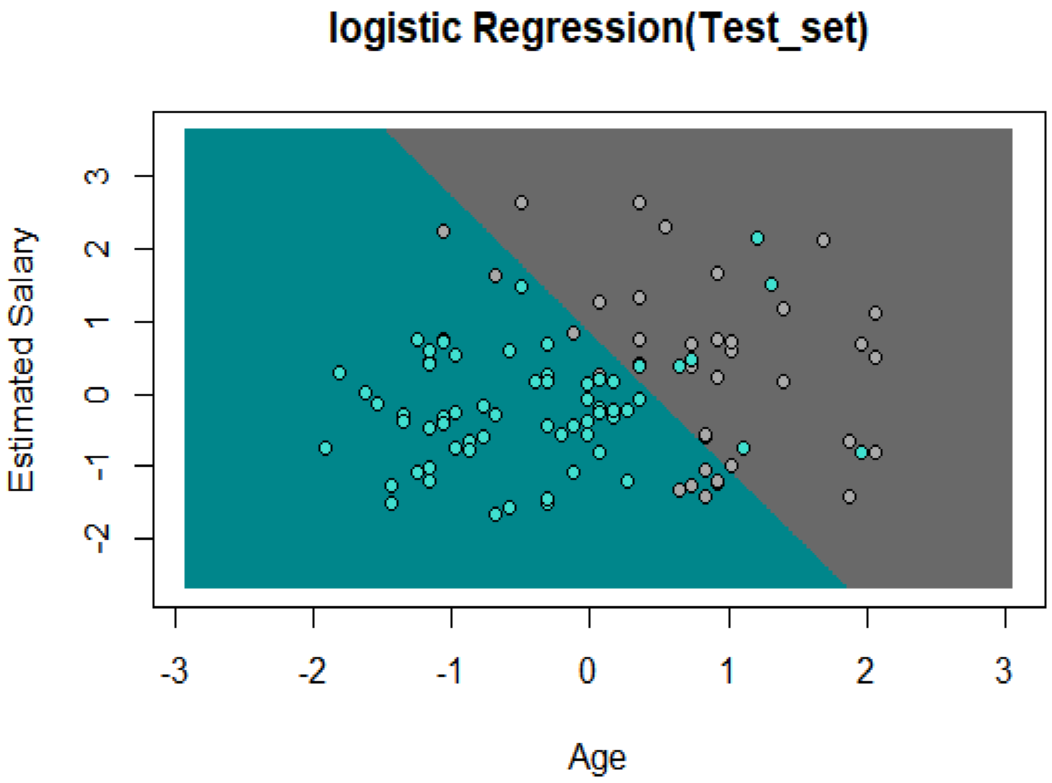
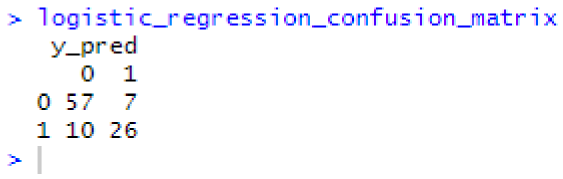


Figure 9.2



```
> logistic_regression_confusion_matrix
   y_pred
     0  1
  0 57  7
  1 10 26
> 
```

9.1.3 Visualisisng The test set Results:

All of these points which are appeared on the above graph are observation points of test set that is all the users that were selected to go to test set blue points are the users who didn't bought the suv and grey points are the users who bought the luxury suv

Here are the consequences of test set with 100 observations we can see that model has done a serious great job by anticipating a large portion of the observations in the right category in the above graph we can see model has predicted 83 observations rightly and 17 incorrect predictions however it seems to have problem in catching some of the grey users who in spite of the lower estimated salary bought the luxurious suv as well as the grey user who also bought the suv as these points are in the region where our classifier predicts that user don't buy the suv and those incorrect predictions are specifically due to the fact that classifier is a linear classifier and the users are not linearly distributed if they all are linearly distributed then we will have all the grey points in the predicted region then linear classifier would perfectly separate them.

9.2 K NEAREST NEIGHBOUR

Below is the results of K nearest Neighbour classifier on training set shown our data of 400observations and divided into 300 observations of training set and 100 observations of test set. 300 observation using function set seed(123) are chosen to learn the model from training set observations and then this learning is applied on predicting the 100 observation of test set

So model will be prepared on training set with 300 observations and afterword's the models performance will be tested on test set with 100 observations however following is the equation of the classifier applied on training set to learn and further used this model to predict the test set observations using R studio **.**

y_grid=knn(train=trainl[,-3],test=grid_set,cl = trainl[,3], k = 5)

where K =5

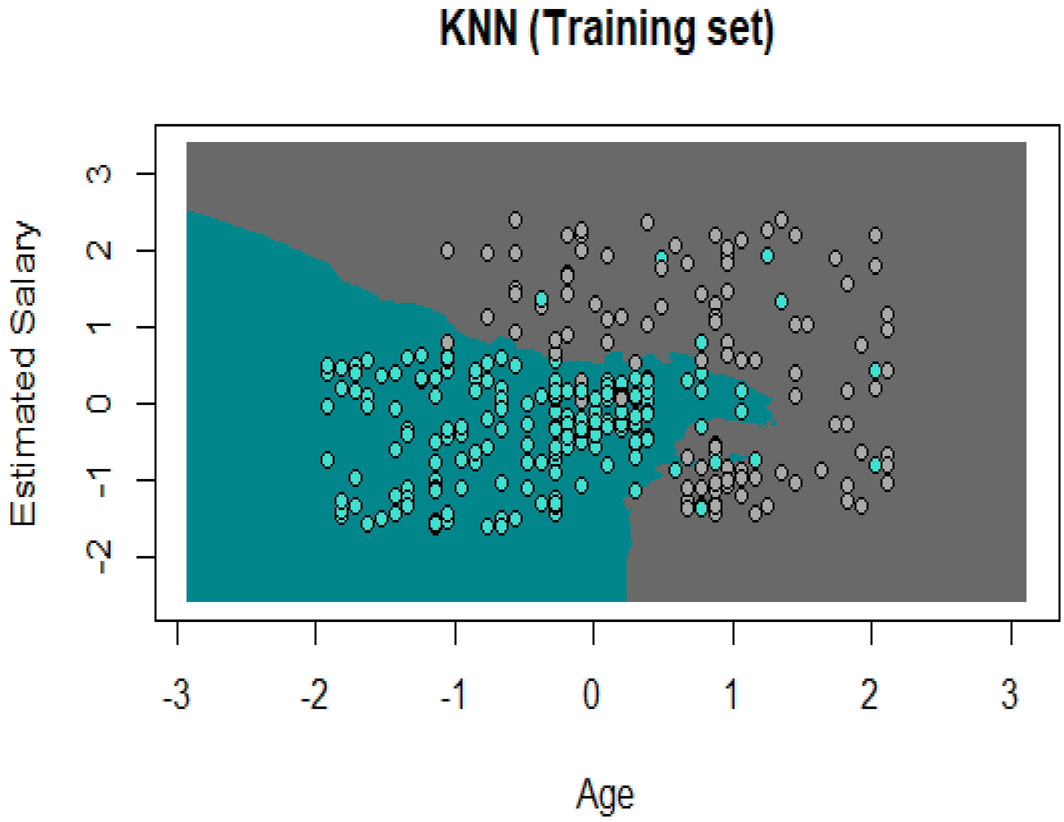


Figure 9.3

9.2.1 Visualize the training set result :

In figure 17 Here we have some blue and grey points so all of these points which are present on the above graph are observation points of training set that is all the users that were selected to go to training set and every one of this client is portrayed by age that is on x hub and evaluated pay that is on y hub .

Presently there are some blue points and some grey points the blue points are training set observations for which dependent variables purchased = 0. The dim grey for which subordinate variable bought = 1. That suggests blue focuses are the clients who didn't bought the SUV and grey points are the customers who bought the indulgence SUV.

Directly the clients who are youthful with low evaluated pay really didn't purchased the suv and the clients who are more seasoned(older) with high assessed pay the majority of them have purchased the sumptuous SUV.

Well that really bode well as SUV is progressively similar to a family vehicle and in this way more interesting for the more seasoned clients with high evaluated pay.

Besides some older people even with lower estimated salary actually bought the SUV those grey points that corresponds to an age above the average age. These older guys although have lower estimated salary bought this luxurious suv probably they were saving up some money or may be they finished paying off their mortgage ,but thing is they couln’t resist buying the very cool luxirious suv offered at high price.

There are additionally a few youths with higher assessed pay have really purchased the suv might be its cool SUV and presumably they need to dazzle their companion or might be they are pulled in towards the suv.

NOW GOAL OF THIS CLASSIFICATION

Well objective here is to arrange right clients into right classifications that is attempting to building a classifier that will get the correct clients into right classes which are yes they purchased the suv and no they haven't purchased the suv.. And the classifier catches the users by plotting which are often known as prediction regions .

.

9.2.2 Visualising the test set

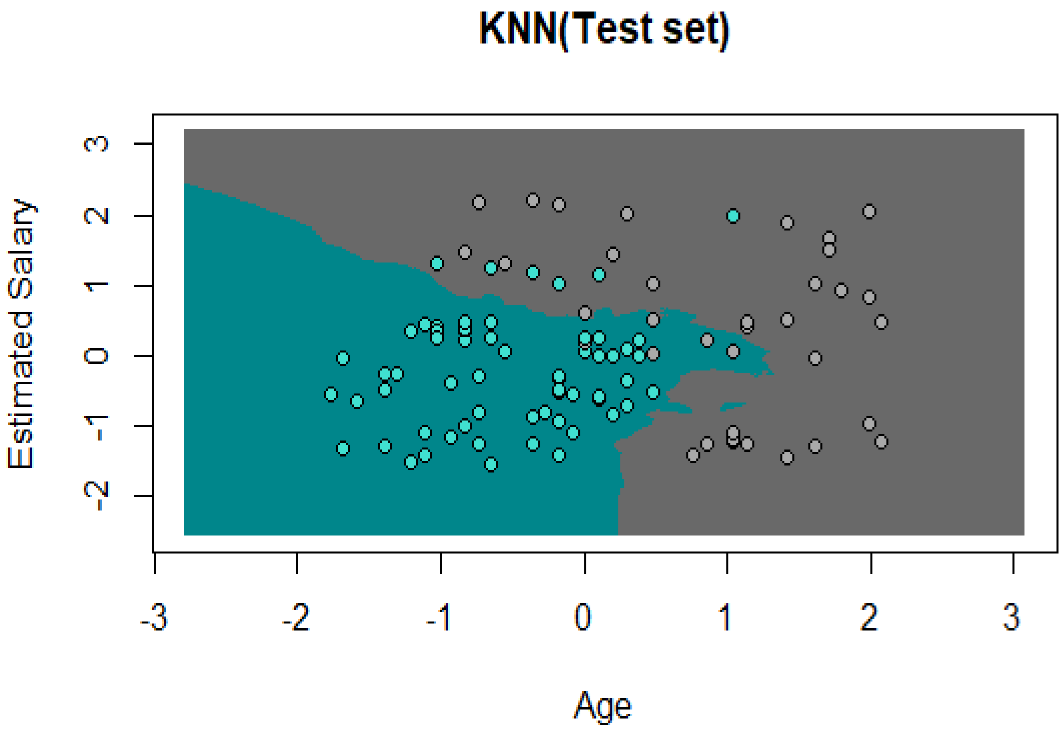
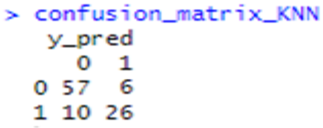


Figure 9.4



```
> confusion_matrix_KNN
   y_pred
     0  1
  0 57  6
  1 10 26
```

Here is the Test set Result

In Figure 18 appeared above is the consequence of test set with 100 observations it tends to be deciphered that model has done a significant great job by predicting almost all of the observations in the right category in the above graph we can see model has predicted 84 observations rightly and 16 incorrect predictions however it seems to have problem in catching some of the blue users who in spite of the higher estimated salary bought the luxurious suv as well as the grey user who also bought the suv as these points are in the region where our classifier predicts that user don't buy the suv and those incorrect predictions overall this classifier has accuracy of 84% which was not achieved by logistic regression classifier, hence KNN has outperformed logistic regression classifier.

9.3 NAIVE BAYES

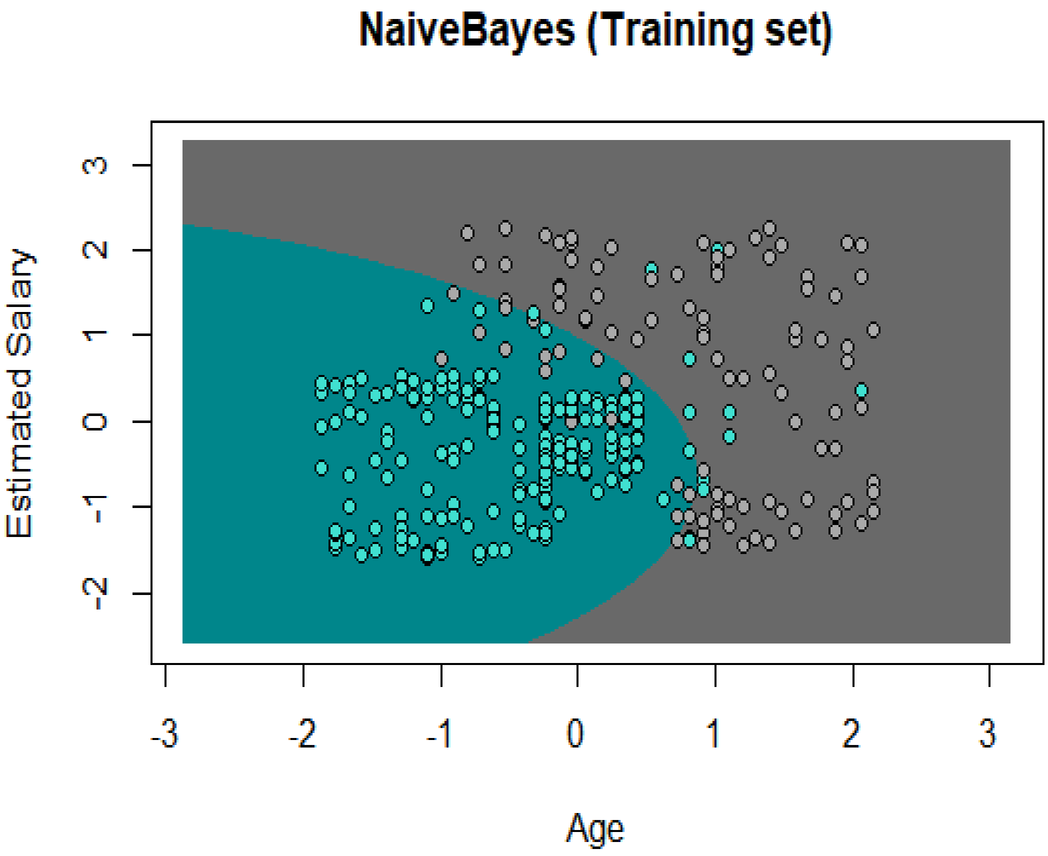


Figure 9.5

9.3.1 Visualising the test set results:

In figure 9.5 Here we have some blue and grey points so all of these points which are present on the above graph are observation points of training set that is all the users that were selected to go to training set and every one of this client is portrayed by age that is on x hub and evaluated pay that is on y hub .

Presently there are some blue points and some grey points the blue points are training set observations for which dependent variables purchased = 0. The dim grey for which subordinate variable bought = 1. That suggests blue focuses are the clients who didn't bought the SUV and grey points are the customers who bought the indulgence SUV.

Directly the clients who are youthful with low evaluated pay really didn't purchased the suv and the clients who are more seasoned(older) with high assessed pay the majority of them have purchased the sumptuous SUV.

Well that really bode well as SUV is progressively similar to a family vehicle and in this way more interesting for the more seasoned clients with high evaluated pay.

Besides some older people even with lower estimated salary actually bought the SUV those grey points that corresponds to an age above the average age. These older guys although have lower estimated salary bought this luxurious suv probably they were saving up some money or may be they finished paying off their mortgage ,but thing is they couln’t resist buying the very cool luxirious suv offered at high price.

There are additionally a few youths with higher assessed pay have really purchased the suv might be its cool SUV and presumably they need to dazzle their companion or might be they are pulled in towards the suv.

NOW GOAL OF THIS CLASSIFICATION

Well objective here is to arrange right clients into right classifications that is attempting to building a classifier that will get the correct clients into right classes which are yes they purchased the suv and no they haven't purchased the suv. And the classifier catches the users by plotting which are often known as prediction regions .

In the aforementioned plot of which was performed using naïve bayes algorithm the prediction boundary is a smooth curve as the naïve bayes classifier classifying the observations quite well . So this smooth curve of naive bayes classifier catching all those grey points that was not caught by the linear classifier known as logistic regression classifier which we have used because linear classifier has a straight prediction boundary but thanks to this naïve byes classifier which is having prediction boundary in shape of smooth curve that’s why making less incorrect predictions

But still there some incorrect predictions in this naïve bayes classifier with the users who are young and having higher estimated salary

Overall naive bayes classifier does quite a good by predicting most of the users correctly

9.3.2 Visualising the test set results:

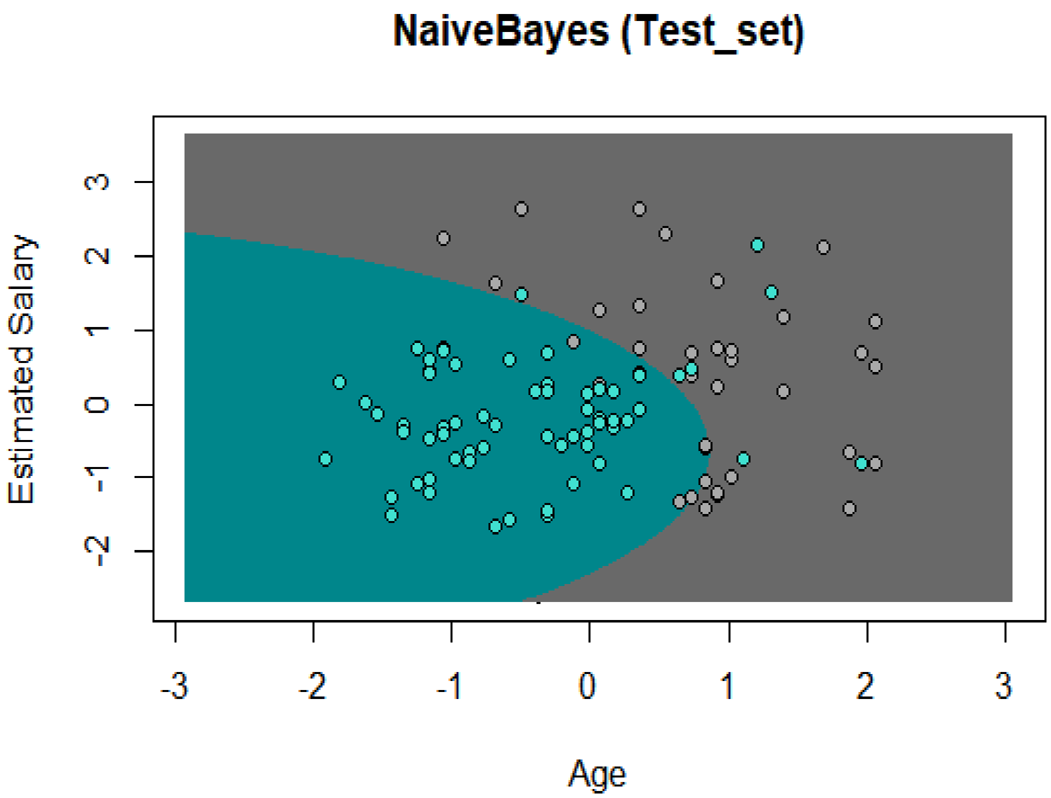


Figure 9.6

In figure 9.6 the plot above which was performed using naïve bayes algorithm on the test set of 100 observations the prediction boundary is a smooth curve as the naïve bayes classifier classifying the observations quite well . So this smooth curve of naive bayes classifier catching all those grey points that could not caught by the linear classifier known as logistic regression classifier which we have used because linear classifier h as a straight prediction boundary but thanks to this naïve byes classifier which is having prediction boundary in shape of smooth curve that's why making less incorrect predictions even on the test set .

Naïve bayes did a terrific job by predicting 86 observations correctly and only 14 incorrect predictions even incorrect predictions are easily visible on the plot but still there some incorrect predictions in this naïve bayes classifier with the users who are young and having higher estimated salary overall naive bayes classifier does a significant decent by anticipating the greater part of the clients accurately with most noteworthy precision among all the models ..

9.4 Decision TREE

9.4.1 Visualising the training and test set results

Following Classifier is used to train the model in R studio

classifier=rpart(formula = Purchased~ . , data= train)

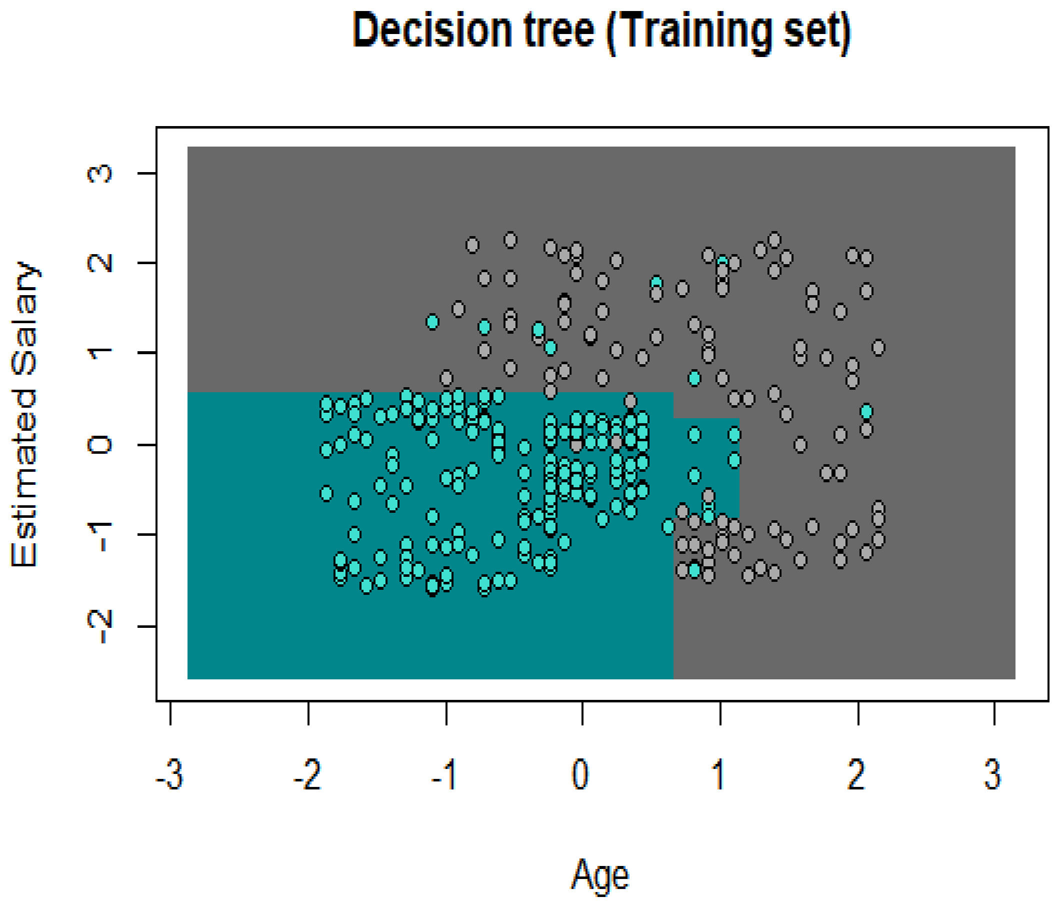


Figure 9.7

In figure 9.7 some blue and grey points so all of these points which we can see on the above graph are observation points of training set that is all the users that were selected to go to training set and each of this user is characterised by age that is on x axis and estimated salary that is on y axis .

Presently there are some blue points and some grey points the blue points are training set observations for which dependent variables purchased = 0. The dim grey for which subordinate variable bought = 1. That suggests blue focuses are the clients who didn't bought the SUV and grey points are the customers who bought the indulgence SUV.

Presently the clients who are youthful with low assessed compensation really didn't purchased the suv and the clients who are more established with high evaluated pay the vast majority of them have purchased the suv . Besides some older people even with lower estimated salary actually bought the suv those grey points 1that corresponds to an age above the average age . These older guys although have lower estimated salary bought this luxurious suv probably they were saving up some money or may be they finished paying off their mortgage , however thing is they couldn't avoid purchasing the cool luxurious suv offered at significant expense.

There are additionally some youngsters higher estimated salary have actually bought the suv may be its very cool suv and probably they want to impress their friend or may be they are attracted towards the suv.

NOW GOAL OF THIS CLASSIFICATION

Well objective here is to arrange right clients into right classifications that is attempting to building a classifier that will get the correct clients into right classes which are yes they purchased the suv and no they haven't purchased the suv. And the classifier catches the users by plotting which are often known as prediction regions.

This decision tree classifier that has predicted boundary that is only horizontal and vertical lines that's because by finding intervals that will make conditions that will classify in some rectangles

Here there is less overfitting that's why there are more number of incorrect predictions that's because **rpart** library which was used selects the right parameter's to prevent overfitting as here in this plot there is no trace of overfitting .

In the aforementioned plot decision tree classifier is doing a terrific job in classing the user whether they have bought the suv , most of the blue points are in blue region and most of the grey points are in grey region . Even after this terrific job decision tree classifier wont be able to predict some of the predictions correctly that's due to the fact that model is pulling every string to prevent overfitting in the data .

9.4.2 Visualising the test set results:

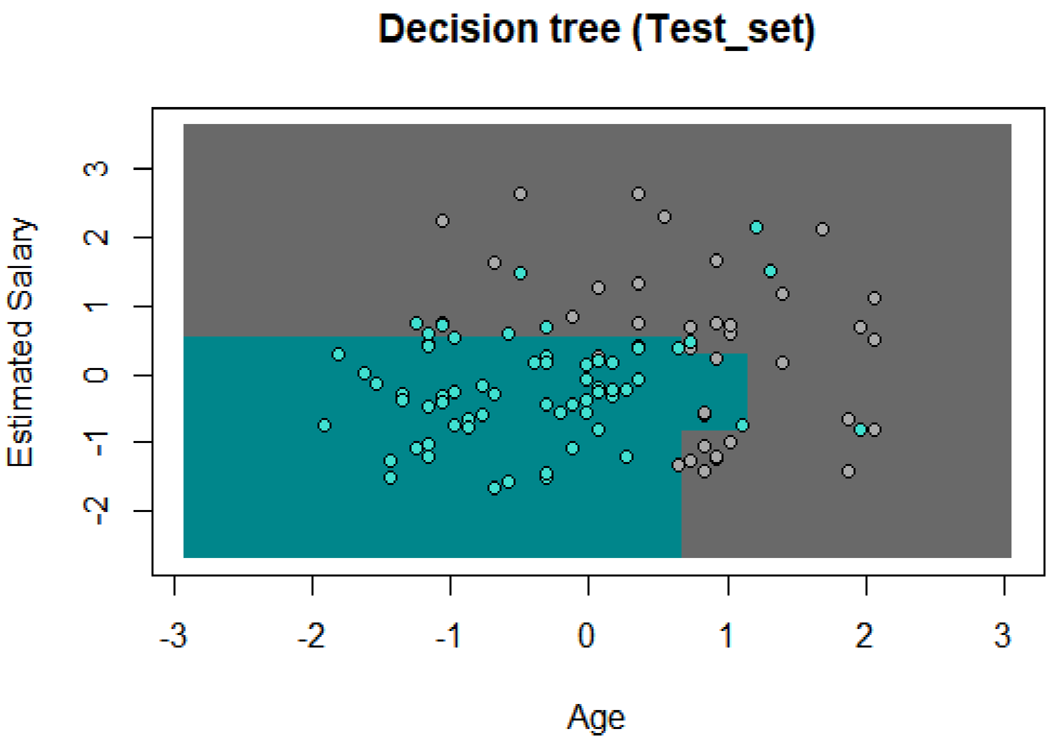
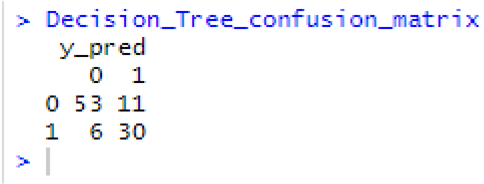


Figure 9.8



```
> Decision_Tree_confusion_matrix
   y_pred
     0  1
  0 53 11
  1  6 30
> 
```

That is the observation of test set this is the set where we have 17 incorrect prediction and it is classifying most of the blue points in blue region and most of grey users in grey region but here in this decision tree classifier we have many blue points in the grey region with higher age and higher estimated salary that’s unlucky this is due to the fact which also mentioned earlier that is here more focus is on preventing overfitting then trying to minimize to zero number of incorrect predictions . overall decision tree classifier has also performed quite well with 83 correct predictions and 17 incorrect predictions.

the above classification of confusion matrix uncovers that around 83 level of the reaction with respect to the buy decision for purchasing an exceptional suv vehicle is accurately arranged dependent on the model turn of events. This extremely high rate (83%) of grouping.

9.4.3 Decision Tree of Decision Tree Classifier:

```
plot(classifier)
text(classifier)
```

Exploring the Decision tree Behind the SCEN

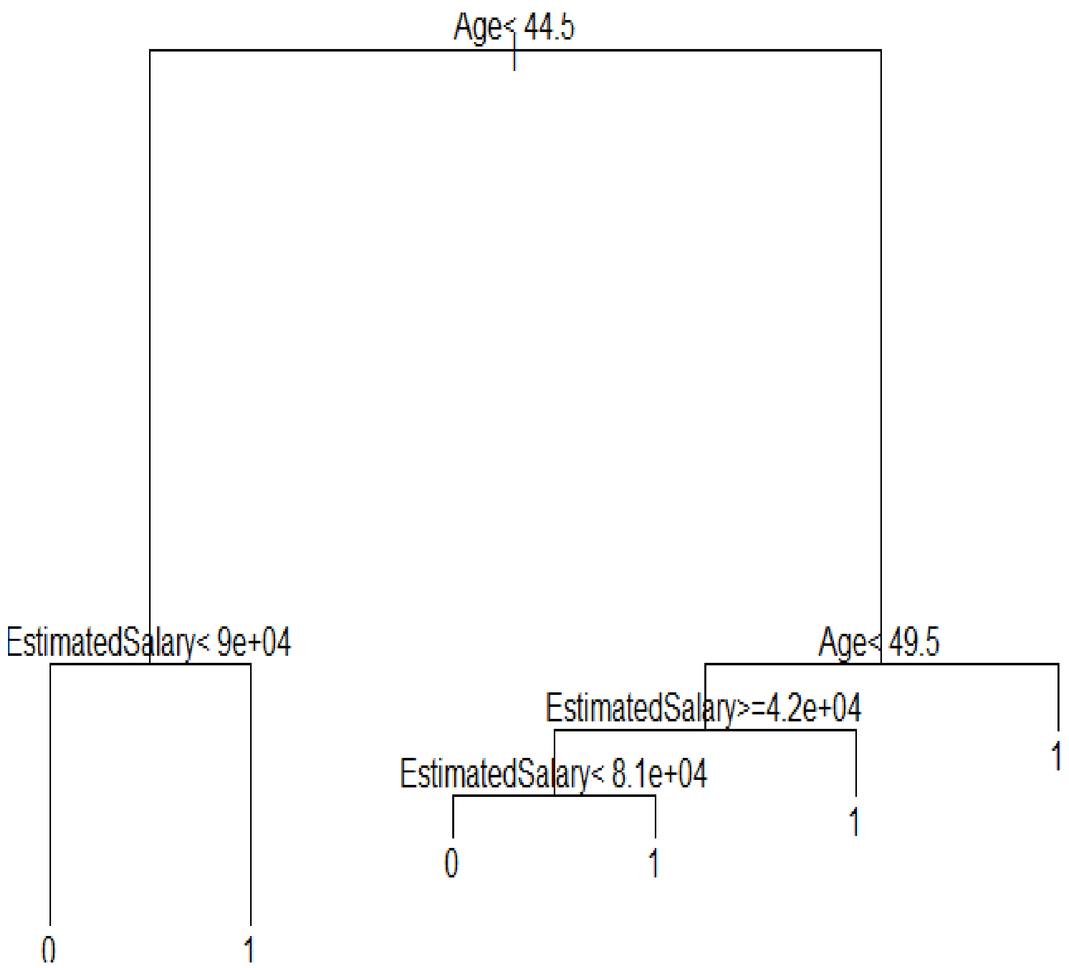


Figure 9.9

9.4.4 Exploring the Decision tree Behind the scene

In figure 9.9 plotting of this decision tree can be done by removing the feature scaling .in this decision tree at each split the condition that is generating the splits .

First split is made based upon the condition of age , if the age is lower than 44.5 years that means user below years of age he /she will get into subcategory after the split and if the user ‘s age is greater than 44.5(>44.5) years will end up in different category ,at 3rd split there are some new condition on the independent variable .

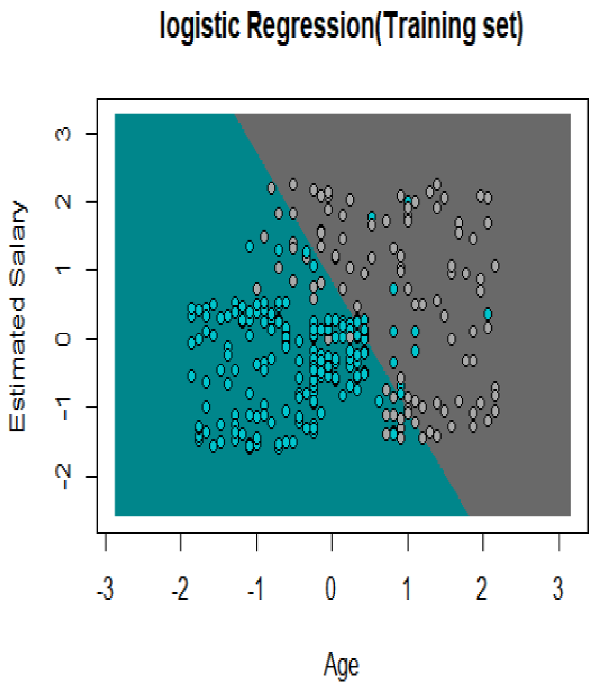
So in 2nd split there is a condition of the estimated salary of the user is below $90,000 that means if the user is younger than 44.5 years and having estimated salary below $90,000 then according to this classifier this user wont buy the suv because the result here is zero.

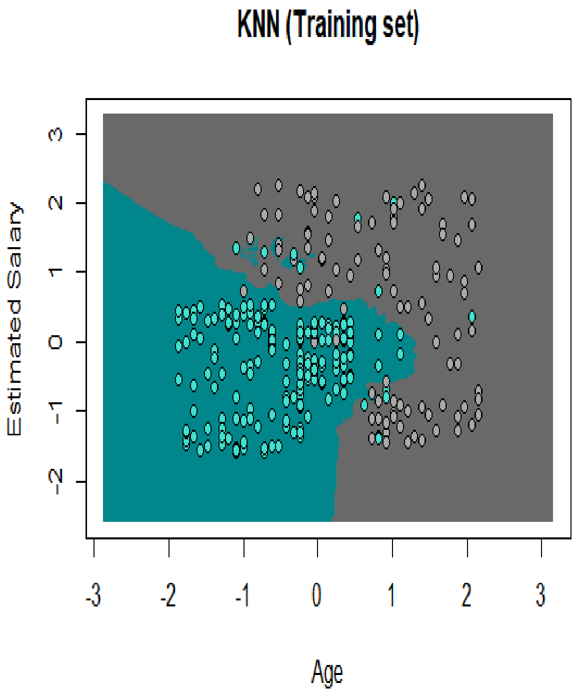
And if user is younger than 44.5 years of age and having an estimated salary above $90,000 then according to the decion tree classifier user will buy the suv because the result is 1.

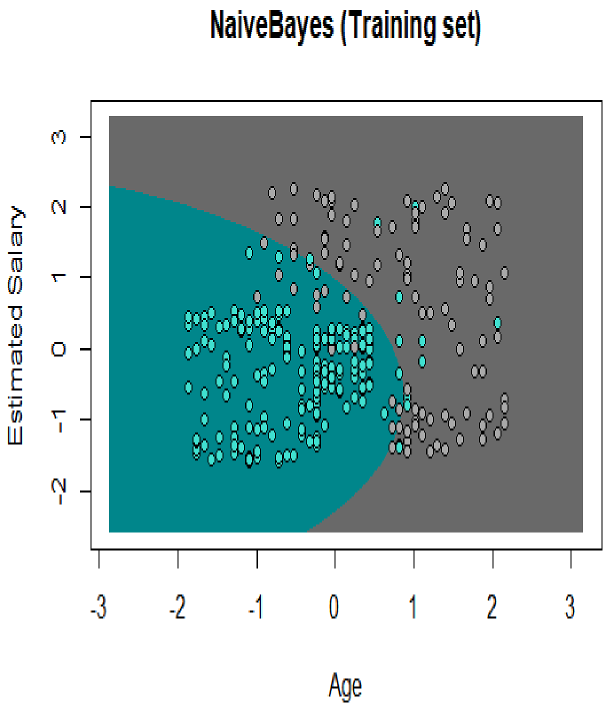
And by looking at the other side of the tree that is the 3 rd split that contained all the users who are older the years of age and some new condition generating new split and further there are two more splits based on the condition of age and estimated salary by following these conditions tree will end up final category the final model of decision tree where user is predicted not to buy the suv and buy the suv so by looking at this decision tree one can have very interpretable results because here on this decision tree plot one can see how decision tree classifier actually working and how decision tree classifier decides weather the user will be predicted to buy or not and how decision tree classifier learned from the data to classify the users into their respective categories

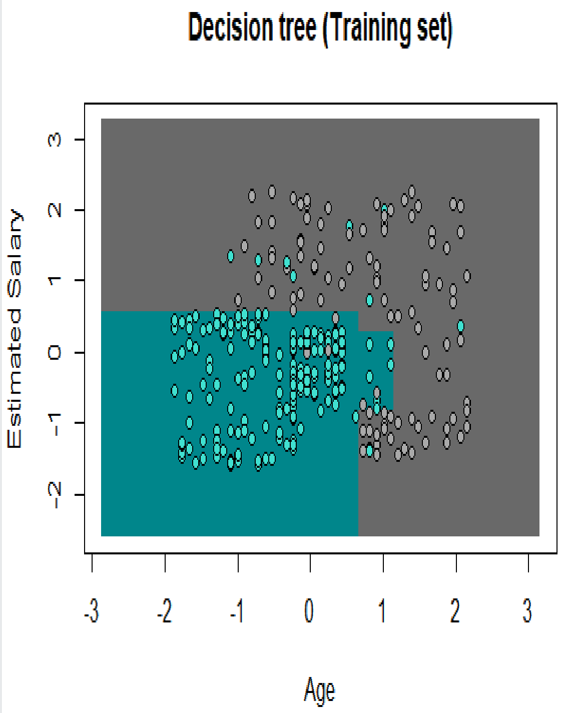
Chapter 10

Visual Comparison of classifiers with 300 training observation each

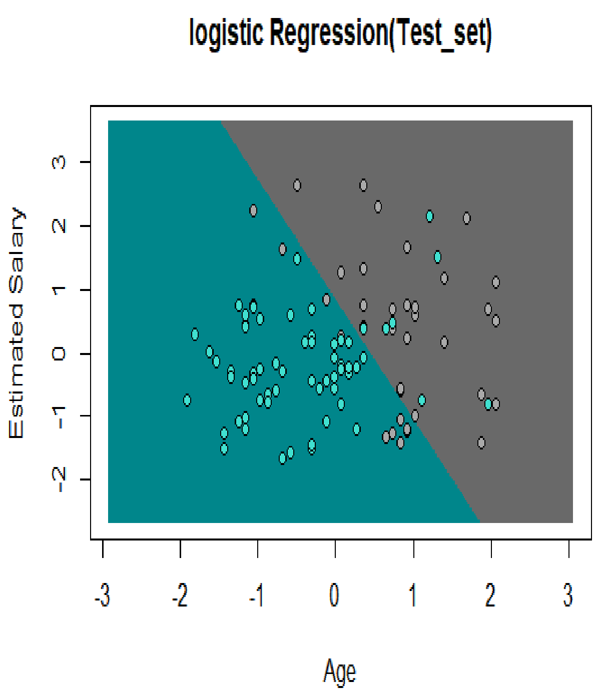

logistic Regression(Training set)

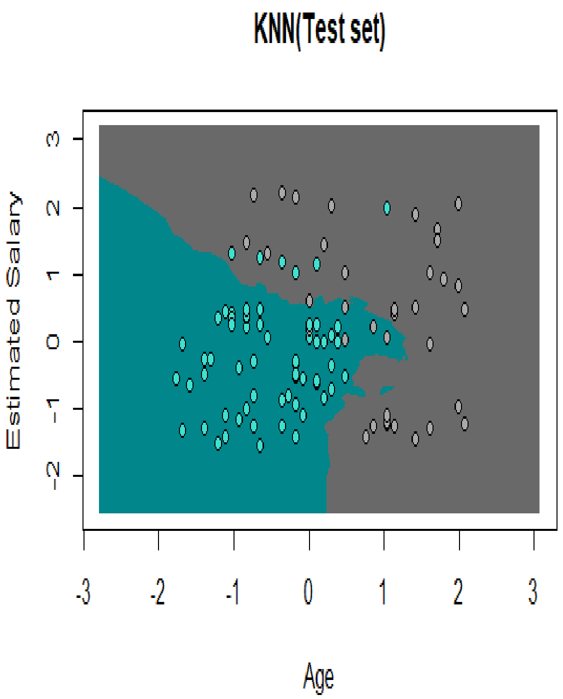

KNN (Training set)

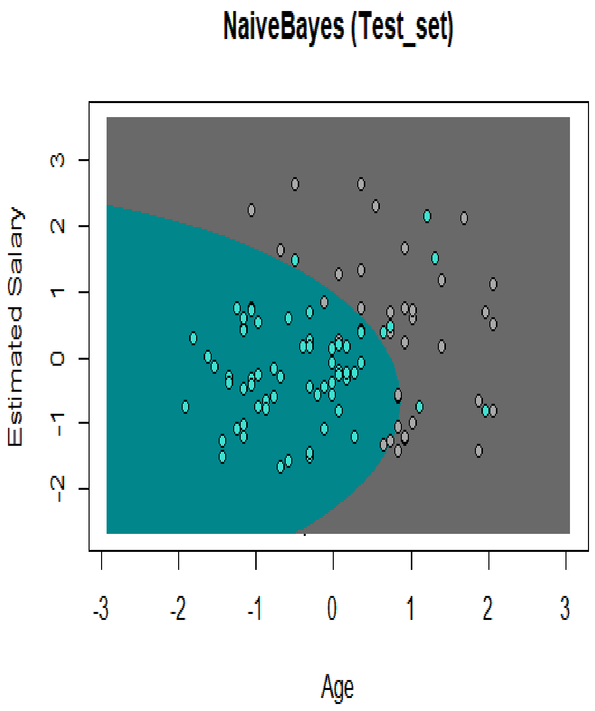

NaiveBayes (Training set)

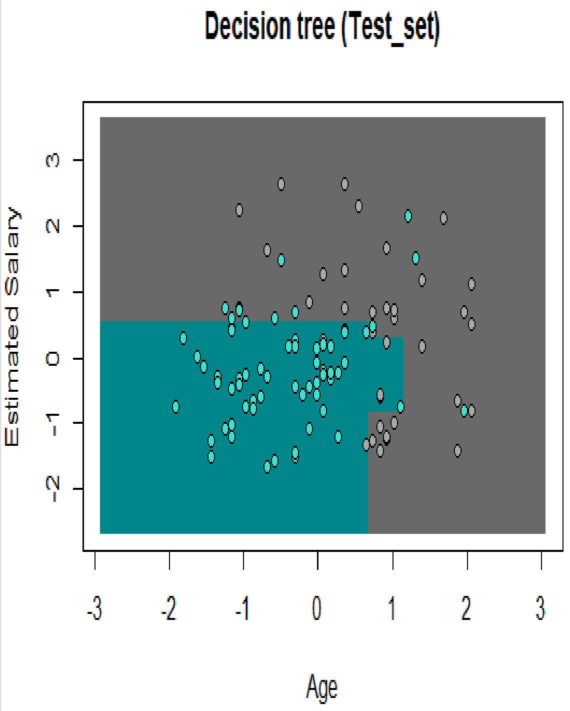

Decision tree (Training set)

10.2 Visual comparison of Test set 100 observations









Chapter 11

Predicted Decisions of Consumers on Test set(100 observations)

11.1 Predicted Decisions of Consumers on Test set(100 observations)

Incorrect prediction of each classifier with their respective id is shown in Yellow

userid	Actual Decision	Logistic Regression Classifier	KNN Classifier	Naïve Bayes Classifier	Decision Tree Classifier
4	0	0	0	0	0
5	0	0	0	0	0
9	0	0	0	0	0
12	0	0	0	0	0
18	1	0	0	0	1
19	1	0	0	1	1
20	1	0	1	1	1
22	1	1	0	0	0
29	0	0	0	0	0
32	1	0	1	1	0
34	0	0	0	0	0
35	0	0	0	0	0
38	0	0	0	0	0
45	0	0	0	0	0
46	0	0	0	0	0
48	0	0	0	0	0
52	0	0	0	0	0
66	0	0	0	0	0
69	0	0	0	0	0
74	0	0	1	0	0
75	0	0	0	0	0
82	0	0	0	0	0
84	0	0	0	1	1
85	0	0	0	0	0
86	1	0	1	1	1
87	0	0	0	0	0
89	0	0	0	0	0
103	0	0	0	1	0

104	1	1	1	1	1
107	0	0	0	0	0
108	0	0	0	1	0
109	0	0	0	1	0
117	0	0	0	0	0
124	0	0	0	0	0
126	0	0	0	0	1
127	0	0	0	0	0
131	0	0	0	0	0
134	0	0	0	0	0
139	0	0	0	0	0
148	0	0	0	0	0
154	0	0	0	0	0
156	0	0	0	0	0
159	0	0	0	0	0
162	0	0	0	1	0
163	0	0	0	0	0
170	0	0	0	0	0
175	0	0	0	0	0
176	0	0	0	0	0
193	0	0	0	0	0
199	0	0	0	0	0
200	0	0	0	0	0
208	0	1	1	1	1
213	0	1	1	1	1
224	1	1	1	1	1
226	0	0	0	0	0
228	1	1	1	1	1
229	0	0	0	0	0
230	1	1	0	0	0
234	1	1	1	1	1
236	1	1	1	1	0
237	0	0	0	0	0
239	0	1	1	1	0
241	1	1	1	1	1
255	0	1	1	0	1
264	0	0	0	0	0
265	1	1	1	1	1

266	1	1	1	1	1
273	1	1	1	1	1
274	1	1	1	1	0
281	1	1	1	1	1
286	1	0	0	1	1
292	1	1	1	1	0
299	0	1	1	0	0
302	1	1	1	1	0
305	0	0	0	0	1
307	0	1	1	1	1
310	0	0	0	0	0
316	1	0	0	0	0
324	1	0	1	1	1
326	0	0	0	0	0
332	1	1	1	1	1
339	0	0	0	0	0
341	1	1	1	1	1
343	0	0	0	0	0
347	1	1	1	1	1
353	1	1	1	1	1
363	1	1	0	0	0
364	0	1	0	0	0
367	1	1	1	1	1
368	1	1	1	1	1
369	0	0	0	0	0
372	1	1	1	1	1
373	0	0	0	0	0
380	1	1	1	1	1
383	1	1	1	1	1
389	1	0	1	1	1
392	1	0	1	1	0
395	0	0	0	0	0
400	1	1	1	1	1

Chapter 12

False Positive & Negative

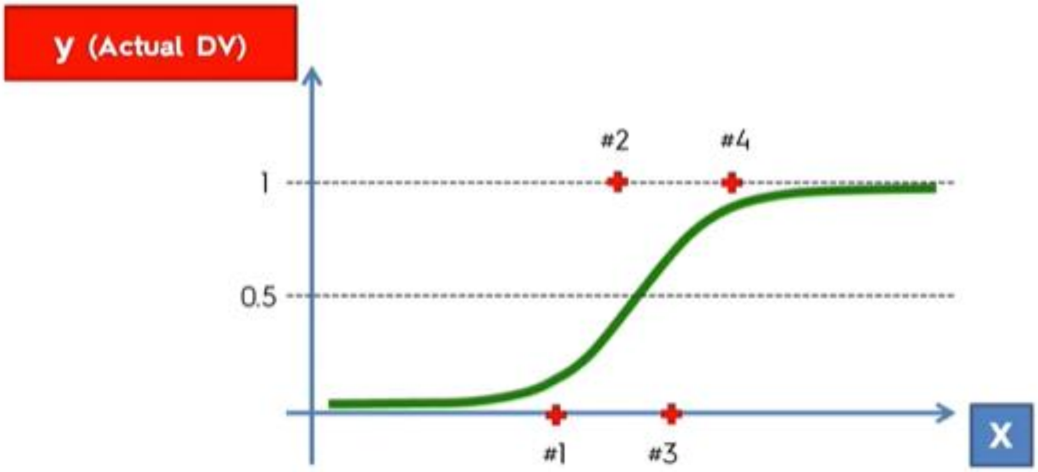


Figure 12.1

Taking 4 data points from the data set and create the logistic regression on the vertical axis in red actual value of dependent variable is shown and observations 1 and 3 didn't bought the suv and top 2 and 4th observations did bought the suv if logistic regression applied on this values of observations are projected on to the curve blue dots that have been modeled by the curve now the so the probablities of the observation 1st may be around 0.15 , observation 2nd : 0.4, observation 3rd : 0.7 and observation 4 : 0.85.

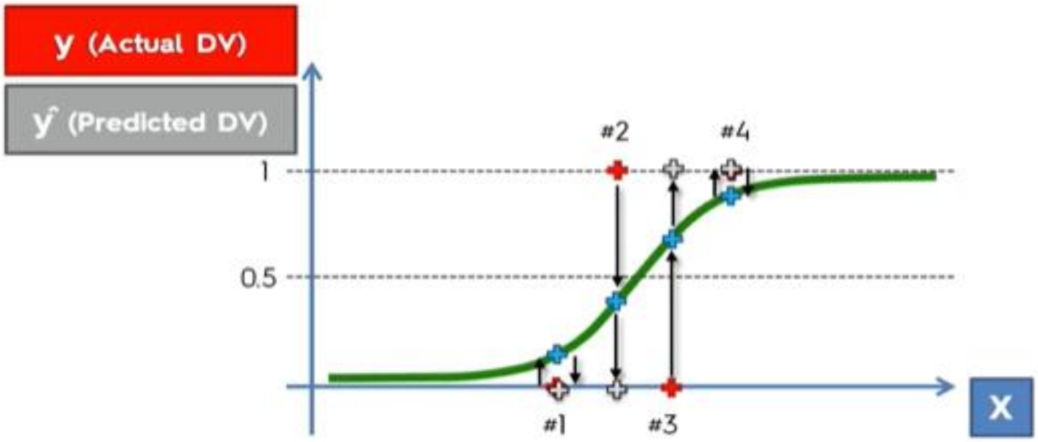


Figure 12.2

However finding probablities is not the class of interest , class of interst is predicted value will the person going to buy the suv or not so much more concerned with the working here by using a horizontal line at 0.50 so anyhing below 0.50 line is going to be projected on the horizontal line which is zero that means customer does buy the suv and anything above the 0.50 line will be projected on to the horizontal line which is 1 or 100 % where it can be concluded those customers end up buying the suv.

Predicted value ycap is in grey so what actually happen is in red and what predicted is going to happen is grey and for observation 1 and 4 where predicted correctly so predicted the observation not buy the suv same with the observation 4 predicted to buy the suv and he actualy buys the suv**.**

Observation 2 grey mark is at bottom means that the model predicted for the person based on the age and estimated salary is not going to buy the suv however it is quite visible that red mark is at top meaning that person did buy the suv that means logistic regression has made an error as at he same type observatiuon then th grey mark is at the top and that means model predicted that person will buy the suv and then for logistic regression made a mistake here and same thing for person 3 the grey mark is at the top that's means model predicted that person will buy the suv but red mark is at bottom meaning person didn't buys the suv and therefore logistic regression made a mistake onces again **.**

These is take they have specific names

That 4th observation is FALSE POSITIVE TYPE 1 ERROR means model predicted a positive outcome but it was false so model predicted something that didn't occur

And at observation 2nd is FALSE NEGATIVE TYPE 2 ERROR we predicted that they wouldn't be an effect but effect actually did occur so other prediction was negative that is type 2 error

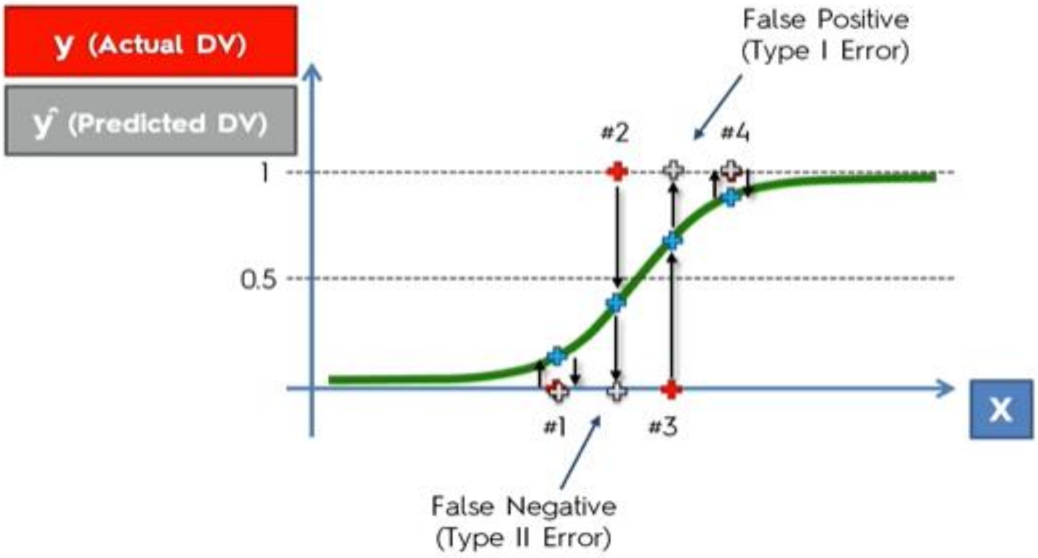
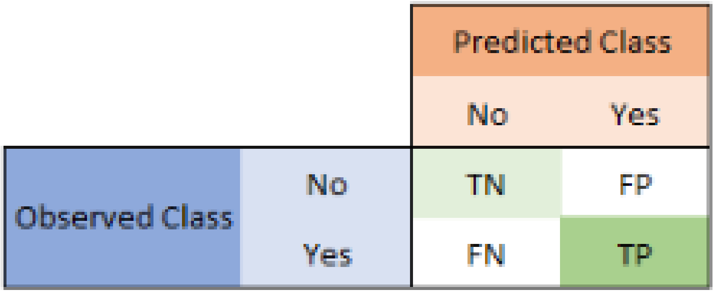


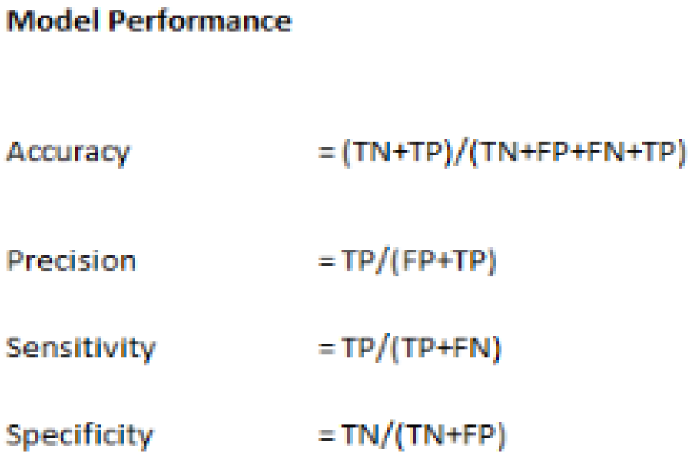
Figure 12.3

Performance Parameters



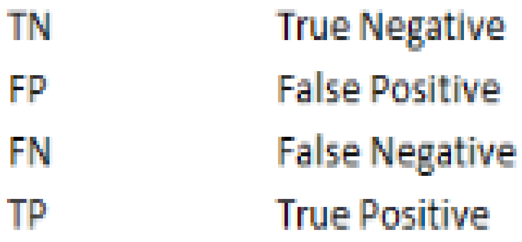
		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

Figure 12.4



Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)



TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Chapter 13

Performance of Each Classifier

Classifier	Correct Predictions	Incorrect Predictions	Accuracy	Precision	Recall
Logistic Regression	83	17	83%	0.78	0.72
KNN	84	16	84%	0.81	0.72
Naïve Bayes	86	14	86%	0.80	0.80
Decision Tree	83	17	83%	0.73	0.83

Table 13.1

The above Performance table reveals that about 88.6 level of the reaction with respect to the buy decision for purchasing a superior vehicle is accurately characterized dependent on the model turn of events. This exceptionally high rate (86%) of grouping further affirms the decency of the model picked for the investigation.

Performance Parameters

Utilizing confusion matrix, performance parameters of a classifier can be determined. The performance parameters include: precision, recall, accuracy, simply using accuracy results can be misleading. They prescribed when evaluating binary decision problems to use Precision and Recall , which show how the quantity of correctly classified positive examples varies with the number of incorrectly classified negative examples.

These Performance Parameters are usual the criterion for identifying the prediction power of different classification methods, and the Recall and Precision are the important evaluation metrics which can be applied for selecting the best classification method.

Chapter 14

Discussion

As communicated in the past area, the analysis did reveals that naïve bayes outperforms logistic regression ,decision tree and knn . It is the best in all exhibition parameters like accuracy, precision and recall (sensitivity) when comparing naïve bayes and decision tree in the classification of test set find that the accuracy, precision and recall of naïve bayes are 86, 0.80, and 0.80 respectively. This is better than decision tree whose accuracy , precision and recall (sensitivity) are: 83,0.73,0.8 , respectively, investigating different procedures for report portrayal: the straightforward bayes classifier, the nearest neighbor classifier, decision trees and determined backslide classifier . Their preliminary outcomes show that the honest bayes classifier technique beat the other three classifiers on the instructive assortments. A Naive Bayes classifier is a straightforward classifier. Be that as it may, despite the fact that it is basic, naive bayes can beat progressively modern order techniques. Other than that it has in like way indicated high exactness and speed when applied to colossal database . Furthermore, it is outstandingly snappy for both learning and anticipating. Its learning time is immediate in the measure of models and its conjecture time is autonomous of the amount of models .Naive bayes classifier is moreover fast, consistent,, easy to keep up and exact in the request for order of trait information . What's more, from calculation perspective, innocent Bayes is progressively proficient both in the learning and in the characterization task than decision tree and logistic. In contrast. The performance of logistic regression classifier in this is the most noticeably awful among different classifiers. Since k-nn uses number of nearest neighbour k as one of the parameter in characterizing a thing, at that point this worth may influence the exhibition of the classifier. In their examination utilizing k-nn to order, find that the best execution of k-nn is when k=5. Utilizing this k esteem, k-nn beats logistic and decision tree classifier. Utilizing greater and littler k value, the k-nn performance is worst. for higher estimations of k, the presentation diminishes. The analysis utilizes 75 % of perceptions for preparing and 25 % for test set The above table shows that shows that naïve reaches the best performance. Adjacent to low execution of knn , The k-nn classifier requires a tremendous memory to store the entire preparing set . Hereafter, the more prominent the preparation set, the greater memory necessity and the bigger separation figurings must be performed. This causes the characterization is moderate. This is the inspiration driving why the arrangement time of k-nn in our investigation is exceptionally enormous the most exceedingly terrible among the three classifiers., The very brisk order time by decision tree is because of the nonappearance of estimation in its grouping procedure. The tree model is made utilizing R studio And the model is converted into rules before being incorporated into the application.

Chapter 15

Conclusion

An epic strategy to look through elective structure in a vitality recreation device is proposed. A classification method is utilized in looking the alternative design. There are Four classifiers used in this assessment to be explicit Naïve Bayes, Decision Tree, and k-Nearest Neighbor and Logistic Regression . This outcome shows that Decision Tree is the fastest and k-Nearest Neighbor is the slowest , The speedy characterization time of Decision Tree considering the way that there is no calculation in its order . The tree model is made outside the application that is utilizing R, Furthermore, the model is changed over into rules before being incorporated into the application. Classification by way of following the tree rules is quicker than the ones that need calculation as in the case of Naïve Bayes and k-NN. In the mean time k-Nearest Neighbour is the slowest classifier on the grounds that the characterization time is straightforwardly identified with the quantity of information. The gigantic the size of information, the bigger separation counts must be performed. This causes the order is incredibly moderate. In spite of it is a basic strategy, Naïve Bayes can outflank increasingly complex arrangement strategies. In this examination, Naïve Bayes beats Decision Tree and k Nearest Neighbor and logistic regression .

The demographic factors have a lot of noteworthiness in advertising. They are used as purpose behind dividing the market and their activity in purchaser's buying decision is striking and major. The significant segment factors are Age, Gender, Income. Along these lines, recognizing the segment profile of the people is huge for the advertisers(promotional organizations) From the discoveries of the examination, the vehicle advertisers will have the option to decide the buy decision of shoppers and to whom clients they should send limited time offers for purchasing vehicle when the segment profile is known.

Chapter 16

Future scope

The demographic variables have a lot of centrality in marketing. They are used as explanation behind dividing the market and their activity in purchaser's buying decision is amazing and essential. The significant segment factors are Age, Gender, Income. In any event, when the objective market is delineated in non segment terms (state, a character type), the association back to segment attributes is required so as to survey the size of the potential market and the media which ought to be utilized to arrive at it proficiently.. Thus, recognizing the segment profile of the people is huge for the advertisers(promotional organizations) From the discoveries of the investigation, the vehicle advertisers will have the option to decide the buy decision of shoppers and to whom clients they should send special proposals for purchasing vehicle when the segment profile is realized This will help the vehicle organizations in following manners:

- Production Analysis
- Establishing Sales Outlets
- Strategic planning
- Targeting potential clients
- Gaining piece of the pie
- Revenue improvement

.

Chapter 17

REFERENCES

- https://www.slideshare.net/hemanthcrpatna/a-study-of-consumer-perception-of-car-market-buying-behavior

- Mahatoo, Winston H (1985), The Dynamics of Consumer Behavior, John Weley& Sons Canada Ltd., Toranto, Ontario, pp 5-8.

- https://towardsdatascience.com/machine-learning-project-17-compare-classification-algorithms-87cb50e1cb60

- https://medium.com/@omairaasim/machine-learning-project-16-random-forest-classifier-414bb558d2c2

- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0219-y

- https://link.springer.com/article/10.1186/s40537-019-0219-y#Sec2

- https://www.youtube.com/watch?v=XycruVLySDg

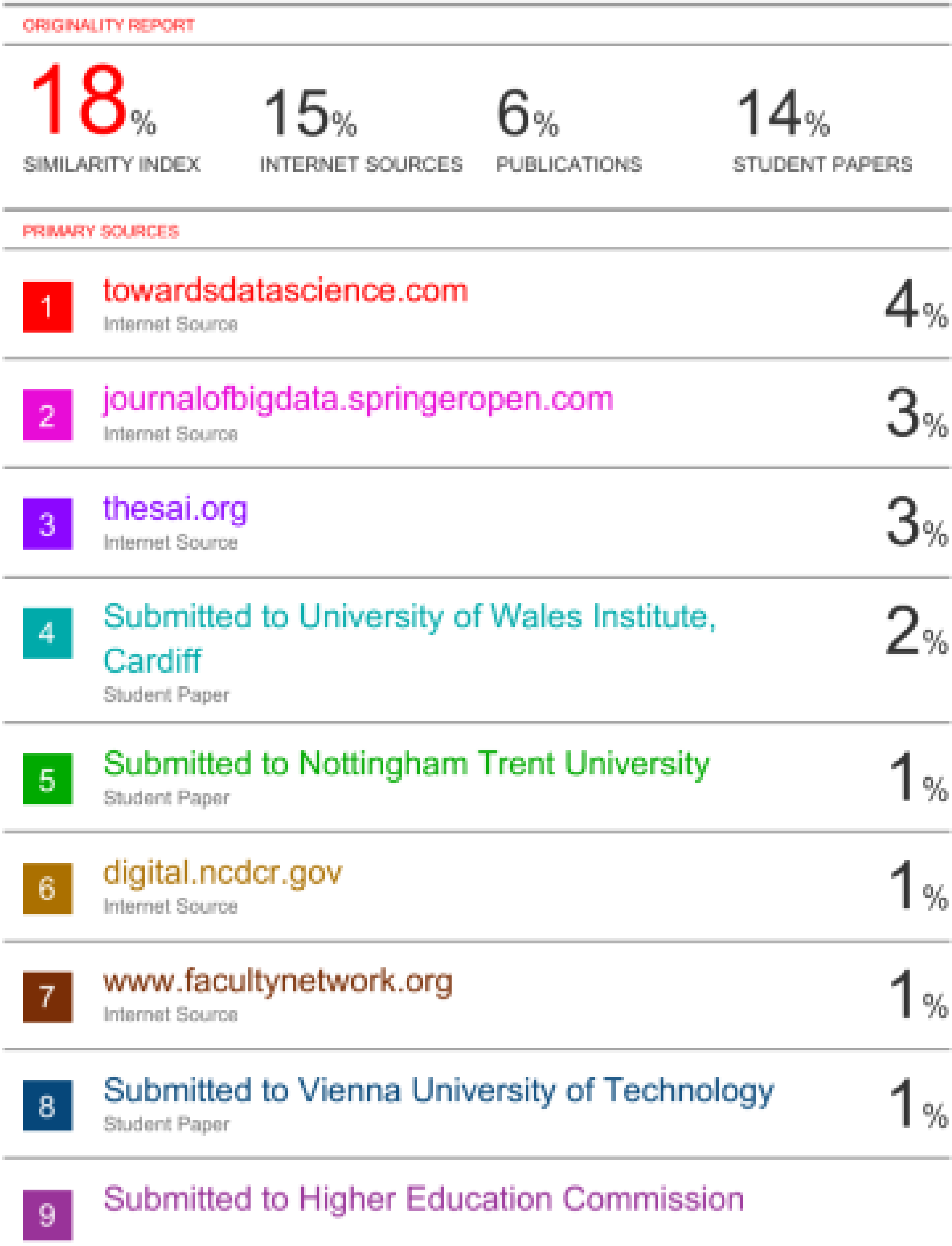
- https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

- G. K. Gupta, Introduction to Data Mining with Case Studies. Prentice Hall of India, New Delhi, 2006

- V. Mohan, “Decision Trees: A comparison of various algorithms for building Decision Trees,” Available at: http://cs.jhu.edu/~vmohan3/document/ai_dt.pdf

Chapter 18

PLAGIARISM REPORT



ORIGINALITY REPORT

18%	15%	6%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	4%
2	journalofbigdata.springeropen.com Internet Source	3%
3	thesai.org Internet Source	3%
4	Submitted to University of Wales Institute, Cardiff Student Paper	2%
5	Submitted to Nottingham Trent University Student Paper	1%
6	digital.ncdcr.gov Internet Source	1%
7	www.facultynetwork.org Internet Source	1%
8	Submitted to Vienna University of Technology Student Paper	1%
9	Submitted to Higher Education Commission	