Machine Learning Report

April 2021

Tay Cheng Jun

190399353

Table of Contents

Part 1: 3

Part 2: 6

- Portuguese Subject 6
- Mathematics Subject 7
- Conclusion 8

Part 3: 10

Part 1:

In part 1, we will be investigating the European working conditions survey. We start out by creating our dataset by importing the data “EWCS_2016.csv”. We will investigate our dataset through the use of principal component analysis by reducing the dimensionality of the large data sets, while retaining most of the information in the data sets that create linear functions of our variables.

```
> apply(data, 2,var)
      Q2a       Q2b      Q87a      Q87b      Q87c      Q87d      Q87e      Q90a      Q90b
 128.3675 1681.0429 3582.8905 3201.7705 3582.8888 3330.2296 5871.7074 4089.4287 4090.3005
     Q90c      Q90f
3325.9265 9393.2895
```

Figure 1.1

As seen in figure 1.1, there is a large amount of variation. Hence, we will make a principal component model named “data1” and plot the first ten principal components into a bar chart.

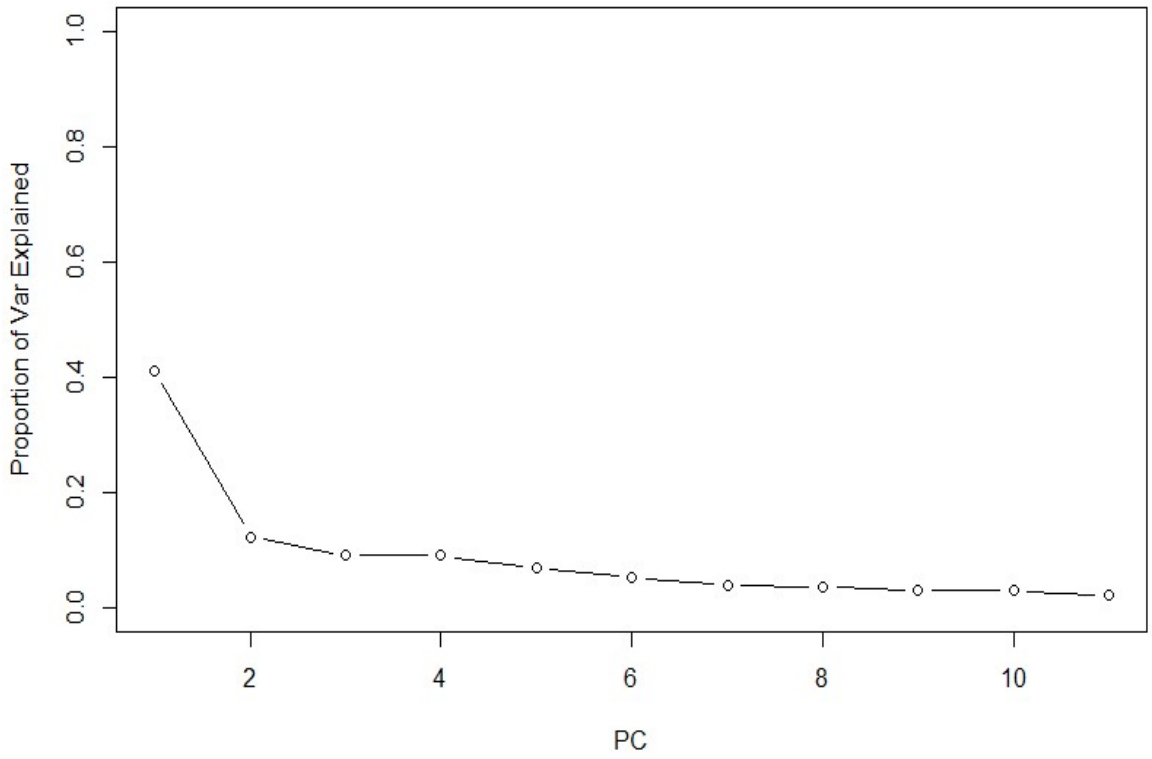

Figure 1.2

In figure 1.2, we can see clearly that principal component 1 (PC1) show the largest variation followed by PC2 and subsequently the rest of the principal components. It shows that PC1 account for 40% of all the variation in the data and the variation it account get lower in subsequent principal components.

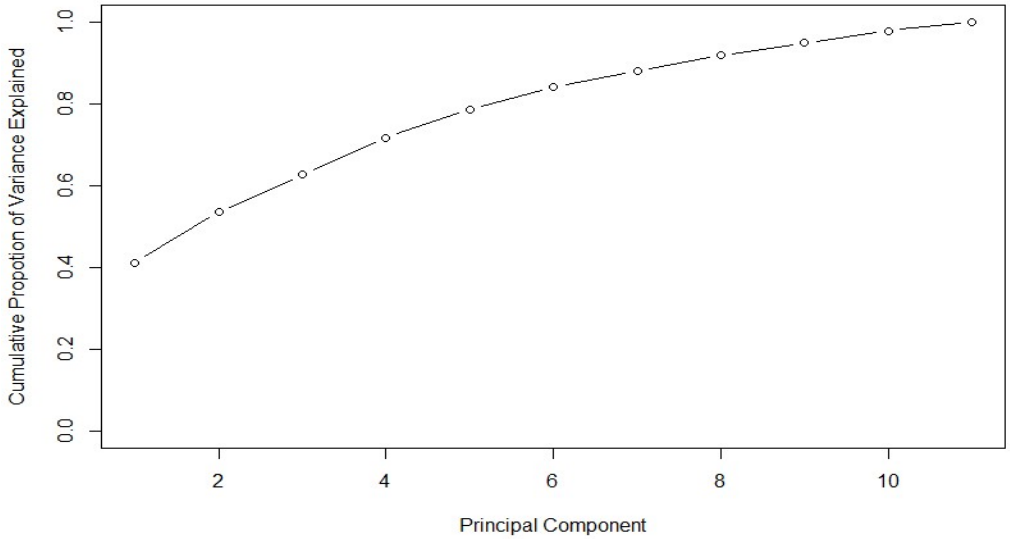

Figure 1.3

In figure 1.3, we can see the cumulative variance plot. In this plot, we can see that having an increase in principal component combined will result in a higher proportion of variance explained. PC1 will account for 40% of the variation, but PC1+PC2 will result in a higher variation and it will slowly increase until all the principal component is added up. As seen in the plot, when all 10 of the principal components are added together, almost 98% of the variation are accounted for.

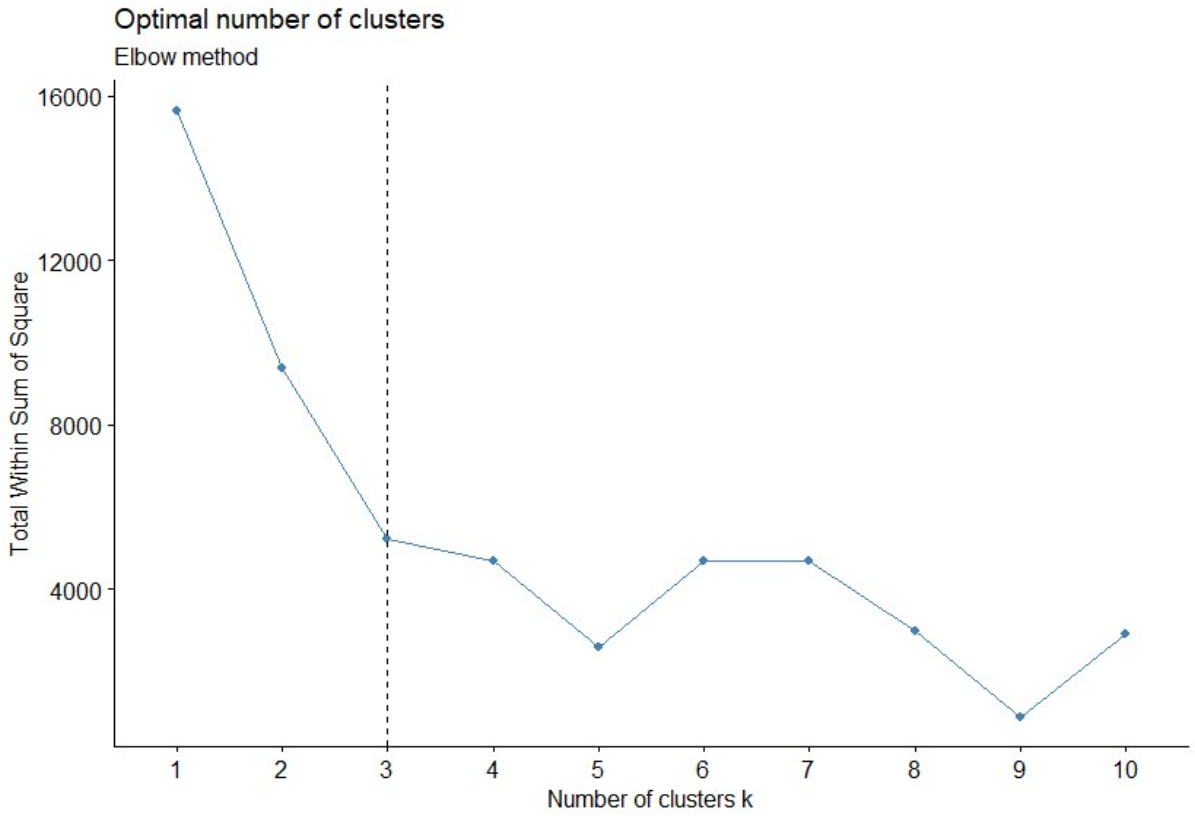

Figure 1.4

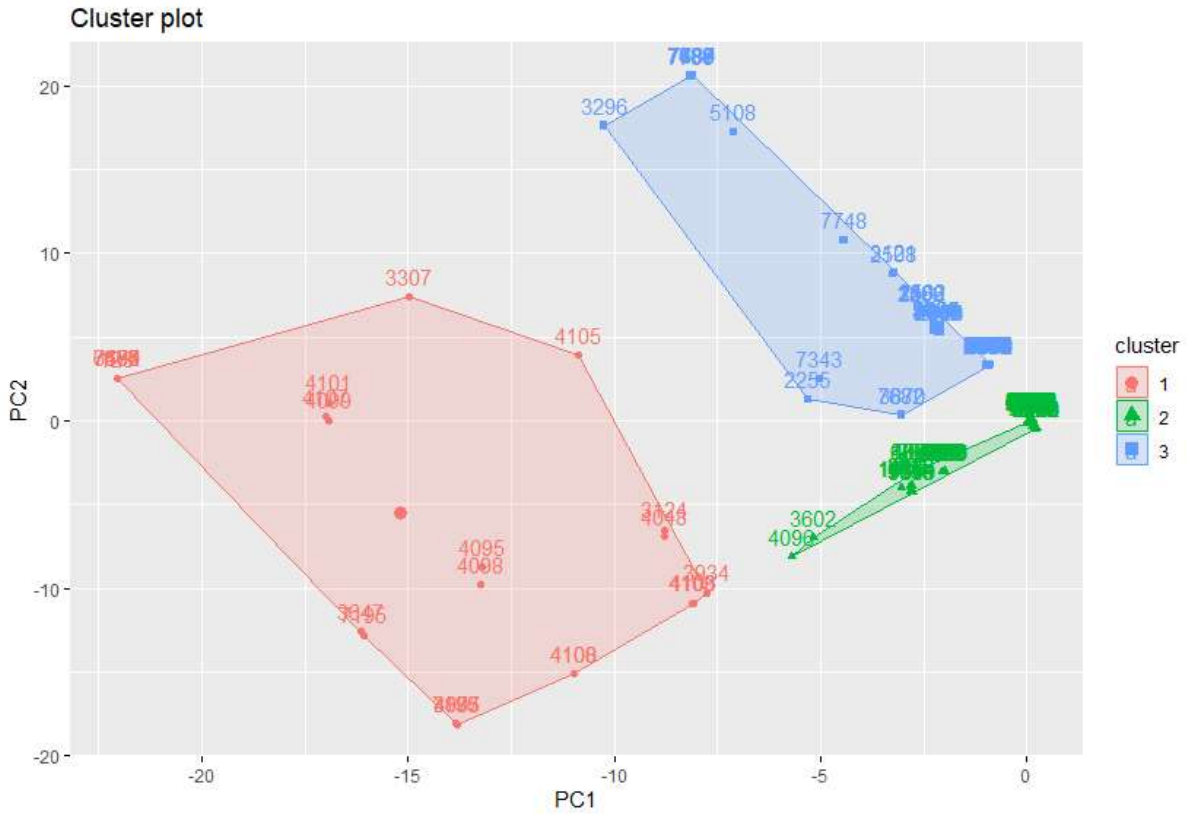

Figure 1.5

A K-mean clustering plot is shown in figure 1.5. We use the “elbow method” to determine the optimal number of clusters to use in our K-mean clustering. As seen in figure 1.4, we can see that the elbow is at 3, so we will equate K=3 for our total number of clusters. Hence, in figure 1.5 we can see clearly that there are 3 cluster in the plot. K-mean clustering is to minimize the variation between intra-cluster points, while also keeping the clusters as different as possible.

Part 2:

- Portuguese Subject

In this part, we will use linear regression to work on data that are being collected. First, we will work on the data of the Portuguese Subject.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.87683770	0.1276353	93.05291715	1.977940e-282
school	-0.56814323	0.1468429	-3.86905608	1.266129e-04
sex	-0.10730473	0.1520303	-0.70581164	4.806966e-01
age	0.10774769	0.1474520	0.73073056	4.653512e-01
address	0.14862987	0.1445894	1.02794438	3.045676e-01
famsize	0.17697020	0.1350896	1.31002094	1.909048e-01
Pstatus	0.04505096	0.1387287	0.32474146	7.455384e-01
Medu	-0.01134772	0.1935969	-0.05861521	9.532865e-01
Fedu	0.35439644	0.1731563	2.04668469	4.131100e-02
Mjob	-0.02358161	0.1536024	-0.15352373	8.780590e-01
Fjob	-0.02984064	0.1393994	-0.21406577	8.305997e-01
reason	0.09913294	0.1323233	0.74917203	4.541729e-01
guardian	0.11037784	0.1358947	0.81223064	4.171194e-01
traveltime	0.01472799	0.1446180	0.10184066	9.189318e-01
studytime	0.33291164	0.1402483	2.37373112	1.805828e-02
failures	-0.77222103	0.1574024	-4.90603161	1.331595e-06
schoolsup	-0.49803945	0.1425685	-3.49333353	5.276918e-04
famsup	-0.01962899	0.1360674	-0.14425931	8.853649e-01
paid	-0.23082017	0.1374651	-1.67911806	9.387277e-02
activities	0.15282257	0.1333972	1.14562044	2.526046e-01
nursery	-0.08310155	0.1339569	-0.62036045	5.353570e-01
higher	0.67128540	0.1419822	4.72795601	3.100711e-06
internet	0.04371701	0.1432060	0.30527353	7.603091e-01
romantic	-0.20462844	0.1317441	-1.55322661	1.211222e-01
famrel	0.06612062	0.1356050	0.48759726	6.260896e-01
freetime	-0.13056589	0.1450300	-0.90026836	3.684934e-01
goout	-0.01321684	0.1531116	-0.08632160	9.312519e-01
Dalc	-0.28328057	0.1622503	-1.74594799	8.155116e-02
Walc	-0.12282096	0.1846963	-0.66498890	5.064226e-01
health	-0.29155622	0.1351064	-2.15797411	3.149605e-02
absences	-0.08835479	0.1392610	-0.63445466	5.261296e-01

Figure 2.1

In total, there are 30 variables shown in the figure above. Variables that are directly correlated to the Portuguese course grade are: age, address, family size, Parental status, Mother education, Father education, reason, etc those that are positive are positively correlated with the course grade, whereas those that are negative like: school, sex, Mother's education, failures, educational support, family educational support, paid extra classes, romantics, etc means the factors are having indirect relation to the grade.

Next, we will use stepwise regression to determine which variables are more significant in affected G3.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.8702     0.1257  94.464  < 2e-16 ***
school       -0.6371     0.1290  -4.939 1.12e-06 ***
Fedu          0.2887     0.1315   2.195 0.028699 *
studytime     0.3934     0.1303   3.019 0.002679 **
failures     -0.7922     0.1441  -5.497 6.56e-08 ***
schoolsup    -0.4983     0.1363  -3.657 0.000286 ***
paid         -0.2164     0.1321  -1.638 0.102064
higher        0.6822     0.1352   5.045 6.63e-07 ***
Dalc         -0.3956     0.1234  -3.206 0.001442 **
health       -0.3280     0.1285  -2.553 0.011013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.2

As seen in Figure 2.2, school, wanting to take higher education, number of past class failures, school and extra education support is the most statistically significant variable in determining G3. The next most statistically significant variable is amount of study time and workday alcohol consumption and at the least, the least statistically significant variable will be father's education and health of the student.

- ## Mathematics Subject

Next, we will look at the Mathematics course subject.

```
                Estimate Std. Error     t value      Pr(>|t|)
(Intercept) 10.521711376  0.2554291 41.19228954 2.205002e-112
school       0.086787725  0.3006868  0.28863165  7.731063e-01
sex          0.593859727  0.3018485  1.96740999  5.025957e-02
age         -0.495244733  0.3256420 -1.52082562  1.295879e-01
address      0.474315867  0.2822384  1.68055067  9.411951e-02
famsize      0.207622122  0.2832284  0.73305550  4.642225e-01
Pstatus     -0.504362709  0.2789108 -1.80832997  7.177626e-02
Medu         0.410734750  0.3826972  1.07326299  2.842048e-01
Fedu         0.142255141  0.3452627  0.41202005  6.806835e-01
Mjob        -0.558928755  0.2985635 -1.87205996  6.238481e-02
Fjob         0.079236777  0.2585967  0.30641057  7.595512e-01
reason       0.412550936  0.2692188  1.53239996  1.267088e-01
guardian     0.262041665  0.2876055  0.91111478  3.631268e-01
traveltime   0.010681583  0.2722257  0.03923797  9.687325e-01
studytime    0.165438406  0.2869226  0.57659584  5.647397e-01
failures    -1.140604810  0.2838132 -4.01885733  7.777868e-05
schoolsup   -0.446498741  0.2858829 -1.56182389  1.196150e-01
famsup      -0.338890709  0.2811037 -1.20557170  2.291410e-01
paid         0.315804634  0.2887744  1.09360327  2.751987e-01
activities   0.138529674  0.2702588  0.51258160  6.087040e-01
nursery     -0.009249268  0.2629081 -0.03518061  9.719643e-01
higher       0.438221054  0.2491118  1.75913395  7.979794e-02
internet    -0.009936859  0.2756299 -0.03605145  9.712706e-01
romantic    -0.482355022  0.2765607 -1.74412009  8.238721e-02
famrel       0.021937602  0.2718533  0.08069646  9.357490e-01
freetime     0.342519919  0.2736725  1.25156883  2.119158e-01
goout       -0.574626804  0.2944829 -1.95130797  5.215549e-02
Dalc        -0.031952422  0.3519540 -0.09078580  9.277367e-01
Walc         0.375664109  0.3710454  1.01244781  3.123188e-01
health      -0.158835756  0.2677130 -0.59330613  5.535214e-01
absences     0.312878919  0.2735074  1.14395049  2.537561e-01
```

Figure 2.3

As seen in figure 2.3, most of the variables have a direct relation with Mathematic subject except, age, Parental status, Mother's job, failures, family education support, attended nursery school, etc those that are negative. Those that are positive like school, sex, address, etc, have a direct relation on the grade of the student in Mathematics course, as they are positively correlated.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.5096     0.2487  42.253  < 2e-16 ***
sex           0.6970     0.2633   2.647   0.0086 **
address       0.4713     0.2451   1.923   0.0556 .
Pstatus      -0.5823     0.2537  -2.295   0.0225 *
Medu          0.5338     0.3026   1.764   0.0789 .
Mjob         -0.5234     0.2809  -1.863   0.0635 .
reason        0.4577     0.2548   1.797   0.0735 .
failures     -1.1939     0.2604  -4.584 7.03e-06 ***
schoolsup    -0.3853     0.2618  -1.471   0.1424
higher        0.5229     0.2338   2.236   0.0262 *
romantic     -0.4203     0.2577  -1.631   0.1041
goout        -0.3701     0.2521  -1.468   0.1433
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2.4

In figure 2.4, we can see that numbers of past class failures are the most significant variable in affecting G3. Subsequently, gender of the student is also one of the variables that is statistically significant followed by Parental status and desire to take higher education in being statistically significant in G3.

- ## Conclusion

As we can see from the dataset, there are many variables affecting each respective subject differently. Gender of the student will affect the Portuguese. The r-square value for the Portuguese is 0.2942924 which indicate only 29.4% of the data fit the regression model as for Mathematics the r-square value is 0.2972521, which means only 29.7% of the data fit the regression model.

In figure 2.5 and 2.6, we use k-fold cross validation to shows the scatter plot of the actual vs predicted values respectively for Mathematic and Portuguese subject. The small symbols are the predicted value while the big symbols are the actual values. Figure 2.5 is more accurate as compare to Figure 2.6 as we can see that in Figure 2.5, there is a stronger correlation between predicted value and G3 as the small symbols are not as spaced out when compared to figure 2.6.

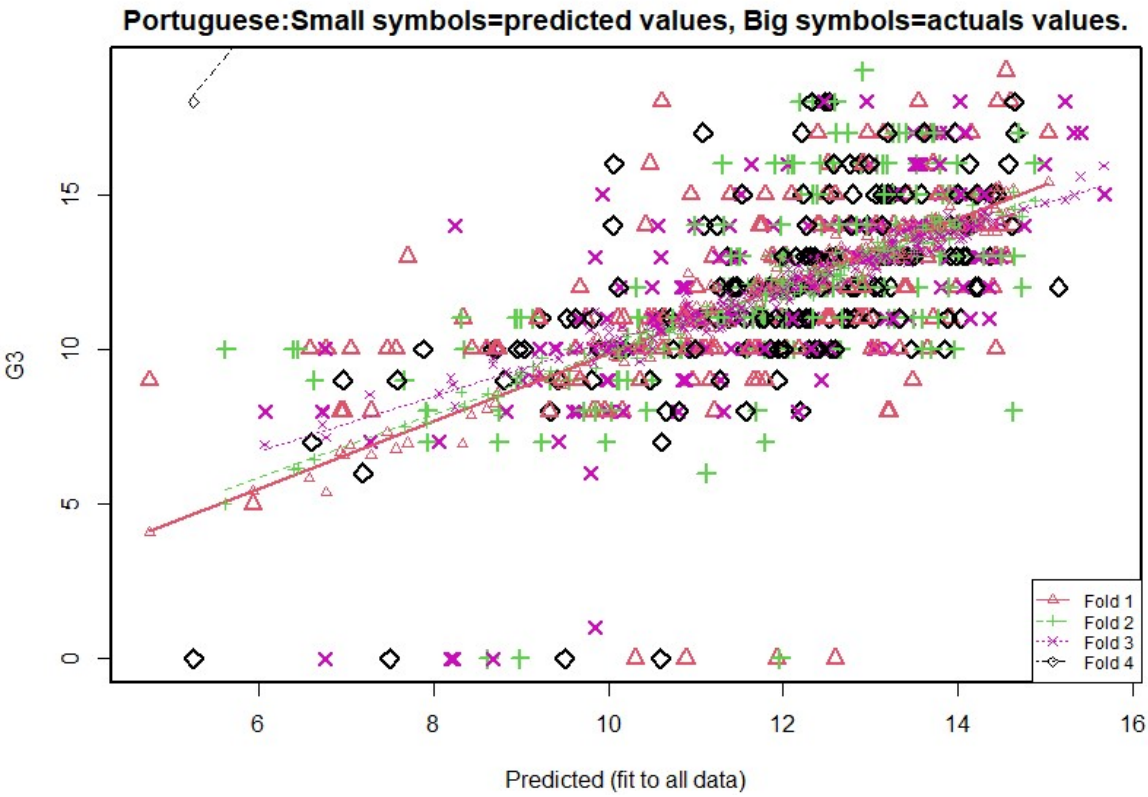

Figure 2.5

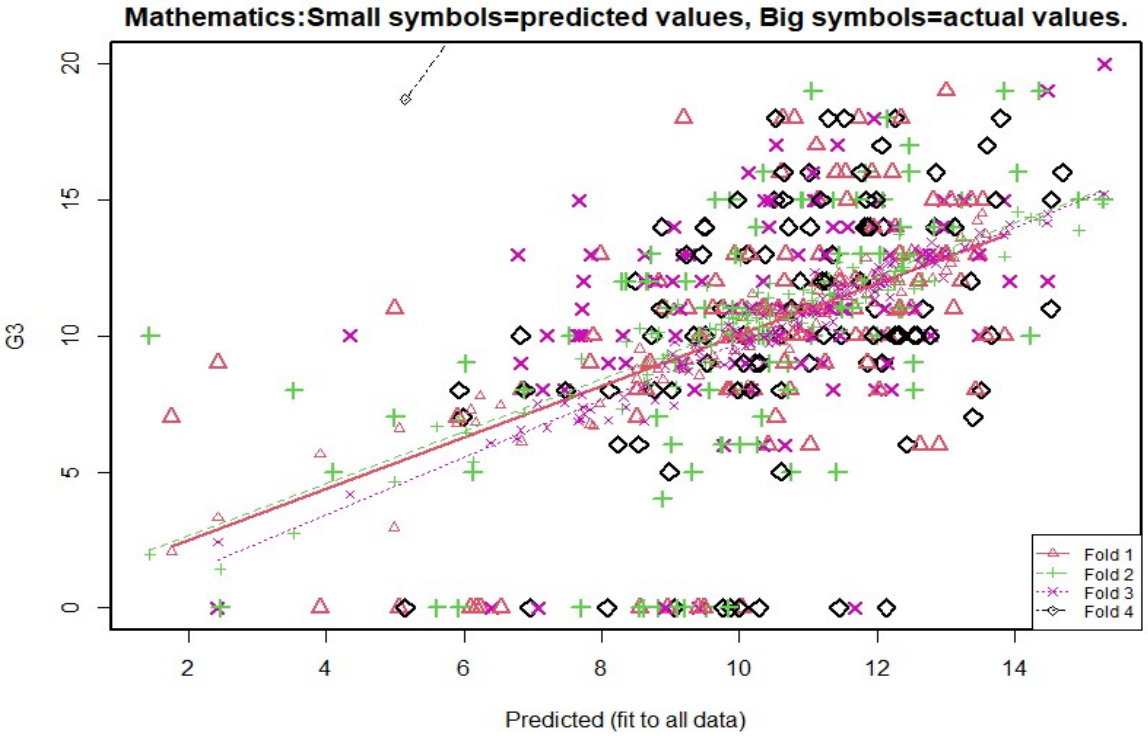

Figure 2.6

Part 3:

In part 3, we will be using the random forest regression to interpret the data. After getting the data, we will able to come up with a confusion matrix as seen in figure 3.1.

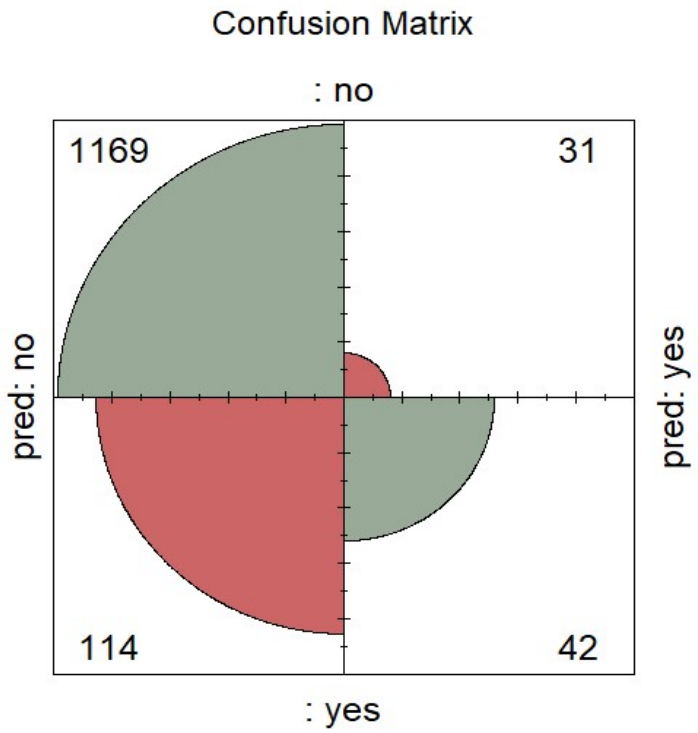

Figure 3.1

So, as seen in figure 3.1, a total of 1,356 samples are taken. Out of those sample, 1,169 are true negative which means that the model correctly predicted that they did not subscribe to the term deposit. Whereas, 42 are true positive which mean that the model accurately predict that 42 customers did subscribe to the term deposit. 114 of the sample are false positive means the model wrongly predict 114 of people subscribing when in fact they did not and 31 are false negative which means that they predict 31 did not subscribe when in fact they did subscribe to the service.

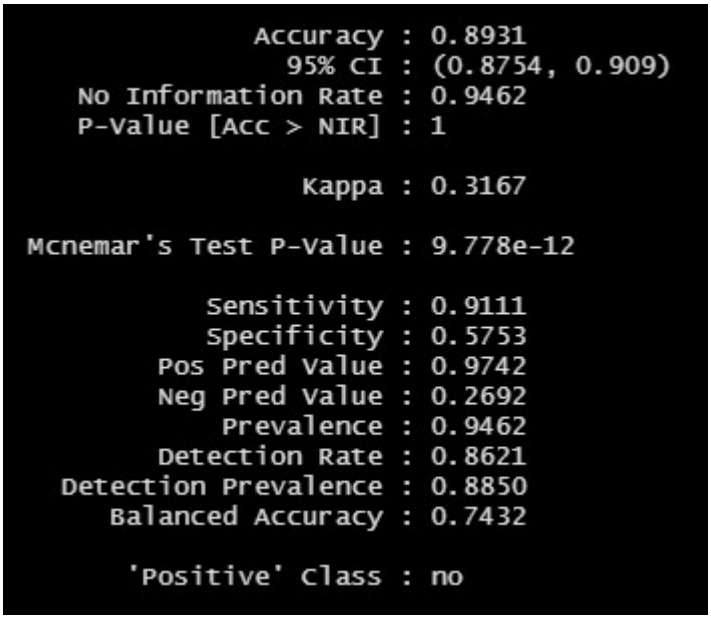

Figure 3.2

In figure 3.2, it is a summary of the calculated R result. It indicated that this model has an accuracy of 0.8931, means that the model will accurately predict 89.3% of the data when compared to the actual values. It has a positive predictive value of 0.9742, meaning it has a 97.2% rate of predicting positives that are actually positive and a negative predictive value of 0.2692, means it has 26.9% chances of predicting negative positives that are actually negative.

References

Jaadi, Z. (4 September, 2019). *Built in*. Retrieved from A step-by-step explanation of Component Analysis: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Sharma, P. (22 July, 2019). *towards datascience*. Retrieved from Decoding the Confusion Matrix: https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb