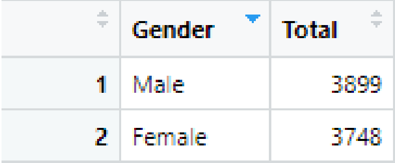
Contents

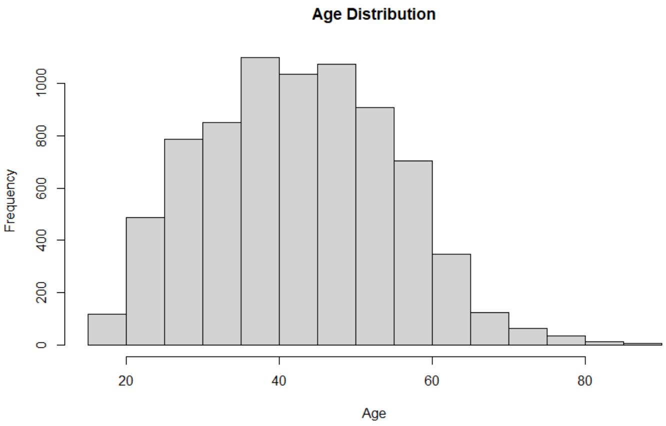
Part 1

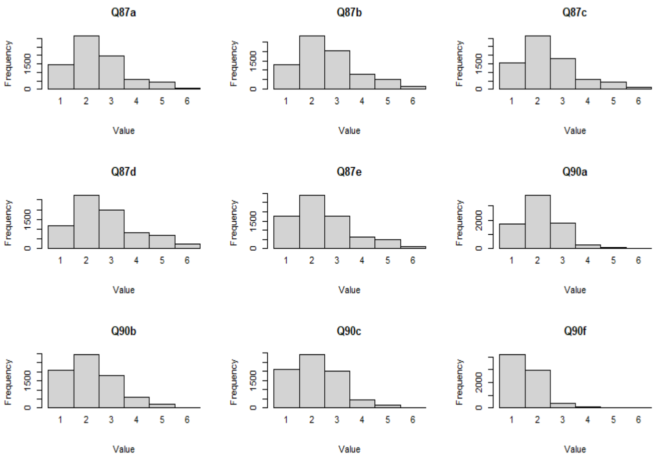
EWCS 2016 is a data set that comprises of 11 variables and 7813 observations. Some of the variables have unusually large negative values of -999 which will be cleaned in order to perform Principal Component Analysis (PCA).



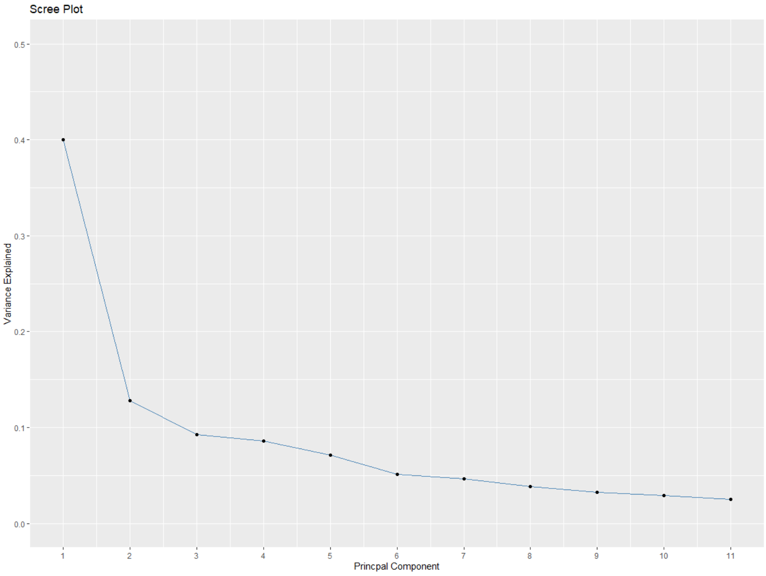
	Gender	Total
1	Male	3899
2	Female	3748

The names of variables Q2a and Q2b will be changed to “Gender” and “Age” respectively for clarification. Next, we will check the distribution for each variable. The Gender table above shows that there are slightly more males than females but the difference is minimal so there should be no dominance by gender. Age is positively skewed, and that majority of the participants in this study are between the ages 35-50.

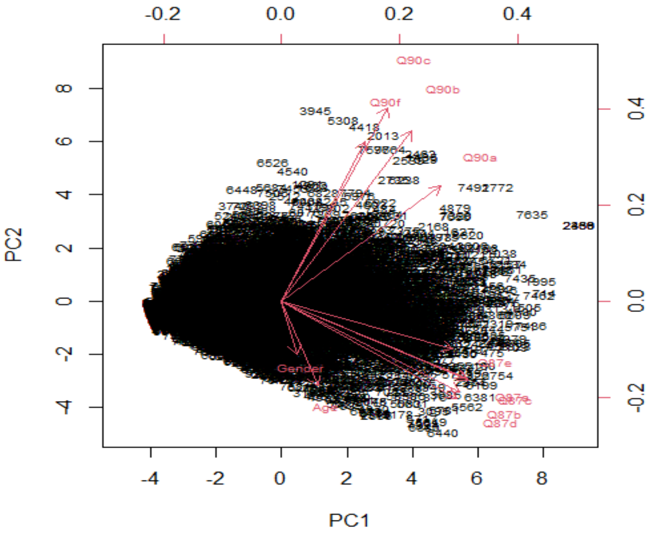


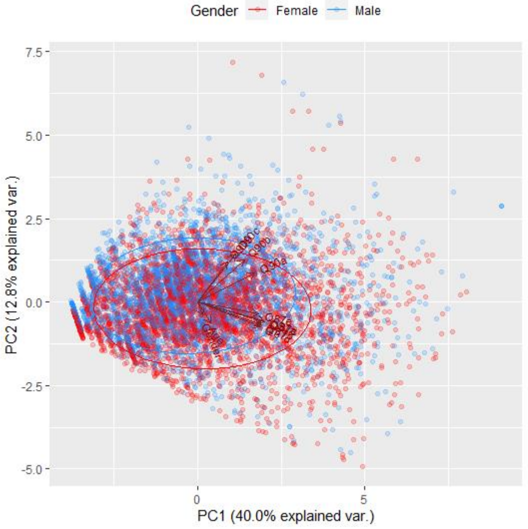


The remaining variables are questions that were answered by the participants. They come in two sets: Q87 and Q90. Set Q87 primarily asks about an individual’s mental and physical well-being, i.e., do they feel cheerful and calm, and do they feel active and rested. Set Q90 focuses on whether they enjoy their job and how they feel about work, e.g., do they feel enthusiastic/energetic at work. For every question except Q90f, many individuals answered with “2: Most of the time.” Many individuals answered with “1: Always.” to “Q90f: In my opinion, I am good at my job.” This suggests that many individuals feel that their physical and mental health is good most of the time, they enjoy their jobs most of the time, and that they are confident at their job.



We will be performing Principal Component Analysis to further explore and analyse this data set. In the scree plot, there is a noticeable drop after the second component. Based on this, we will pick the first two principal components which account for roughly 63% of the variance in the data.

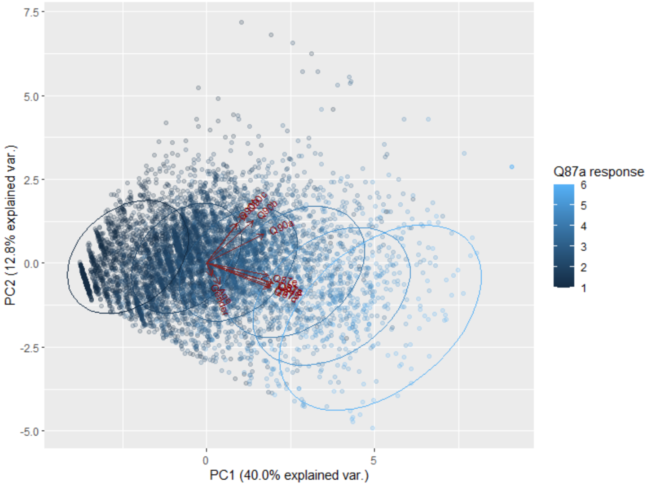




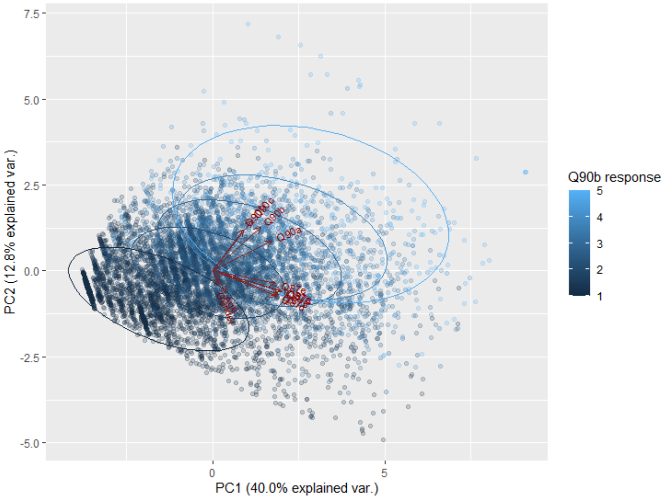
Biplot 1. Standard biplot

Biplot 2. Gender dispersion and ellipses.

To examine both score and loading plots, we will draw a biplot. We can see that all Q87 variables have large positive loadings on component 1. Q90a also has a positive loading on component 1. Recall that while this question was related to how an individual feels about work, it also asked about their energy levels. Taking these into account, we can say that this component focuses on an individual's state of mind and energy levels. Q90 variables have large positive loadings on component 2. Therefore, this component focuses on an individual's feelings about their job. In *Biplot 2* we can see that there are no clusters for the Gender variable. This suggests that gender has no relation to the responses given to the questions.



Biplot 3. Ellipses for Q87a responses.



Biplot 4. Ellipses for Q90b responses.

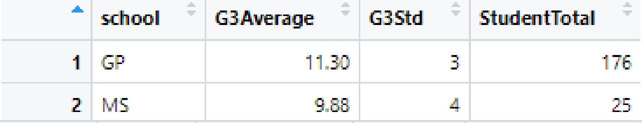
In *Biplot 3* and *Biplot 4*, individuals who responded with “1: Always” to either set of questions are mostly concentrated in the left half of the graphs. Positive responses for either set of questions overlap; individuals who respond positively for one set of questions tend to also respond positively for the other set. However, as the responses become more negative specifically at response “4: Less than half the time” ellipse, the overlap becomes less pronounced. This suggests that while generally people tend to respond positively/neutrally to one set of questions if they respond positively/neutrally to the other set, individuals who respond negatively to one set of questions will not necessarily respond negatively for the other set.

Part 2: Regression

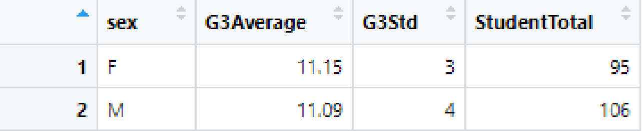
We have two datasets: “student-mat.csv” and “student-por.csv”. They contain data on student performance in Mathematics and Portuguese. Both datasets contain information on the same students, so we will merge them into a single dataset named “grades”. Not all students are in both data sets, so only those that have matching values will be filtered into the new dataset. This new dataset has 35 attributes and 201 observations. There are two response variables in this new dataset: G3.x and G3.y, both of which represent the final grade for Math and Portuguese, respectively. In addition, variables that share the same name are differentiated with either *.x* or *.y*. These variables are absences and failures.

Data preparation

Before merging the two datasets, we will remove the variables G1 and G2 because we want to predict G3 without them. We will also transform variables schoolsup, famsup, paid, activities, nursery, higher, internet and romantic from character type “yes” and “no” to numeric “1” and “0” respectively. This is to allow the creation of a correlation matrix with these variables. I will also merge G3.x and G3.y into a single variable called “G3avg” which will represent the student’s average final grade. G3.x and G3.y will be removed thereafter. This will leave us with 33 variables. Finally, we will look at the relationship between all the variables and the response variable G3avg.



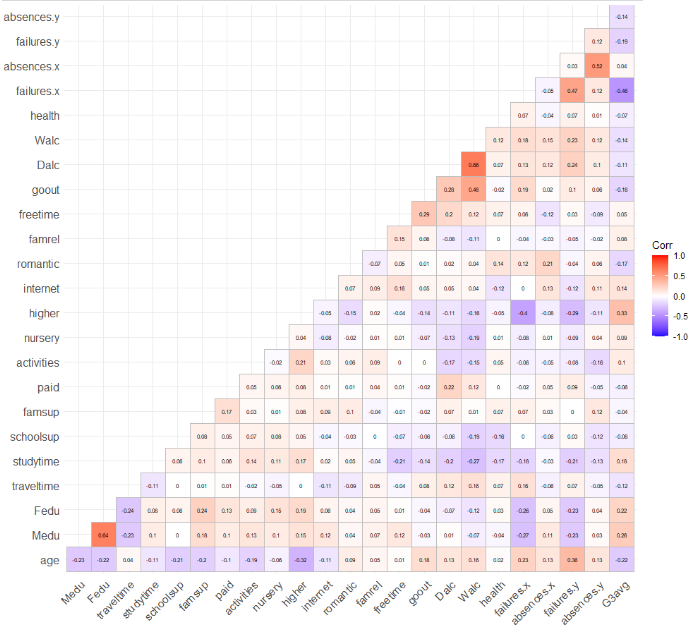
	school	G3Average	G3Std	StudentTotal
1	GP	11.30	3	176
2	MS	9.88	4	25



	sex	G3Average	G3Std	StudentTotal
1	F	11.15	3	95
2	M	11.09	4	106

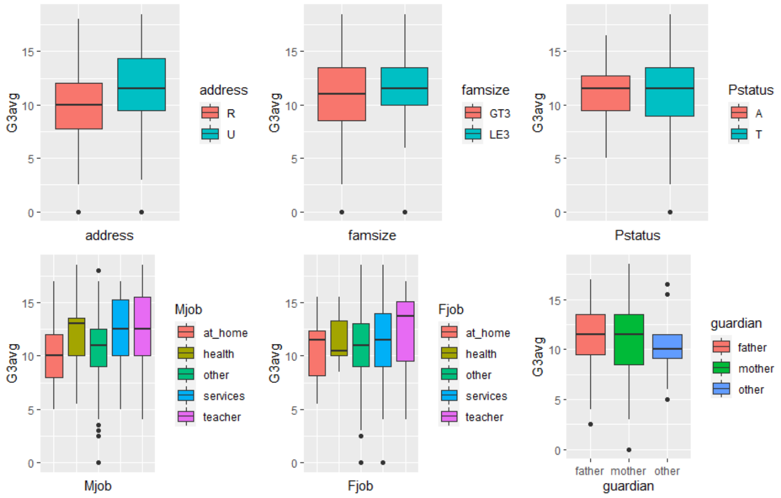
Data analysis

Students from the school GP perform better academically because they have a better final grade on average. There is also lesser variation as evident by the standard deviation. However, there are substantially more students from GP than there are from MS. Females, although fewer in number than males, also perform marginally better on average.

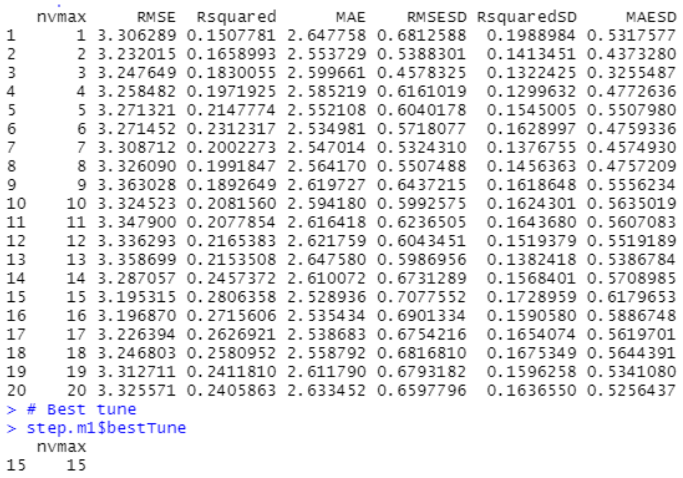


	Medu	Fedu	traveltime	studytime	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	failures.x	absences.x	failures.y	absences.y	G3avg
absences.y																							-0.14
failures.y																						0.12	-0.19
absences.x																					0.03	0.52	0.04
failures.x																				-0.05	0.47	0.12	-0.46
health																			0.07	-0.04	0.07	0.01	-0.07
Walc																		0.12	0.18	0.15	0.23	0.12	-0.14
Dalc																	0.66	0.07	0.13	0.12	0.24	0.1	-0.11
goout																0.28	0.46	-0.02	0.19	0.02	0.1	0.06	-0.18
freetime															0.29	0.2	0.12	0.07	0.06	-0.12	0.03	-0.09	0.05
famrel														0.15	0.06	-0.08	-0.11	0	-0.04	-0.03	-0.05	-0.02	0.08
romantic													-0.07	0.05	0.01	0.02	0.04	0.14	0.12	0.21	-0.04	0.06	-0.17
internet												0.07	0.09	0.16	0.05	0.05	0.04	-0.12	0	0.13	-0.12	0.11	0.14
higher											-0.05	-0.15	0.02	-0.04	-0.14	-0.11	-0.16	-0.05	-0.4	-0.08	-0.29	-0.11	0.33
nursery										0.04	-0.08	-0.02	0.01	0.01	-0.07	-0.13	-0.19	0.01	-0.08	0.01	-0.09	0.04	0.09
activities									-0.02	0.21	0.03	0.06	0.09	0	0	-0.17	-0.15	0.05	-0.06	-0.05	-0.08	-0.16	0.1
paid								0.05	0.06	0.08	0.01	0.01	0.04	0.01	-0.02	0.22	0.12	0	-0.02	0.05	0.09	-0.05	-0.06
famsup							0.17	0.03	0.01	0.08	0.09	0.1	-0.04	-0.01	-0.02	0.07	0.01	0.07	0.07	0.03	0	0.12	-0.04
schoolsup						0.08	0.05	0.07	0.08	0.05	-0.04	-0.03	0	-0.07	-0.06	-0.06	-0.19	-0.16	0	-0.06	0.03	-0.12	-0.08
studytime					0.06	0.1	0.06	0.14	0.11	0.17	0.02	0.05	-0.04	-0.21	-0.14	-0.2	-0.27	-0.17	-0.18	-0.03	-0.21	-0.13	0.18
traveltime				-0.11	0	0.01	0.01	-0.02	-0.05	0	-0.11	-0.09	0.05	-0.04	0.08	0.12	0.16	0.07	0.16	-0.06	0.07	-0.05	-0.12
Fedu			-0.24	0.08	0.06	0.24	0.13	0.09	0.15	0.19	0.06	0.04	0.05	0.01	-0.04	-0.07	-0.12	0.03	-0.26	0.05	-0.23	0.04	0.22
Medu		0.64	-0.23	0.1	0	0.18	0.1	0.13	0.1	0.15	0.12	0.04	0.07	0.12	-0.03	0.01	-0.07	-0.04	-0.27	0.11	-0.23	0.03	0.26
age	-0.23	-0.22	0.04	-0.11	-0.21	-0.2	-0.1	-0.19	-0.06	-0.32	-0.11	0.09	0.05	0.01	0.18	0.13	0.16	0.02	0.23	0.13	0.36	0.13	-0.22

The correlation matrix shows that there is a moderately strong negative relationship between G3avg and failures.x (math). This tells us that the more failures a student has in math, the lower their final grade. Other notable variables are higher, Medu and Fedu, all of which have a positive relationship with G3avg. A student is more likely to have a higher final grade if they are pursuing higher education or if their parents were highly educated. Variable schoolsup has a negative relationship with G3avg, which is quite interesting as it suggests that students tend to have a lower final grade if they receive educational support. However, we can look at this as students who perform badly in their academics receive educational support to improve their grades. This would mean that the support is not the cause of their final grade and it is instead a classifier where students who receive the support tend to have a lower final grade.



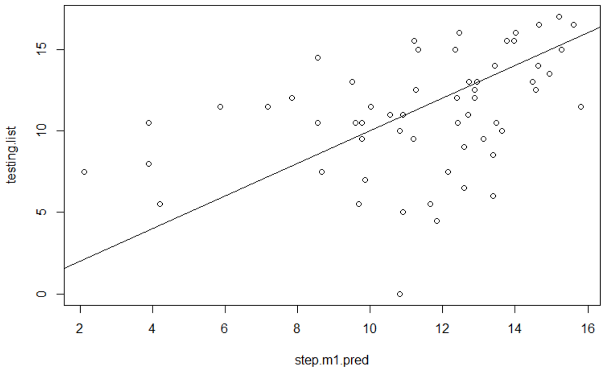
For the other variables not included in the matrix, we will draw boxplots to examine the relationship. We see that urban students tend to have a higher final grade. The type of profession a parent has influences the student’s final grade as well. Students with a guardian listed as “other” typically score lower in their final grade compared to students with mothers and fathers as guardians, but there are some outliers where they scored high.



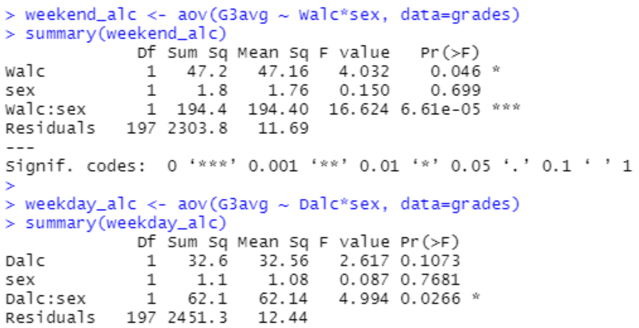
```
   nvmax     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
1      1 3.306289 0.1507781 2.647758 0.6812588  0.1988984 0.5317577
2      2 3.232015 0.1658993 2.553729 0.5388301  0.1413451 0.4373280
3      3 3.247649 0.1830055 2.599661 0.4578325  0.1322425 0.3255487
4      4 3.258482 0.1971925 2.585219 0.6161019  0.1299632 0.4772636
5      5 3.271321 0.2147774 2.552108 0.6040178  0.1545005 0.5507980
6      6 3.271452 0.2312317 2.534981 0.5718077  0.1628997 0.4759336
7      7 3.308712 0.2002273 2.547014 0.5324310  0.1376755 0.4574930
8      8 3.326090 0.1991847 2.564170 0.5507488  0.1456363 0.4757209
9      9 3.363028 0.1892649 2.619727 0.6437215  0.1618648 0.5556234
10    10 3.324523 0.2081560 2.594180 0.5992575  0.1624301 0.5635019
11    11 3.347900 0.2077854 2.616418 0.6236505  0.1643680 0.5607083
12    12 3.336293 0.2165383 2.621759 0.6043451  0.1519379 0.5519189
13    13 3.358699 0.2153508 2.647580 0.5986956  0.1382418 0.5386784
14    14 3.287057 0.2457372 2.610072 0.6731289  0.1568401 0.5708985
15    15 3.195315 0.2806358 2.528936 0.7077552  0.1728959 0.6179653
16    16 3.196870 0.2715606 2.535434 0.6901334  0.1590580 0.5886748
17    17 3.226394 0.2626921 2.538683 0.6754216  0.1654074 0.5619701
18    18 3.246803 0.2580952 2.558792 0.6816810  0.1675349 0.5644391
19    19 3.312711 0.2411810 2.611790 0.6793182  0.1596258 0.5341080
20    20 3.325571 0.2405863 2.633452 0.6597796  0.1636550 0.5256437
> # Best tune
> step.m1$bestTune
   nvmax
15    15
```

Stepwise Regression

With 33 variables in our data set, we want to reduce this number and use only the variables that are relevant to our regression model. We will use Stepwise Regression with backward selection to conduct this because the number of observations is larger than the number of predictors. First, we will split the grades dataset: 70% will be used for training and the remaining 30% will be for testing. We will also use K-fold cross-validation where k=10 and set the tuning parameter *nvmax* to 20 which will search for the best models up to 20 variables. The algorithm has found the model with 15 variables is the best. This model with 15 variables has a RMSE of 3.6651 which means that this model is not wholly accurate, evidently shown on the



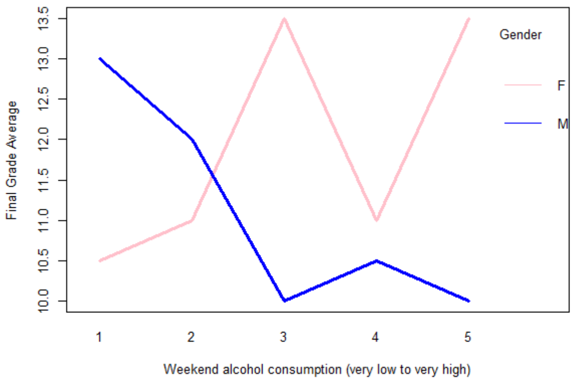
scatterplot above. Next, we will use a simple multiple linear regression model that uses the same variables yielded from the stepwise regression model but with interaction terms.

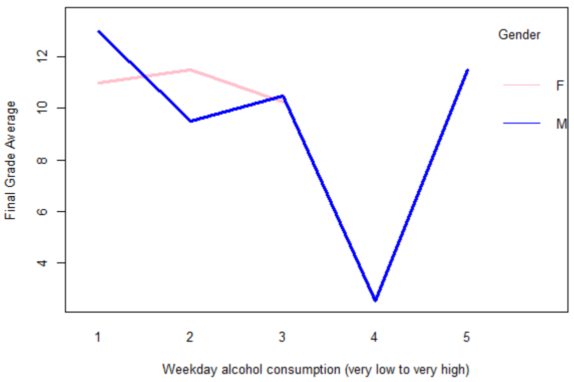


```
> weekend_alc <- aov(G3avg ~ Walc*sex, data=grades)
> summary(weekend_alc)
             Df Sum Sq Mean Sq F value   Pr(>F)    
Walc          1   47.2   47.16   4.032    0.046 *  
sex           1    1.8    1.76   0.150    0.699    
Walc:sex      1  194.4  194.40  16.624 6.61e-05 ***
Residuals   197 2303.8   11.69                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
> weekday_alc <- aov(G3avg ~ Dalc*sex, data=grades)
> summary(weekday_alc)
             Df Sum Sq Mean Sq F value Pr(>F)  
Dalc          1   32.6   32.56   2.617 0.1073  
sex           1    1.1    1.08   0.087 0.7681  
Dalc:sex      1   62.1   62.14   4.994 0.0266 *
Residuals   197 2451.3   12.44                 
```

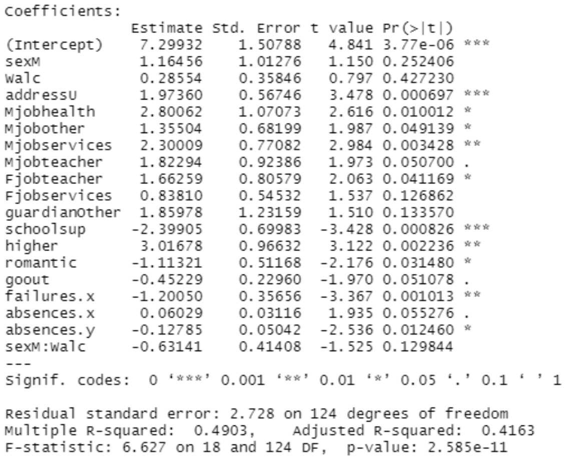
Multiple Linear Regression

From the correlation matrix above, we know that alcohol consumption has a negative relationship with a student’s final grade albeit weak. Using ANOVA, we find that the interaction between sex and alcohol consumption is significant.





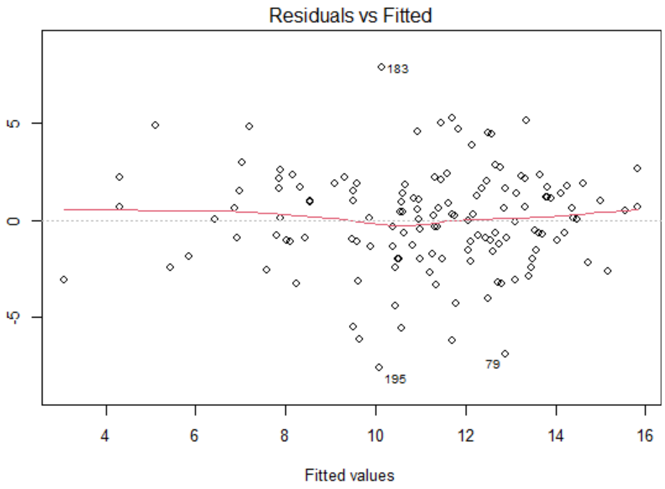
The interaction plot shows that the lines are not parallel, indicating an interaction effect. Interestingly, girls who drink more on the weekend tend to have varying results: they may not have higher grades. Boys, on the other hand, get lower grades the more they drink. Not many girls drink on weekdays as compared to boys as shown on the right interaction plot. As such, it will be disregarded.

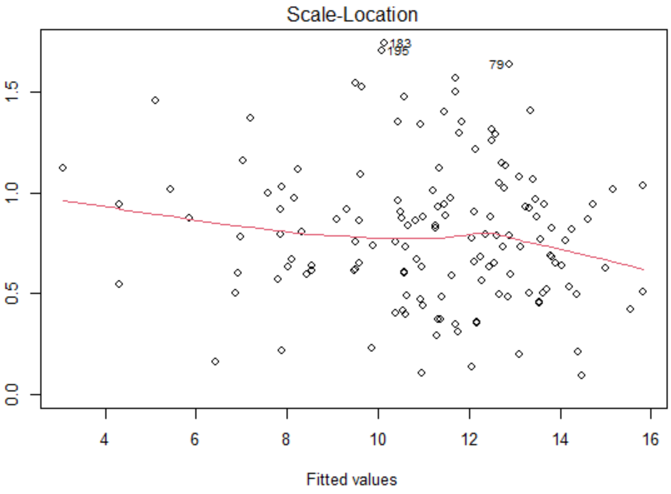


```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     7.29932    1.50788   4.841 3.77e-06 ***
sexM            1.16456    1.01276   1.150 0.252406    
Walc            0.28554    0.35846   0.797 0.427230    
addressU        1.97360    0.56746   3.478 0.000697 ***
Mjobhealth      2.80062    1.07073   2.616 0.010012 *  
Mjobother       1.35504    0.68199   1.987 0.049139 *  
Mjobservices    2.30009    0.77082   2.984 0.003428 ** 
Mjobteacher     1.82294    0.92386   1.973 0.050700 .  
Fjobteacher     1.66259    0.80579   2.063 0.041169 *  
Fjobservices    0.83810    0.54532   1.537 0.126862    
guardianother   1.85978    1.23159   1.510 0.133570    
schoolsup      -2.39905    0.69983  -3.428 0.000826 ***
higher          3.01678    0.96632   3.122 0.002236 ** 
romantic       -1.11321    0.51168  -2.176 0.031480 *  
goout          -0.45229    0.22960  -1.970 0.051078 .  
failures.x     -1.20050    0.35656  -3.367 0.001013 ** 
absences.x      0.06029    0.03116   1.935 0.055276 .  
absences.y     -0.12785    0.05042  -2.536 0.012460 *  
sexM:Walc      -0.63141    0.41408  -1.525 0.129844    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.728 on 124 degrees of freedom
Multiple R-squared:  0.4903,	Adjusted R-squared:  0.4163 
F-statistic: 6.627 on 18 and 124 DF,  p-value: 2.585e-11
```

The final model with sex and Walc as an interaction term has adjusted R^2 value of 41.6% and a slightly better RMSE of 3.5353. However, if we were to check this model for heteroscedasticity on a plot, we will find that the line is not completely horizontal and there are more points as fitted values increase (not equally distributed). The scatterplot on the right with standardized residuals has a line that goes downwards as more fitted values are added, indicating that heteroscedasticity in this model exists.

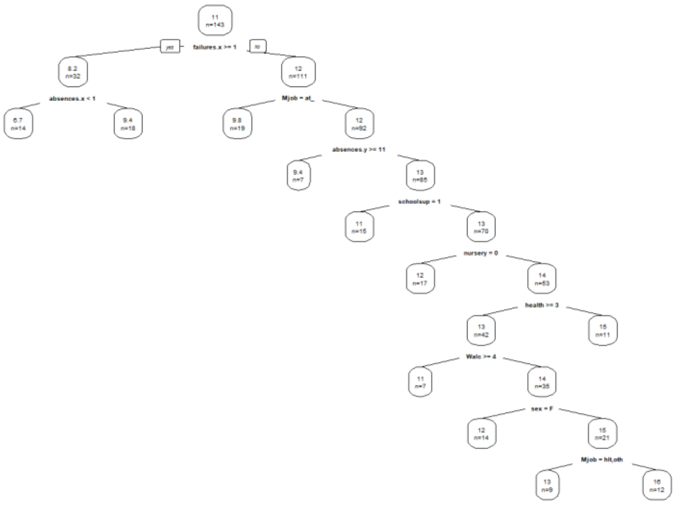




Regression Trees (CART and Random Forest)

Finally, we will try building models with regression trees. Using the rpart package, we build a simple CART model. This model has a RMSE of 2.821, which is much better than the previous two models.

Next, we will build a more complex model with Random Forest. We will use K-fold cross validation to train this model where k=10. 500 trees are trained, and the optimal number of variables sampled at each split is 21. Although the RMSE for this model is 2.9241, only 21.86% of the variance in the model is explained by the predictor variables.



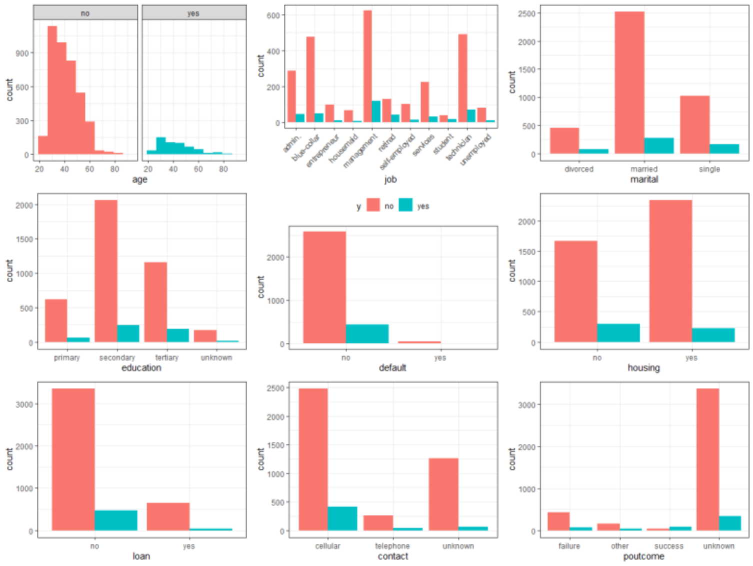
Conclusion

Among all the models, the CART model has the lowest RMSE of 2.821. In terms of RMSE, it is the best model to predict G3, followed by Random Forest.

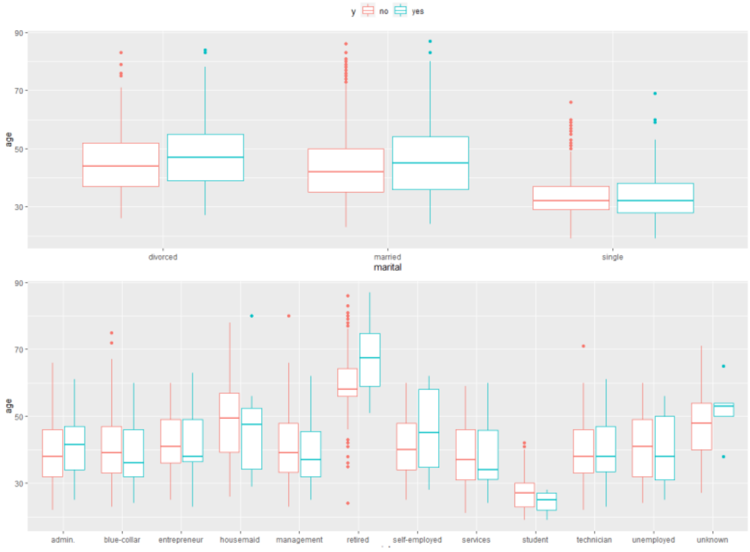
Part 3: Classification

The bank.csv dataset comprises of 4521 observations with 17 variables. Before we begin analysing the data, we will remove 3 variables from the table. The duration variable is removed because it highly affects the outcome variable y. We will also remove the month and day_of_week variables due to their relation to duration. This leaves us with 14 variables.

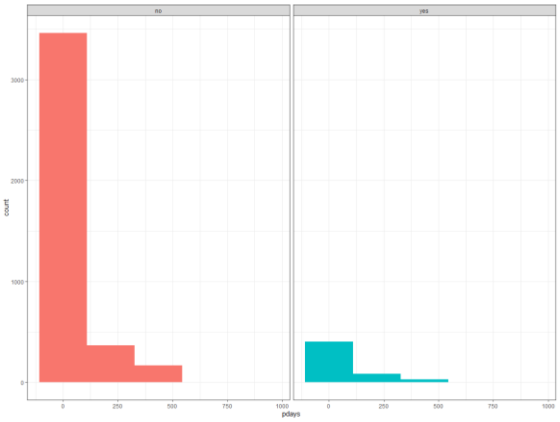
Data analysis



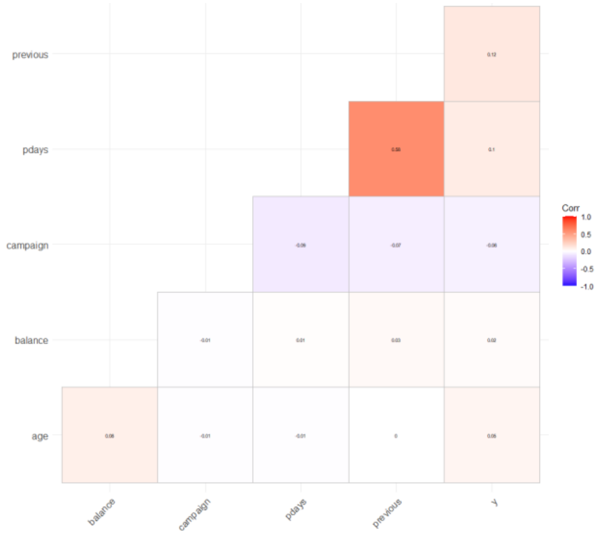
Bar charts were drawn for each categorical variable to find insights into their relationship with y. We found that younger people, specifically those at the ages between 30-50 are more likely to subscribe to a term deposit. The people more likely to subscribe are people working in management and services jobs and retired clients. The more educated a person is, the more likely they are to subscribe. People with housing and people with loans are less likely to subscribe. One thing to note here is the existence of unknown values in education, contact and poutcome. There are many unknown values in poutcome and contact and it does not appear that these variables have a significant influence on y so they will be removed for a more complete dataset. There are a few unknown variables in job and education so we will remove observations that have unknown values in these variables. We can glean more insights by splitting marital and job by age.



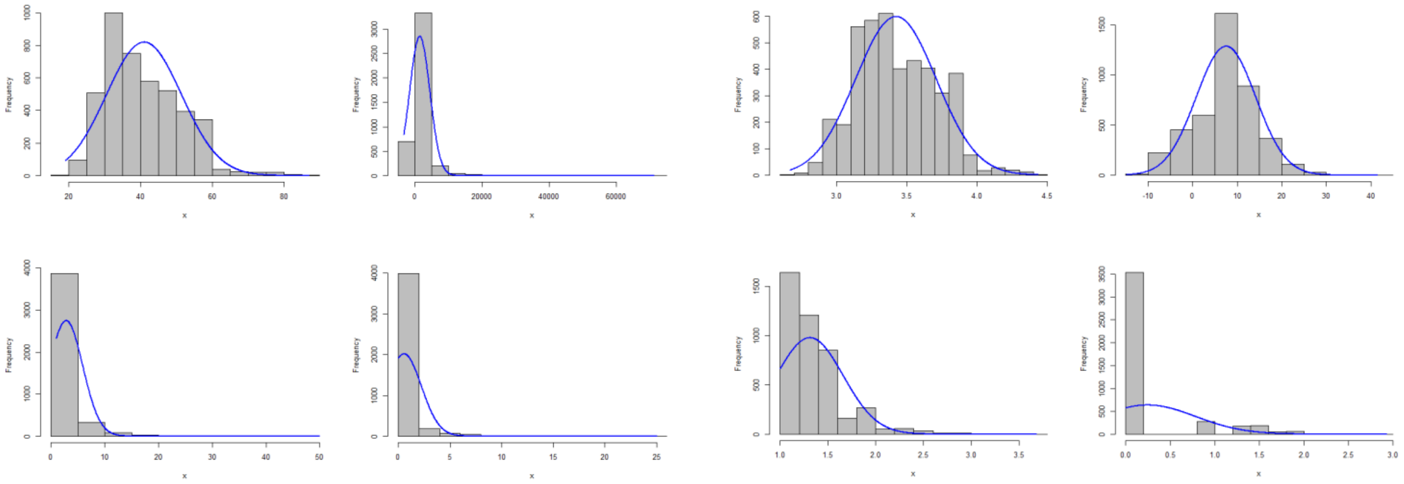
There are many older married clients as well as older single clients who did not subscribe and these are labeled as outliers. The people more likely to subscribe are: Older people in admin jobs, younger people in management blue-collar jobs and older people who are retired. Interestingly, there are a few people younger than 40 who are retired, and they are labeled as outliers. The “retired but did not subscribe” category also has many outliers. Next, we will look at the correlation between numeric variables and y. Before that, we must look at pdays. There is an unusual value of -1 in this variable and we are not told what this value signifies. There are no values of 999 that signifies that people were not previously contacted. The histogram for this variable on the right shows there are many observations with -1 value.



The correlation matrix tells us that the numeric variables all share a weak relationship with y. However, pdays and previous have a moderately strong relationship with each other where the more days that have passed since last contact from the previous campaign, the more contacts performed before the current campaign.



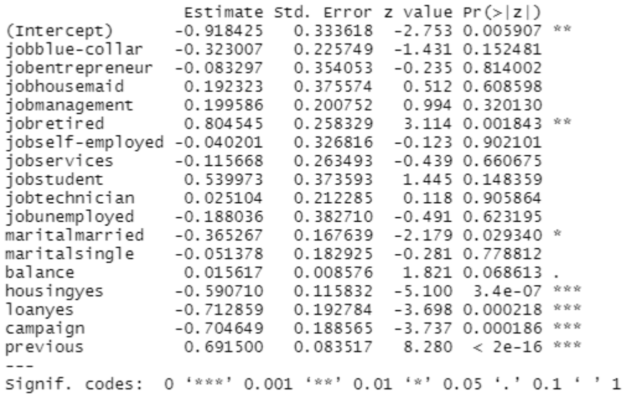
In summary, contact, pdays and poutcome will be removed from the dataset whereas observations with unknown values in job and education will be removed. The outliers found will be kept as they are not rare and are not incorrectly entered according to common sense.



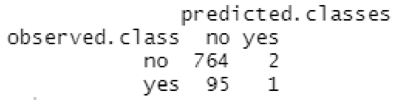
Every numeric variable (age, balance, pdays, previous) aside from age is heavily skewed to the right. Before training our models, we need to have normally distributed predictor variables. Since negative and non-zero values exist, we cannot apply log transformation so cube-root transformation will be used instead. The transformed values look much better, although pdays and previous are still skewed to the right.

Logistic Regression

We will split the dataset: 80% will be used for training and the remaining 20% will be for testing. We will train a logistic regression model with backward selection. The final model uses job, marital, balance, housing, loan, campaign and previous as predictor variables. The coefficient estimate of housing is -0.590710. This would mean that if a client has housing, the probability of them subscribing decreases. The classification prediction accuracy is 89% and the misclassification error rate for this model is 11%, which is good. The confusion matrix also shows that it has at least predicted a true-positive correctly. However, this model has weak sensivity.



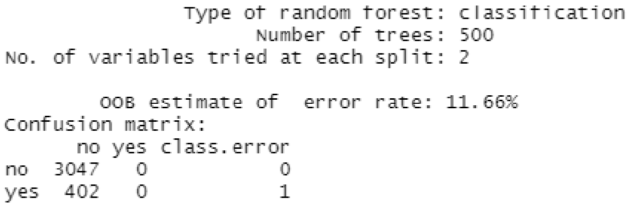
```
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -0.918425   0.333618  -2.753 0.005907 ** 
jobblue-collar   -0.323007   0.225749  -1.431 0.152481    
jobentrepreneur  -0.083297   0.354053  -0.235 0.814002    
jobhousemaid      0.192323   0.375574   0.512 0.608598    
jobmanagement     0.199586   0.200752   0.994 0.320130    
jobretired        0.804545   0.258329   3.114 0.001843 ** 
jobself-employed -0.040201   0.326816  -0.123 0.902101    
jobservices      -0.115668   0.263493  -0.439 0.660675    
jobstudent        0.539973   0.373593   1.445 0.148359    
jobtechnician     0.025104   0.212285   0.118 0.905864    
jobunemployed    -0.188036   0.382710  -0.491 0.623195    
maritalmarried   -0.365267   0.167639  -2.179 0.029340 *  
maritalsingle    -0.051378   0.182925  -0.281 0.778812    
balance           0.015617   0.008576   1.821 0.068613 .  
housingyes       -0.590710   0.115832  -5.100  3.4e-07 ***
loanyes          -0.712859   0.192784  -3.698 0.000218 ***
campaign         -0.704649   0.188565  -3.737 0.000186 ***
previous          0.691500   0.083517   8.280  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
             predicted.classes
observed.class  no yes
           no  764   2
           yes  95   1
```

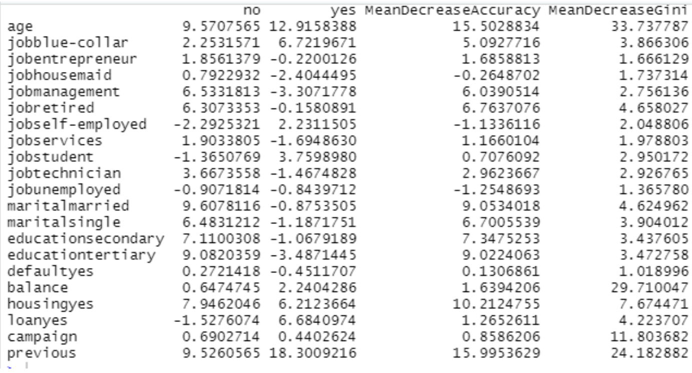
Random Forest Classifier

We will also use K-fold cross validation to train this model where k=10. 500 trees are trained, and the optimal number of variables sampled at each split is 2. This model has a classification prediction accuracy is 89% and the misclassification error rate 11%, which is the same as our logistic regression model. However, if we look at the confusion matrix, this model has extremely weak sensitivity compared to our logistic regression model. The two most important variables in this model are age and previous, both of which reduce the prediction accuracy when excluded from the model.



```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 11.66%
Confusion matrix:
      no yes class.error
no  3047   0           0
yes  402   0           1
```



```
                         no        yes MeanDecreaseAccuracy MeanDecreaseGini
age               9.5707565 12.9158388           15.5028834        33.737787
jobblue-collar    2.2531571  6.7219671            5.0927716         3.866306
jobentrepreneur   1.8561379 -0.2200126            1.6858813         1.666129
jobhousemaid      0.7922932 -2.4044495           -0.2648702         1.737314
jobmanagement     6.5331813 -3.3071778            6.0390514         2.756136
jobretired        6.3073353 -0.1580891            6.7637076         4.658027
jobself-employed -2.2925321  2.2311505           -1.1336116         2.048806
jobservices       1.9033805 -1.6948630            1.1660104         1.978803
jobstudent       -1.3650769  3.7598980            0.7076092         2.950172
jobtechnician     3.6673558 -1.4674828            2.9623667         2.926765
jobunemployed    -0.9071814 -0.8439712           -1.2548693         1.365780
maritalmarried    9.6078116 -0.8753505            9.0534018         4.624962
maritalsingle     6.4831212 -1.1871751            6.7005539         3.904012
educationsecondary 7.1100308 -1.0679189           7.3475253         3.437605
educationtertiary 9.0820359 -3.4871445            9.0224063         3.472758
defaultyes        0.2721418 -0.4511707            0.1306861         1.018996
balance           0.6474745  2.2404286            1.6394206        29.710047
housingyes        7.9462046  6.2123664           10.2124755         7.674471
loanyes          -1.5276074  6.6840974            1.2652611         4.223707
campaign          0.6902714  0.4402624            0.8586206        11.803682
previous          9.5260565 18.3009216           15.9953629        24.182882
```

Conclusion

Both models share the same misclassification error rate of 11% and they both have weak sensitivity but strong specificity. The Logistic Regression model is better in the sense that its sensitivity is slightly stronger, most likely due to it using more predictor variables.