University of the Philippines
Open University

Fundamentals of Analytics Modelling

A Business Analytics Course

Dr. Eugene Rex Jalao
Course Writer

University of the Philippines
OPEN UNIVERSITY

COMMISSION ON HIGHER EDUCATION

University of the Philippines
OPEN UNIVERSITY

UNIVERSITY OF THE PHILIPPINES OPEN UNIVERSITY

Fundamentals of Analytics Modelling
A Business Analytics Course

Welcome to the Fundamentals of Analytics Modelling course. This course provides students with an overview of the business concepts, frameworks, and algorithms in predictive analytics as it relates to decision making and its relationship with other analytics types and various software tools. The course will also provide understanding on predictive analytics tools/forecasting techniques to build and validate predictive models.

COURSE OBJECTIVES

At the end of the course, you should be able to:

1. Provide an overview of the latest industry trends of predictive analytics modelling through an introduction of predictive analytics, its relationship with other analytics types and various software tools.
2. Discuss key high level business concepts, frameworks, and algorithms in predictive analytics as it relates to decision making.
3. Identify appropriate forecasting techniques for different business problems.
4. Utilize predictive analytics tools/forecasting techniques to build and validate predictive models.
5. Design optimization and risk management models to provide the best decision
6. Evaluate the performance of the predictive model.
7. Develop an awareness of ethical norms as required under policies and applicable laws governing confidentiality and non-disclosure of data

COURSE OUTLINE

MODULE 1. Introduction to Predictive Analytics and Analytics Modelling Process

A. Definitions of Predictive Analytics
B. Predictive Analytics Framework

MODULE 2: Data Pre-Processing

MODULE 3: Supervised Learning

A. Classification
B. Regression

MODULE 4: Unsupervised Learning

A. Association Rule Mining
B. Sequential Pattern Mining
C. Clustering
D. Text Mining
E. Social Media Sentiment Analysis

MODULE 5: Introduction to Prescriptive Analytics

MODULE 6: Risk Analysis

MODULE 7: Ethics

COURSE MATERIALS

Your learning package for this course consists of:

1. Course guide;
2. Study guide for each module, which includes the lecture notes and learning activity guides;
3. Video lectures; and
4. Additional reading materials in digital form available on the course site.

These course materials are uploaded in the course site and can be downloaded for your reference.

STUDY GUIDE

The study schedule below will guide you on your pacing as you go through each part of the course/lesson and in doing the course requirements:

Date/ Period	Topic/s	Activity
Week 1	Introduction to Predictive Analytics	WATCH: • Introduction to Predictive Analytics (Dr. Jalao) • Predictive Analytics (Mr. Ligot) • Supervised Learning vs Unsupervised Learning (aProf. Pugoy) • Tools of Data Mining (Dr. Jalao) DO: • Discussion Forum
Week 2-3	Data Pre-Processing	WATCH: • Data Preprocessing (Dr. Jalao) DO: Case Study 1: Data Preprocessing using R and R Studio
Week 4-7	Supervised Learning: Classification	WATCH: • Introduction to Classification (Dr. Jalao) • Naive Bayes (Dr. Jalao) • Decision Trees (Dr. Jalao) • Nearest Neighbours (Dr. Jalao) • Artificial Neural Networks (Dr. Jalao) • Support Vector Machines (Dr. Jalao) • Ensembles (Dr. Jalao) • Random Forests (Dr. Jalao) • Model Evaluation (Dr. Jalao) DO: • Case Study 2: Classification using R and RStudio
Week 8-9	Supervised Learning: Regression	WATCH: • Regression (Dr. Jalao) • Regression Model Evaluation (Dr. Jalao) • Indicator Variables (Dr. Jalao) • Multicollinearity (Dr. Jalao)

		• Logistic Regression (Dr. Jalao) DO: • Case Study 3: Regression using R and RStudio
Week 10-12	Unsupervised Learning	WATCH: • Association Rule Mining (Dr. Jalao) • Sequential Pattern Mining (Dr. Jalao) • K-Means Clustering (Dr. Jalao) • Hierarchical Clustering (Dr. Jalao) • Text Mining (Dr. Jalao) • Social Media Sentiment Analysis (Dr. Jalao) DO: • Case Study 4: Text Mining using R and R Studio
Week 13	Introduction to Prescriptive Analytics and Operations Research	WATCH • "Introduction to Operations Research" by Prof. Ramon Miguel Panis, UPD
Week 14	Introduction to Risk Management	WATCH • Video on Analytics Application – Risk Management by Claire San Juan
Week 15	Applications and Deployment of Analytics Methodologies Ethics	WATCH: • Ethical Issues (Atty. Banez) • "Ethical Implications in Business Analytics" by Mr. Dominic Ligot
Week 16	Final Assessment	

COURSE REQUIREMENTS

To earn a certificate of completion for this course, you are required to do the following:

1. Submit 4 case studies.

2. Submit the final assessment.

MODULE 1: INTRODUCTION TO PREDICTIVE ANALYTICS AND ANALYTICS MODELLING

Introduction

This is the first module in the course. As such, it gives an introduction on what the students can expect to learn all throughout the learning period. Specifically, an overview on the relevant concepts and principles of predictive analytics is defined and identified.

Learning Objectives

After working on this module, you should be able to:

1. Define what predictive analytics is.
2. Discuss fundamental ideas, concepts, and techniques associated with Predictive Analytics.
3. Describe the Predictive Analytics Framework.

Self Study/Learning Resources

The student is expected to study the following resources:

1. Video on “Introduction to Predictive Analytics” by Dr. Eugene Rex Jalao.
2. Video on “Predictive Analytics” by Mr. Dominic Ligot.
3. Slides on “Introduction to Predictive Analytics” by Dr. Eugene Rex Jalao.
4. Slides on “Knowledge Discovery in Databases” (pp. 7-20) by Prof. Yang Hsin-Chang.

Activity 1-1

Discussion forum to spur exchange of ideas on what the students perceive as Predictive Analytics and how it could be applied to the student’s field of work and in other situations as well.

MODULE 2: DATA PRE-PROCESSING

Introduction

The task of data pre-processing is discussed in Module 2. Before proceeding to the predictive analytics proper, this module first discusses the significance and the methods to ensure that unprocessed data are complete, error-free, and consistent. After all, quality decisions must be based on quality data.

Learning Objectives

At the end of the module, students are expected to:

1. Explain the significance of data pre-processing.
2. Discuss and perform appropriate data pre-processing methods.

Self Study/Learning Resources

The student is expected to study the following resources:

1. Video on "Data Pre-Processing" by Dr. Eugene Rex Jalao
2. Video on "Predictive Analytics" by Mr. Dominic Ligot.
3. Slides on "Introduction to Predictive Analytics" by Dr. Eugene Rex Jalao.

Activity 2-1

Case 1: Preprocessing a Dataset using R and R Studio

- Materials:
  - R and R Studio Installed in a laptop
  - Bank Data in a CSV file

MODULE 3: SUPERVISED LEARNING

Introduction

After learning what predictive analytics/data mining and data pre-processing are, it is about time to proceed to the prediction process. There are two main categories of predictive analytics methodologies i.e. supervised learning and unsupervised learning. This module discusses supervised learning. Take note that supervised learning methodologies can be further classified to classification and regression.

Learning Objectives

At the end of the module, students are expected to:

1. Define supervised learning.
2. Differentiate classification from regression.
3. Identify and discuss appropriate supervised learning methodologies for various scenarios and business problems.
4. Build and validate predictive models by utilizing supervised learning methodologies.
5. Evaluate the performance of the predictive model.

3.1. Classification

Given a collection of past records or training data, the goal of classification is to predict the class variable (in other words, the actual class or category) by finding an appropriate model.

Self Study/Learning Resources

Based on the methodologies discussed by the faculty-in-charge, these are the resources available to the student:

1. Video on "Artificial Neural Networks" by Dr. Eugene Rex Jalao
2. Video on "Naive Bayes" by Dr. Eugene Rex Jalao
3. Video on "Support Vector Machines" by Dr. Eugene Rex Jalao
4. Video on "Nearest Neighbours" by Dr. Eugene Rex Jalao
5. Video on "Logistic Regression" by Dr. Eugene Rex Jalao
6. Video on "Ensembles" by Dr. Eugene Rex Jalao

7. Video on “Random Forests” by Dr. Eugene Rex Jalao
8. Video on “Model Evaluation” by Dr. Eugene Rex Jalao
9. Slides on “Classification Methodologies” by Dr. Eugene Rex Jalao
10. Slides on “Decision Trees” (pp. 40-53) by Prof. Yang Hsin-Chang

Activity 3-1

Case 2: Classification Using R and R Studio: “Churn Analysis”

- Materials:
  - R and R Studio Installed in a laptop
  - Churn Dataset in a CSV file

3.2. Regression

On the other hand, instead of predicting a class or category, regression predicts the actual value of a target based on one or more predictors.

Self Study/Learning Resources

The student is expected to study the following resources:

1. Video on “Regression” by Dr. Eugene Rex Jalao
2. Video on “Indicator Variables” by Dr. Eugene Rex Jalao
3. Video on “Multicollinearity” by Dr. Eugene Rex Jalao
4. Video on “Regression Model Evaluation” by Dr. Eugene Rex Jalao
5. Slides on “Regression Methodologies” by Dr. Eugene Rex Jalao

Activity 3-2

Case 3: Regression Using R and R Studio: “Predicting TV Advertising Revenue”

- Materials:
  - R and R Studio Installed in a laptop
  - TV Dataset in a CSV file

MODULE 4: UNSUPERVISED LEARNING

Introduction

This module discusses the other category of predictive analytics methodologies i.e. unsupervised learning.

Learning Objectives

At the end of the module, students are expected to:

1. Define unsupervised learning and differentiate it from supervised learning.
2. Identify and discuss appropriate unsupervised learning methodologies for various scenarios and business problems.
3. Build and validate predictive models by utilizing supervised learning methodologies.

Self Study/Learning Resources

The student is expected to study the following resources:

1. Video on "Association Rule Mining" by Dr. Eugene Rex Jalao
2. Video on "K-Means Clustering" by Dr. Eugene Rex Jalao
3. Video on "Hierarchical Clustering" by Dr. Eugene Rex Jalao
4. Video on "Text Mining" by Dr. Eugene Rex Jalao
5. Video on "Social Media Sentiment Analysis" by Dr. Eugene Rex Jalao
6. Video on "Sequential Pattern Mining" by Dr. Eugene Rex Jalao
7. Slides on "Unsupervised Learning Methodologies" by Dr. Eugene Rex Jalao.

Activity 4-1

Case 4: Text Mining Using R and R Studio

- Materials:
  - R and R Studio Installed in a laptop
  - Reviews Dataset in a CSV file

MODULE 5: INTRODUCTION TO PRESCRIPTIVE ANALYTICS AND OPERATIONS RESEARCH

Introduction

This module tackles prescriptive analytics which is operationalized by the multidisciplinary field of Operations Research and Optimization. Hence, this module in particular introduces topics relating to optimization models and algorithms, queueing and simulation.

Learning Objectives

At the end of the module, students are expected to:

1. Design optimization and risk management models to provide the best decision

Self Study/Learning Resources

The student is expected to study the following resources:

1. Watch: "Introduction to Operations Research" by Prof. Ramon Miguel Panis, UPD
2. Slides on "Introduction to Predictive Analytics" by Dr. Eugene Rex Jalao.

MODULE 6: RISK ANALYSIS

Introduction

Let us now learn how we can post-process and visualize the data inside the data warehouse.

Learning Objectives

After working on this module, you should be able to:

1. Discuss the various techniques used for post-processing of discovered structures and visualization.

Self Study/Learning Resources

The student is expected to study the following resources:

1. Watch: Video on Analytics Application – Risk Management by Claire San Juan
2. Slides on Risk Management by Claire San Juan

MODULE 7: ETHICS

Introduction

Finally, let us discuss the opportunities and ethics surrounding data warehousing.

Learning Objectives

At the end of the module, students are expected to:

1. Discuss the ethical norms as required under policies and applicable laws governing confidentiality and non-disclosure of data/information/documents and proper conduct in the learning process and application of business analytics.

Self Study/Learning Resources

The student is expected to study the following resources:

- Watch: Ethical Issues (Atty. Banez)
- Watch: "Ethical Implications in Business Analytics" by Mr. Dominic Ligot

Case Study 1
Data Preprocessing

1. Bank Data

The Bank Dataset contains 11 independent variables specifically age, region, income, sex, married, children, car, save_act, current_act, and mortgage and one response variable which answers the question: “Did the customer buy a PEP (Personal Equity Plan) after the last mailing?” with a yes/no response. We will analyze this data beforehand using descriptive analytics and preprocess the data for use in various data mining algorithms.

1. Generating Descriptive Analytics

2.1 Install R and R Studio

2.1.1. Download the R (Use the following link or download Latest Version)
2.1.2 https://cran.r-project.org/bin/windows/base/
2.1.3. Install R First.
2.1.4. Download R Studio: (Use the following link or download Latest Version)
2.1.5. https://www.rstudio.com/products/rstudio/download/
2.1.6. Install R Studio.

2.2 Initialize R: Setting the Working Directory

The working directory is the main directory in which R does analysis. Usually before starting any analysis with R, we set the working directory to a folder where all the data is stored.

2.2.1. Open R Studio

2.2.2. On the file explorer tab click on Files. 

2.2.3. Click on Explore 

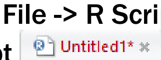
2.2.4. Go to the Desktop Folder -> Module 3 Datasets

2.2.5. Click on More. 

. Click on Set as Working Directory.

2.3 Load the Bank Dataset into R.

2.3.1. Click on File-> New File -> R Script.

2.3.2. In the new tab script 

, type the following code:

```
• bankdata = read.csv("bankdata.csv")
```

2.3.3. Put the cursor at the end of the code and click on Run 

. As a result, the data is loaded in the Environment

2.4 Descriptive Analytics and Visualization

We want to analyze the all attribute columns in terms of Mean, Standard Deviation, Median, Mode, Variance, Range, Minimum, Maximum, Sum and Count.

2.4.1. To calculate for the descriptive statistics, type the following lines of code.

```
• library(pastecs)
• options(scipen=100, digits=2)
• write.csv(stat.desc(bankdata), file =
  "NumericalStatistics.csv")
• write.csv(summary(bankdata), file =
  "CategoricalStatistics.csv")
```

2.4.2. Highlight all lines that were typed in step 2.3.1 and click on Run. As a result, the descriptive statistics results are saved in the Desktop -> Module 3 Datasets -> Case 1 Folder.

2.4.3. Answer the following questions:

2.4.3.1. What is the range of values of the Age variable? What is the minimum, maximum and middle value?

____

2.4.3.2. How many customers have a savings account? Current account?

____

2.4.4. Let's say we want a CrossTab Report for the relationship of the variable "Married" and the Number of Children. Type the following line of code.

```
• xtabs(~married+children,data=bankdata)
```

2.4.5. Highlight the line that was typed in step 2.3.4 and click on Run. Verify the result as follows:

```
         children
married   0  1  2  3
    NO   83 46 50 25
    YES 180 89 84 43
```

2.4.6. Now, we would like to calculate the means of Age, Income and Children by PEP, Married and has Car. To do this, type the following lines of code

```
• install.packages("reshape")
• library(reshape)
• bankdata.m = melt(bankdata, id=c("pep","married", "car"),
  measure=c("age", "income", "children"))
• bankdata.c  = cast(bankdata.m, pep + married + car ~
  variable, mean)
• write.csv(bankdata.c , file = "bankdataByPepStatusCar.csv")
```

2.4.7. Highlight all lines that where typed in step 2.3.6 and click on Run. As a result, the pivot analysis results are saved in the Desktop -> Module 3 Datasets -> Case 1 Folder. Verify the result by opening the file bankdataByPepStatuscar.csv.

	pep	married	car	age	income	children
1	NO	NO	NO	36.54762	21758.49	1.5
2	NO	NO	YES	39.2619	23635.47	1.5
3	NO	YES	NO	40.12698	25031.17	0.833333
4	NO	YES	YES	41.65517	26355.49	1.008621
5	YES	NO	NO	44.38333	30565.43	0.8
6	YES	NO	YES	45.98333	31752.53	0.783333
7	YES	YES	NO	43.39474	28293.14	1.052632
8	YES	YES	YES	46.73077	32145.53	1.076923

2.4.7.1. As Age increases, what pattern do you see in terms of buying a PEP?

____

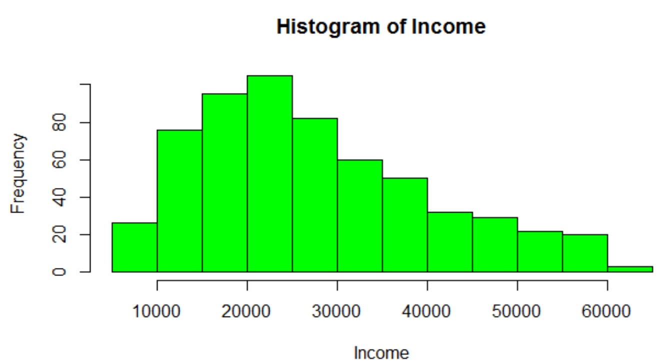
2.4.7.2. In terms of the number of children what pattern do you see in terms of buying a PEP? For Being Married?

____

2.4.8. Now, we would like to calculate for a histogram of the Income variable. Type the following code:

```
• hist(bankdata$income,breaks=15,
  col="green",xlab="Income",main="Histogram of Income")
```

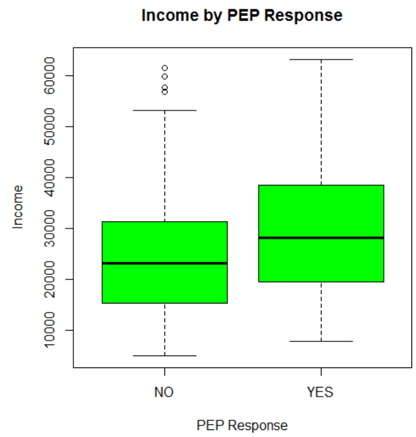
2.4.9. Highlight the line that where typed in step 2.3.8 and click on Run. Click on Zoom to view the result. The histogram should look like this:



2.4.10. We would like to calculate for a box plot of the Income variable by PEP. Type the following code:

```
• boxplot(income~pep,data=bankdata, main="Income by PEP
  Response", xlab="PEP Response", ylab="Income", col="green")
```

2.4.11. Highlight the line that where typed in step 2.3.10 and click on Run. Click on Zoom to view the result. The box plot should look like this:



2.4.11.1. What can you generalize from this Box Plot?

____

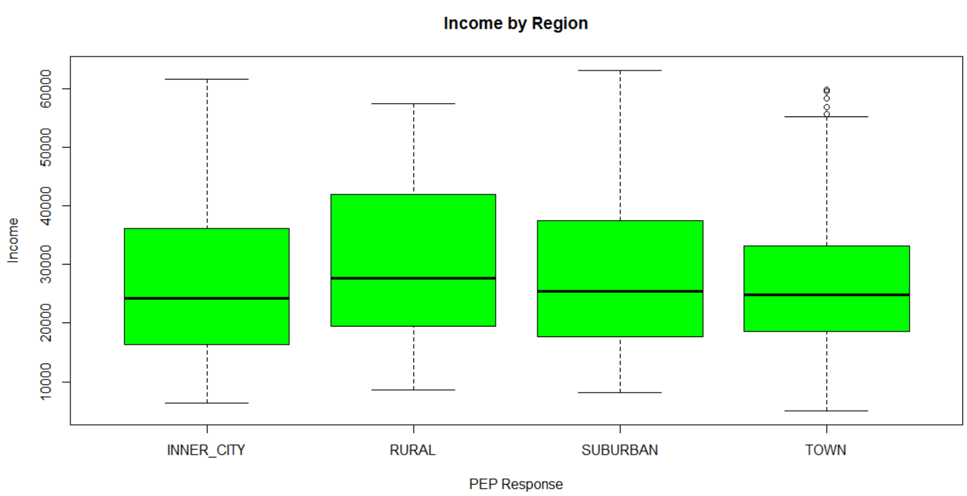
2.4.11.2. Are there any outliers?

____

2.4.12. We would like to generate a box plot of the Income variable by region. Type the following code:

```
• boxplot(income~region,data=bankdata, main="Income by
  Region", xlab="Region", ylab="Income", col="green")
```

2.4.13. Highlight the line that where typed in step 2.3.12 and click on Run Run. Click on Zoom Zoom to view the result. The box plot should look like this:



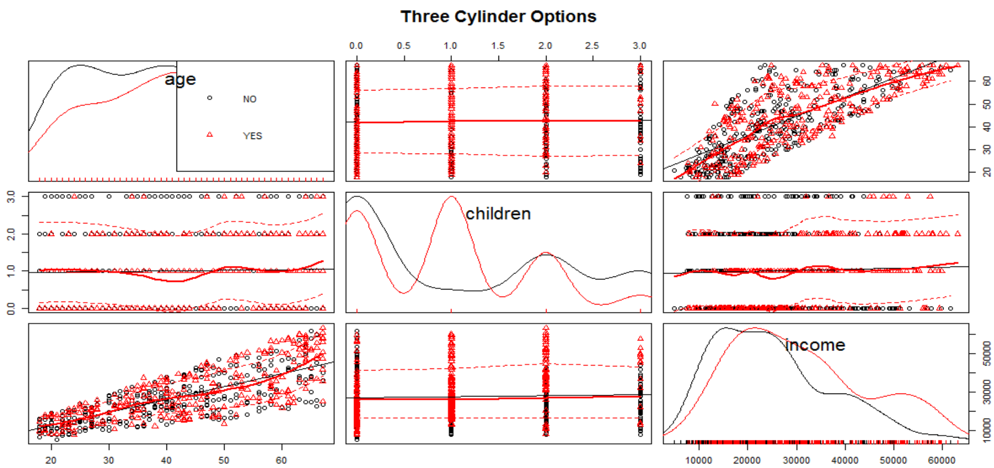
2.4.13.1. What can you generalize from this Box Plot?

____

2.4.14. To plot a scatter plot matrix of the Income, Age and # of Children, type the following code:

```
install.packages(“car”)
library(car)
scatterplotMatrix(~age+children+income|pep,data=bankdata,main="Age Children and Income by PEP")
```

2.4.15. Highlight the line that where typed in step 2.3.14 and click on Run . Click on Zoom to view the result. The box plot should look like this:



2.4.15.1. What can you generalize from this Scatter Plot?

____

2.5.Data Transformation

We want to transform certain variables into a different format as an input to the various data mining methodologies.

2.5.1. To Normalize the Income Column into a [0,1] scale, type the following code:

```
IncomeData = bankdata[,5]
NormalizedIncomeData = (IncomeData-min(IncomeData))/(max(IncomeData)-min(IncomeData))
bankdata = cbind(bankdata,NormalizedIncomeData )
View(bankdata)
```

2.5.2. Highlight the line that where typed in step 2.4.1 and click on Run . The result of the code from step 2.4.1 is the same bankdata but with a new column NormalizedIncomeData at the End.

2.5.3. Suppose that we want to create an equal depth(frequency) variable for Income where the new variable could take in “Low”, “Medium” and “High.” Type the following code:

```
• bins=3
• cutpoints=quantile(IncomeData,(0:bins)/bins)
• DiscreteIncome =cut(IncomeData,  cutpoints,
  include.lowest=TRUE, dig.lab=5, labels=c("Low", "Med",
  "High"))
• bankdata = cbind(bankdata, DiscreteIncome)
• View(bankdata)
```

2.5.4. Highlight the line that where typed in step 2.4.3 and click on Run 

. The result of the code from step 2.4.3 is the same bankdata but with a new column DiscreteIncome at the end representing the discretized Income.

2.5.5. Suppose that we want to create dummy variables for the four values of Region. Type the following code:

```
• indicators=model.matrix( ~ region - 1, data = bankdata)
• bankdata = cbind(bankdata,indicators)
• View(bankdata)
```

2.5.6. Highlight the line that where typed in step 2.4.5 and click on Run 

. The result of the code from step 2.4.5 is the same bankdata but with four new columns specifically regionINNER_CITY, regionRURAL, regionSUBURBAN, regionTOWN.

2.6. Data Sampling

2.6.1. To sample 100 rows of the bank data without replacement type the following:

```
• Samplebankdata = bankdata[sample(nrow(bankdata),100,replace
  = FALSE),]
• View(Samplebankdata)
```

2.6.2. The result is a subset sample of the bankdata dataset.

Case 2
Selecting the Best Classification Model Using R

1. Churn Data Introduction

Churn rate is also sometimes called attrition rate. It is one of two primary factors that determine the steady-state level of customers a business will support. In its broadest sense, churn rate is a measure of the number of individuals or items moving into or out of a collection over a specific period of time. This data set contains a total of 3333 mobile subscriber mobile plans. There are a 17 attributes that might affect churn. The classification goal is to predict if the customer will churn(y) or not (n) as well as to identify business rules that can help minimize customer attrition.

1. Predictor Variables

- Account Length: length of time in days the customer is using the plan.
- Int'l Plan: plan has an international promo.
- VMail Plan: plan has a voicemail booster.
- VMail Message: number of voice mail messages received
- Day Mins: number of day minutes called (6am – 6pm)
- Day Calls: number of calls made
- Day Charge: total cost of day calls in USD
- Eve Mins: number of eve minutes called (6pm-12 midnight)
- Eve Calls: number of eve calls made
- Eve Charge: total cost of eve calls in USD
- Night Mins: number of night minutes called (12 midnight-6am)
- Night Calls: number of night calls made
- Night Charge: total cost of night calls in USD
- Intl Mins: number of international minutes called
- Intl Calls: number of international calls
- Intl Charge: total cost of international calls
- CustServ Calls: number of calls to call center for service support

1. Modeling

For each of the models here in this section, create the specified model and utilize 10-fold cross validation to fill in the requested information about the model.

2.1. Modeling a Decision Tree

Create a decision tree for the Churn Dataset using the J48 command. Summarize the needed information as follows:

2.1.1. Accuracy: ________________________

2.1.2. Confusion Matrix:

	Predicted False.	Predicted True.
Actual False.		
Actual True.		

2.1.3. True Positive Rate of Churn=True Class: ____________
2.1.4. Precision of Churn=True Class: _______________
2.1.5. ROC Area of Churn=True Class: _______________

2.2. Creating a Rule Based Classifier

Create a rule-based classifier for the Churn Dataset using the JRip command. Summarize the needed information as follows:

2.2.1. Accuracy: _________________________
2.2.2. Confusion Matrix:

	Predicted False.	Predicted True.
Actual False.		
Actual True.		

2.2.3. True Positive Rate of Churn=True Class: ____________
2.2.4. Precision of Churn=True Class: _______________
2.2.5. ROC Area of Churn=True Class: _______________

2.3. Creating an ANN

Create an ANN classifier for the Churn Dataset using the MLP command. Summarize the needed information as follows:

2.3.1. Accuracy: _________________________
2.3.2. Confusion Matrix:

	Predicted False.	Predicted True.
Actual False.		
Actual True.		

2.3.3. True Positive Rate of Churn=True Class: ____________
2.3.4. Precision of Churn=True Class: _______________
2.3.5. ROC Area of Churn=True Class: _______________

2.4. Creating an Adaboost Learner with Rule Classifiers

Create an Adaboost Classifier with Rule Based Classifiers for the Churn Dataset using the AdaboostM1 + JRip command. Summarize the needed information as follows:

2.4.1. Accuracy: ____________________________

2.4.2. Confusion Matrix:

	Predicted False.	Predicted True.
Actual False.		
Actual True.		

2.4.3. True Positive Rate of Churn=True Class: _____________

2.4.4. Precision of Churn=True Class: ________________

2.4.5. ROC Area of Churn=True Class: ________________

2.5. Creating a Random Forest Model

Create a Random Forest Classifier for the Churn Dataset using the RF command. Summarize the needed information as follows:

2.5.1. Accuracy: ____________________________

2.5.2. Confusion Matrix:

	Predicted False.	Predicted True.
Actual False.		
Actual True.		

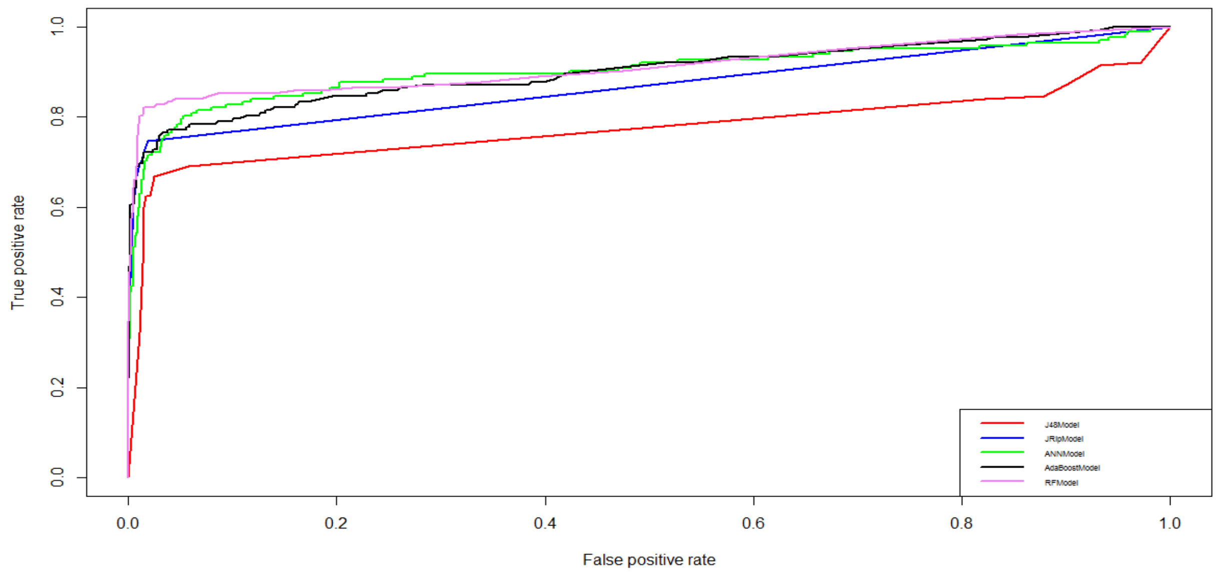
2.5.3. True Positive Rate of Churn=True Class: _____________

2.5.4. Precision of Churn=True Class: ________________

2.5.5. ROC Area of Churn=True Class: ________________

1. ROC Curves

Create an ROC curve for the 5 models using Training and Testing Data. Utilize 67%-33% mix of the data. Choose the Churn = True class. For replicable results, please utilize set.seed(123). The ROC curve should look something like this:



While model will you choose?

____

Case Study 3
Regression Modelling

1. TV Dataset

Jalao (2012) proposed a regression model to predict the revenue of advertising for a 30 second primetime TV show slot. Significant factors that affect the revenue of advertising were also determined. Data was obtained and compiled from multiple websites that provide information that could potentially affect the revenue of advertising. Moreover, the effect of several social media websites on the revenue of advertising was also studied.

1. Data Set Description

Table 1: Data Description and Modelling

Variable	Description	Source	Model
Revenue (Response)	Average Revenue of Advertising in a 30 second primetime advertisement slot in USD	adage.com	Continuous (Response)
Length	Either 30 minutes or 1 hour Broadcast time	Show official website site	Continuous
Viewers	Nielsen Average Number of Viewers for 2011-2012 Season	deadline.com	Continuous
18-49 Rating	Nielsen Average 18-49 Demographic Rating Share in % for 2011-2012 Season	deadline.com	Continuous
Facebook	Number of Facebook Likes from official show Facebook page	Show's official Facebook Page	Continuous
Facebook Talking About	Number of Active Social Media users talking about the show on Facebook	Show's official Facebook Page	Continuous
Twitter	Number of Tweeter Followers from official tweeter pages	Show's official Twitter Page	Continuous
Age	Number of Episodes Aired	Show official website	Continuous
Network	Network that broadcasts the show: ABC, CBS, CW, Fox or NBC. Baseline is CW since it has the lowest average revenue of advertising for all shows.	Show official website	Network_ABC={1 if show is in ABC 0 o/w Network_CBS={1 if show is in CBS 0 o/w Network_Fox={1 if show is in Fox 0 o/w Network_NBC={1 if show is in Fox 0 o/w
Day	Day of show broadcast, Sunday through Friday. No data points for Saturday. Baseline is Friday since it has the lowest average revenue of advertising for all shows.	Show official website	Day_Su={1 if show is on Sunday 0 o/w Day_M={1 if show is on Monday 0 o/w

			Day_T={1 if show is on Tuesday 0 o/w Day_W={1 if show is on Wednesday 0 o/w Day_Th={1 if show is on Thursday 0 o/w
Type	Type of Show: Drama, Sit-com, Sports or Reality TV. Baseline is Reality TV.	Show official website	Type_D={1 if show is a Drama 0 o/w Type_C={1 if show is a sitcom 0 o/w Type_S={1 if show is Sport event 0 o/w

1. Loading Data to R Studio

3.1. Initialize R: Setting Working Directory

3.1.1. Open R Studio

3.1.2. On the file explorer tab click on Files.

3.1.3. Click on Explore

3.1.4. Go to the Desktop Folder -> Module 3 Datasets -> Case 3

3.1.5. Click on More. . Click on Set as Working Directory.

3.2. Load Bank Dataset into R.

3.2.1. Click on File-> New File -> R Script.

3.2.2. In the new tab script , type the following code:

```
• options(scipen=999,digits=2)
• tvdataset = read.csv("tvdataset.csv")
```

3.3. Highlight the two lines and click on Run . As a result, the data is loaded in the Environment

3.4. Fitting the Full Model

3.4.1. In the new tab script , type the following code:

```
• tvdataset.fit =lm(cost~network + day + length + d1849rating
  + facebooklikes + facebooktalkingabout + twitter+ age +
  type, data= tvdataset)
• summary(tvdataset.fit)
```

3.4.2. Highlight the two lines of code and click on Run .

3.4.3. The result of the linear regression fit would be as follows:

```
Call:
lm(formula = Cost ~ Network + Day + Length + D1849Rating + FacebookLikes +
    FacebookTalkingAbout + Twitter + Age + Type, data = TvDataSet)

Residuals:
   Min     1Q Median     3Q    Max
-67361 -22593    471  19219  89728

Coefficients:
                          Estimate    Std. Error t value    Pr(>|t|)
(Intercept)           28562.183987  28638.086708    1.00      0.3235
NetworkCBS           -35573.726314  14046.278717   -2.53      0.0146 *
NetworkCW              3105.565017  23991.005744    0.13      0.8975
NetworkFOX            43801.861329  16454.035539    2.66      0.0105 *
NetworkNBC            12614.802349  19063.074859    0.66      0.5112
DayM                  40142.310809  19080.878228    2.10      0.0406 *
DaySU                 59872.262811  20165.986209    2.97      0.0046 **
DayT                  38785.982953  18658.280925    2.08      0.0429 *
DayTH                 52198.450242  17776.564069    2.94      0.0050 **
DayW                  49756.761436  17266.720826    2.88      0.0059 **
Length                 -785.750218    450.461212   -1.74      0.0874 .
D1849Rating           17979.931421   2919.571073    6.16 0.00000013 ***
FacebookLikes             0.001872      0.000977    1.92      0.0613 .
FacebookTalkingAbout     -0.192229      0.103268   -1.86      0.0687 .
Twitter                   0.042084      0.018394    2.29      0.0265 *
Age                      91.473764     55.252155    1.66      0.1042
TypeD                -23500.818469  17952.191859   -1.31      0.1966
TypeN                -43507.191268  34377.164223   -1.27      0.2116
TypeR                -18827.492731  25326.774305   -0.74      0.4608
TypeS                160657.211898  67941.452663    2.36      0.0221 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36900 on 49 degrees of freedom
Multiple R-squared:  0.868, Adjusted R-squared:  0.817
F-statistic: 16.9 on 19 and 49 DF,  p-value: 0.00000000000000218
```

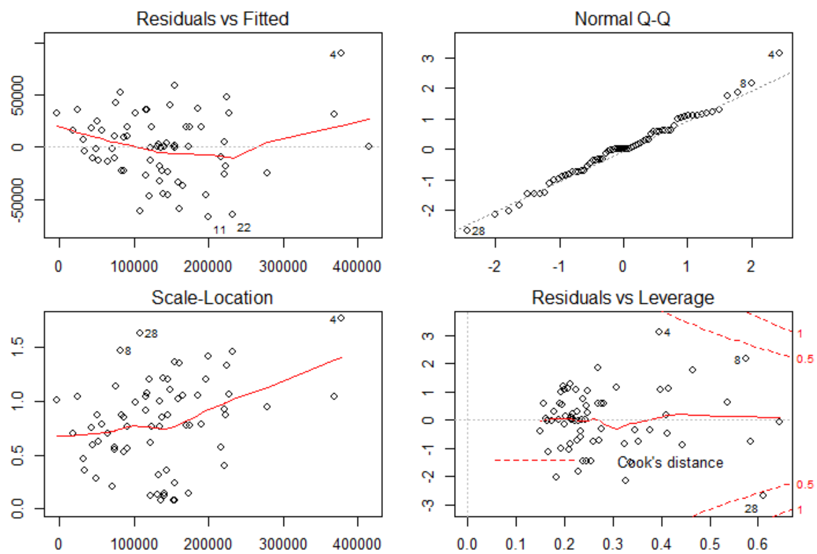
4. Model Adequacy Checking

4.1. To check for diagnostics as well as studentized residuals and Leverage (hat values) we type the following.

- `par(mfrow =c(2,2),mar=c(2,2,2,2))`
- `plot(tvdataset.fit)`
- `rstudent(tvdataset.fit)`
- `hatvalues(tvdataset.fit)`

4.2. Highlight these lines of code and click on Run .

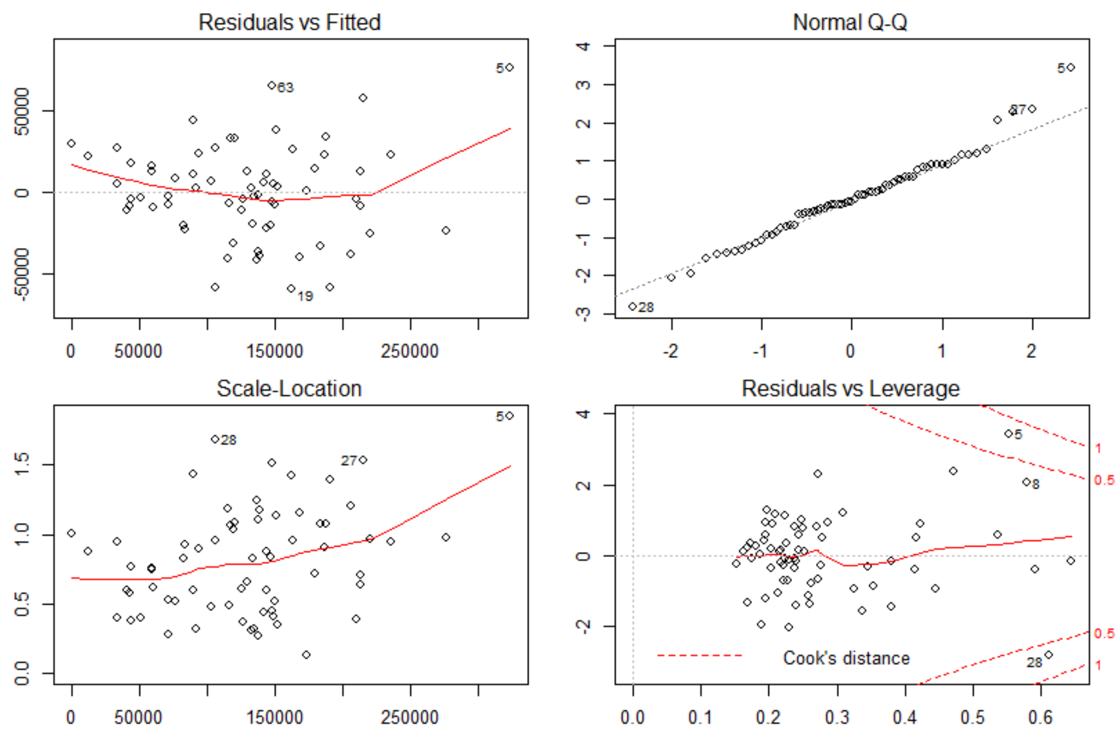
```
> rstudent(TvDataSet.fit)
     1      2      3      4      5      6      7      8      9     10     11     12     13     14     15     16     17     18
 0.208  0.609  0.094  3.461  1.101  0.021 -0.323  2.256 -0.041 -1.472 -2.088  0.016 -0.006  0.278 -0.826  0.014 -1.115  0.121
    19     20     21     22     23     24     25     26     27     28     29     30     31     32     33     34     35     36
-1.877  0.300 -0.761 -2.232 -0.480  0.605  1.142 -0.014  1.804 -2.850  1.117  0.985 -0.126 -0.747  0.018  0.161 -0.715 -1.502
    37     38     39     40     41     42     43     44     45     46     47     48     49     50     51     52     53     54
-0.993 -0.349  1.079  1.221  0.601 -1.461 -0.006  0.557 -0.382  0.309  0.577 -1.033  1.093 -0.572 -0.758  0.491    NaN  1.020
    55     56     57     58     59     60     61     62     63     64     65     66     67     68     69
 0.006 -0.726 -1.475 -0.854  0.598  0.572 -0.370  0.056  1.926 -0.897  1.303 -0.326  1.167  0.764 -0.077
> hatvalues(TvDataSet.fit)
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32
0.2 0.5 0.2 0.4 0.4 0.2 0.3 0.6 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.6 0.3 0.4 0.3 0.4 0.2 0.5 0.6 0.2 0.2 0.2 0.3
 33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
0.2 0.4 0.2 0.3 0.2 0.3 0.2 0.2 0.3 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.4 0.2 1.0 0.2 0.2 0.3 0.3 0.3 0.3 0.2 0.1 0.2 0.3 0.4
 65  66  67  68  69
0.2 0.4 0.3 0.2 0.6
```



4.3. It seems that observations 4, and 53 are outliers. We thus eliminate these rows, refit the regression model and plots as follows:

```
• reducedtvdataset=tvdataset[-c(4, 53), ]
• reducedtvdataset.fit =lm(cost~network + day + length +
  d1849rating + facebooklikes + facebooktalkingabout +twitter+
  age + type, data= reducedtvdataset)
• summary(reducedtvdataset.fit)
• par(mfrow =c(2,2),mar=c(2,2,2,2))
• plot(reducedtvdataset.fit)
```

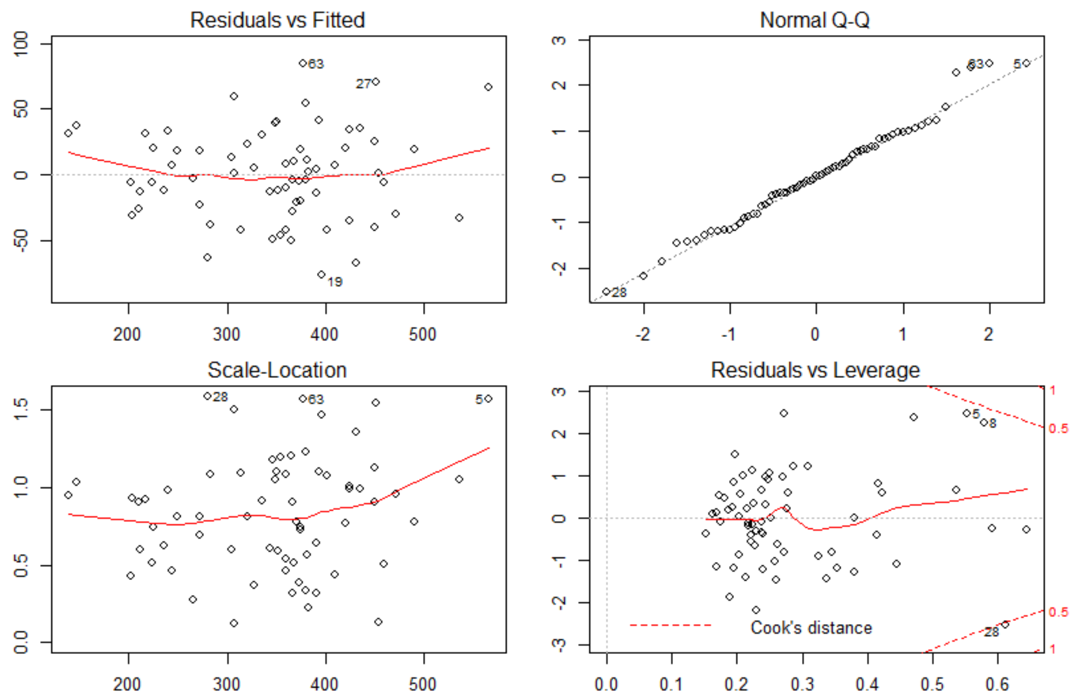
4.4. Highlight these lines of code and click on Run Run.



4.5. Based on the Residuals vs. Fitted graph, the constant variance assumption does not hold. We then transform the Cost variable as follows:

```
• #Transform Data Squareroot
• reducedtvdataset.fit =lm(cost^0.5~network + day + length +
  d1849rating + facebooklikes + facebooktalkingabout +twitter
  + age + type, data= reducedtvdataset)
• par(mfrow =c(2,2),mar=c(2,2,2,2))
• plot(reducedtvdataset.fit)
```

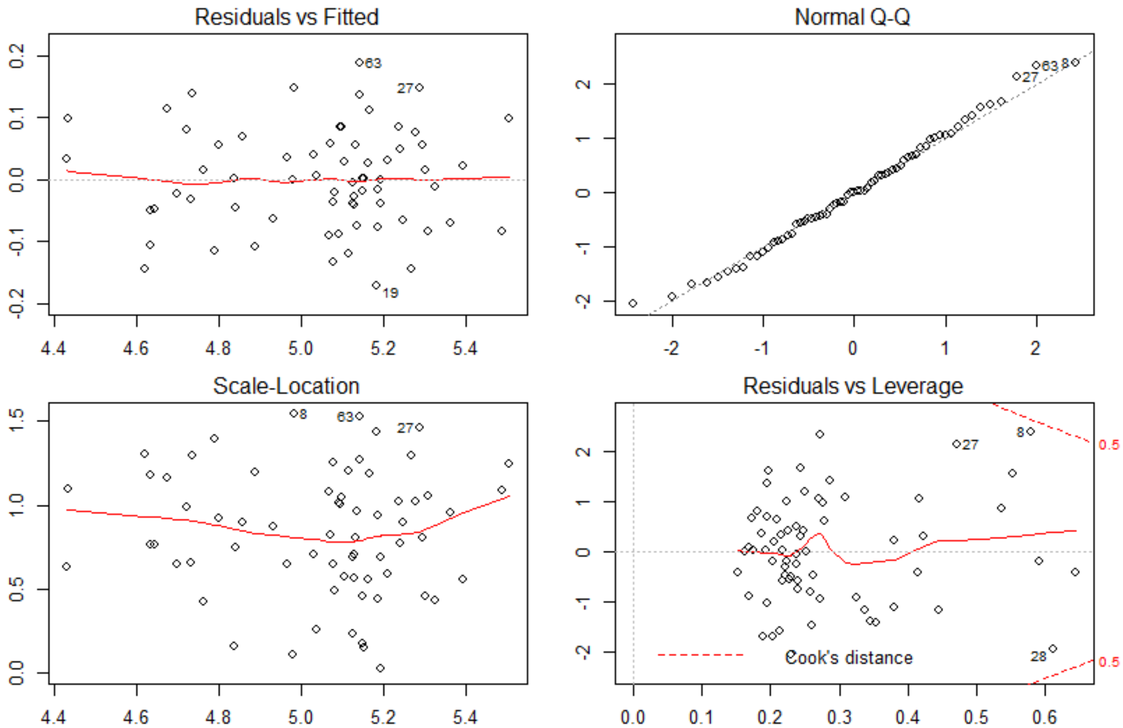
4.6. The Residuals vs Fits plot for Square Root transformation is as follows:



4.7. Based on the Residuals vs. Fitted graph, the constant variance assumption still does not hold. We further transform the Cost variable further as follows:

```
• #Transform Data Log10
• reducedtvdataset.fit =lm(log10(cost)~network + day + length
  + d1849rating + facebooklikes + facebooktalkingabout +
  twitter + age + type, data= reducedtvdataset)
• par(mfrow =c(2,2),mar=c(2,2,2,2))
• plot(reducedtvdataset.fit)
```

4.8. **The plot for Log10 transformation is as follows:**



4.9. **Based on the graph, the constant variance assumption holds.**

5. Variable Selection

5.1. We now choose the most relevant variables for the regression model. Type the following code and run it.

```
1.    base.fit =lm(log10(cost)~1, data= reducedtvdataset)
1.    forward = step(base.fit, scope =
list(lower=~1,upper=~network + day + length + d1849rating +
facebooklikes + facebooktalkingabout +twitter+ age + type),
direction = "both", trace=1)
1.    summary(forward)
```

5.2. The result of the regression model is as follows:

```
1.    Start:  AIC=-183
1.    log10(cost) ~ 1
1.
1.                               Df  Sum of Sq  RSS   AIC
1.    + network                   4      2.307 1.93 -228
1.    + d1849rating               1      1.557 2.68 -212
1.    + day                       5      1.357 2.88 -199
1.    + facebooklikes             1      0.763 3.48 -194
```

```
+ facebooktalkingabout  1      0.648 3.59 -192
+ type                  3      0.704 3.54 -189
+ twitter               1      0.355 3.89 -187
+ age                   1      0.173 4.07 -184
+ length                1      0.138 4.10 -183
<none>                               4.24 -183

#Deleted Results Here…

#Final Model Results:
Step:  AIC=-304
log10(cost) ~ network + day + facebooklikes + d1849rating + 
length + 
    twitter

                        Df Sum of Sq    RSS   AIC
<none>                                0.470 -304
- twitter               1      0.017 0.487 -304
+ age                   1      0.008 0.462 -304
- length                1      0.027 0.497 -302
+ facebooktalkingabout  1      0.000 0.470 -302
+ type                  3      0.023 0.448 -302
- facebooklikes         1      0.065 0.535 -298
- d1849rating           1      0.162 0.632 -286
- network               4      0.591 1.061 -258
- day                   5      0.671 1.141 -255
```

6. Fitting the Final Model:

6.1. Type the following code to determine the final regression model:

```
• Finaltvdataset.fit =lm(log10(cost)~network + day
  +length+  d1849rating + facebooklikes + twitter, data=
  reducedtvdataset)
• summary(Finaltvdataset.fit)
```

6.2. Run these lines of code and the results of the regression modelling would be as follows:

```
Call:
lm(formula = log10(cost) ~ network + day + length + d1849rating +
    facebooklikes + twitter, data = reducedtvdataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17946 -0.05854 -0.00237  0.06284  0.19115

Coefficients:
                   Estimate     Std. Error t value            Pr(>|t|)
(Intercept)    4.75278215191  0.06164145390   77.10 < 0.0000000000000002 ***
networkCBS    -0.07073656502  0.03463270668   -2.04               0.0461 *
networkCW     -0.33298729975  0.05984208290   -5.56        0.00000088458 ***
networkFOX     0.03982952602  0.04235347786    0.94               0.3513
networkNBC    -0.05526145744  0.04590072167   -1.20               0.2340
dayM           0.26465160544  0.04280396946    6.18        0.00000009251 ***
daySU          0.29540731482  0.04668271676    6.33        0.00000005421 ***
dayT           0.23723645687  0.04206255559    5.64        0.00000067261 ***
dayTH          0.30246690633  0.04006455902    7.55        0.00000000059 ***
dayW           0.27407535455  0.03953369718    6.93        0.00000000579 ***
length        -0.00145359477  0.00083127230   -1.75               0.0861 .
d1849rating    0.03062687592  0.00717434530    4.27        0.00008163732 ***
facebooklikes  0.00000000418  0.00000000155    2.70               0.0094 **
twitter        0.00000006288  0.00000004591    1.37               0.1766
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09 on 53 degrees of freedom
Multiple R-squared:  0.889,     Adjusted R-squared:  0.862
F-statistic: 32.7 on 13 and 53 DF,  p-value: <0.0000000000000002
```

7. Fitting the Final Model with Standardized Coefficients:

7.1. Type the following code to determine the final regression model:

```
• #Convert To Numerical
• #Network
• networkind =model.matrix( ~ network - 1, data =
  reducedtvdataset)
• #Set CW as Baseline
• networkind = subset(networkind, select = -c(networkCW) )
• #Day
• dayind =model.matrix( ~ day - 1, data = reducedtvdataset)
• dayind = subset(dayind, select = -c(dayF) )
• x = cbind(subset(reducedtvdataset, select =
  c(3,6,8,9,11)),networkind,dayind)
• z = data.frame(scale(x, center = TRUE, scale = TRUE))
• z$cost = scale(log10(x$cost), center = TRUE, scale = TRUE)
• standardizedfinaltvdataset.fit =lm(cost~., data= z)
• summary(standardizedfinaltvdataset.fit)
```

7.2. **Run these lines of code and the results of the regression modelling would be as follows:**

```
> summary(standardizedfinaltvdataset.fit)

Call:
lm(formula = cost ~ ., data = z)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7080 -0.2309 -0.0093  0.2479  0.7541 

Coefficients:
                         Estimate           Std. Error t value      Pr(>|t|)    
(Intercept)    -0.00000000000000206 0.04539110817968406    0.00        1.0000    
length         -0.10039397880233666 0.05741265422259709   -1.75        0.0861 .  
d1849rating     0.35540819462595830 0.08325436513685587    4.27 0.00008163732 ***
facebooklikes   0.17832075023559130 0.06612794859581098    2.70        0.0094 ** 
twitter         0.07932897569036329 0.05792764102414330    1.37        0.1766    
networkABC      0.52339166391349790 0.09406018596782995    5.56 0.00000088458 ***
networkCBS      0.47699213263113960 0.11123712173298030    4.29 0.00007659521 ***
networkFOX      0.60245974368967625 0.08237909330971163    7.31 0.0000000141 ***
networkNBC      0.39335465681816678 0.07013399723016814    5.61 0.00000075387 ***
dayM            0.37483707158197810 0.06062504150561469    6.18 0.00000009251 ***
daySU           0.40039354941947136 0.06327351328298184    6.33 0.00000005421 ***
dayT            0.34930303641890076 0.06193221135108153    5.64 0.00000067261 ***
dayTH           0.47541950549368939 0.06297374172367007    7.55 0.0000000059 ***
dayW            0.43079347459403744 0.06213932952311195    6.93 0.0000000579 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4 on 53 degrees of freedom
Multiple R-squared:  0.889,	Adjusted R-squared:  0.862 
F-statistic: 32.7 on 13 and 53 DF,  p-value: <0.0000000000000002
```

7.2.3. **Which variable is the most influential in terms of predicting revenue?**

____

Case 4
Text Mining using R

1. Introduction

- We are to dentify the most common words in a sample of 1000 reviews of popular fee apps from the iTunes Store

2. Text Mining Using R

2.1. Open RStudio from the programs menu.

2.2. Click on New Script, then type the following lines of code.

```
• library(wordcloud)
• library(tm)
• reviews <- read.csv("reviews.csv", stringsAsFactors=FALSE)
• review_source <- VectorSource(reviews$text)
• corpus <- Corpus(review_source)
• summary(corpus)
• corpus <- tm_map(corpus, content_transformer(tolower))
• corpus <- tm_map(corpus, removePunctuation)
• corpus <- tm_map(corpus, stripWhitespace)
• corpus <- tm_map(corpus, removeWords, stopwords("english"))
• corpus <- tm_map(corpus, removeWords, c("game"))
• dtm <- DocumentTermMatrix(corpus)
• dtm2 <- as.matrix(dtm)
• frequency <- colSums(dtm2)
• frequency <- sort(frequency, decreasing=TRUE)
• head(frequency,14)
• words <- names(frequency)
• wordcloud(words[1:100],
  frequency[1:100],colors=brewer.pal(8, "Dark2"))
```


graphics
much
make
doesnt
best
play
just
addictive
spend
people
one
easy
keep
please
lot
playing
cool
see
hard
love
thing
gold
problem
way
money
things
addicting
stop
get
casino
day
energy
many
got
even
little
still
fix
subway
needs
without
takes
app
work
every
real
far
stars
try
update
ever
times
level
add
recommend
better
getting
nice
now
addicted
can
think
never
want
able
gems
great
also
farm
fun
always
new
will
facebook
played
long
use
back
hours
free
buy
good
time
ive
crashes
amazing
take
enjoy
like
awesome
give
crashing
coins
games
cant
dont
keeps
first
really
need

Data Warehousing Assessment Exam

Data Warehousing Assessment Exam

1. A Data Warehouse has the following properties except:

a. Used for decision making b. Contains a lot of history

c. Integrates many data into one d. **All of the above are properties**

2. Which of the following organizational groups has the primary responsibility of maintaining a Data Warehouse?

a. Managers **b.** Analysts

c. **IT** **d.** Executives

3. Which of the following is not a valid goal when implementing a Data Warehouse?

a. Empowers analysts to do reporting b. Excellent Return of Investment

c. To have one version of the truth d. **All of the above are valid goals**

4. Which of the following is the top complaint when implementing a Data Warehouse?

a. Dirty Data b. **Lack of Data**

c. Not leveraging enough sources d. Data ownership

5. What is the most important feature that a Data Warehouse must have?

a. Integration to current IT environment b. **Query Performance**

c. Scalability d. Support for Open Source

6. Which of the following are valid sources of data for a Data Warehouse?

a. Point-of-Sale database b. A Data Mart

c. Another Data Warehouse d. **All of the above**

7. When comparing a data warehouse to a source database, which of the following is true?

a. Size of a Data Warehouse < Size of a source database b. **Query Speed of a Data Warehouse > Query Speed of a source database**

c. Number of users of a Data Warehouse > Number of users of a source database d. Amount of redundancy in a source database < amount of redundancy in a Data Warehouse

8. What is a Project Management concept where the original deliverables increase over the duration of the project?

a. Track Issues b. Requirements Analysis

c. **Scope Creep** d. None of the Above

9. In which life-cycle phase do you define the scope of the DW?

a. Project Management b. **Program Planning**

c. Dimensional Modeling d. Design of the ETL

10. The following are properties of a source-to-target map except:

a. On a high level, it lists down the source database tables and points to DW tables b. On the detail level it shows the transformation of each column from source to the DW

c. **It should be written in Excel** d. All of the above are valid properties

11. In which life-cycle phase do you do the user acceptance testing (UAT) of the DW?

a. ETL b. **Deployment**

c. Business Definitions d. Maintenance

12. Which of the following are the cost components of implementing a DW?

a. On-going maintenance expenses b. Consultants costs

c. Expenses to support growth d. **All of the above are valid cost components**

13. What do you call a logical group of activities of processes done in a step-by-step manner?

a. Business Objectives b. **Business Processes**

c. Business Mission d. None of the above

14. What is something of lasting interest in an enterprise in which data can be stored about?

a. Business Processes b. **Entities**

c. OLTP d. None of the above

15. Which of the following is true about the staging area/back room database?

a. This is where the ETL is done c. Comes after analytics

b. Delivery the transformed data to the DW d. **Both a and c are true**

16. What is the primary advantage of Kimball model over Inmon's dimensional model?

a. Easy to Use and Fast b. Big Bang Approach

c. Normalization d. None of the above are advantages

17. Which of the following describes dimensional modelling except?

a. Divides the world into measurements and context b. **Uses normalized models**

b. Measurements are known as facts d. Contexts are known as dimensions

18. Which of the following examples would qualify as a dimension of a dimensional model?

a. Sales of Product X c. Attendance of Students

b. **Suppliers of Product X** d. Count of Subscribers in a Plan

19. What is the main reason for not using normalized models in data warehouses?

a. Easy to interpret b. Hard Disk Space

c. **Joins** d. Easy to use

20. Which of the following best describe normalized databases?

a. Designed to eliminate redundancies b. Objective is 2nd Normal Form

c. Lots of repeating information d. Small number of tables

21. The SELECT statement in SQL can be used for:

a. Querying data from source databases b. Aggregate datasets

c. Both a and b. d. None of the above are valid uses

22. Given that the Customers table has 5 columns and 150 rows of data, how many rows of data will be selected from the following command: SELECT * FROM Customers:

a. 5 columns and 150 rows of data b. 150 columns of data and 5 rows of data

c. 4 columns and 150 rows of data d. None of the above

23. Designing dimension tables consist of four steps, which of the following is not part of the steps in the design?

a. Choose Business Process b. Declare the Grain

c. Identify Dimensions and Facts d. **Identify Problems**

24. Which of the following is true about facts?

a. Known in advance b. Textual Data

c. **Performance measure of a dimension** d. None of the above

25. A 4 Byte sized Primary Key Column can handle at most how many unique rows?

a. 4 Million Rows b. **4 Billion Rows**

c. 2 Million Rows d. 2 Billion Rows

26. On average, a supermarket has 1000 transactions daily with each transaction containing 5 sales items. Given the granularity statement: "One row for each sales item" how many rows are expected to be added in 7 days?

a. 7000 rows b. 35 rows

c. 35000 rows d. None of the above

27. When determining the grain of a dimension table, from the point of view of the business, what question does it answer?

a. What is the ETL population guideline? b. **What is the meaning of each row?**

c. What is the primary key of the data? d. None of the above

28. Which of the following is an example of a detail fact table?

a. Account Balance Fact b. **Sales Transaction Fact**

c. Year to Date Sales d. None of the above

29. A perfect cube fact table happens when:

a. Granularity matches dimension keys b. Granularity is more descriptive as compared to dimension keys

c Data is denormalized d. None of the above

30. We design fact tables based on:

a. Business Processes b. Business Entities

c. Business Questions d. None of the above

31. Which of the following is not recommended to be included in a fact table?

a. Dimension Keys b. Measures

c. **Indicators** d. All of the above

32. Which of the following is an example of an additive fact?

a. Current Inventory Status b. **Unit Price**

c. Net Income d. None of the above

33. Which of the following is false about dimension tables?

a. Usually contains multiple surrogate keys to uniquely identify each row b. Usually verbose

c. Usually bigger in size as compared to fact tables d. None of the above are false

34. What is the size of a date dimension table that tracks historical data for the past 10 years?

a. 356 rows b. **3560 rows**

c. 52 rows d. 356 times 24 rows

35. If we want to track whether a customer purchases different types of magazines, what type of dimension table should be used?

a. Correlation Dimension b. Version Dimension

c. Degenerate Dimension d. **None of the above**

36. Which of the following is an advantage of using degenerate dimensions?

a. Removes the transaction ID from the fact table b. Can be reused by different fact tables

c. Protects against reuse of IDs d. **All of the above**

37. What is the primary purpose of designing dimension tables with dimension families?

a. If new table is required with a different granularity, other tables within the family can be used b. Interpretability

c. Lessens ETL effort d. None of the above are valid purposes

38. How do we implement the case when a purchase order has a Purchase Order Date and a Delivery Date field?

a. Create two physical tables PO -Date-Dim and Del-Date-Dim b. **Create a Date-Dim and Create two Views named PO -Date-Dim and Del-Date-Dim**

c. Do not create date tables since Date-Dim is sufficient d. None of the above

39. Which of the following is not part of the advantages of having surrogate keys?

a. Improve performance b. **Interpretability**

c. Make the DW manage keys d. All of the above are advantages

40. Which of the following is not a valid reason to avoid a complete dimension refresh?

a. **Loose some part of history** b. Takes a lot of time

c. Only applicable to small isolated DWs d. All of the above

41. Which part of the development of a DW takes 70% of the time on average?

a. Modelling b. Problem Analysis

c **ETL** d. Denormalization

42. When doing ETL, it is recommended to do intermediate tables. Which of the following is not an advantage of having intermediate tables?

a. Restart at certain steps not at the very start b. **Longer Development Time**

c. Handles different arrivals of data d. All of the above are advantages

43. In terms of data extraction, which of the following are methodologies to extract incremental data from source systems?

a. Use logs of source systems b. Capture using date and time stamp

c. Use a customized source application to send new data as it comes in d. **All of the above**

44. Which of the following is not a valid transformation task?

a. Summarization of transactions per month b. Standardization of fields to a single format

c. **Deduplication** d. None of the above are invalid tasks

45. Which of the following should you consider in choosing the Dimension Update Type?

a. Type of Update: Correction or Change b. Frequency of Changes

c. Size of Changes d. **All of the above are considerations**

46. What Dimension Update type should you choose if you want to track all the history of changes that does not happen often?

a. Type 1 b. **Type 2**

c. Type 3 d. Type 5

47. Which intermediate Dimension Table holds new rows ready to be inserted into the DW?

a. **I Table** b. U Table

c. C Table d. D Table

48. When loading data, which of the following is not a valid technique in loading incremental data?

a. Destructive Merge b. Append

c. **Delete history** d. None of the above

49. Which Dimension Update type should you consider when you want to track only the immediate past value of the attribute?

a. Type 1 b. Type 2

c. **Type 3** d. Type 5

50. What types of SQL statements are used for the Initial Loading ETL processes?

a. **INSERT** b. UPDATE

c. APPEND d. LOAD

Analytics Modelling Exam

1. In the Predictive Analytics Framework, we usually start with?

a. Data Preparation
b. Modelling
c. **Problem Definition**
d. None of the above

2. Which of the following is not an advantage of having a standardized Predictive Analytics framework?

a. Ease of use for new adopters
b. Aids Project Management
c. Allows projects to be replicated
d. **Increases Dependency on Experts**

3. Which of the following best describes the Predictive Analytics framework?

a. Attached to a single tool
b. **Non-proprietary**
c. Attached to a single industry type
d. None of the above

4. Which Predictive Analytics framework phase do we do data preprocessing?

a. Modelling
b. Business Understanding
c. Data Understanding
d. **Data Preparation**

5. For categorical data, which descriptive statistics can we use?

a. Mean
b. **Mode**
c. Median
d. None of the above

6. What is the conversion of data into a visual or tabular format?

a. Descriptive Statistics
b. Representation
c. **Visualization**
d. Encoding

7. What type of plot shows the relationship of two numerical variables?

a. Histogram
b. Boxplot
c. **Scatterplot**
d. Excel Plot

8. Which of the following is a valid cause of having incomplete data?

a. Different Sources of Data
b. Duplicate Records
c. Not a required field upon data entry
d. None of the above are valid causes

9. In Min-Max normalization, how do we compute for the new data scale?

a. Subtract Min Divide By Old Range Multiply by New Range then Add New Min
b. Random Assign New Scale
c. Subtract Mean then Divide by Standard Deviation
d. None of the Above

10. Suppose that we need to track the variable "Traffic Light Colors" and we want to transform this to numerical variables for regression, how many indicator (Dummy) variables do we need to declare?

a. 1
b. 3
c. **2**
d. 4

11. Which of the following is not a valid reason to reduce the number of variables of a dataset?

a. Avoid the curse of dimensionality
b. Reduce Noise
c. Reduce processing time of data mining algorithms
d. **Reduce the predictive power of data mining algorithms**

12. Which classification model just predicts the majority class of the historical data when new data comes in?

a. ZeroR
b Naïve Bayes
c. OneR
d. Decision Tree

13. A dataset has 100 rows each representing a customer in which, 60 are male customers and 40 are female customers. Of the 60 male customers, 2/3 bought Product X while of the 40 female customers, only ½ bought Product X. Calculate the probability P(Male| Did not Buy Product X)**.**

a. 40/60
b. 60/100
c. 40/100
d. **20/40**

14. When comparing Decision Trees (DT) and K Nearest Neighbors (KNN), which of the following is true?

a. Interpretability of KNN is better as compared to the Interpretability of DT
b. **DT handles missing data better than KNN handles missing data**
c. DT handles numerical variables better than KNN handles numerical variables
d. None of the Above are True

15. When comparing Decision Trees (DT) and Support Vector Machines (SVM), which of the following is false?

a. Interpretability of DT is better as compared to SVMs
b. DTs can handle categorical data better while SVMs handle numerical data better
c. SVM provides an optimal model as compared to a sub optimal DT
d. **None of the above are false**

16. The following perceptron issues are solved using an ANN except:

a. Non separable data
b. Missing Data
c. Non numerical data
d. **Non unique solution**

17. When a predictive model is said to be overfitting the training data, which of the following is false?

a. We have a complex model
b. We have lots of errors in the testing data
c. We have lots of errors in the training data
d. **Both b and c are false.**

18. A confusion matrix for a prediction model has Actual True and Predicted True = 20, and Actual False and Predicted False = 10. If the dataset has 50 rows, what is the accuracy of the prediction model?

a. 10/50
b. **30/50**
c. 20/50
d. 40/50

19. If the testing dataset has predictor variables that are all categorical, which pair of classification models would be good for this data?

a. Decision Trees and Perceptrons
b. ANN and SVM
c. **Rule Based Classifiers and Naïve Bayes**
d. Decision Tress and KNN

20. When the data is expected to have lots of outliers, which of the following classifications can handle this dataset well?

a. SVM
b. Naïve Bayes
c. KNN
d. None of the above

21. On average, how many rows are not selected in a bootstrap sample?

a. 1/10
b. 1/5
c. **1/3**
d. ¼

22. When comparing boosting and bagging, which of the following is true?

a. Boosting is a parallel ensemble while bagging is a serial ensemble
b. **In bagging, all classifiers vote for the prediction equally while in boosting, a weighted average is used**
c. Probability of being selected never changes in boosting as compared to bagging
d. None of the above are true

23. Which of the following is false about Random Forests?

a. It is an ensemble of decision trees
c. **It is a serial ensemble**

b. They are combined by average for regression and voting for classification

d. Random attributes are chosen to inject randomness

24. What is the first thing you should look at in the R Output to validate a regression model?

a. R Squared

c. **P-Value/F Statistic**

b. T Values

d. Residual Error

25. Adjusted R Squared is better as compared to R Squared because of:

a. R Squared is harder to compute

c. **R Squared can be inflated by adding nuisance variables**

b. R Squared Adjusted is higher for models that have more variables

d. R is just adjusted.

26. Which plot checks for non-constant variance?

a. Residuals Versus Order Plot

c. Normal Probability Plot

b. Histogram

d. None of the above

27. When looking for outliers in regression, which of the following can be considered true?

a. If a point is far from the center mass of the data, then it is considered an outlier

c. Outliers tend to greatly influence the slope of the regression line

b. We usually remove outliers then rerun the regression model

d. **All of the above**

28. If the regression model fails the non-constant variance assumption, what transformations are needed?

a. Transform all x variables to x

c. **Transform the** y **variable to** y

b. Stop the regression analysis and gather more data

d. All of the above

29. If you are planning to do all possible regressions on a model with five predictor variables, how many models do we need to check?

a. 15

c. 16

b. 32

d. **31**

30. All possible regressions is done on a model with 5 predictor variables. Which of the following models would you choose?

a. A model with all 5 variables with Radj2=86%

c. **A model with 2 variables with** Radj2=85%

b. A model with 1 variable with Radj2=70%

d. None of the above

31. Which variable selection methodology starts with all variables in the model and one at the time removes each variable until all that remains are the significant ones?

a. Forward Regression

c. **Backward Elimination**

b. Stepwise Regression

d. Ridge Regression

32. What is the primary purpose of removing the dimensions of each variable by standardizing all predictor variables in a regression model?

a. To hasten the computation of linear regression coefficients

c. **To determine the relative importance of each variable as compared to all others**

b. To select a subset of the variables

d. None of the above are valid purposes

33. We wish to model a categorical variable named Tool Type in a regression model on the tool life of a certain Lathe. There are two Tool types, Type A which is selected as a baseline and Type B. If the coefficient of Type B is 30 minutes, what can we infer?

a. The rate of change when changing from type A to type B is 30 minutes

c. **From the baseline of type A, an additional 30 minutes tool life is observed when switching to B**

b. 30 minutes is the baseline for Type B

d. None of the above.

34. We want to predict the age of a person based on several factors. Two of which is height and weight. We know that height and weight are related to each other since in general, a tall person is heavier. What can we expect from the regression model?

a. VIF of these variables would be large

b. Principal Components will identify height and weight as related variables

c. There is a problem on Multicollinearity

d. **All of the above**

35. In a multiple linear regression model where we predict delivery time in minutes (y) from number of cases (x1) and distance travelled in feet (x2), we would have a regression model y=β0+β1x1+β2x2. Based on the t-Table, x2 failed the t test. What can be inferred?

a. x2 is not needed at all

b. The model fails the overall test of regression

c. **x2 is not needed in a model that contains x1**

d. All of the above

36. Which of the following is not a correct reason for not using linear regression to predict a binary response variable like from 0 and 1?

a. **Linear regression will predict values outside the range of 0 and 1**

b. Errors will not have equal variance in 0 and 1.

c. Errors will not be normally distributed

d. None of the above are correct reasons

37. When differentiating between supervised learning and unsupervised learning which of the following is false?

a. **You can calculate errors on both types of learning**

b. Supervised learning has a response variable while unsupervised has none

c. You are not guaranteed to get any model or pattern from unsupervised learning

d. None of the above are false

38. If the support of itemset Milk, Coke and Diapers is 5/10 which of the following is true?

a. The itemset Milk, Code, Diapers and Beer cannot have a support greater than 5/10

b. If minimum support is 40% then Milk, Coke and Diapers is a frequent itemset

c. **If there are 1000 transactions, Milk, Coke and Diapers appears in 500 of them**.

d. All of the above are true

39. If the confidence of the rule Milk, Coke → Diapers is 2/3 which of the following is true?

a. 1/3 of the data contains Diapers

b. If the minimum confidence is 70% then Milk, Coke → Diapers is a frequent rule

c. **If there are 3 transactions of Milk and Coke, then 2 of them also contain Diapers**

d. None of the above are true

40. Given a sequence, <{a, b, c}, {d,e}>which of the following is a valid subsequence?

a. <{a},{f}>

b. <{a},{d}>

c. <{a},{b}>

d. <{a,b}>

41. Which of the following is false about K-Means clustering?

a. Initial centroids are randomly chosen

b. **K-Means is an example of agglomerative clustering**

c. Centroids are updated iteratively

d. None of the above are false

42. Which of the following is not a valid post processing step after running K-Means?

a. **Eliminate Outliers**

b. Merge Similar Clusters

c. Eliminate small clusters

d. Split loose clusters

43. In hierarchical clustering, which of the following is false about a dendogram?

a. **It is randomly created**

b. Any number of clusters can be generated by cutting the dendogram at certain levels

c. Corresponds to a hierarchy like animal taxonomies (kingdom, phylum…)

d. All of the above are false

44. Which of the following calculates the similarity of two clusters by selecting the point closest to each other's cluster?

a. MAX c. **MIN**
b. Average d. Centroid

45. Which of the following is not a limitation of Hierarchical Clustering?

a. Once two clusters are combined it cannot be undone
b. **Sensitive to noise and outliers**
c. No function is minimized in general
d. None of the above

46. The words, "the", "a", "them", "of", "you", are called:

a. Stem Words
b. **Stop Words**
c. Tokens
d. Corpha

47. What would be the size of a term-by-document matrix if there are 10 documents and 1000 unique words in all 10 documents?

a. 1000 rows by 10 columns
b. 10 rows by 10 columns
c. **10 rows by 1000 columns**
d. 1000 rows by 1000 columns

48. Which of the following cannot be detected by Social Media Sentiment Analysis?

a. Sentiment of a product
b. term-by-document matrix
c. List of negative words
d. **Sarcasm**

49. Google uses the page-rank algorithm to rank "credible websites." This is an example of

a. Web content mining
b. Web usage mining
c. Web structure mining
d. None of the above

50. If a sentiment of a product is on average +2, what can be said about it?

a. On average, there are 2 more negative words than positive words in each document
b. We have a positive sentiment
c. On average, there are 2 more positive words than positive words in each document
d. **Both b and c**