

# Configuration Manual

MSc Research Project  
FinTech

Steven Kawala  
X15018121

School of Computing  
National College of Ireland

Supervisor: Victor Del Rosal

**National College of Ireland**  
**MSc Project Submission Sheet**  
**School of Computing**



**Student Name:** STEVEN KAWALA  
**Student ID:** X15018121  
**Programme:** MSc FINTECH **Year:** 2020  
**Module:** MSc RESEARCH PROJECT  
**Lecturer:** VICTOR DEL ROSAL  
**Submission Due Date:** 18/08/2020  
**Project Title:** Non-Technical Electricity Loss: Predicting and Defining correlation of Electricity Theft Determinants Using Machine Learning Algorithms  
**Word Count:** 1168 **Page Count:** 6

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:** *Steven Kawala*  
**Date:** 18/08/2020

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission,</b> to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project,</b> both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

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


<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Configuration Manual

Steven Kawala  
Student ID: x15018121

The reason for this research is to check how well Artificial Intelligent Algorithms can be used to predict power theft considering the social and economic determinants behind electricity theft. The research has answered the research question by training seven models and compare how well they performed in predicting the power theft. This configuration summary gives every step carried out to get the results for analysis. It is divided into four section as follows:

- System Requirement
- Data Exploration
- Data preparation
- Model development and evaluation

**NOTE:** The Full code for the whole project is submitted in a separate file called “ Project Code ”, which is followed by a Video presentation of the code.

## 1 System Requirements

System requirements may depend on how complex and size of the data analysis. Table 1 summarizes the system requirements used for the project.

SUGGESTED	USED
Laptop or Notebook	Laptop
Wi-Fi card	Wi-Fi card
64-bit x86 Multi Core processor	Intel Core i5
Current Ubuntu, MacOS or Windows version	Windows 10
SSD with 40 GB available space.	250 available disk space
At least 256 MB of RAM because most data are stored in RAM	8 Gb RAM
At least 3.5.2 R Version or new for computation and analysis	R Version 4.0.2

Table 1: System Requirements

## 2 Data Exploratory

- Step1: Extracting data from State Grid Corporation of China <https://www.sgcc.com.cn>.
- Step2: Setting working directory on ‘R’ studio.
- Step3: Upload dataset in R studio for analysis. (Table 1)

```
## Loading Data To "R"  
>electricity = read.csv ("electricityNew.csv", header = TRUE, sep = ",")  
>view(electricity)  
>dim(electricity)  
>summary(electricity)
```

Table 1 Data Loading & structure

- Step4: Check the data distribution for data understanding through data visualization by creating histogram (Table 2)

```
## Data Distribution
>hist(electricity)
>hist (electricity, freq = FALSE, col=" Electricity", main = "Distribution Curve")
>curve (dnorm (x, mean = mean(electricity)), add = TRUE,col= "red")
```

Table 2:Data Exploration

- Step5: Missing Values. There are 7663 missing vales in the form of blank spaces and zeros for customers who did not complete the survey (Table 3).

```
## Checking for missing values
>sum(is.na(electricity))
>colSums(is.na(electricity))
>any(is.na(electricity))
```

Table 3: Missing Data

- Step6: Descriptive statistics of data to quantitatively describe the main features of the data for simple understanding and interpretation of the data.

```
## Descriptive analysis of Data
>electricity <- rnorm (6525, mean = 0, sd = 1)
>mean(electricity)
>sd(electricity)
  >descriptives <- function(electricity) {
  >u <- length(unique(electricity))
  >missing <- sum(is.na(electricity)) / length(electricity) * 100
  >quantiles <- quantile (electricity, na.rm =T)
  >av <- mean (electricity, na.rm = T)
  >stdev <- sd (electricity, na.rm = T)
  >df <- data. frame (u, missing, quantiles[1], quantiles[2], quantiles[3],
  quantiles[4], quantiles[5],av,stdev)
  >names(df) <- c("no.unique","%missing ", "min","first
  quartile",median","third quartile","max","sd")
  >rownames(df) <- NULL
  >return(df)
  }
df <- descriptives(electricity)
```

Table 4:Discriptive Analysis

### 3 Data Preparation

- Step7: Removing missing data and correct encoding of variables.

```
## Removing missing data
>newdata <- na.omit(electricity)

## Checking encoding correct string to variables
>str(newdata)
>newdata$flag <- as. factor(newdata$flag)
```

Table 5: Omitting Missing Data & Variable encoding

- Step 8: Install package and recall library (Boruta)
- Step 9: Checking variable importance to remove unimportant variables. One variable identified as unimportant (Meter ID)
- Step 10: Create a new data for important variables along with response variable.

```
## Feature Selection
>set.seed(111)
>boruta<-Boruta (flag~., data=newdata, doTrace=2,maxRuns=100)
>print(boruta)
>plot(boruta)

##Removing unimportant columns
>newData <-newdata
>newData$Meter.ID<-NULL
>str(newData)
```

Table 6: Removing unimportant columns.

- Step 11: Checking variable levels
- Step 12: Data Normalization to scale numeric values between 0 and 1
- Step 13: Install package and recall library (kohonen)
- Step 14: Data visualization and mapping to see data pattern using SOM. This reduces dimensionality and makes it easy to read the dataset patterns.
- Step 15: Sample data 70% train data ,30% test data.

```
## Data Split
>set.seed(123)
>modell<-sample (2, nrow(newData),replace = T , prob = c(0.7,0.3))
>train<- newData [modell== 1,]
>test<- newData [modell== 2,]
>dim(train)
>dim(test)
```

Table 7: Data Sampling /Split

## 4 Model development and evaluation

### ##Experiment 1

- Step 16: Install packages and recall library for (kohonen)
- Step 17: Training Self Organizing Mapping on train dataset.
- Step 18: Validating results on test dataset then produce confusion matrix. For model evaluation we install and recall caret package to generate confusion matrix.

### ## Experiment 2

- Step 19: Install packages and recall library for (tree, caret, e1071)
- Step 20: Training Random Forest on train dataset.
- Step 21: Validating results on test dataset then produce confusion matrix
- Step 22: Recall 'ROCR' and 'pROC' libraries to plot ROC and calculate AUC.

### ##Experiment 3

- Step 23: Install packages and recall library for (naivebayes, dplyr)
- Step 24: Training Naïve Bayes on train dataset.
- Step 25: Result validation on test dataset the produce confusion matrix.

- Step 26: Recall ‘ROCR’ and ‘pROC’ libraries to plot ROC and calculate AUC.

```
## Evaluation
>library(ROCR)
>p3 = predict(model3,type = "prob",newdata = test)
>p3 = prediction(p3[,2],test$flag)
>naiveperf = performance(p3,"tpr","fpr")
>plot(naiveperf,colorize=T,
      >main = "Naive ROC Curve",
      >ylab = "Sensitivity",
      >xlab = "Specificity")
Abline(a = 0, b = 1)

## Area under the curve
>auc<- performance(p3,"auc")
>auc<-unlist(slot(auc,"y.values"))
>auc<-round(auc,4)
```

Table 8: Naïve Bayes Evaluation

#### ##Experiment 4

- Step 27: Install packages and recall library for (rpart, rpart. plot)
- Step 28: Training Decision Tree on train dataset.
- Step 29: Result validation on test dataset then produce confusion matrix.
- Step 30: Recall ‘ROCR’ and ‘pROC’ libraries to plot ROC and calculate AUC.

```
## Evaluation
>install.packages("ROCR")
>library(pROC)
>library(ROCR)
>d_pred<-predict(model, test,type = "prob" )
>AUC<-auc(test$flag, pred[,2])
>plot(roc(test$flag, pred[,2]), colorize = T,main="DT ROC Curve" )
```

Table 9: Decision Tree Evaluation

#### ##Experiment 5

- Step 31: Install packages and recall library for (tidyverse)
- Step 32: Training kNN on train dataset.
- Step 33: Result validation on test data the produce confusion matrix.
- Step 34: Recall ‘ROCR’ and ‘pROC’ libraries to plot ROC and calculate AUC.

```
## Evaluation
>install.packages("ROCR")
>library(pROC)
>library(ROCR)
>k_pred<-predict(mod_fit, test,type = "prob" )
>AUC<-auc(test$flag, k_pred[,2])
>plot(roc(test$flag, k_pred[,2]), colorize = T,main="kNN ROC Curve" )
```

Table 10: kNN Evaluation

#### ## Experiment 6

- Step 35: Install packages and recall library for (caret, e1071)
- Step 36: Training Support Vector Machine (SVM) using train dataset.
- Step 37: Result validation on test dataset then produce confusion matrix.
- Step 38: Recall ‘ROCR’ and ‘pROC’ libraries to plot ROC and calculate AUC.

```

## SVM Evaluation
>s_pred<-prediction (s_pred, test$flag)
>roc<-performance (s_pred," tpr"," fpr")
>plot (roc, colorize = T, main = "SVM ROC Curve", ylab = "sensitivity", xlab = "1-
specificity")
>abline (a=0, b=1 )
## Area Under Curve
>auc<-performance (s_pred, "auc")
>auc<- unlist (slot (auc, "y. values"))
>auc<-round (auc, 4)

```

Table 11: SVM Evaluation

## ##Experiment 7

- Step 39: Install packages and recall library for (nnet)
- Step 40: Apply Logistic Regression on train dataset.
- Step 41: Result validation on test dataset then produce confusion matrix.
- Step 42: Recall library ‘ROCR’ and ‘pROC’ to plot ROC and calculate AUC.

```

## Logistic Regression Evaluation
>roc<- performance (pred, "tpr", "fpr")
>plot (roc, colorize = T, main = "LP ROC Curve", ylab = "sensitivity", xlab =
"specificity")
>abline (a=0, b=1)
## Area Under Curve
>auc<-performance (pred, "auc")
>auc<- unlist (slot (auc, "y. values"))
>auc<-round (auc, 4)

```

Table 12: Logistic Regression Evaluation

## References

Datavedas.com. 2020. *MODEL EVALUATION IN R | Data Vedas*. [online] Available at: <<https://www.datavedas.com/model-evaluation-in-r/>> [Accessed 2 August 2020].

Datatofish.com. 2020. *How to Import A CSV File into R (Example Included) - Data to Fish*. [online] Available at: <<https://datatofish.com/import-csv-r/>> [Accessed 2 August 2020].

R, D., 2020. *11 Most Useful Steps to Create Data Exploration in R | Methods | Example | Definition*. [online] EDUCBA. Available at: <<https://www.educba.com/data-exploration-in-r/>> [Accessed 2 August 2020].