RAMAIAH
Institute of Technology

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

20 Marks Non CIE Component

CSE01: DATA MINING TERM: Aug - Dec 2019

Title of the Project: PROPORTIONALITY OF WEATHER CAPACITY

PROJECT TEAM MEMBERS

Sl. No	Semester Section	USN	Name
1	A	1MS15CS028	ASHOK KAMATHAM

INTRODUCTION

CLASSIFICATION:

Classification is one of the data mining functionalities and the functionalities also includes Characterization, Discrimination, Association Analysis, Prediction, Clustering, Outlier Analysis, Evolution and Deviation Analysis

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

WHY IS CLASSIFICATION IMPORTANT?

After starting a credit policy, the Our Video Store managers could analyze the customer behaviours, their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky".

The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

PREDICTION:

Prediction is another important data mining functionality. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes.

WHY IS PREDICTION USED?

Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

CLASSIFICATION METHODS

 Classification according to the type of data source mined:
According to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
 Classification according to the data model drawn on:
Based on the data model involved such as relational database, object-oriented database, data warehouse, transactional, etc.
 Classification according to the king of knowledge discovered:
Based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.
 Classification according to mining techniques used:
Categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database-oriented or data warehouse-oriented, etc.
degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

PREDICTION :

The prediction of continuous values can be modelled by statistical techniques of regression. Many problems can be solved by linear regression.
Multiple regression is an extension of linear regression involving more than one predictor variable. In statistics nonlinear regression is a form of regression analysis in which observational data are modelled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation.

DISADVANTAGES OF NON-LINEAR REGRESSION:

Nonlinear regression can be a powerful alternative to linear regression but there are disadvantages. nonlinear regression allows a nearly infinite number of possible functions, it can be more difficult to setup. It is easier to use linear regression model.

DATASET DESCRIPTION:

The Source of our Data Set
The weather data set, was sourced from the internet.

Attribute Description

• TIME: The Attribute time, helps us determine the expected temperature capacity of a place at that time, which is then used to compare with the overall temperature capacity we are recording in our data set.

• DAY/DATE: Suggests to us, the temperature of a place recorded(past) or Present and future of a particular.

• TEMPERATURE: Provides us, with the weather conditions or temperature of the place. Who's data has been recording Data such as temperature of a place if it is sunny, fog, etc.

• PRECIPITATION: This Attribute shows the precipitation. This is one of the main Factor that is responsible in determining the adverse effects in the weather of a place.

SIZE OF DATASET:

No of bytes : 60 KB
No of tuples : 655

INFERENCES DRAWN:

1. The first inference we are drawing from our analysis on the weather data set is, temperature is directly proportional to the temperature of the male/female.

2. The second inference we are drawing is, the amount of temperature is inversely proportional to the male/female's body capacity.

3. The third inference we are making is, the age of a male/female is inversely proportional to their temperature capacity.

4. From weather dataset, we can approximate average temperature in that country, find out the lifestyle, etc.

5. From temperature data, we can use it in Medical Fields .

6. From the temperature values, we can predict earth's weather condition and predict if there may or may not occur a natural calamity.

Non-Linear Regression:

- Polynomial Regression can be modelled by adding Polynomial terms to the basic Linear model

- By applying transformations to the variables, we can convert the non-linear model into a linear one that can be solved by the method of least squares.

Transformation of a polynomial regression model to a linear regression model:

Consider a cubic polynomial relationship given by:

$y=w_0+w_1x+w_2x^2+w_3x^3$

To convert this equation to form, we define new variables:

$x_1=x$

$x_2=x^2$

$x_3=x^3$

First equation can then be converted to linear form by applying the above assignments, resulting in the equation

$y=w_0+w_1x+w_2x^2+w_3x^3$

which is easily solved by the method of least squares.

CODE:

```
#PREDICTING TEMPERATURE WRT PRECIPITATION BY RUNNING LINEAR REGRESSION MODEL###


rm(list = ls(all=TRUE))

weatherData= read.csv(file="~/Desktop/dm/ben.csv",header=T)


colnames(weatherData) #column names of the dataset
weatherData
summary(weatherData)  #shows minimum,max, mean , median of each columns


library(ggplot2)  #to access graphs

ggplot()+
```

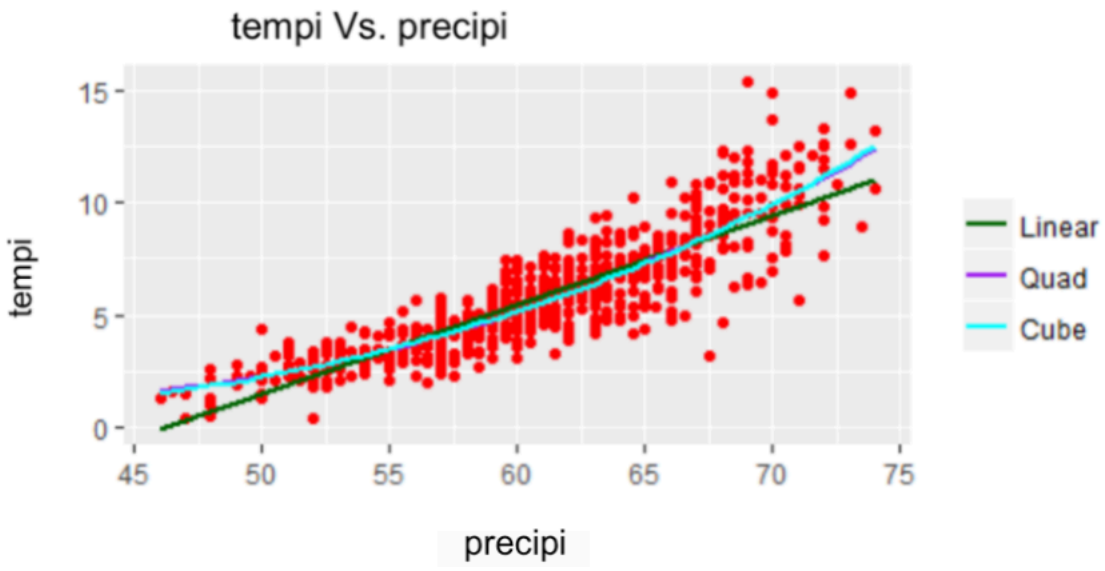
```
  geom_point(aes(x=weatherData$tempi, y=weatherData$precipi), color='red')+
  ggtitle('temperature Vs. precipitation')+
  xlab('temperature')+
  ylab('precipitation')   #basic plot between temp and prec

linear_regressor = lm(weatherData$tempi ~ weatherData$precipi) #code for linear regression
linear_regressor #predicting temp with respect to precipi
summary(linear_regressor) #after running the linear model, checking for r square value to check accuracy of the model

plot(linear_regressor) #plotting tempi against precipi after linear

ggplot()+
  geom_point(aes(x=weatherData$tempi, y=weatherData$precipi), color='red')+
  geom_line(aes(x=weatherData$tempi,y=predict(linear_regressor)),
  color='darkgreen',lwd=1)+
  ggtitle('temperature Vs. precipitation')+
  xlab('tempi')+
  ylab('precipi')

quad_regressor = lm(weatherData$tempi ~ poly(weatherData$precipi, degree = 2, raw = T))
quad_regressor
summary(quad_regressor)

plot(quad_regressor)

ggplot()+
  geom_point(aes(x=weatherData$tempi, y=weatherData$precipi), color='red')+
```

```
  geom_line(aes(x=weatherData$tempi, y=predict(linear_regressor), color = "Linear"), lwd=1)+
  geom_line(aes(x=weatherData$tempi, y=predict(quad_regressor), color = "Quad"), lwd=1)+
  scale_color_manual("", breaks = c("Linear", "Quad"),values = c("Linear"="darkgreen", "Quad"="purple"))+
  ggtitle('tempi Vs. precipitation')+
  xlab('tempi')+
  ylab('precipi')

cube_regressor = lm(weatherData$tempi~ poly(weatherData$precipi, degree = 3, raw = T))
cube_regressor
summary(cube_regressor)

ggplot()+
  geom_point(aes(x=weatherData$precipi, y=weatherData$tempi), color='red')+
  geom_line(aes(x=weatherData$precipi, y=predict(linear_regressor), color = "Linear"),lwd=1)+
  geom_line(aes(x=weatherData$precipi, y=predict(quad_regressor), color = "Quad"), lwd=1)+
  geom_line(aes(x=weatherData$precipi, y=predict(cube_regressor), color = "Cube"), lwd=1)+
  scale_color_manual("", breaks = c("Linear", "Quad", "Cube"),values = c("Linear"="darkgreen", "Quad"="purple", "Cube"="cyan"))+
  ggtitle('tempi Vs. precipi')+
  xlab('precipi')+
  ylab('tempi')

anova(quad_regressor,cube_regressor)
```

```
library(caTools)

#set.seed(47)
split <- sample.split(weatherData$tempi, SplitRatio = 0.8) #splitting the data into train and
test
split
training_set <- subset(weatherData,split==T)
training_set
testing_set <- subset(weatherData,split==F)
testing_set


poly_regressor = lm(formula = tempi ~ poly(precipi, degree = 2, raw = T), data =
training_set)
poly_regressor
summary(poly_regressor)


ggplot()+
  geom_point(aes(x=training_set$precipi,y=training_set$tempi),color='red')+
  geom_line(aes(x=training_set$precipi,y=predict(poly_regressor,newdata =
training_set)),color='blue', lwd=1)+
  ggtitle('tempi vs precipi(trainingset)')+
  xlab('tempi')+
  ylab('precipi')

ggplot()+
  geom_point(aes(x=testing_set$precipi,y=testing_set$tempi),color='green')+
  geom_line(aes(x=training_set$precipi,y=predict(poly_regressor,newdata =
training_set)),color='blue',lwd=1)+
  ggtitle('tempi vs precipi(testingset)')+
```

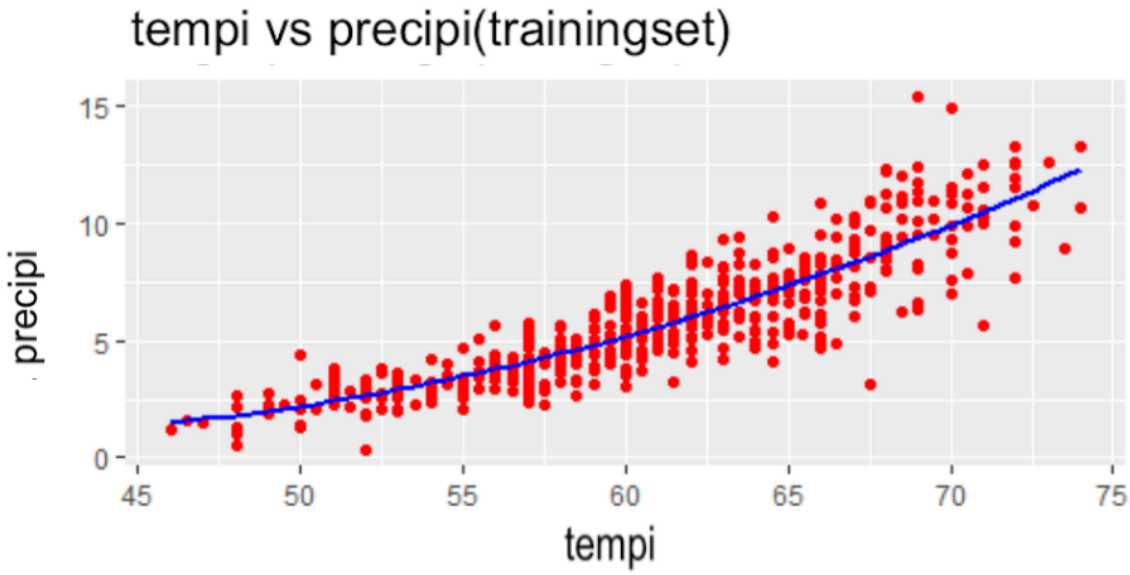
```
xlab('tempi')+
ylab('precipi')
```

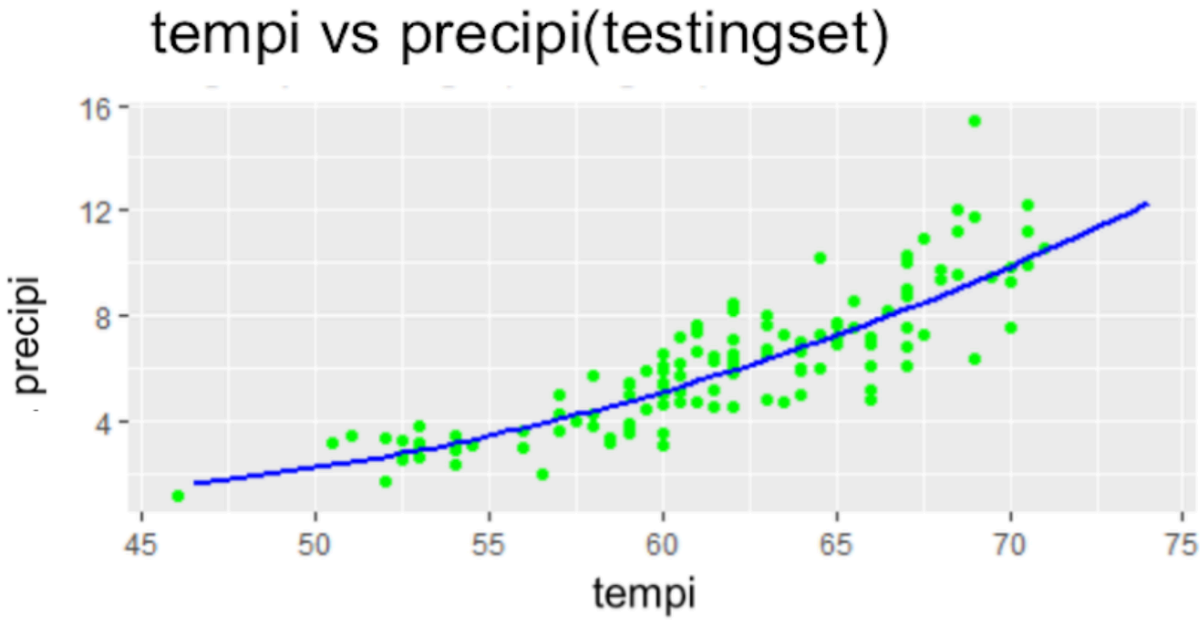
SNAPSHOT OF CODE WITH GRAPH:



Linear, Quadratic and Cubic graph:

Cubic equation gives a best fitted line than quadratic equation gives a best fitted line than linear equation. Hence we can conclude that





IMPACT:

Exploratory analysis has highlighted interesting relationship across variables. Such preliminary results give us hints for predictors selection and will be considered in the prosecution of this analysis.