National College of Ireland

# Configuration Manual

MSc Research Project
FinTech

Clara Onunkwo
Student ID: X18122558

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Clara Onunkwo |
| **Student ID:** | X18122558 |
| **Programme:** | FinTech **Year:** 2019 |
| **Module:** | Research Project |
| **Lecturer:** | Noel Cosgrave |
| **Submission Due Date:** | 12 August 2019 |
| **Project Title:** | A classification-based approach for modelling disputed responses based on consumer complaint on financial products. |
| **Word Count:** | 878 **Page Count:** 9 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** .....

**Date:** 12 August 2019

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Configuration Manual

Clara Onunkwo
Student ID: x18122558

## 1 Introduction

This paper gives the overview used to produce the study. This implies that everything needed to reproduce this research will be shown in this paper. This includes the R version used to carry out the analysis, packages used, laptop used during the study, code implemented for each of the required analysis, memory of the laptop upon which the analysis was carried out, plots derived from the analysis, system processor, speed usage, the allocated heap size and every instrument or facility used for the study. This help to show how well the study was carried out and enable the study to be reproduceable if need be.

## 2 Details of Device Specification

- **System used**: HP Pavilion laptop 15-cclxx.
- **RAM:** 12.0GB (11.9GB usable)
- **Processor:** Intel(R) Core™ i5-8250U CPU@ 1.60GHz.
- **System type:** 64-bit operating system, x64-based processor.
- **Speed:** 1.80GHz

## 3 Software Tool Used

- **Name of Programming Environment:** R studio
- **Name of programming language:** R
- **Version of programming Environment:** R x64 3.6.1

## 4 Libraries Installed

```
library(naivebayes)
library(randomForest)
library(dplyr)
library(ggplot2)
library(psych)
library(caret)
library(forecast)
library(nnet)
library(ROCR)
library(labeling)
```

# 5 Codes used for Study Analysis

```
#### Reading the dataset into R

complaints_data = read.csv("~/consumer_complaints.csv", na.strings = c("","NA"))

#### View the data

View(complaints_data)

#### Structure of the data

str(complaints_data)
dim(complaints_data)


#### Remove irrelevant attributes

comp_data<-complaints_data[c(2,3,4,5,7,8,9,11,12,13,15,16,17)]


#### Omitting Empty values

comp_data<-comp_data %>% na.omit()
str(comp_data)

#### removing more attributes

comp_data1<-comp_data[c(1,2,5,8,9,10,11,12,13)]
str(comp_data1)


################################################################################
##########
#### Exploratory Data Analysis
################################################################################
#########

################## DATA SUMMARY

summary(comp_data1)
pairs.panels(comp_data1)

#################### Data Visualization with plot

#### Product

plot(comp_data1$product,comp_data$consumer_disputed, ylab= "Consumer_disputed",
    xlab="Products", main="Graph of Consumer_disputed Against
products",col=c("blue","red"))
```

```
#### Sub product

plot(comp_data$sub_product,comp_data$consumer_disputed, ylab=
"Consumer_disputed",
    xlab="Sub_Products",main="Graph of Consumer_disputed Against
Sub_products",col=c("blue","red"))

####

plot(comp_data$issue,comp_data$consumer_disputed, ylab= "Consumer_disputed",
    xlab="Issues", main="Graph of Consumer_disputed Against
Issues",col=c("blue","red"))

plot(comp_data$sub_issue,comp_data$consumer_disputed, ylab= "Consumer_disputed",
    xlab="Sub_issues",main="Graph of Consumer_disputed Against
Sub_issues",col=c("blue","red"))


##########################################################################
###################################
######## USING GGPLOT
##########################################################################
##############
######## Product

ggplot(comp_data1,aes(x=product,fill=consumer_disputed.))+
  geom_bar()+
  ggtitle("BOXPLOT OF The Product for consumer disputed")

######### Sub Product

ggplot(comp_data1,aes(x=sub_product,fill=consumer_disputed.))+
  geom_bar()+
  ggtitle("BoxPlot of The SubProduct for consumer_disputed")

######### Tags

ggplot(comp_data1,aes(x=tags,fill=consumer_disputed.))+
  geom_bar()+
  ggtitle("BoxPlot of The tags for consumer_disputed")

######### Timely Response

ggplot(comp_data1,aes(x=timely_response,fill=consumer_disputed.))+
  geom_bar()+
  ggtitle("BoxPlot of The timely responsee for consumer_disputed")


hist(table(comp_data1$product))
```

```
#########################################################################
##### Data partitioning
#########################################################################

set.seed(1234)
ind1 <- sample(2,nrow(comp_data1),replace =T ,prob=c(0.7,0.3))
Baye_train <- comp_data1[ind1==1,]
Bayes_test <- comp_data1[ind1 ==2,]


#########################################################################
###### Naive Bayes' Classifier
#########################################################################

Bayes_model <-naive_bayes(consumer_disputed. ~. , data = Baye_train)
Bayes_model

summary(Bayes_model)

plot(Bayes_model)



############################## predictions#########################

########## using train data

pred_bayes_train<-predict(Bayes_model,Baye_train,type = "prob")
tab_bayes_train<-(cbind(pred_bayes_train,Baye_train))

#### confusion matrix for train dataset

pred1_bayes_train<-predict(Bayes_model,Baye_train)
train_confusion_matrix<-table(pred1_bayes_train,Baye_train$consumer_disputed.)
print(train_confusion_matrix)

#### misclassification for train data

1-sum(diag(train_confusion_matrix))/sum(train_confusion_matrix)
sensitivity(tab_bayes_train$consumer_disputed.,Baye_train$consumer_disputed.)
kappa(train_confusion_matrix)


########## using test dataset

bayes_pred_test<-predict(Bayes_model,Bayes_test)
bayes_test_confusion_matrix<-table(bayes_pred_test,Bayes_test$consumer_disputed.)
print(bayes_test_confusion_matrix)
```

```
## misclassification rate for test data

1-sum(diag(bayes_test_confusion_matrix))/sum(bayes_test_confusion_matrix)

###### Importance


################################################################################
##########
############ Random Forest classifier
################################################################################
#######

##### Training set: VALIDATION SET =70:30(random)


set.seed(1234)

ind2<- sample(nrow(comp_data1),0.7*nrow(comp_data1),replace = F)
Forest_train<-comp_data1[ind2,]
Forest_test<-comp_data1[-ind2,]

summary(Forest_train)
summary(Forest_test)

### creating a random forest model

Forest_model1<-randomForest(consumer_disputed. ~., data = Forest_train)
Forest_model1


Forest_model2<-randomForest(consumer_disputed. ~., data =Forest_train , ntree=500,
mtry=8)
Forest_model2

##########################################Predictions#######################
###
########predicting on train set

pred_forest_train<-predict(Forest_model1,Forest_train, type = "class")


#### checking classification accuracy

table(pred_forest_train,Forest_train$consumer_disputed.)

#####predicting on test set

pre_forest_test<-predict(Forest_model1,Forest_test, type = "class")
```

```
### checking classification accuracy

mean(pre_forest_test==Forest_test$consumer_disputed.)

### confusion Matrix

forest_test_confusion_matrix<-table(pre_forest_test,Forest_test$consumer_disputed.)
print(forest_test_confusion_matrix)
confusionMatrix(pre_forest_test,pred_forest_train)
sensitivity(pred_forest_train,Forest_train$consumer_disputed.)
kappa(forest_test_confusion_matrix)

## misclassification rate for test data

1-sum(diag(forest_test_confusion_matrix))/sum(forest_test_confusion_matrix)

#### To check for important variable

importance(Forest_model1)
varImpPlot(Forest_model1)

#Logistic regression

logreg<-multinom(consumer_disputed.~.,data=comp_data1)

#Misclassification

predlogreg<-predict(logreg,comp_data1)
tablogreg<-table(predlogreg,comp_data1$consumer_disputed.)
tablogreg
sum(diag(tablogreg))/sum(tablogreg)
1-sum(diag(tablogreg))/sum(tablogreg)
table(comp_data1$consumer_disputed.)

#Performance Evaluation

predlogreg2<-predict(logreg,comp_data1,type = "prob")
predlogreg3<-prediction(predlogreg2,comp_data1$consumer_disputed.)
hist(predlogreg3)
performance(predlogreg3,"acc")->eval
plot(eval)
abline(h=0.80,v=0.36)

#Identify Best Values

which.max(slot(eval,"y.values")[[1]])->max
acc<-slot(eval,"y.values")[[1]][max]
acc
slot(eval,"x.values")[[1]][max]->cut
cut
```

```
print(c(Accuracy=acc,Cutoff=cut))

#Receiver Operating Characteristic curve (ROC)

predlogreg4<-prediction(predlogreg2, comp_data1$consumer_disputed.)
roc<-performance(predlogreg4,"tpr","fpr")
plot(roc,colorize=T,main="ROC Curve",ylab="Sensitivity",xlab="1-Specificity")
abline(a=0,b=1)

#Area Under the Curve (AUC)
performance(predlogreg4,"auc")->auc
```

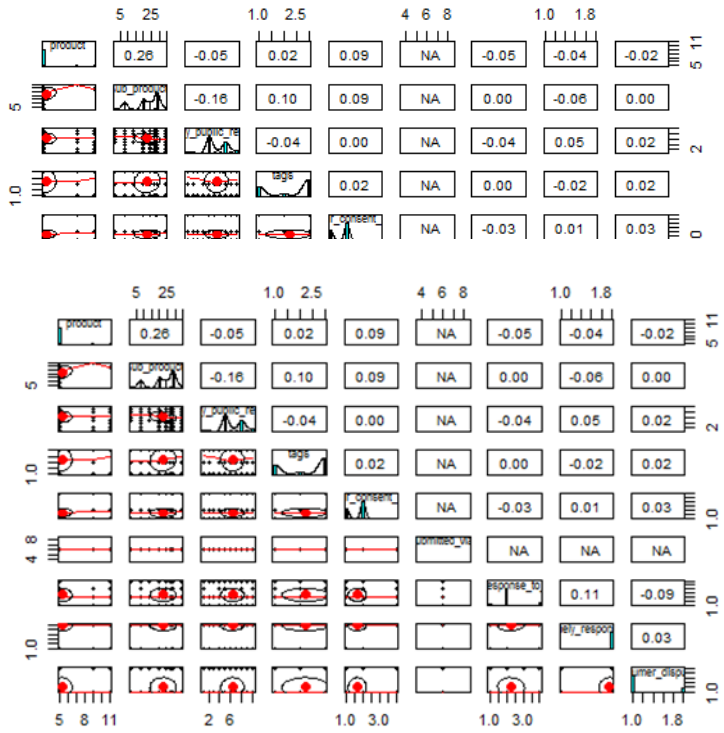


Figure 1: Exploratory Data Analysis (EDA)

This graph above Is the correlation matrix between the variables. It shows the strength of the relationship between the variables/attributes under study. This is done using the pairs panel.