**NAME**

**COLLEGE NAME**

**Abstract**

Data analysis and statistical tools have been applied for a long time over the years and from the ages and times of the mediaeval to the current advance and modified technologies and tools. With this kind of advancements, developers, research and development teams together with engineers have gone ahead to develop tools, and algorithms to help solve some of the statistical problems in today industry. Some of these tools include SATA, STRATA, MATLAB and R. There are different modules and packages that have also been built inside these applications that support the development and maintenance of such tools and applications have enabled students of data science and even professional engineers to rapidly bring out of the box the abilities and capabilities of strong data understanding and interpretation. Python as a programming languages has a set of very powerful tools and packages that to date are still applied to help users make the best programs and explore the capabilities underlying in data science, machine learning and predictive algorithms. Almost all sectors of the industry need to make sense of the kind of data and information that they are producing, the sectors range from Finance, Banking, Health and Educational sector. With the introduction of big data as part of the data science industry, organizations are shifting more focus into coming up a more robust and technical means of analysing and making sense of the data that is currently consumed and used.

**Keywords:** Big data,Empirical, Predictive, Linearity, Skew

**Introduction**

One of the sectors that relies on constant communication in order to make sense from its information system is the Finance sectors. This may be also attributed to the risky nature in which this industry operates, it deals with money so loses and profits must be kept in check at all times. Sequentially, R as a powerful tool and package has been specifically developed to help users get a quick summary from their data, explore positions and moreover make predictions. Prior to making any specific forecasts using the tool, lets discus below some of the models and classifications ions that can be employed on the tool to help achieve the required objectives and significances:

**Random Forest**

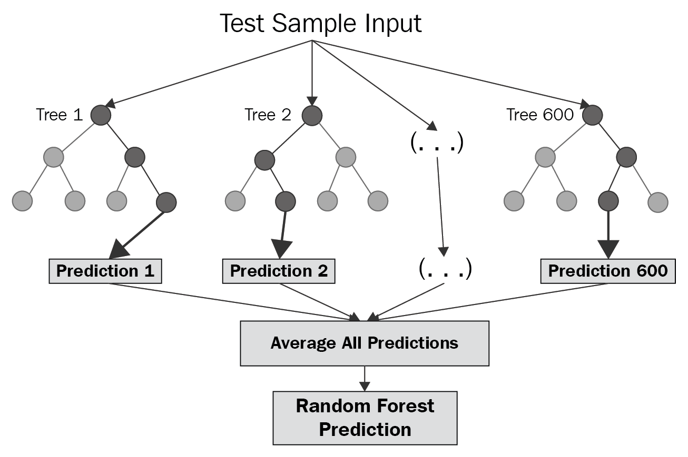


Figure 1 Random forest illustration in R

This model is most effective in machine learning. The model works for both data classification and prediction. Usually the root close of the problem is identified and then the problem is solved based on the classification of nodes also known as a bag. The higher node is also known as the parent and the node that is passed down to it becomes the child node. The model is highly accurate in predicting data where classification is a major issues.

**Linear regression**

Is efficient in measuring the degree or extent to which the object in question are closely related to each other. Usually the standard point of measurement being the 1 reference upon which values that are negatively skewed will be considered as non- linear to the centered of references. On the hand, values closer to the1 are considered as relating. Suppose wed draw a linear graph along the axis, the straight line graph/ line of best fit will show concentration s around the same line.

**KNN MODEL**

The KNN model measure the closeness of classifications among each other, the model tries to predict the next closer classification class on the series.

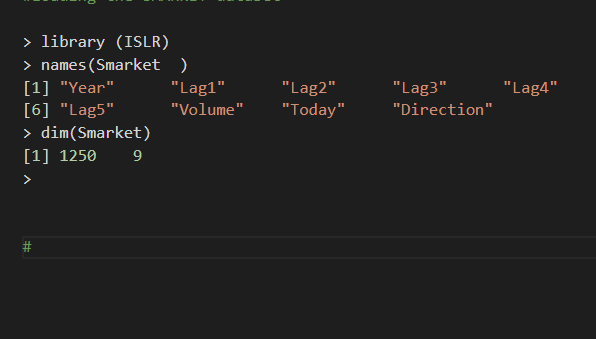
**Dataset description and problem statement**

The dataset description given here is the dataset obtained from the S&P 500 market dataset of over years of 2001 to 2005, the values represent the past 5 revenues and the projected direct sales in the incoming months

**Objectives and aims**

* To determine the linear regression based on this dataset
* To determine the KNN statics based on this dataset
* To determine the random distribution of this dataset

Loading the SMARKET dataset;



summary (Smarket)

      Year           Lag1                Lag2

 Min.   :2001   Min.   :-4.922000   Min.   :-4.922000

 1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500

 Median :2003   Median : 0.039000   Median : 0.039000

 Mean   :2003   Mean   : 0.003834   Mean   : 0.003919

 3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750

 Max.   :2005   Max.   : 5.733000   Max.   : 5.733000

      Lag3                Lag4                Lag5

 Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.92200

 1st Qu.:-0.640000   1st Qu.:-0.640000   1st Qu.:-0.64000

 Median : 0.038500   Median : 0.038500   Median : 0.03850

 Mean   : 0.001716   Mean   : 0.001636   Mean   : 0.00561

 3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.59700

 Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.73300

     Volume           Today           Direction

 Min.   :0.3561   Min.   :-4.922000   Down:602

 1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648

 Median :1.4229   Median : 0.038500

 Mean   :1.4783   Mean   : 0.003138

 3rd Qu.:1.6417   3rd Qu.: 0.596750

 Max.   :3.1525   Max.   : 5.733000

#Getting the lags  of from the dataset

> head(Smarket)

: Year   Lag1   Lag2   Lag3   Lag4   Lag5 Volume  Today Direction

1 2001  0.381 -0.192 -2.624 -1.055  5.010 1.1913  0.959        Up

2 2001  0.959  0.381 -0.192 -2.624 -1.055 1.2965  1.032        Up

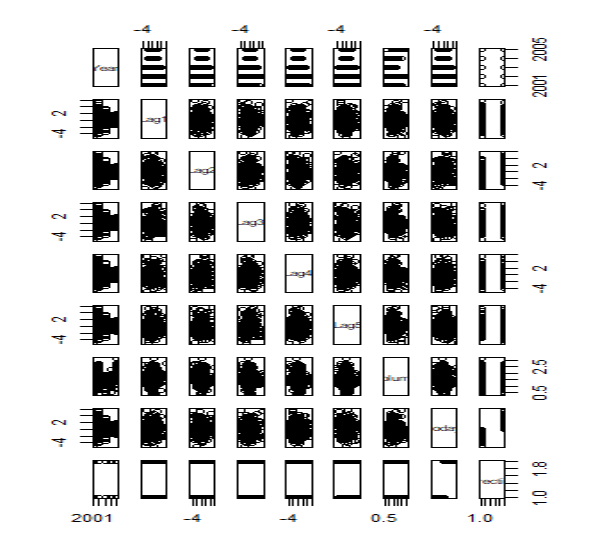
3 2001  1.032  0.959  0.381 -0.192 -2.624 1.4112 -0.623      Down

4 2001 -0.623  1.032  0.959  0.381 -0.192 1.2760  0.614        Up

5 2001  0.614 -0.623  1.032  0.959  0.381 1.2057  0.213        Up

6 2001  0.213  0.614 -0.623  1.032  0.959 1.3491  1.392        Up

>



**Getting the correlation matrix**

> cor(Smarket[,-9])

             Year         Lag1         Lag2         Lag3         Lag4         Lag5

Year   1.00000000  0.029699649  0.030596422  0.033194581  0.035688718  0.029787995

Lag1   0.02969965  1.000000000 -0.026294328 -0.010803402 -0.002985911 -0.005674606

Lag2   0.03059642 -0.026294328  1.000000000 -0.025896670 -0.010853533 -0.003557949

Lag3   0.03319458 -0.010803402 -0.025896670  1.000000000 -0.024051036 -0.018808338

Lag4   0.03568872 -0.002985911 -0.010853533 -0.024051036  1.000000000 -0.027083641

Lag5   0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641  1.000000000

Volume 0.53900647  0.040909908 -0.043383215 -0.041823686 -0.048414246 -0.022002315

Today  0.03009523 -0.026155045 -0.010250033 -0.002447647 -0.006899527 -0.034860083

            Volume        Today

Year    0.53900647  0.030095229

Lag1    0.04090991 -0.026155045

Lag2   -0.04338321 -0.010250033

Lag3   -0.04182369 -0.002447647

Lag4   -0.04841425 -0.006899527

Lag5   -0.02200231 -0.034860083

Volume  1.00000000  0.014591823

Today   0.01459182  1.000000000

>

Logistic regression

> glm.fits=glm(Direction∼Lag1+Lag2+Lag3+Lag4+Lag5+Volume , data=Smarket ,family =binomial)

> summary(glm.fits)

Call:

glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +

    Volume, family = binomial, data = Smarket)

Deviance Residuals:

   Min      1Q  Median      3Q     Max

-1.446  -1.203   1.065   1.145   1.326

Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.126000   0.240736  -0.523    0.601

Lag1        -0.073074   0.050167  -1.457    0.145

Lag2        -0.042301   0.050086  -0.845    0.398

Lag3         0.011085   0.049939   0.222    0.824

Lag4         0.009359   0.049974   0.187    0.851

Lag5         0.010313   0.049511   0.208    0.835

Volume       0.135441   0.158360   0.855    0.392

(Dispersion parameter for binomial family taken to be 1)

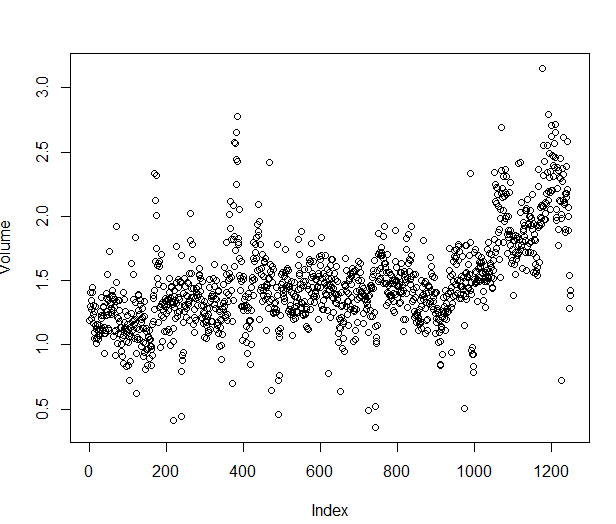
    Null deviance: 1731.2  on 1249  degrees of freedom

Residual deviance: 1727.6  on 1243  degrees of freedom

AIC: 1741.6

Number of Fisher Scoring iterations: 3

>



Correlation model

> coef(glm.fits)

 (Intercept)         Lag1         Lag2         Lag3         Lag4         Lag5

-0.126000257 -0.073073746 -0.042301344  0.011085108  0.009358938  0.010313068

      Volume

 0.135440659

From the above model it can be seen that the model sets the correlational at 0.13, meaning that, there is a weak relationship between the lags and the volume of production of the return across the lags of 1,2, 3 and 4

The next step in this dataset is to determine the coefficients of relationship

> summary(glm.fits)$coef

 Estimate Std. Error    z value  Pr(>|z|)

(Intercept) -0.126000257 0.24073574 -0.5233966 0.6006983

Lag1        -0.073073746 0.05016739 -1.4565986 0.1452272

Lag2        -0.042301344 0.05008605 -0.8445733 0.3983491

Lag3         0.011085108 0.04993854  0.2219750 0.8243333

Lag4         0.009358938 0.04997413  0.1872757 0.8514445

Lag5         0.010313068 0.04951146  0.2082966 0.8349974

Volume       0.135440659 0.15835970  0.8552723 0.3924004

>

Predicting the direction of the market

##PREDICT

> contrasts(Direction)

     Up

Down  0

Up    1

> glm.pred=rep("Down" ,1250)

> glm.pred[glm.probs >.5]=" Up"

> table(glm.pred ,Direction)

        Direction

glm.pred Down  Up

     Up   457 507

    Down  145 141

> mean(glm.pred== Direction)

[1] 0.116

>

**Random forest prediction**

##RANDOM FOREST TRAIN AND PREDICTION

> library(randomForest)

randomForest 4.6-14

Type rfNews() to see new features/changes/bug fixes.

Warning message:

package ‘randomForest’ was built under R version 4.1.1

> library(ISLR)

> table(Smarket$Year)

2001 2002 2003 2004 2005

 242  252  252  252  252

> train<-Smarket[Smarket$Year<2004,]

> test<-Smarket[Smarket$Year>=2004,]

> fit<-randomForest(Direction~Lag1+Lag2+Lag3,data=train,

+                   ntree=1000,importance=T)

> test.pred<-predict(fit,test,type='class')

> table(test.pred,test$Direction)

test.pred Down  Up

     Down  107 142

     Up    116 139

>

From the dataset already provided, we can see that the random model predicted that the market revenue direction would go up, based on the supplied dataset, the test data indicates a positively skewed direction of the dataset as shown above.

**REFERENCES**

<https://rpubs.com/ramsatpm/235748>

https://builtin.com/data-science/random-forest-algorithm