Do you tend to drink more?

I. Motivation

Suppose we are an event-holder who want to provide some drinks in our events, it is important to buy the optimal number in order to leave no drink. Another situation is if we plan to run a bar, finding the best location will be a top priority for us. Both situations need to know whether their guests or customers' drinking habit. Hence, we think that building a suitable model to predict a person's drink habit would be a proper approach for these situations.

II. Data Explanation

There are total 30 variables in our dataset, which are shown and explained as below.

i. Alc - alcohol consumption (numeric: from 1 - very low to 5 - very high)

ii. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

iii. sex - student's sex (binary: 'F' - female or 'M' - male)

iv. age - student's age (numeric: from 15 to 22)

v. address - student's home address type (binary: 'U' - urban or 'R' - rural)

vi. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

vii. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

viii. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

ix. guardian - student's guardian (nominal: 'mother', 'father' or 'other')

x. traveltime - home to school travel time (numeric: 1 - 1 hour)

xi. studytime - weekly study time (numeric: 1 - 10 hours)

xii. failures - number of past class failures (numeric: n if 1<=n<3, else 4)

xiii. schoolsup - extra educational support (binary: yes or no)

xiv. famsup - family educational support (binary: yes or no)

xv. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

xvi. activities - extra-curricular activities (binary: yes or no)

xvii. nursery - attended nursery school (binary: yes or no)

xviii. higher - wants to take higher education (binary: yes or no)

xix. internet - Internet access at home (binary: yes or no)

xx. romantic - with a romantic relationship (binary: yes or no)

xxi. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

xxii. freetime - free time after school (numeric: from 1 - very low to 5 - very high)

xxiii. goout - going out with friends (numeric: from 1 - very low to 5 - very high)

xxiv. health - current health status (numeric: from 1 - very bad to 5 - very good)

xxv. absences - number of school absences (numeric: from 0 to 93)

xxvi. mathG1 - first period grade for math (numeric: from 0 to 20)

xxvii. mathG2 - second period grade for math (numeric: from 0 to 20)

xxviii. mathG3 - final grade for math (numeric: from 0 to 20, output target)

xxix. porG1 - first period grade for Portuguese (numeric: from 0 to 20)

xxx. porG2 - second period grade for Portuguese (numeric: from 0 to 20)

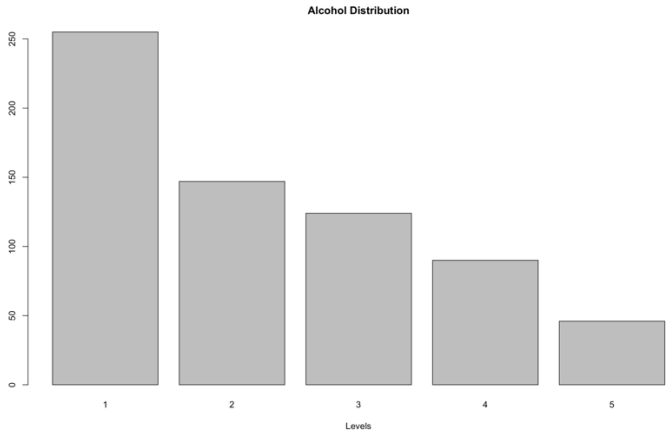
xxxi. porG3 - final grade for Portuguese (numeric: from 0 to 20, output target)

III. Data Preprocessing

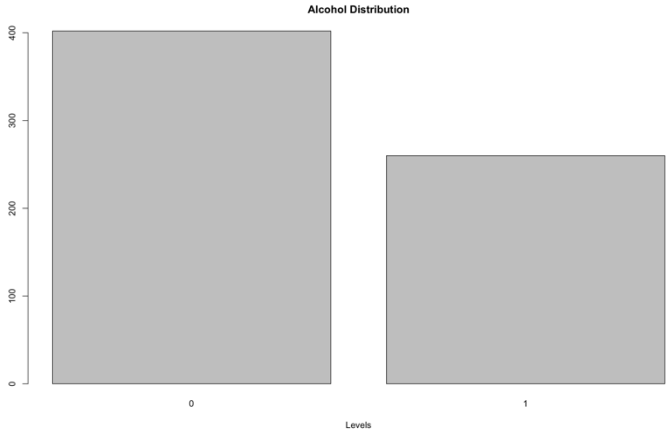
Since there are multiple variables that are category variables, we decided to transform them into dummy variable. For those variables that only have two non-numeric value, such as ‘school’, ‘sex’ or ‘schoolsup’, we use binary value, 0 and 1, to represent them. Then for those variables with more than 2 value, we created some new variables such as ‘r_course’ and ‘r_home’ to represent the reason that the student chose that school.

Moreover, we found that the grades for math and Portuguese are highly correlated to each other in same subject by plotting the correlation graph. Hence, we decided to calculate average grades for both subjects, which resulted in two variables, ‘math_avg_grade’ and ‘por_avg_grade’.

The last thing we did in preprocessing is dealing with the output variable, ‘Alc’. We wanted the result to be binary in order to fit in Logistic Regression and also could use ROC curve to show how well the model is, so we turned Alc’s value to 0 if the original value was 1 or 2, and the other value were turned into 1.



(Before Transforming)



(After Transforming)

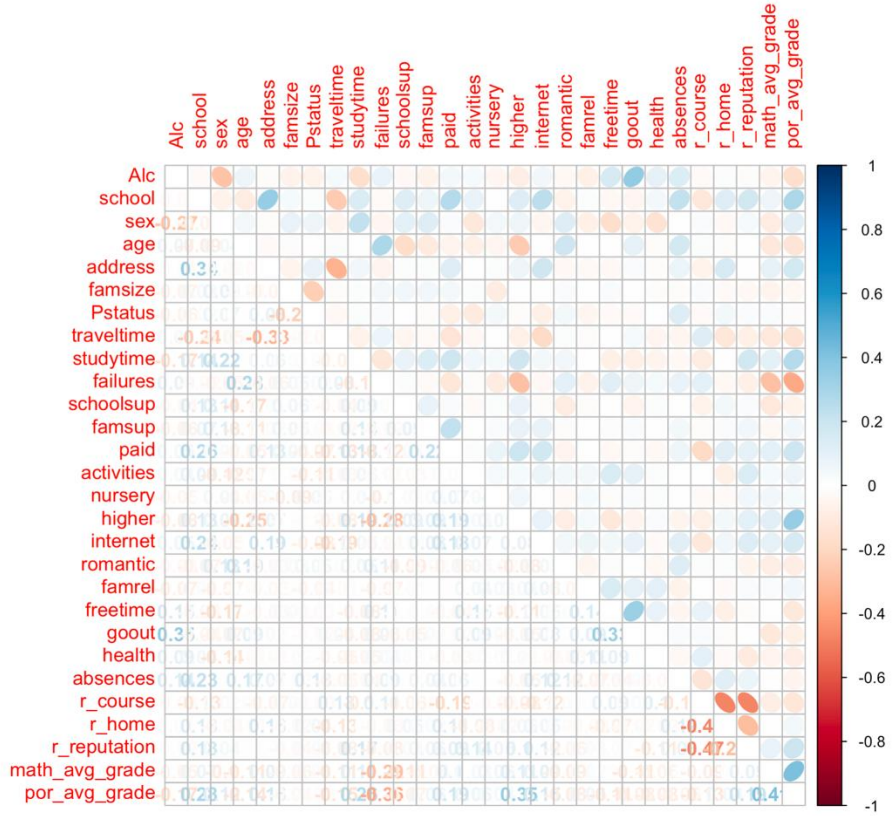
As a result, we have 28 variables in total. And the correlation between variables is shown as below.

	Alc	school	sex	age	address	famsize	Pstatus
Alc	1.00000000	-0.022478340	-0.271422451	0.065121845	-0.013839771	-0.067654330	-0.0616317644
school	-0.02247834	1.000000000	-0.069006559	-0.094516906	0.344107897	0.033000939	0.0152264508
sex	-0.27142245	-0.069006559	1.000000000	0.041280282	-0.017267860	0.089566936	0.0658548161
age	0.06512185	-0.094516906	0.041280282	1.000000000	-0.024995327	-0.012043772	0.0091605521
address	-0.01383977	0.344107897	-0.017267860	-0.024995327	1.000000000	-0.057359623	0.0813566434
famsize	-0.06765433	0.033000939	0.089566936	-0.012043772	-0.057359623	1.000000000	-0.2360976085
Pstatus	-0.06163176	0.015226451	0.065854816	0.009160552	0.081356643	-0.236097609	1.0000000000
traveltime	0.03380443	-0.244944283	-0.051013512	0.014529750	-0.333913376	-0.019669493	-0.0362063933
studytime	-0.16890770	0.140221060	0.220133768	0.001453688	0.057320130	0.013612781	0.0166521167
failures	0.08903615	-0.023235796	-0.049809636	0.280021256	-0.056027098	0.049576838	0.0217849739
schoolsup	-0.03513001	0.125180173	0.107447739	-0.173869291	0.013394291	0.056126920	0.0033164505
famsup	-0.06388311	0.069976608	0.127771701	-0.109789443	0.016459692	0.050048831	-0.0107373226
paid	0.04899923	0.262107456	0.022287985	-0.054641090	0.127218314	0.040477065	-0.0722469473
activities	0.03536547	0.088224253	-0.116321831	-0.072102586	-0.007518046	0.012205652	-0.1095101733
nursery	-0.04728727	-0.001446131	0.043274804	-0.048986677	0.011379608	-0.094037942	0.0470302051
higher	-0.08111728	0.128901130	0.067748357	-0.247504062	0.068398666	-0.005917727	-0.0161894900
internet	0.05118861	0.242993158	-0.049831609	0.001131184	0.185169168	0.007916197	-0.0769457812
romantic	-0.03452008	-0.069527369	0.126214414	0.185674339	-0.031905466	0.029178327	0.0462393459
famrel	-0.07196147	0.019856929	-0.074524909	-0.008453035	-0.032978459	0.001924402	-0.0429379422
freetime	0.15493604	-0.032941835	-0.167665510	-0.002414169	-0.033966807	0.021829784	-0.0271003593
goout	0.35227023	-0.040351378	-0.066535408	0.094744172	0.021541747	-0.001094184	-0.0207962313
health	0.09439028	0.044860743	-0.142132038	-0.013347170	-0.014031487	-0.016322766	-0.0004229342
absences	0.13859887	0.225523257	-0.001747465	0.168312690	0.074319589	-0.019638585	0.1269850978
r_course	-0.02565408	-0.127025480	0.020637482	0.019089568	-0.069947526	0.022822032	-0.0286045419
r_home	0.01249379	0.128839797	-0.036276293	-0.010548206	0.155387077	-0.024681698	0.0442241100
r_reputation	-0.01866141	0.177488682	0.038380017	-0.003110483	-0.005976786	-0.037441056	0.0253266784
math_avg_grade	-0.05472679	0.068094841	-0.097277104	-0.113644365	0.090969894	-0.059360950	0.0251894425
por_avg_grade	-0.16507101	0.284248900	0.119752396	-0.137923829	0.160216977	-0.038550770	-0.0181852938

	traveltime	studytime	failures	schoolsup	famsup	paid	activities
Alc	0.033804426	-0.168907698	0.08903615	-0.035130009	-0.063883107	0.048999231	0.035365469
school	-0.244944283	0.140221060	-0.02323580	0.125180173	0.069976608	0.262107456	0.088224253
sex	-0.051013512	0.220133768	-0.04980964	0.107447739	0.127771701	0.022287985	-0.116321831
age	0.014529750	0.001453688	0.28002126	-0.173869291	-0.109789443	-0.054641090	-0.072102586
address	-0.333913376	0.057320130	-0.05602710	0.013394291	0.016459692	0.127218314	-0.007518046
famsize	-0.019669493	0.013612781	0.04957684	0.056126920	0.050048831	0.040477065	0.012205652
Pstatus	-0.036206393	0.016652117	0.02178497	0.003316450	-0.010737323	-0.072246947	-0.109510173
traveltime	1.000000000	-0.066171910	0.07881989	-0.043328936	-0.024895080	-0.134504763	-0.033482288
studytime	-0.066171910	1.000000000	-0.12524989	0.089572142	0.138250535	0.184903863	0.058979326
failures	0.078819885	-0.125249889	1.00000000	0.011926987	-0.021286893	-0.121429498	-0.012726985
schoolsup	-0.043328936	0.089572142	0.01192699	1.000000000	0.087434482	0.028018336	-0.026850115
famsup	-0.024895080	0.138250535	-0.02128689	0.087434482	1.000000000	0.220109379	0.007732275
paid	-0.134504763	0.184903863	-0.12142950	0.028018336	0.220109379	1.000000000	0.019945872
activities	-0.033482288	0.058979326	-0.01272698	-0.026850115	0.007732275	0.019945872	1.000000000
nursery	-0.008043868	0.036379239	-0.09626275	0.018023756	0.028818192	0.073357298	0.037435750
higher	-0.078189926	0.188071755	-0.28250679	0.087434710	0.088850358	0.190120277	0.031390785
internet	-0.189718635	0.045192243	-0.03710912	-0.014901160	0.083105879	0.179834604	0.071884674
romantic	0.016361339	0.052777599	0.10722871	-0.093609713	-0.018181723	-0.057010603	0.046815358
famrel	-0.010458559	0.003935105	-0.06813389	-0.013907217	0.009102141	-0.003696747	0.040974040
freetime	-0.007536398	-0.077845396	0.11446251	-0.008808388	0.015401564	-0.036280725	0.148938598
goout	0.039606949	-0.078264309	0.06021737	-0.051002550	0.021598117	-0.017851449	0.091307499
health	-0.045229979	-0.054129755	0.03537227	0.019801749	0.004979426	-0.033898364	0.017671056
absences	-0.049896088	-0.039014723	0.08758440	0.008504524	0.044081976	0.061797146	0.008013972
r_course	0.128461457	-0.096253214	0.10770059	-0.063942458	-0.012574588	-0.187480551	0.001344864
r_home	-0.127599355	0.004130599	-0.03110137	0.047217443	0.020328092	0.110718084	-0.077543169
r_reputation	-0.078838189	0.174133004	-0.07844151	0.026403234	0.058948442	0.092005718	0.138990623
math_avg_grade	-0.112311975	0.108391662	-0.28647920	-0.113947955	-0.035783578	0.102356977	0.030359824
por_avg_grade	-0.147444354	0.262434528	-0.36203693	-0.068630483	0.056675093	0.187385687	0.060409133

	nursery	higher	internet	romantic	famrel	freetime	goout
Alc	-0.047287268	-0.081117282	0.051188614	-0.034520080	-0.071961469	0.154936036	0.352270231
school	-0.001446131	0.128901130	0.242993158	-0.069527369	0.019856929	-0.032941835	-0.040351378
sex	0.043274804	0.067748357	-0.049831609	0.126214414	-0.074524909	-0.167665510	-0.066535408
age	-0.048986677	-0.247504062	0.001131184	0.185674339	-0.008453035	-0.002414169	0.094744172
address	0.011379608	0.068398666	0.185169168	-0.031905466	-0.032978459	-0.033966807	0.021541747
famsize	-0.094037942	-0.005917727	0.007916197	0.029178327	0.001924402	0.021829784	-0.001094184
Pstatus	0.047030205	-0.016189490	-0.076945781	0.046239346	-0.042937942	-0.027100359	-0.020796231
traveltime	-0.008043868	-0.078189926	-0.189718635	0.016361339	-0.010458559	-0.007536398	0.039606949
studytime	0.036379239	0.188071755	0.045192243	0.052777599	0.003935105	-0.077845396	-0.078264309
failures	-0.096262750	-0.282506791	-0.037109118	0.107228708	-0.068133890	0.114462513	0.060217374
schoolsup	0.018023756	0.087434710	-0.014901160	-0.093609713	-0.013907217	-0.008808388	-0.051002550
famsup	0.028818192	0.088850358	0.083105879	-0.018181723	0.009102141	0.015401564	0.021598117
paid	0.073357298	0.190120277	0.179834604	-0.057010603	-0.003696747	-0.036280725	-0.017851449
activities	0.037435750	0.031390785	0.071884674	0.046815358	0.040974040	0.148938598	0.091307499
nursery	1.000000000	0.051896058	0.002642690	-0.020620433	0.033733363	-0.016720315	0.027594245
higher	0.051896058	1.000000000	0.080788613	-0.084738751	0.039432613	-0.110163705	-0.053708241
internet	0.002642690	0.080788613	1.000000000	0.045321276	0.061047702	0.047207134	0.083045671
romantic	-0.020620433	-0.084738751	0.045321276	1.000000000	-0.045228614	0.012651128	0.008851754
famrel	0.033733363	0.039432613	0.061047702	-0.045228614	1.000000000	0.140164897	0.091316603
freetime	-0.016720315	-0.110163705	0.047207134	0.012651128	0.140164897	1.000000000	0.332688119
goout	0.027594245	-0.053708241	0.083045671	0.008851754	0.091316603	0.332688119	1.000000000
health	0.012602679	0.022990746	-0.027158010	-0.034537998	0.109476151	0.088608539	0.002163089
absences	-0.008444397	-0.053380212	0.118429192	0.124432418	-0.067647120	-0.042331672	0.025046167
r_course	-0.025562925	-0.079876555	-0.116246134	0.013274781	-0.003765392	0.088910672	0.024681750
r_home	-0.024443933	0.048803845	0.051003509	0.018088433	-0.029252335	-0.070750256	-0.018613490
r_reputation	0.052099428	0.104972884	0.123560997	-0.058427384	0.027776765	-0.015219143	-0.008282467
math_avg_grade	0.058301450	0.113643809	0.091759276	-0.085319108	0.021975680	0.009826935	-0.110164920
por_avg_grade	0.035296604	0.348466686	0.152577130	-0.083639075	0.059662276	-0.112136380	-0.078079346

	health	absences	r_course	r_home	r_reputation	math_avg_grade	por_avg_grade
Alc	0.0943902842	0.1385988667	-0.025654076	0.012493791	-0.018661413	-0.0547267869	-0.16507101
school	0.0448607430	0.2255232565	-0.127025480	0.128839797	0.177488682	0.0680948406	0.28424890
sex	-0.1421320384	-0.0017474650	0.020637482	-0.036276293	0.038380017	-0.0972771037	0.11975240
age	-0.0133471699	0.1683126897	0.019089568	-0.010548206	-0.003110483	-0.1136443651	-0.13792383
address	-0.0140314873	0.0743195893	-0.069947526	0.155387077	-0.005976786	0.0909698943	0.16021698
famsize	-0.0163227662	-0.0196385851	0.022822032	-0.024681698	-0.037441056	-0.0593609499	-0.03855077
Pstatus	-0.0004229342	0.1269850978	-0.028604542	0.044224110	0.025326678	0.0251894425	-0.01818529
traveltime	-0.0452299787	-0.0498960880	0.128461457	-0.127599355	-0.078838189	-0.1123119751	-0.14744435
studytime	-0.0541297551	-0.0390147226	-0.096253214	0.004130599	0.174133004	0.1083916617	0.26243453
failures	0.0353722736	0.0875844032	0.107700585	-0.031101367	-0.078441513	-0.2864791986	-0.36203693
schoolsup	0.0198017488	0.0085045243	-0.063942458	0.047217443	0.026403234	-0.1139479551	-0.06863048
famsup	0.0049794256	0.0440819760	-0.012574588	0.020328092	0.058948442	-0.0357835776	0.05667509
paid	-0.0338983639	0.0617971457	-0.187480551	0.110718084	0.092005718	0.1023569773	0.18738569
activities	0.0176710561	0.0080139716	0.001344864	-0.077543169	0.138990623	0.0303598238	0.06040913
nursery	0.0126026792	-0.0084443967	-0.025562925	-0.024443933	0.052099428	0.0583014501	0.03529660
higher	0.0229907460	-0.0533802120	-0.079876555	0.048803845	0.104972884	0.1136438090	0.34846669
internet	-0.0271580098	0.1184291918	-0.116246134	0.051003509	0.123560997	0.0917592758	0.15257713
romantic	-0.0345379978	0.1244324175	0.013274781	0.018088433	-0.058427384	-0.0853191080	-0.08363908
famrel	0.1094761515	-0.0676471198	-0.003765392	-0.029252335	0.027776765	0.0219756800	0.05966228
freetime	0.0886085390	-0.0423316721	0.088910672	-0.070750256	-0.015219143	0.0098269354	-0.11213638
goout	0.0021630887	0.0250461673	0.024681750	-0.018613490	-0.008282467	-0.1101649204	-0.07807935
health	1.0000000000	-0.0399025876	0.103844583	-0.020815346	-0.109049974	-0.0588337783	-0.08170029
absences	-0.0399025876	1.0000000000	-0.133382455	0.116880844	0.075662972	-0.0004871207	-0.05845076
r_course	0.1038445826	-0.1333824545	1.000000000	-0.477022063	-0.472927855	-0.0858559627	-0.12627895
r_home	-0.0208153460	0.1168808440	-0.477022063	1.000000000	-0.292962798	0.0046548671	0.04375926
r_reputation	-0.1090499739	0.0756629720	-0.472927855	-0.292962798	1.000000000	0.0861119868	0.19021730
math_avg_grade	-0.0588337783	-0.0004871207	-0.085855963	0.004654867	0.086111987	1.0000000000	0.41003443
por_avg_grade	-0.0817002901	-0.0584507557	-0.126278950	0.043759256	0.190217302	0.4100344268	1.00000000

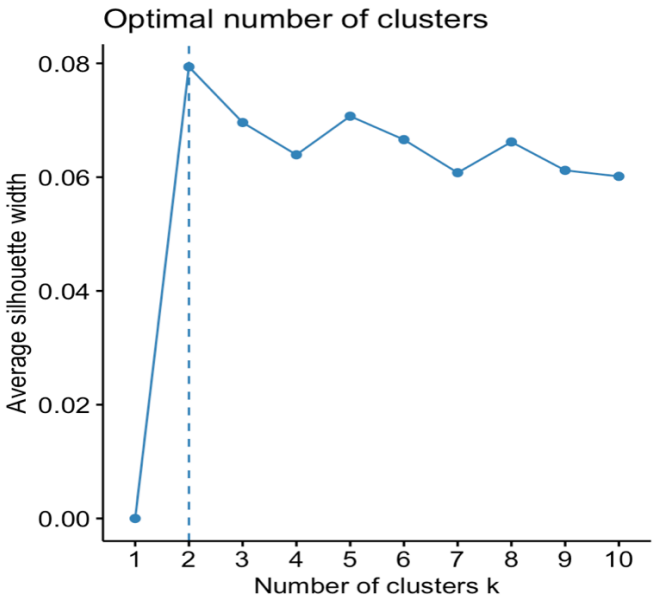


IV. Clustering

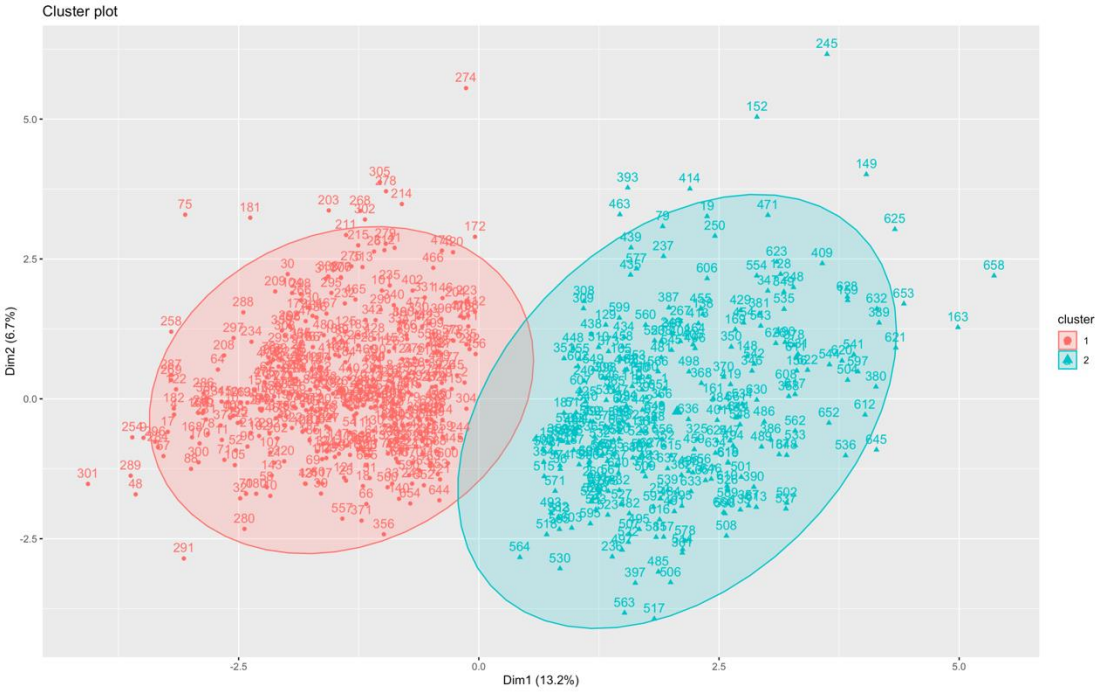
i. K-Means

After preprocessed the data. We applied k-means for clustering. However, due to the different scale of variables, the result could be impacted, so we scaled the data before clustering. Then, we used

‘Average Silhouette Method’ to find the optimal number of clusters.



As the graph showed, it recommended us to divided students into two groups. Then, we set center equals to 2, nstart equals to 100 and iteration as 200 for the k-means model. After plotting the result, we are quite satisfied to it since there is no huge overlapping between two groups.

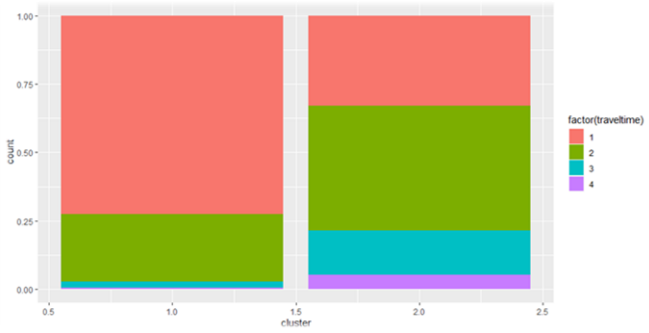


After that, we named and grouped the cluster with color red as ‘cluster1’, the cluster with color tiffany blue as ‘cluster2’. Next, we tried to label them.

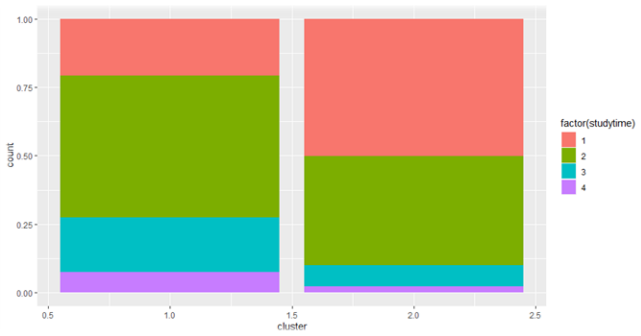
ii. Labeling

For labeling, we plotted the continuous variables with boxplot. For category variables, we used bar chart to show their proportion ratio.

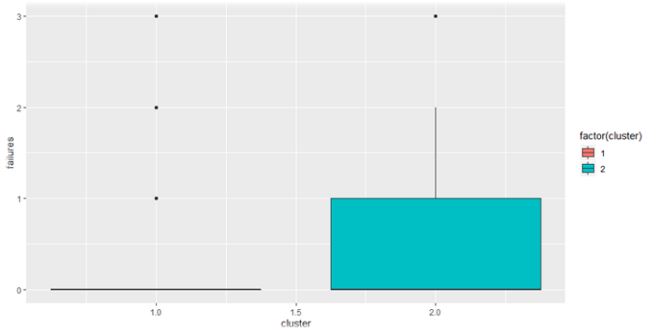
After plotting all the variables for two clusters. We found five variables that had significant differences in two clusters, ‘traveltime’, ‘studytime’, ‘failures’, ‘higher’ and ‘school’.



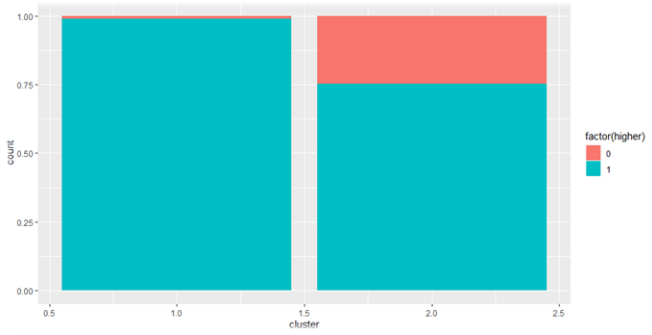
(traveltime)



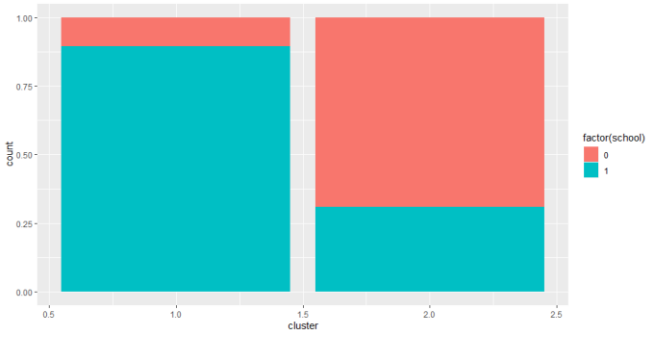
(studytime)



(failures)



(higher)



(school)

We then labeling cluster1 as the students come from Mousinho da Silveira who tend to spend more time on study and will like to pursue higher education. On the other hand, cluster2 is labeled as the students come from Gabriel Pereira who tend to spend more time on commuting and are easily to get fail. Moreover, we found that the students from cluster2 are also more likely to get a drink then those from cluster1.

Cluster1	Cluster2
1.From GP 2.Spend More Time on Study 3.Pursue higher education **4.Drink Less**	1.From MS 2.Spend more time commuting 3.More students fail in class **4.Drink More**

V. Before Training Model

Before we started training our models, we hoped our final models could fit in two cluster well respectively and try to predict new data as accurate as possible. Hence, we will train several models for both clusters. Here, we split data in to training set and validation set with 70% and 30% of the total population. As a result, we got two training set for cluster 1 and2, and two validation sets for cluster 1 and 2 as well.

VI. Model Training

Here, we will use Logistic Regression, Random Forest, Gradient Boosting Machine and Neural Network as the prediction model for each cluster. Since the concept of applying these four model in two cluster are the same, we will explain them together.

i. Logistic Regression

For Logistic Regression, we used all of the variables to as independent variables to predict dependent variable, Alc in our

analysis.

```
Call:
glm(formula = Alc ~ ., family = binomial(link = "logit"), data = train.df1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9134  -0.7814  -0.5079   0.8467   2.6855

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.65995    3.11170  -0.533  0.59372
school          0.74651    0.55526   1.344  0.17880
sex            -0.43679    0.34354  -1.271  0.20358
age             0.10919    0.13727   0.795  0.42638
address        -0.08405    0.49289  -0.171  0.86460
famsize        -0.43657    0.33684  -1.296  0.19494
Pstatus        -0.03611    0.46496  -0.078  0.93810
traveltime     -0.17170    0.30324  -0.566  0.57124
studytime      -0.33022    0.19700  -1.676  0.09369 .
failures       -0.41297    0.34849  -1.185  0.23601
schoolsup      -0.17109    0.45226  -0.378  0.70521
famsup          0.00353    0.32662   0.011  0.99138
paid            0.70985    0.30966   2.292  0.02189 *
activities     -0.16254    0.31227  -0.521  0.60269
nursery        -0.45598    0.38306  -1.190  0.23390
higher          1.24051    1.52623   0.813  0.41633
internet        0.38532    0.54453   0.708  0.47919
romantic       -0.06368    0.33833  -0.188  0.85070
famrel         -0.58319    0.19167  -3.043  0.00234 **
freetime        0.10090    0.16556   0.609  0.54222
goout           0.73788    0.15440   4.779 1.76e-06 ***
health          0.02855    0.11106   0.257  0.79713
absences        0.03700    0.01908   1.939  0.05248 .
r_course       -1.02931    0.65022  -1.583  0.11342
r_home         -0.48609    0.62359  -0.780  0.43568
r_reputation   -0.62216    0.64023  -0.972  0.33116
math_avg_grade  0.01024    0.06085   0.168  0.86637
por_avg_grade  -0.10733    0.07941  -1.352  0.17653
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(cluster1)

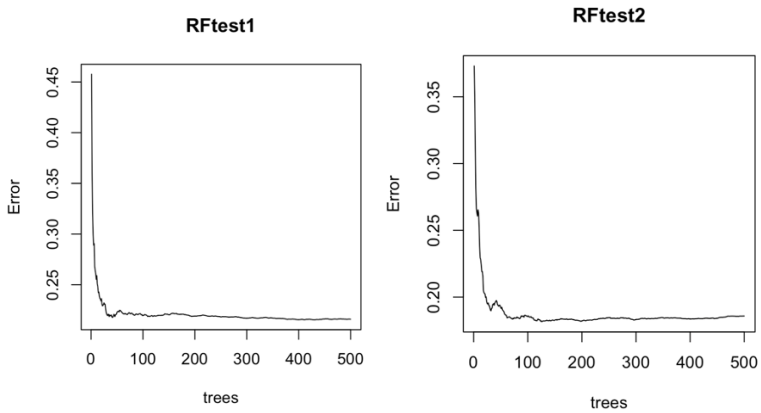
```
Call:
glm(formula = Alc ~ ., family = binomial(link = "logit"), data = train.df2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1678  -0.7249   0.2332   0.7030   2.5932

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.42677    3.82307  -0.896  0.37007
school         -0.78231    0.54297  -1.441  0.14964
sex            -1.73505    0.44206  -3.925 8.68e-05 ***
age             0.09154    0.18623   0.492  0.62305
address         0.41781    0.41977   0.995  0.31958
famsize        -0.56494    0.47530  -1.189  0.23460
Pstatus        -0.69986    0.70371  -0.995  0.31996
traveltime      0.19549    0.27714   0.705  0.48058
studytime      -0.09168    0.33013  -0.278  0.78124
failures        0.12348    0.27795   0.444  0.65684
schoolsup       1.18793    0.86941   1.366  0.17183
famsup         -0.89780    0.43635  -2.058  0.03964 *
paid           -0.12051    0.84679  -0.142  0.88683
activities      0.06480    0.44584   0.145  0.88443
nursery         0.40413    0.50580   0.799  0.42429
higher         -0.22613    0.52986  -0.427  0.66955
internet        0.22202    0.43306   0.513  0.60818
romantic       -0.37047    0.43045  -0.861  0.38943
famrel          0.04634    0.19372   0.239  0.81095
freetime       -0.15320    0.20450  -0.749  0.45378
goout           0.81354    0.21363   3.808  0.00014 ***
health          0.17413    0.14443   1.206  0.22796
absences        0.10556    0.05524   1.911  0.05604 .
r_course        0.65530    0.57498   1.140  0.25442
r_home          0.43503    0.79758   0.545  0.58545
r_reputation    0.06406    0.85491   0.075  0.94027
math_avg_grade  0.01834    0.10450   0.175  0.86071
por_avg_grade  -0.08503    0.09360  -0.908  0.36368
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(cluster2)

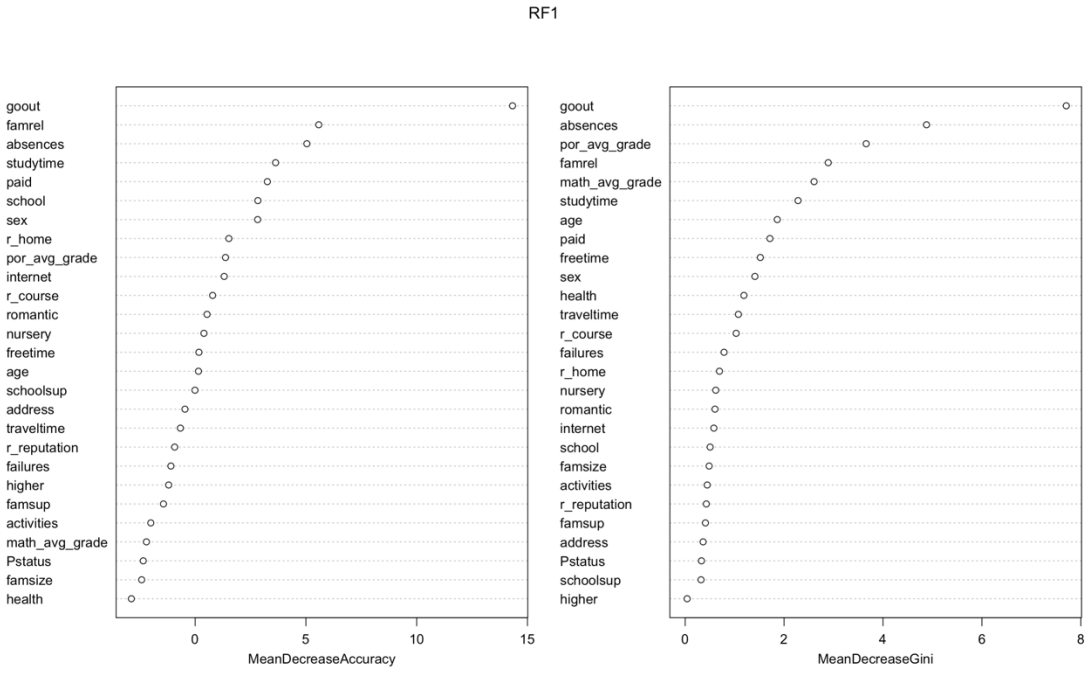
ii. Random Forest

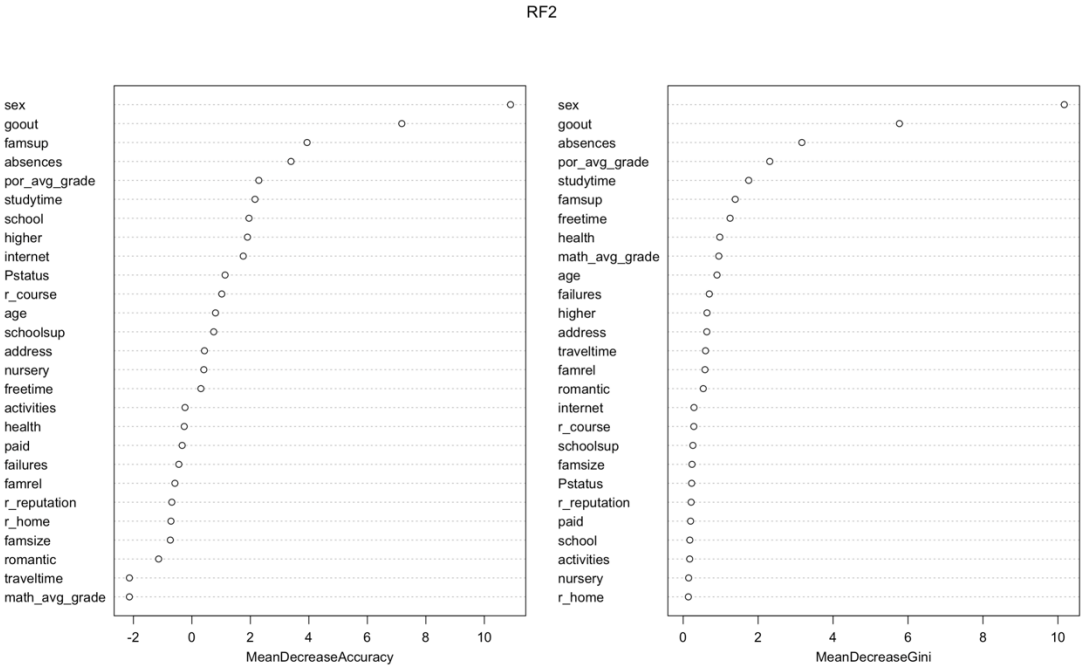
Before we built forests for two clusters, we wanted to find the optimal number of trees for two forest. Consequently, we plotted two graph with number of trees and error for two clusters. Then, we found the minimum MSE for each cluster, 431 for cluster1 and

126 for cluster2.



After finding the optimal number of trees, we used the result to build two Random Forest with nodesize equals to 25 for both forests.



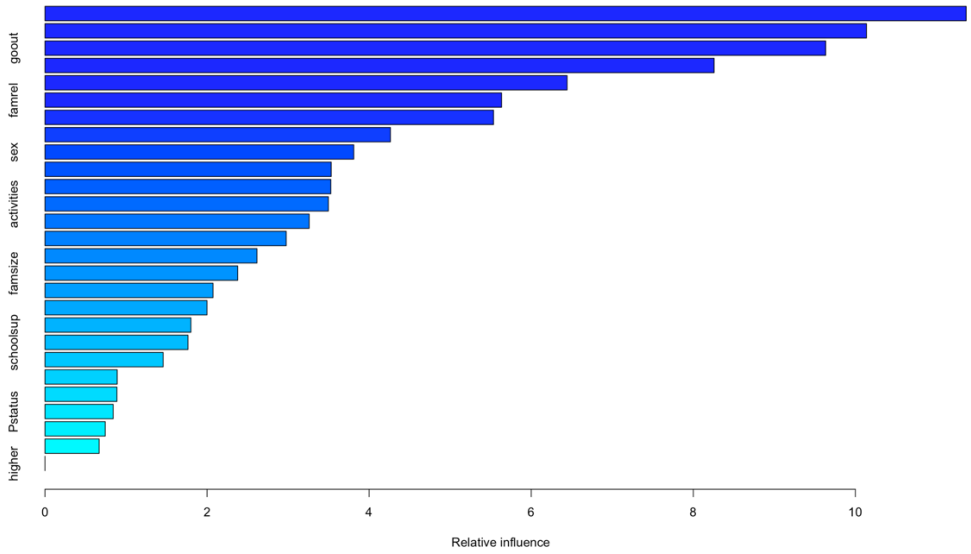


As the results plotted, the top important variables for different clusters are slightly similar. However, it is still acceptable for us.

iii. Gradient Boosting Machine

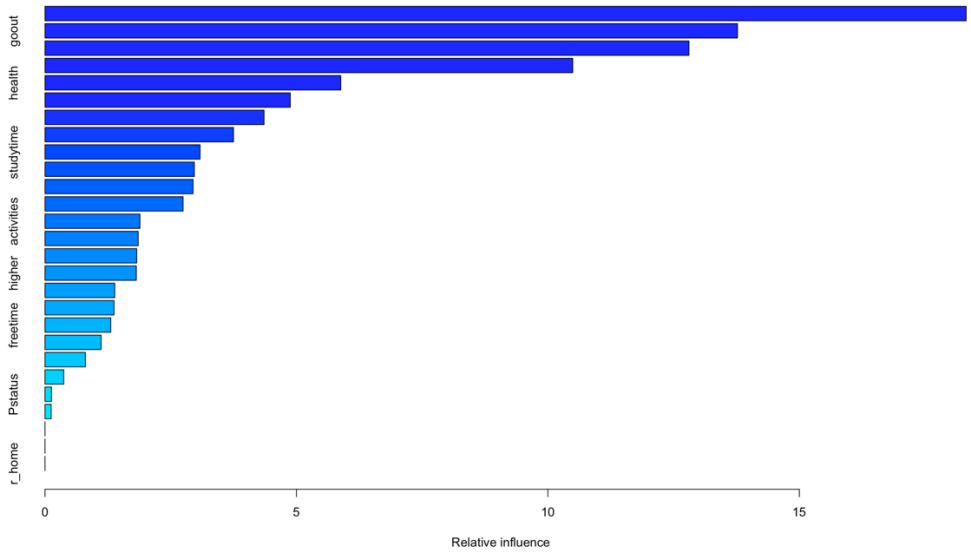
For GBM, we set number of trees same as Random Forest for each cluster. Moreover, the interaction depth and shrinkage are 4 and 0.05 respectively.

```
> summary(gbm.UB1)
                             var    rel.inf
absences                absences 11.3684804
por_avg_grade     por_avg_grade 10.1376897
goout                      goout  9.6337313
math_avg_grade math_avg_grade  8.2569610
age                          age  6.4430778
famrel                    famrel  5.6345247
health                    health  5.5351653
paid                        paid  4.2627350
sex                          sex  3.8107627
freetime              freetime  3.5322802
studytime            studytime  3.5269585
activities          activities  3.4968354
r_reputation      r_reputation  3.2602688
romantic              romantic  2.9769193
traveltime          traveltime  2.6160606
famsize                famsize  2.3780055
nursery                nursery  2.0738412
r_home                  r_home  1.9993919
famsup                  famsup  1.8009036
schoolsup            schoolsup  1.7646148
r_course              r_course  1.4586886
failures              failures  0.8909009
internet              internet  0.8875942
Pstatus                Pstatus  0.8415484
school                  school  0.7435369
address                address  0.6685234
higher                  higher  0.0000000
```



These two graphs are the result of GBM whose input data was cluster1. The graphs below are the result of GBM as well while input data was cluster2.

```
> summary(gbm.UB2)
                           var    rel.inf
sex                        sex 18.3183732
goout                    goout 13.7701141
absences              absences 12.8051258
por_avg_grade    por_avg_grade 10.4942767
health                  health  5.8817525
famsup                  famsup  4.8784200
famrel                  famrel  4.3571492
age                        age  3.7482665
studytime            studytime  3.0845751
romantic              romantic  2.9700766
traveltime          traveltime  2.9452474
r_course              r_course  2.7458938
activities          activities  1.8887745
internet              internet  1.8546445
famsize                famsize  1.8243321
higher                  higher  1.8164354
school                  school  1.3876277
address                address  1.3740177
freetime              freetime  1.3077672
math_avg_grade math_avg_grade  1.1166857
nursery                nursery  0.8046298
failures              failures  0.3735899
Pstatus                Pstatus  0.1290695
r_reputation      r_reputation  0.1231550
schoolsup            schoolsup  0.0000000
paid                      paid  0.0000000
r_home                  r_home  0.0000000
```

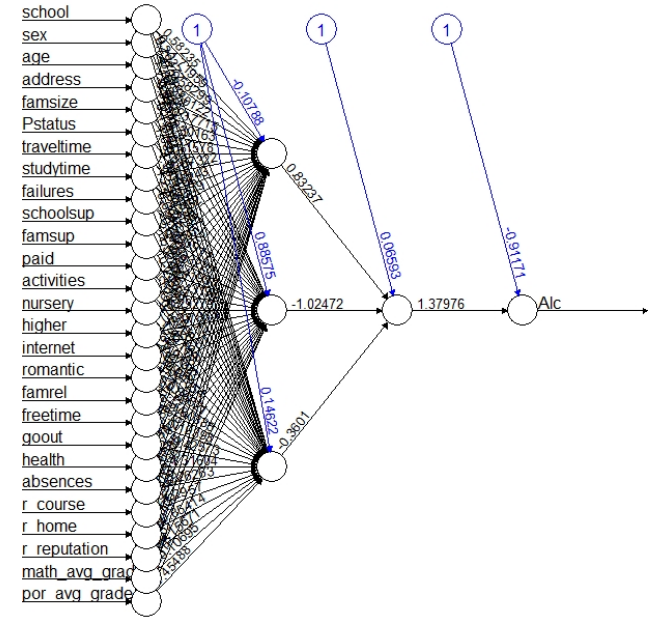


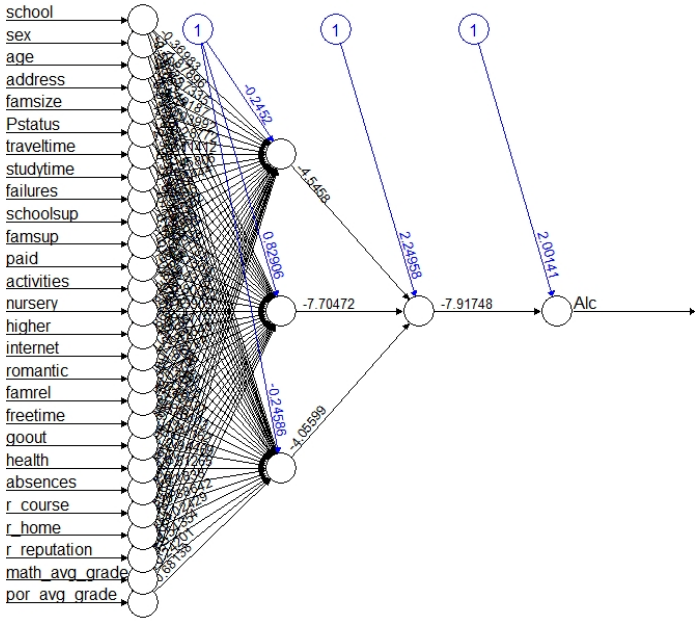
From the result of GBM, we can found that there are some differences from Random Forest. Nevertheless, it is still too early to determine which model is better since we neither calculate the True Positive Rate nor plot the ROC curve.

iv. Neural Network

The last model we used was Neural Network. Unlike those three models above, we set different number of hidden layers and different numbers of neural in each hidden layer for two clusters. First, we tried to find the best model for Neural Network; however, our hardware were not good enough to finish the calculation. Hence, we adjusted the parameters manually and the final parameters with minimum MSE.

The two Neural Network architectures are shown below. The first one is for cluster1 and the other is for cluster2.





VII. Full Dataset Training

With an aim to know whether building model for every cluster is better, we also trained those four models above for the full dataset. Note that the whole process in this section is the same as the process above, and the percentage of training and validation set as well.

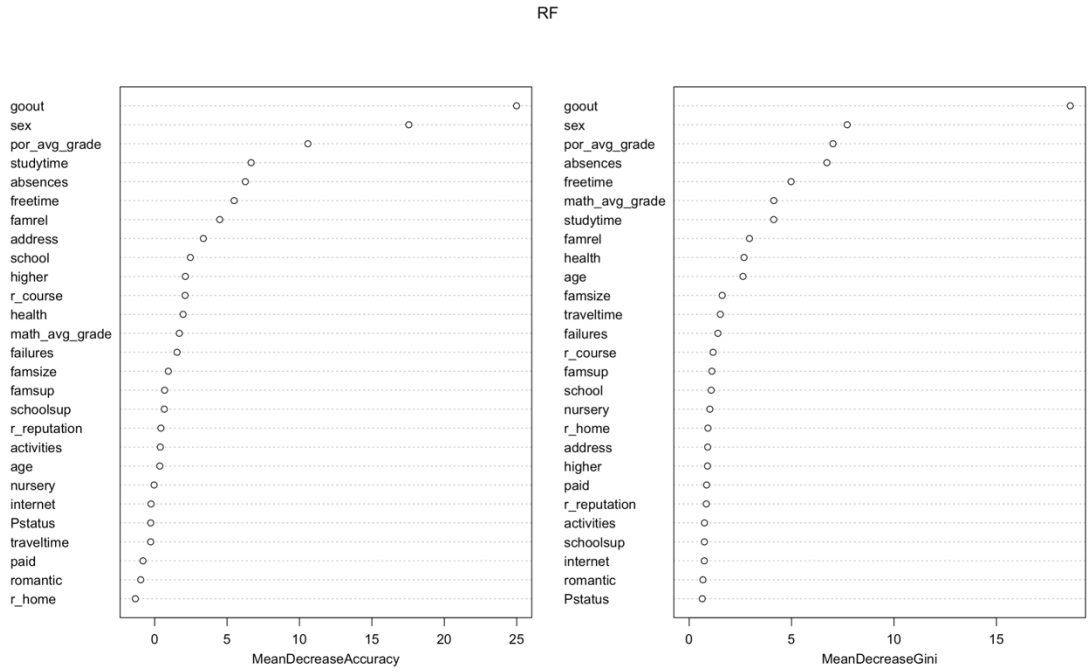
i. Logistic Regression

```
Call:
glm(formula = Alc ~ ., family = binomial(link = "logit"), data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1268  -0.7931  -0.4332   0.8056   2.6886

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.83518    2.08536  -0.880 0.378843
school         -0.37289    0.30258  -1.232 0.217809
sex            -0.98345    0.26201  -3.753 0.000174 ***
age             0.09859    0.10363   0.951 0.341436
address         0.18996    0.28525   0.666 0.505440
famsize        -0.53659    0.26099  -2.056 0.039787 *
Pstatus        -0.47527    0.39070  -1.216 0.223805
traveltime     -0.09034    0.16703  -0.541 0.588577
studytime      -0.34166    0.15730  -2.172 0.029854 *
failures        0.09716    0.17440   0.557 0.577444
schoolsup       0.16355    0.41640   0.393 0.694486
famsup         -0.40674    0.24855  -1.636 0.101741
paid            0.66232    0.28428   2.330 0.019818 *
activities      0.19355    0.25052   0.773 0.439765
nursery        -0.28512    0.28515  -1.000 0.317361
higher          0.03394    0.41419   0.082 0.934699
internet        0.07830    0.30286   0.259 0.796002
romantic       -0.17558    0.25704  -0.683 0.494564
famrel         -0.32879    0.12834  -2.562 0.010412 *
freetime       -0.04277    0.12444  -0.344 0.731038
goout           0.85972    0.12129   7.088 1.36e-12 ***
health          0.18502    0.08546   2.165 0.030386 *
absences        0.04498    0.01727   2.605 0.009199 **
r_course       -0.26730    0.38832  -0.688 0.491229
r_home         -0.32165    0.42064  -0.765 0.444473
r_reputation   -0.13733    0.44008  -0.312 0.754993
math_avg_grade  0.05661    0.04848   1.168 0.242906
por_avg_grade  -0.09458    0.05264  -1.797 0.072372 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For Logistic Regression, the significant variables are slightly similar with the combination of those in cluster1 and cluster2.

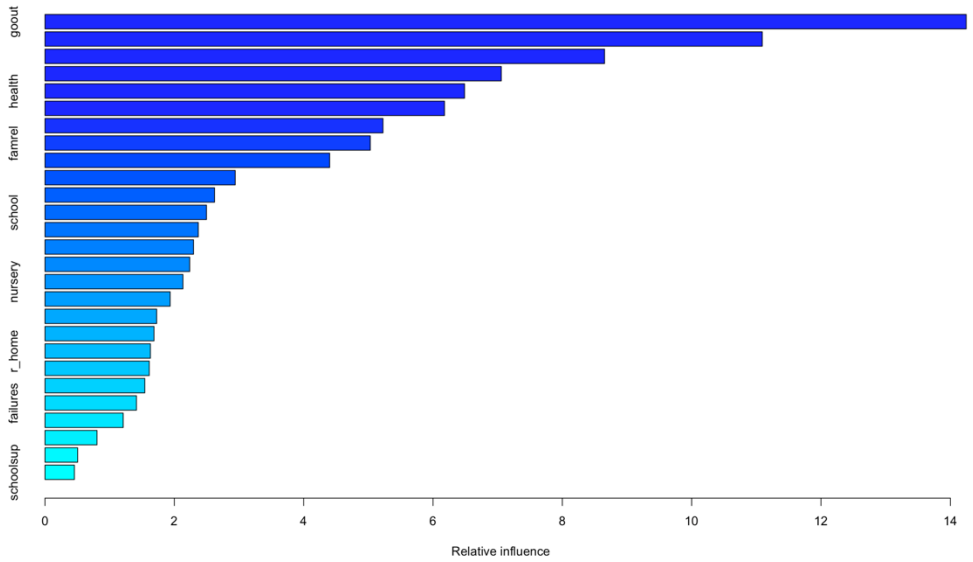
ii. Random Forest

Here we also ran the optimal ntree, and the result suggested us to set ntree to 491.



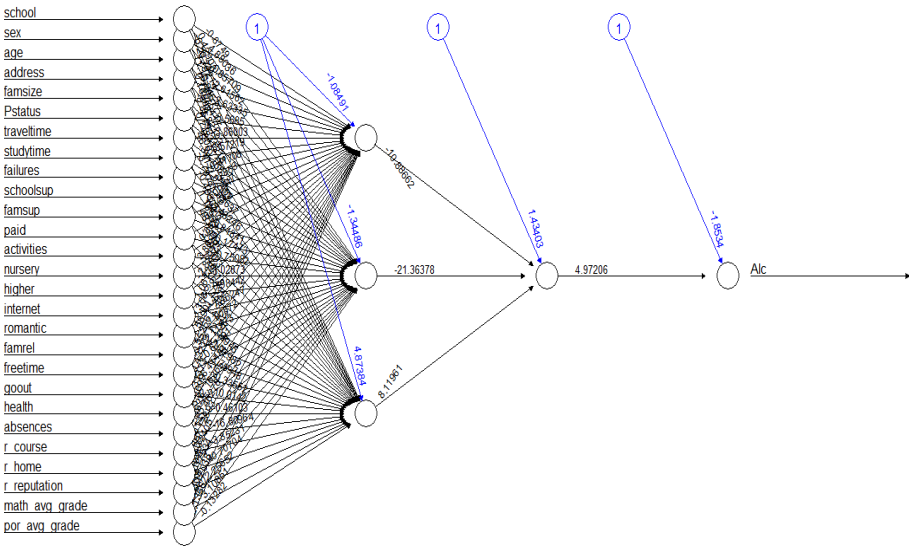
iii. GBM

	var	rel.inf
goout	goout	14.2463661
por_avg_grade	por_avg_grade	11.0921494
absences	absences	8.6532581
math_avg_grade	math_avg_grade	7.0564601
health	health	6.4885928
sex	sex	6.1780581
age	age	5.2269257
famrel	famrel	5.0302981
studytime	studytime	4.4024424
freetime	freetime	2.9421948
famsup	famsup	2.6223319
school	school	2.4976204
famsize	famsize	2.3707090
traveltime	traveltime	2.2971349
paid	paid	2.2382110
nursery	nursery	2.1334156
r_course	r_course	1.9353195
activities	activities	1.7274203
r_reputation	r_reputation	1.6864398
r_home	r_home	1.6305437
address	address	1.6126581
internet	internet	1.5434710
failures	failures	1.4145827
romantic	romantic	1.2089382
Pstatus	Pstatus	0.8044722
higher	higher	0.5052249
schoolsup	schoolsup	0.4547610



iv. Neural Network

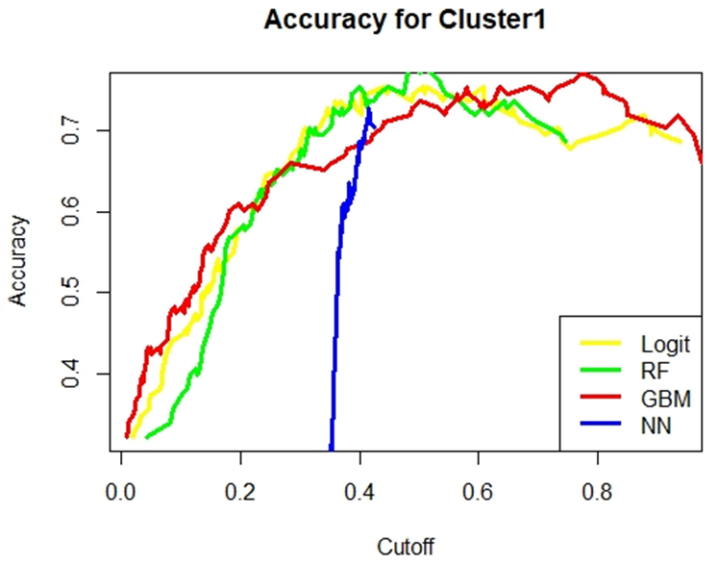
For Neural Network, we also set (3,1) for the full dataset.

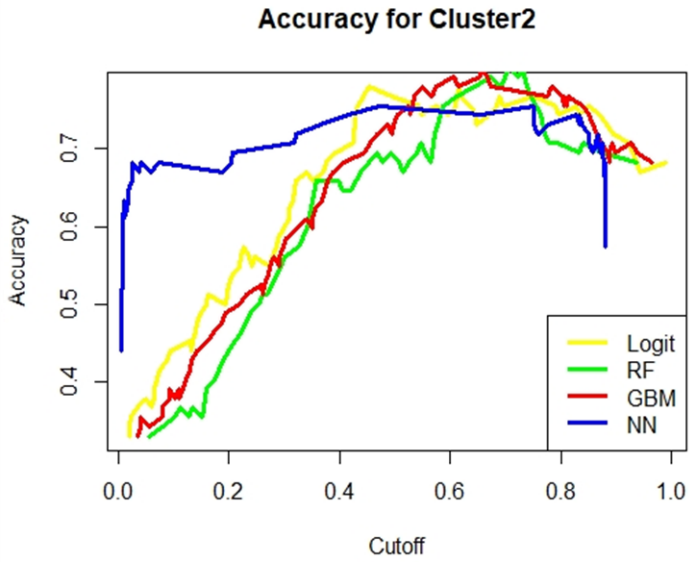


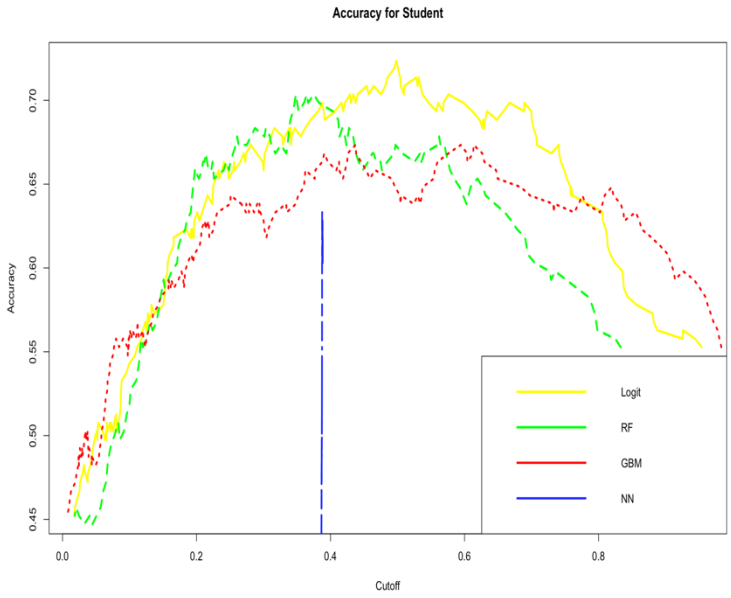
VIII. Prediction

After building all the models, we used validation sets to find out how well the models are.

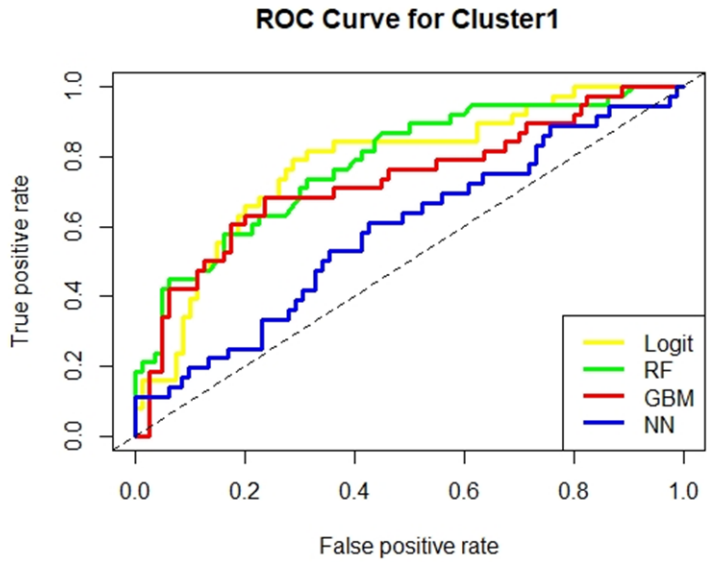
i. Accuracy

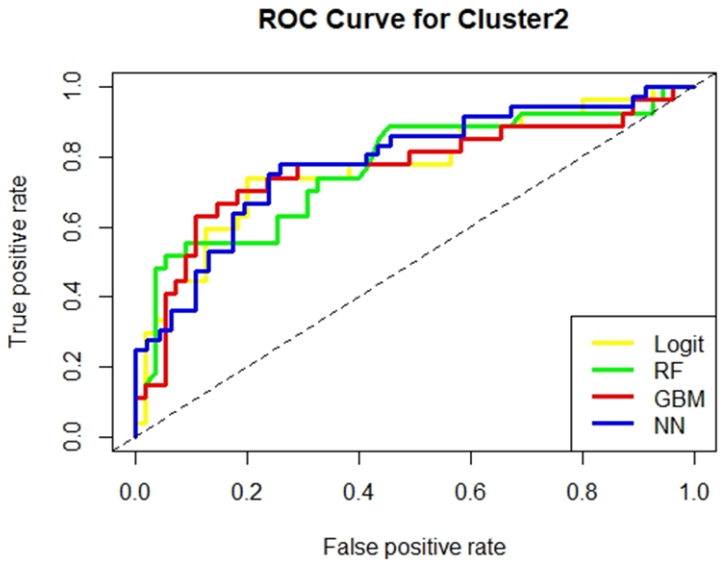



Accuracy for Cluster2
Accuracy
0.4
0.5
0.6
0.7
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff
Logit
RF
GBM
NN

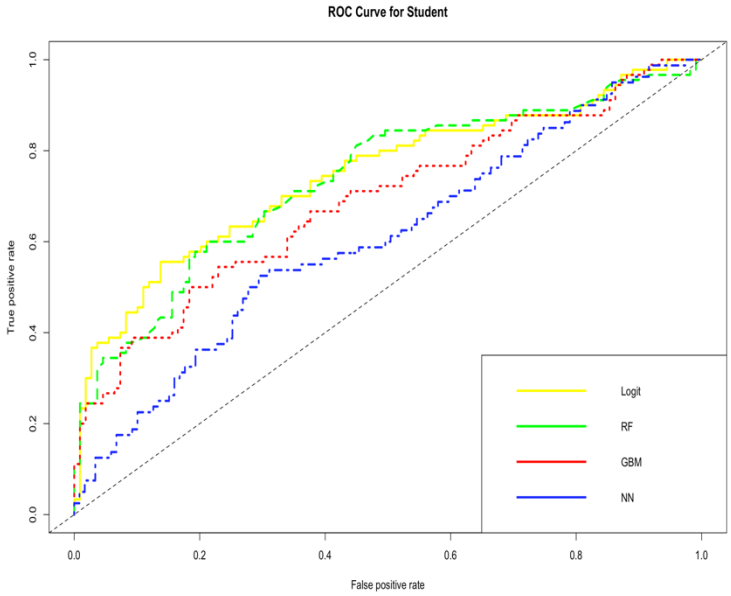

Accuracy for Student
Accuracy
0.45
0.50
0.55
0.60
0.65
0.70
0.0
0.2
0.4
0.6
0.8
Cutoff
Logit
RF
GBM
NN

ii. ROC Curve





For all of the models of both clusters, all of them are useful since they are all better than just random guess. However, it is clear to find that Neural Network is the worst model among for cluster1 . As a result, we can rank Neural Network is the last one for cluster1 based on its low performance on ROC curve. Nevertheless, we are not able to tell the ranking of top 3 models since the ROC curves of Logistic Regression, Random Forest and GBM are too similar. In addition, there is no good reason to rank all of the models for cluster2.



Comparing two clusters' ROC Curve, it is still not much convincing to say that building model for each cluster is better than simply using the full dataset.

iii. AUC

With a view to find the best model for each cluster, we calculated AUC.

For cluster1, it is obvious that Neural Network is the worst one, just like the conclusion after plotting ROC Curve. However, AUC helped us finding out that Random Forest is the best prediction model for cluster1 with 0.7809, which is the highest AUC among all.

logistics regression	0.7757
Random Forest	0.7809
GBM	0.7303
Neuralnet	0.586

(AUC for cluster1)

For cluster2, unlike the other cluster, Neural Network got 0.7856 and became the best prediction model.

logistics regression	0.7717
Random Forest	0.7704
GBM	0.7670
Neuralnet	0.7856

(AUC for cluster2)

Moreover, we calculated the AUC for full data set. The best model for full dataset is Logistic Regression

logistics regression	0.7453
Random Forest	0.7360
GBM	0.6882
Neuralnet	0.6075

(AUC for full dataset)

Finally, we compared the best model from each cluster and compared them with the best model from full data.
set.

Cluster1 – Random Forest	0.7809
Cluster2 – Neuralnet	0.7856
Full Dataset – logistic regression	0.7453

As we can see from the table, either model's AUC from each cluster is better than the AUC from full dataset, which shows that building model for each cluster is truly better.

IX. Conclusion

In conclusion, if we want to hold an extra-curricular activity for students, we can use the data to predict how much drink we have to prepare for the activity.

Another circumstance is that if we want to run a bar, we can predict where the bar should be set or what marketing plan we can implement.

However, if we want to better predict the best location where the new bar should be set, maybe we will need a larger dataset, which contains students' information from more schools.

X. Conclusion

There are some variables are hard to get, such as 'studytime', 'freetime' and so on. It might be a problem for building a full model because of missing values. Consequently, building a reduce model would be an effective way to solve this problem in this situation.