ANALYSIS OF STUDENT PERFORMANCE DATA

Tarun Chhabra

GROUP 2

INTRODUCTION

Student performance data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Datasets for Portuguese and Mathematics have been analysed by Group 2. We as a group analysed various parameters and tried to gain more insight into how some of these factors affect the student’s performances and tried to predict various metrics relevant to the data set.

I have worked on gauging academic performance of students in both subjects. Three performance indicators namely G1, G2 and G3 which are grade in period 1, grade in period 2 and grade in period 3 respectively have been provided for the both courses. I have made an assumption that total score which is the sum of grades of all the three periods is a performance indicator.

In the courses, instructors have to decide the cut-off for each letter grade to be given to students. My performance metrics deals with similar scenario in which we are gauging the student’s performance based on comparison of total marks with a certain predefined cut-off. Cut off can vary from one course to another and may vary have different meaning in different scenarios. A lot of competitive exams use percentile as an indicator to define cut- off for performance and use it as an admission criterion. For example, Indian Institute of Management(IIM) conducts one of the most competitive exams in the world in which the performance is gauged based on certain cut-off defined using percentiles.

My work is divided 3 segments for each course. First segment involves Exploratory data analysis which is useful for getting familiar with data set and gaining insight into various variables and their interactions. Second segment involves finding significant predictors responsible for determining performance class. This analysis gives us insight into the some of the significant factors which affect student’s performance. Third segment involves using prediction models to predict the performance class of the students based on the various features.

ANALSYIS OF STUDENTS PERFORMANCE IN PORTUGUESE COURSE

<u>SPECIFIC AIM 1: DATA EXPLORATION AND VISUALIZATION</u>

Dataset Summary:

No. of Observations:649

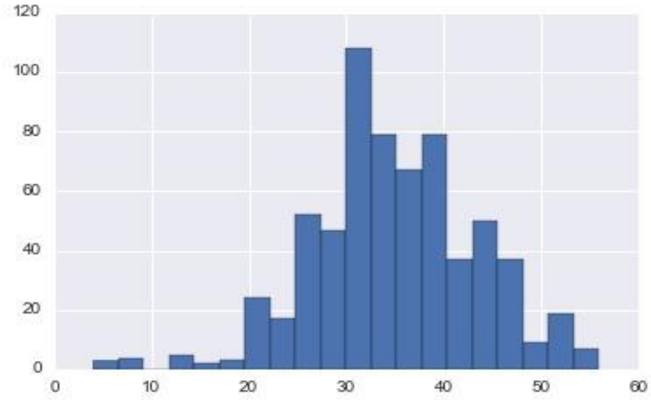
Continuous Variables: 5 – Age, Absences, G1, G2, G3 (grades at different periods)

Categorical Variables: 29 categorical variables based on demographic, social, etc. features.

Feature Creation:

Total Score (total)

Variable total is defined as the sum of grades in 1st term, 2nd term and final term.
Total=G1+G2+G3



Histogram of total score.

Obs	P40	P44	P45	P46
1	32	33	33	34

Above table shows different values corresponding to different percentiles for total scores. For example, 45th percentile is a score of 33.

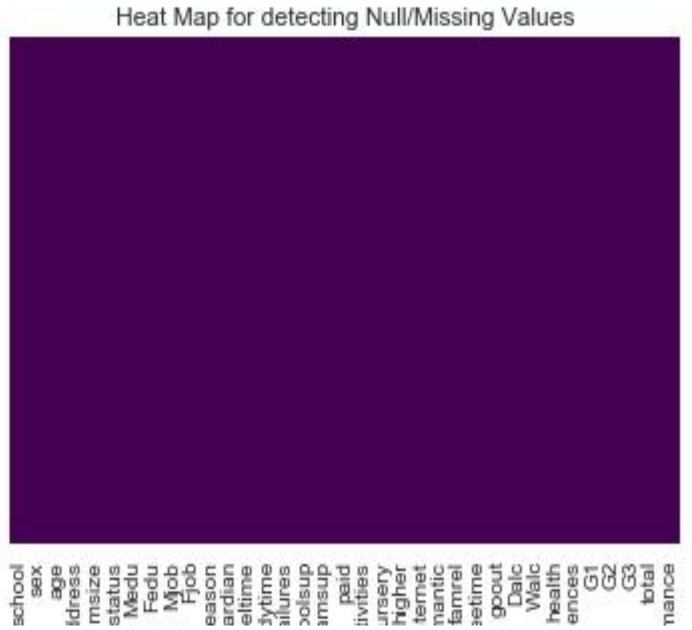
Performance*:* Performance indicator divides the students into two groups based on their total score percentile. From the analysis of total score, we choose cut off between 45th and 46th percentile and hence define **cut-off = 33.5.** As the total scores are all integers, keeping 33.5 cut off will divide the groups appropriately. So I have defined level for performance as below

Level 1: Below cut-off (total <33.5)

Level 2: Above cut-off (total>33.5)

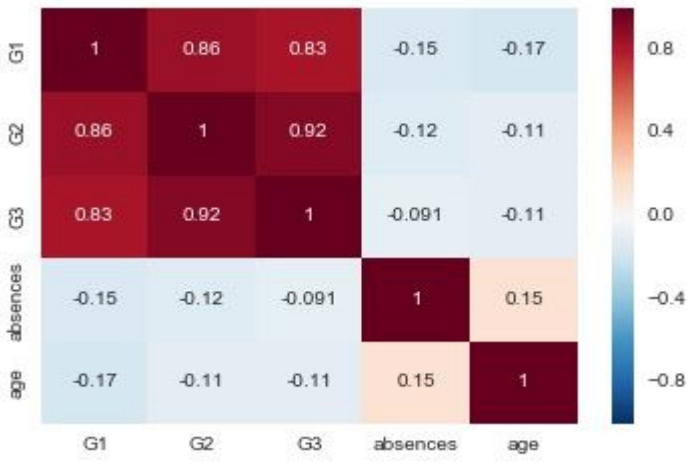
Data Visualization:

Detecting Null or Missing Values:



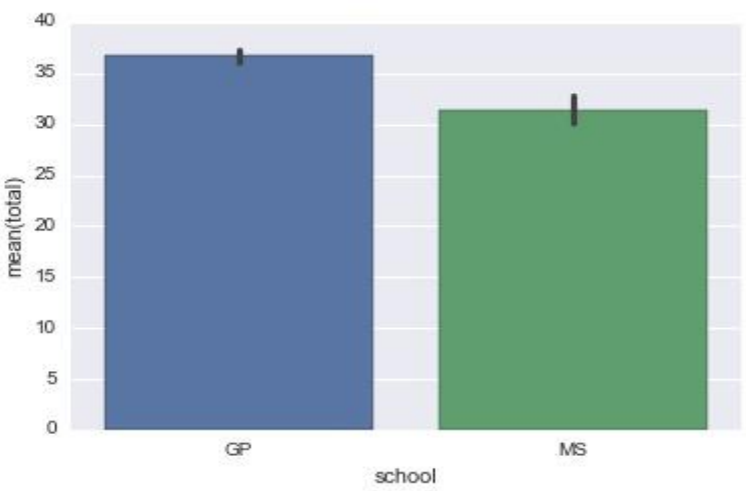
In order to detect any missing or null values visually, heat map has been used. Since the map colour is consistent, we conclude that there are no missing values for any of the variables.

CORRELATION:



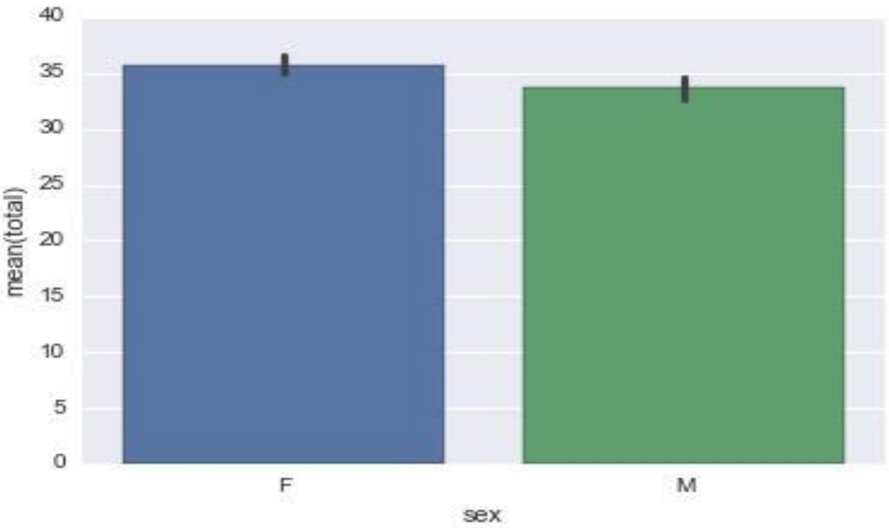
We can observe that G1, G2 and G3 are strongly correlated. It was expected as G1, G2, G3 represent grades in first, second and third period. So performance of students has generally been consistent. Number of absences have a negative correlation with all of G1, G2 and G3, this is understandable as missing classes generally has impact on the grades. Age has a negative correlation with the grades, it is understandable as young students learn especially languages really quickly and hence that reflects in their grades as well.

TOTAL GRADE vs SCHOOL



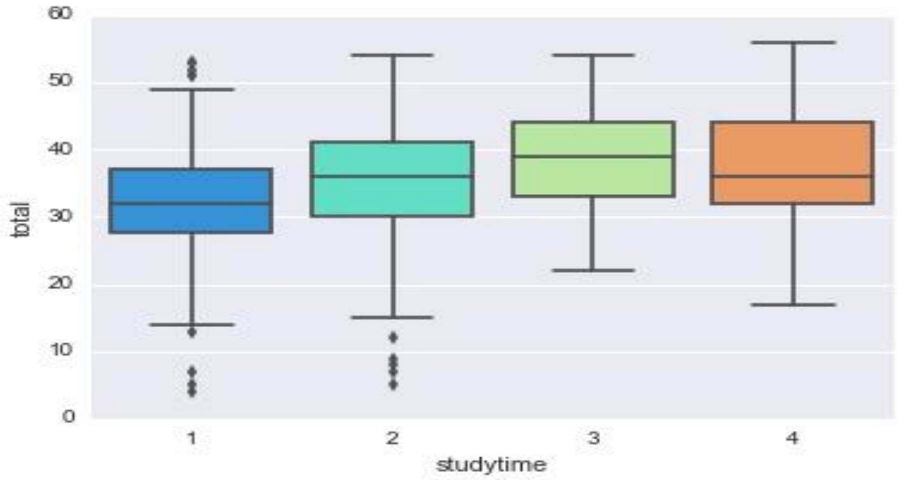
Average total for students in GP school is greater than students in MS school. It would be interesting to see if this factor is one of the significant factors in performance of students.

TOTAL GRADE vs SEX



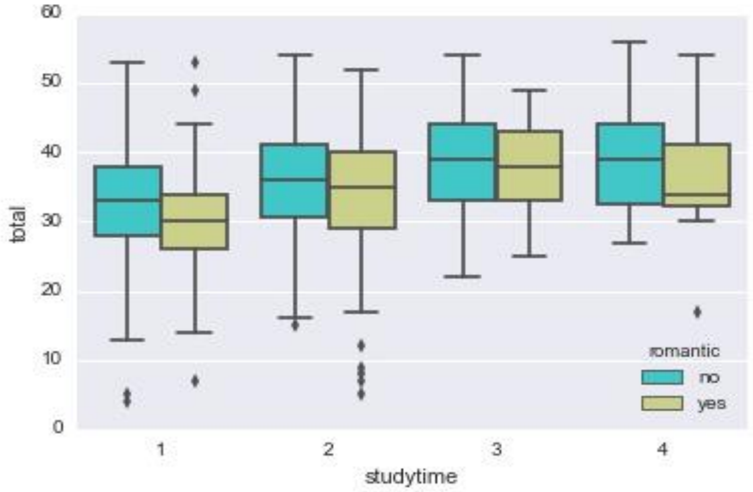
Mean total for female students is greater than male students. The difference seems to be around 3 points, which is significant. So gender has potential to be one of the crucial factors for prediction of the grade or overall performance.

TOTAL GRADE vs STUDYTIME



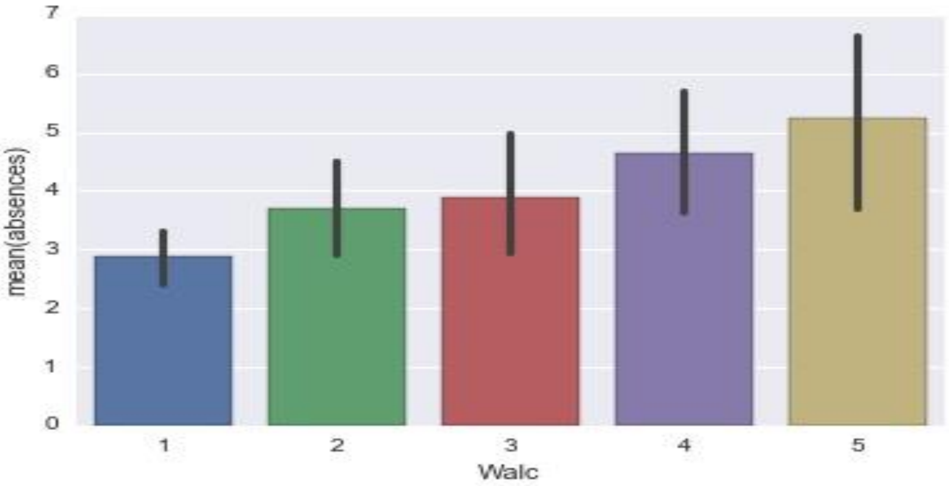
As study time increases, the mean value of total has generally increased which is expected as the students work harder tend to get better grades. One interesting observation from box plot is that there are some students who have scored extremely well but are in category 1 of study time and still have excellent marks. All the grades can't be attributed to hard work, IQ of the students, and pre course familiarity of Portuguese could be potential reason for the outliers.

TOTAL GRADE vs ROMATIC RELATIONSHIP



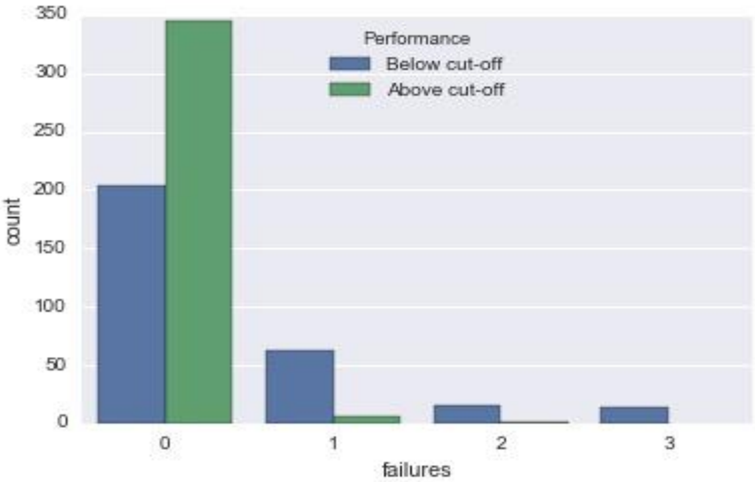
In previous boxplot we observed how study time affects the grades, now we observe how total scores vary for students involved in romantic relationship in comparison to students who aren't given their study time. Something interesting that we observe, that students who are in relationship, tend to get lesser marks in comparison to those who are not, if they study equal amounts of time. So romantic relationships end up being a distraction. It would be interesting to see, if this factor is significant in student's performance.

ABSENCE vs Weekend Alcohol Consumption Intensity



As the intensity of weekend alcohol consumption increases, we observe that mean number of absences increases as well. It would be interesting to see how much of this can be attribute to hangover by checking number of absences on Monday. Next we would observe how this alcohol consumption affects their performance.

PERFORMANCE Count Plot vs failures



In the Performance count plot with respect to past failure in courses, we can observe that most of the students haven’t failed in any course before. There are very few students who have failed a course before and we can observe that there is not a single observation for ‘Above cut-off’ performance in category 3.

SPECIFIC AIM 2: SIGNIFICANT FACTORS FOR PERFORMANCE

Objective: To find the significant factors impacting the performance of students. Performance has two levels: Below cut-off, Above cut-off. Analysis will give insight into how changes in significant factors impact the Performance level.

Model used: Logistic Regression

Logistic Regression model is fit with event as performance as "Below cut-off,' as we are more interested in what factors are significant in prediction of 'Below cut-off' performance class of students.

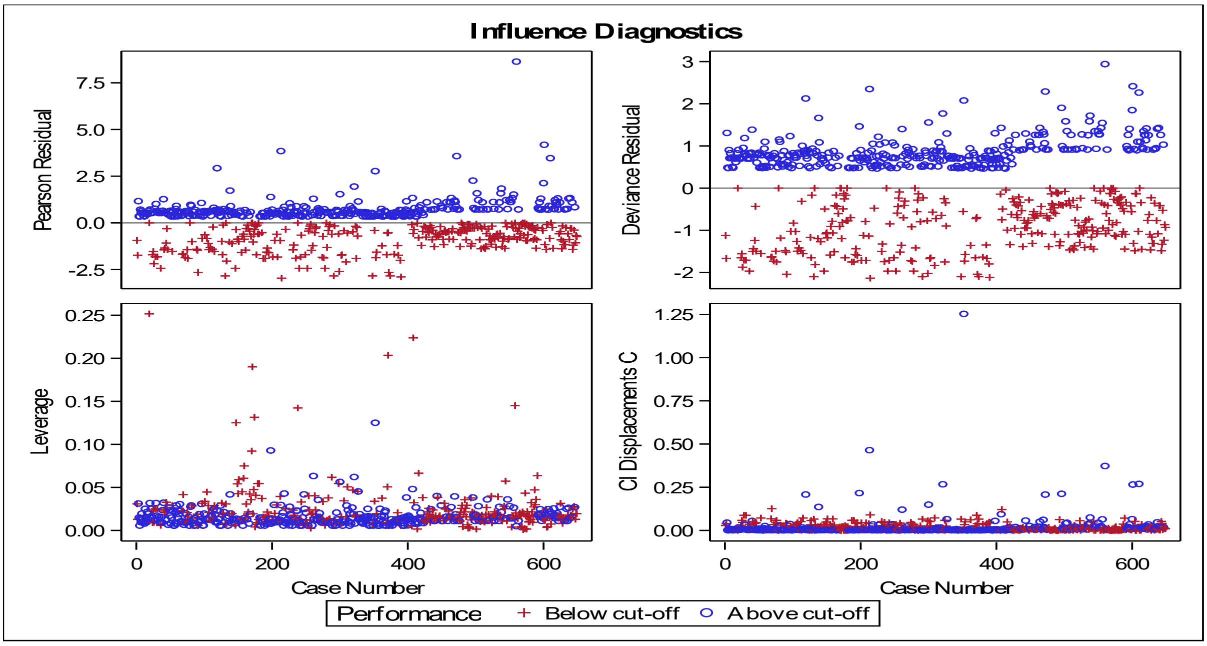
Feature Selection: Stepwise Selection

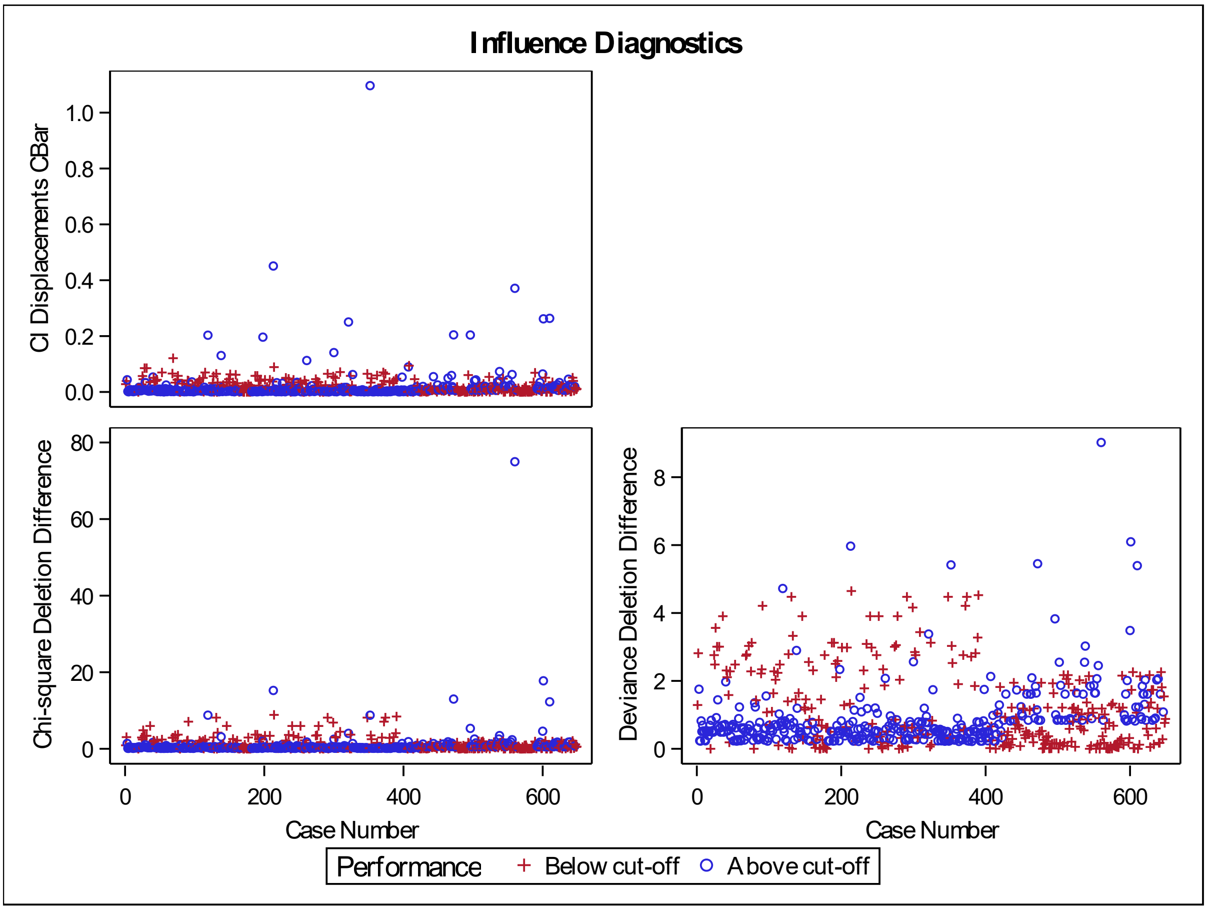
We fit a model with all the variables and use stepwise selection to come up with the significant factors.

Summary of Stepwise Selection

	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Failures		3	1	103.0519		<.0001
2	School		1	2	47.0211		<.0001
3	Higher		1	3	34.3462		<.0001
4	Absences		1	4	13.8741		0.0002
5	Sex		1	5	12.4225		0.0004
6	Schoolsup		1	6	12.2497		0.0005
7	Mjob		4	7	11.8502		0.0185

We observe that *number of past failures, school, preference for higher education, number of absences, sex, school support and Mother's job* are most significant factors in prediction of performance of students.


Influence Diagnostics
Pearson Residual
7.5
5.0
2.5
0.0
-2.5
Deviance Residual
3
2
1
0
-1
-2
Leverage
0.25
0.20
0.15
0.10
0.05
0.00
CI Displacements C
1.25
1.00
0.75
0.50
0.25
0.00
0
200
400
600
Case Number
Performance + Below cut-off ○ Above cut-off


Influence Diagnostics
CI Displacements CBar
1.0
0.8
0.6
0.4
0.2
0.0
Chi-square Deletion Difference
80
60
40
20
0
Deviance Deletion Difference
8
6
4
2
0
0
200
400
600
Case Number
Performance + Below cut-off ○ Above cut-off

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
school GP vs MS	0.230	0.150	0.352
Absences	1.091	1.044	1.140
sex F vs M	0.414	0.277	0.621
Mjob at_home vs teacher	2.451	1.207	4.980
Mjob health vs teacher	1.097	0.453	2.655
Mjob other vs teacher	1.046	0.556	1.965
Mjob services vs teacher	1.085	0.545	2.163
failures 0 vs 3	<0.001	<0.001	>999.999
failures 1 vs 3	<0.001	<0.001	>999.999
failures 2 vs 3	<0.001	<0.001	>999.999
schoolsup no vs yes	0.348	0.190	0.637
higher no vs yes	9.876	3.640	26.798

As we can observe that confidence interval for odds ratio for failures categories in comparison with 3rd category vary from less than 0.001 to greater than 999.99. Hence we refit the model **remove 'failure' as predictor.**

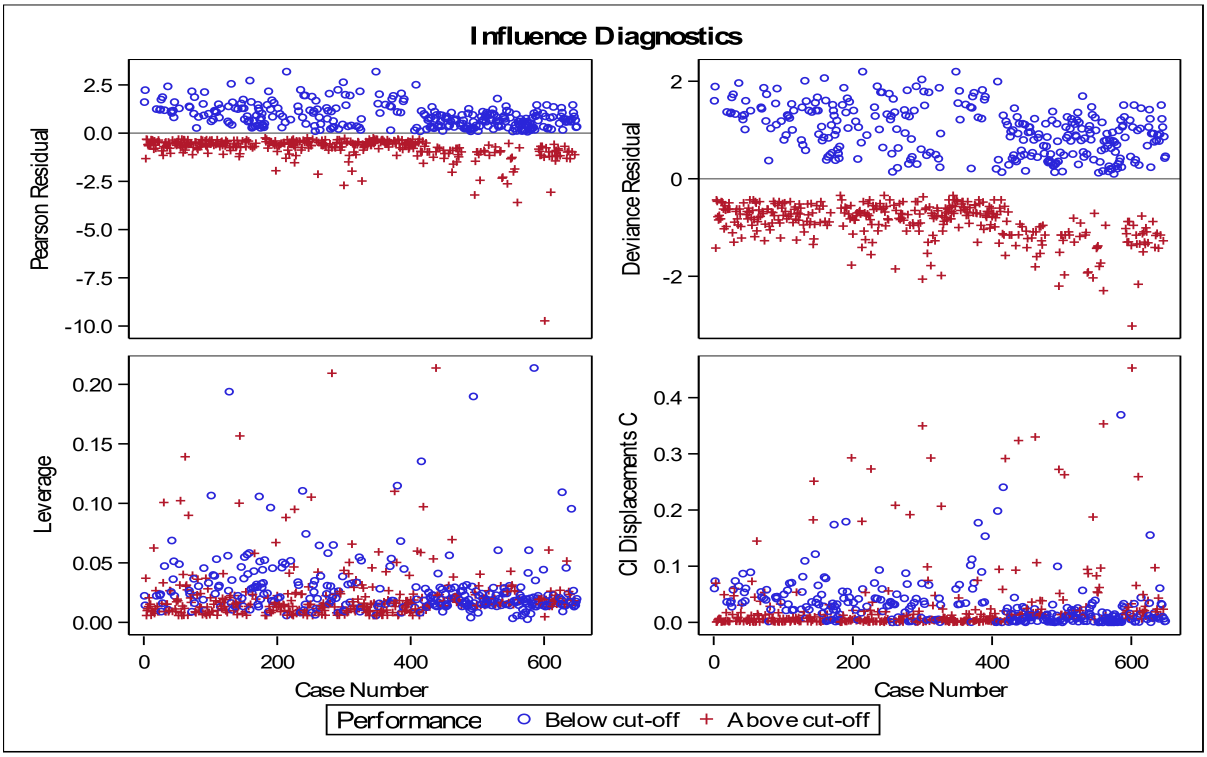
A new model is fitted by removing failures as one of the categorical variables

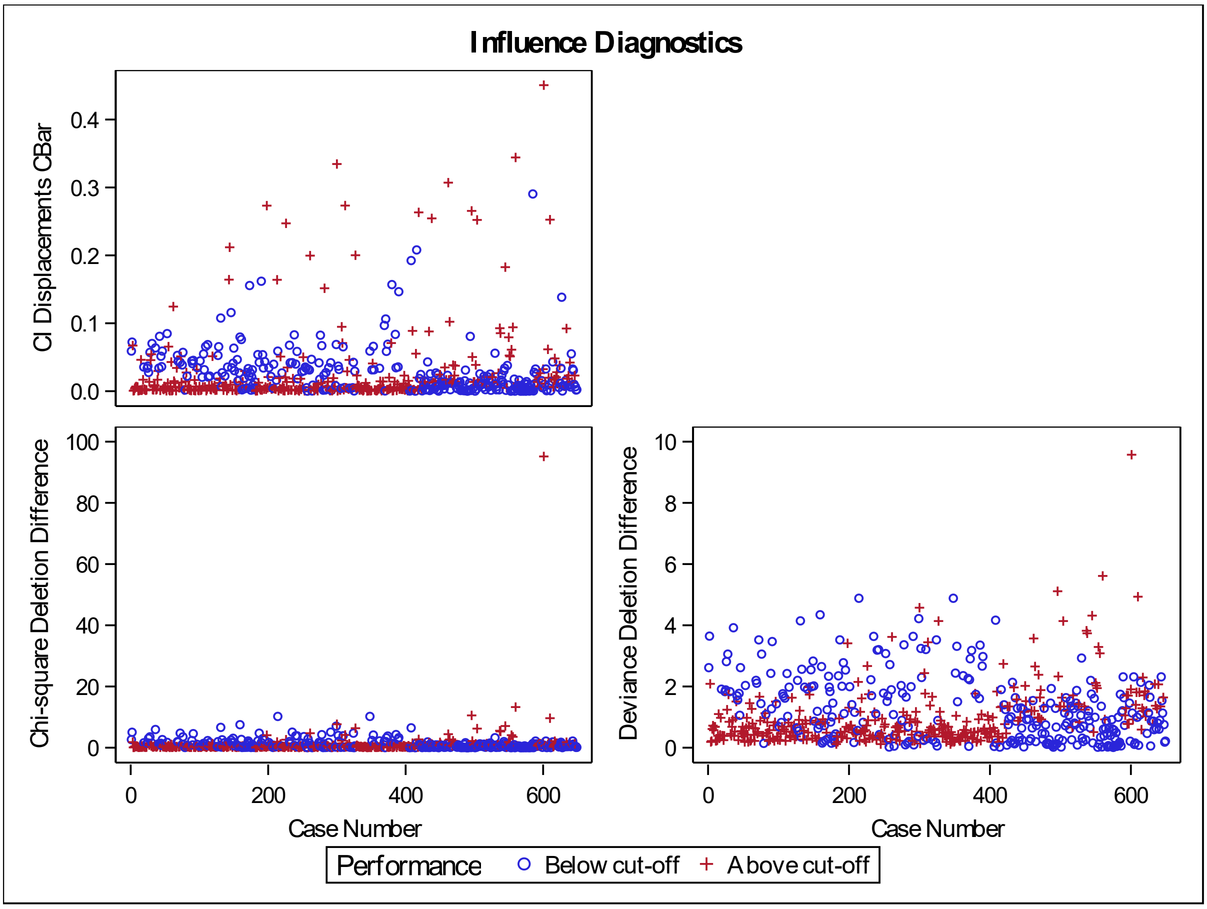
Feature Selection for New Model:

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Higher		1	1	69.1767		<.0001
2	School		1	2	50.6265		<.0001
3	Absences		1	3	22.3733		<.0001
4	Sex		1	4	13.7898		0.0002
5	Schoolsup		1	5	13.4280		0.0002
6	Medu		4	6	14.5302		0.0058
7	Dalc		4	7	11.5563		0.0210
8	Guardian		2	8	6.4832		0.0391

We observe that *school, preference for higher education, number of absences, sex, school support, Daily Alcohol Consumption intensity, Mother's education and guardian* are most significant factors in prediction in our new model.

Influence Diagnostics:





It can be observed that one or two points on Cbar plot have higher value compared to others so I refit the model after removing that influential point.

REFIT MODEL AFTER REMOVING INFLUENTIAL POINTS

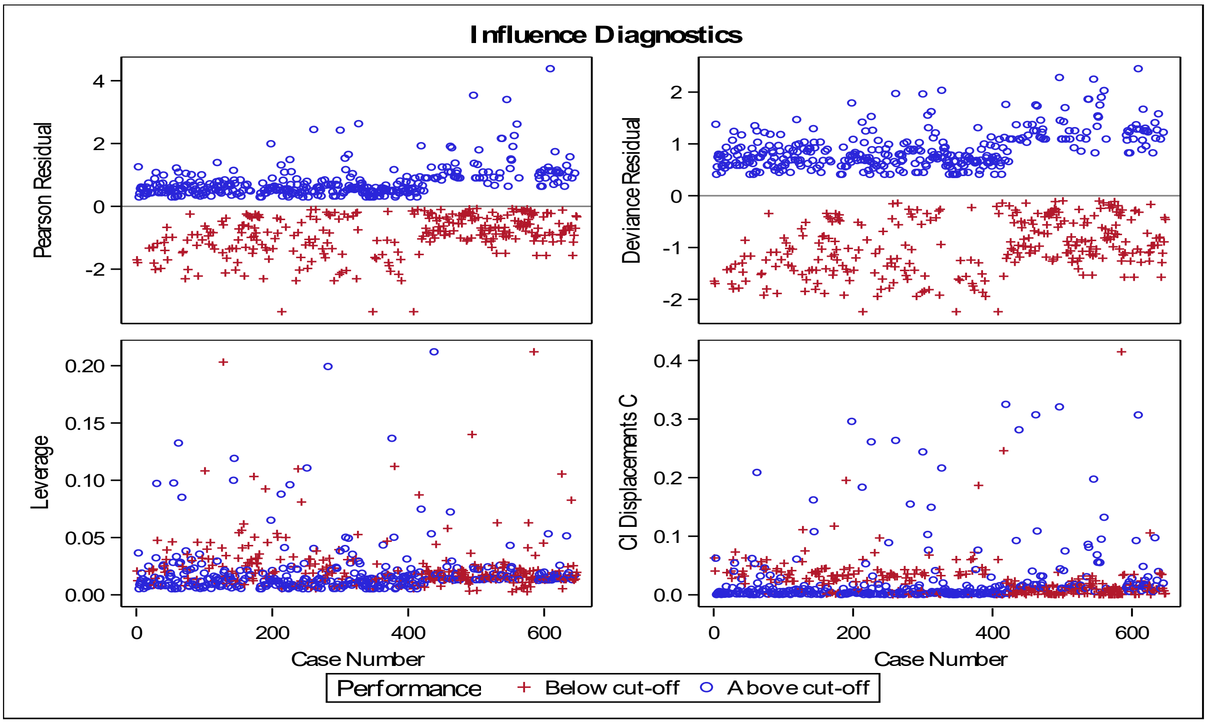
Feature Selection:

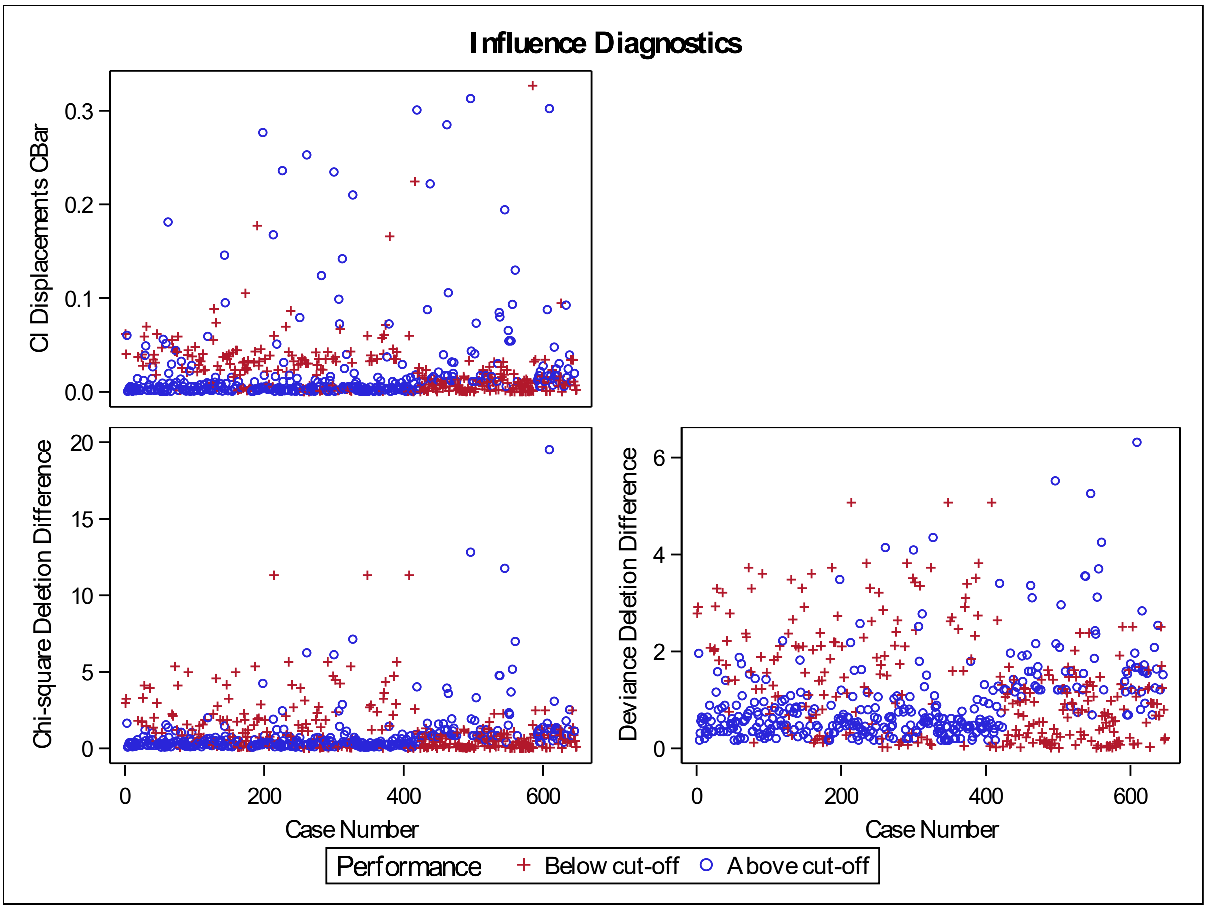
Summary of Stepwise Selection

	Effect						
Step	**Entered**	**Removed**	**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**
1	**Higher**		1	1	71.8382		<.0001
2	**school**		1	2	52.0883		<.0001
3	**absences**		1	3	22.3998		<.0001
4	**Sex**		1	4	14.6206		0.0001
5	**schoolsup**		1	5	13.6189		0.0002
6	**Medu**		4	6	15.0676		0.0046
7	**Dalc**		4	7	13.3975		0.0095

We observe that *school, preference for higher education, number of absences, sex, school support, Daily Alcohol Consumption intensity and Mother's education* are most significant factors in prediction in our new model.

INFLUENCE DIAGNOSTICS





At this stage we do not want to remove any more points since the largest point in terms of Cbar measure, does not have a very high Cbar as compared to rest of the points. So we continue with the model.

MODEL STATISICS

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	895.473	709.032
SC	899.947	771.666
-2 Log L	893.473	681.032

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	212.4416	13	<.0001
Score	181.3991	13	<.0001
Wald	125.7697	13	<.0001

The AIC is noticeably lower for our model than for the intercept only model, and the global tests all agree that at least one of the parameters are different from 0 at 5% statistical significance level.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
6.4242	8	0.5998

The Hosmer-Lemeshow test, is again highly insignificant indicating no problem with lack of fit in the model.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8762	0.3916	22.9572	<.0001
School	GP	1	-0.7627	0.1061	51.7126	<.0001
Higher	no	1	1.3881	0.2709	26.2474	<.0001
schoolsup	no	1	-0.4977	0.1477	11.3574	0.0008
Medu	0	1	0.0781	0.7404	0.0111	0.9160
Medu	1	1	0.4511	0.2561	3.1025	0.0782

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Medu	2	1	0.0727	0.2403	0.0916	0.7622
Medu	3	1	0.0217	0.2528	0.0074	0.9315
Sex	F	1	-0.3752	0.1029	13.2840	0.0003
Dalc	1	1	-0.6495	0.2262	8.2422	0.0041
Dalc	2	1	-0.3781	0.2634	2.0615	0.1511
Dalc	3	1	0.6890	0.3735	3.4022	0.0651
Dalc	4	1	0.4670	0.5124	0.8306	0.3621
Absences		1	0.0901	0.0216	17.3458	<.0001

The parameter estimates for GP as school, female as sex, absences, higher education(no), school support and daily alcohol consumption level as 1 are significant. While the estimates are positive for number of absences, no as a preference for higher education and mother’s job at home, they are negative for rest of them.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
school GP vs MS	0.218	0.144	0.330
higher no vs yes	16.058	5.552	46.444
schoolsup no vs yes	0.370	0.207	0.659
Medu 0 vs 4	2.017	0.318	12.779
Medu 1 vs 4	2.929	1.673	5.131
Medu 2 vs 4	2.007	1.207	3.337
Medu 3 vs 4	1.907	1.116	3.257
sex F vs M	0.472	0.315	0.707
Dalc 1 vs 5	0.594	0.161	2.193
Dalc 2 vs 5	0.779	0.202	3.005
Dalc 3 vs 5	2.265	0.503	10.204
Dalc 4 vs 5	1.814	0.316	10.422
Absences	1.094	1.049	1.142

For a unit increase in number of absences, odds get multiplied by 1.094. In comparison to male students, females have lesser odds for poor performance. In comparison to preference for higher education, odds increase significantly if student is higher education. If mother’s education is less than higher education, odds are greater for ‘Below cut-off’ performance.

SPECIFIC AIM 3: CLASSIFICATION POWER

We now, predict the class of performance based on the given predictors. To increase the prediction accuracy, we use all the predictors and will split the data into training and testing data.

Data is randomly split into 70:30 as training and testing. We would train our models on same training set and test it onto the test data. The model selection would be done on the basis of test data.

Definitions of some of the metrics that we use with respect to classifying to Poor category are:

- **Recall** indicates what proportion of students who actually performed poorly have been captured.

- **Precision** indicates what proportion of students whose performance has been classified as Poor have actually performed poorly.

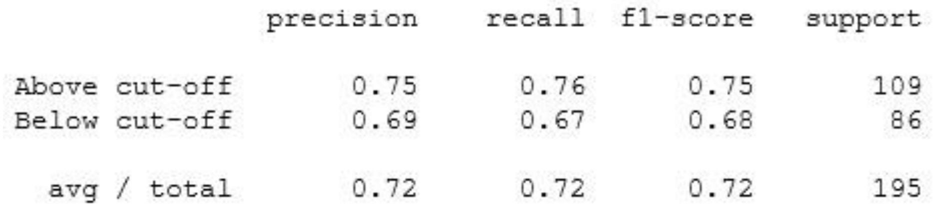
Priority: Correct classification of ‘Below cut-off’

MODEL 1: LOGISTIC REGRESSION

INTRODUCTION:

Logistic regression model is fit using the selected features from our analysis. In order to tune the model to give best prediction results, various combinations like full feature model were used to check precision and recall values. Of all the models, logistic regression with features selected from stepwise selection model had maximum recall value for ‘Below cut-off’ performance.

Model Results:



```
              precision    recall  f1-score   support

Above cut-off       0.75      0.76      0.75       109
Below cut-off       0.69      0.67      0.68        86

  avg / total       0.72      0.72      0.72       195
```

CONCLUSION

We can observe that Precision for Below cut-off is 69% which means out of all the students whose performance has been classified as Below cut-off, 69% of them in reality fall in Below cut-off category. Since Recall value for Below cut-off category is 67% which means we have only been able to capture 67% of total students whose total marks are Below cut-off.

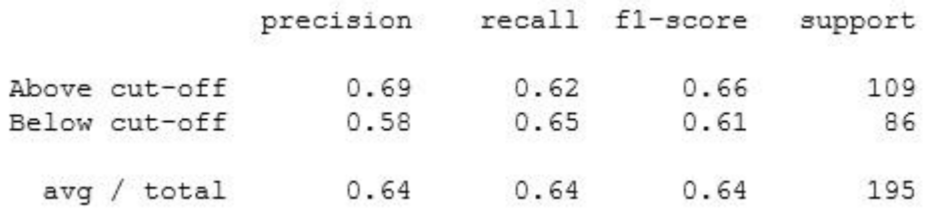
For different business needs, precision and recall values hold different importance. For this problem, we focus on increasing the recall value and keep an eye on precision. Hence I tried other models like Support Vector Machine and Random Forest Algorithm.

MODEL 2: SUPPORT VECTOR MACHINE

INTRODUCTION

In machine learning, **support vector machines** (**SVMs**, also **support vector networks**) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

MODEL RESULTS



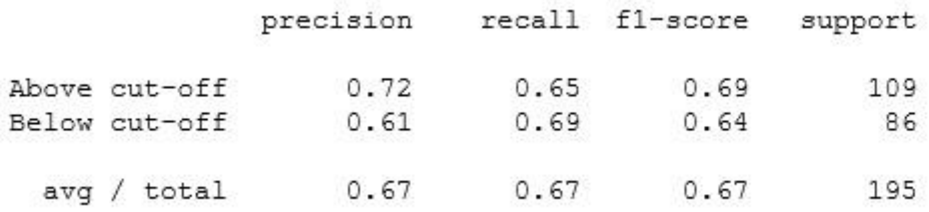
```
              precision    recall  f1-score   support

Above cut-off       0.69      0.62      0.66       109
Below cut-off       0.58      0.65      0.61        86

  avg / total       0.64      0.64      0.64       195
```

SVM has performed poorly than Logistic regression with selected features as the recall value is 65% for Below cut-off category and 62 % for Above cut-off and accordingly precision has decreased as well. Let’s tune the parameters of SVM to get better recall results.

AFTER TUNING RESULTS:



```
              precision    recall  f1-score   support

Above cut-off       0.72      0.65      0.69       109
Below cut-off       0.61      0.69      0.64        86

  avg / total       0.67      0.67      0.67       195
```

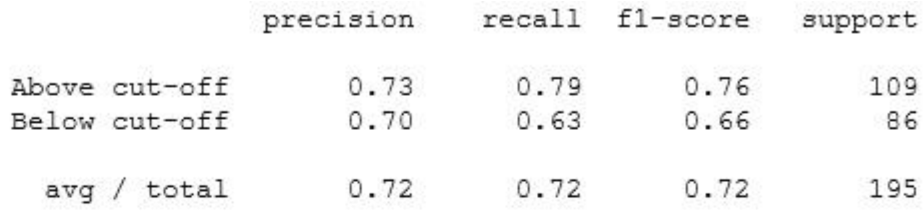
In comparison to un-tuned SVM model, the model has much better results. In comparison to logistic regression, recall value for Below cut-off has increased by 2% though recall value of Above cut-off has come down but our priority was to capture Below cut off students and hence this model is slightly better than logistic.

MODEL 3: RANDOM FOEST

INTRODUCTION

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set

RESULTS



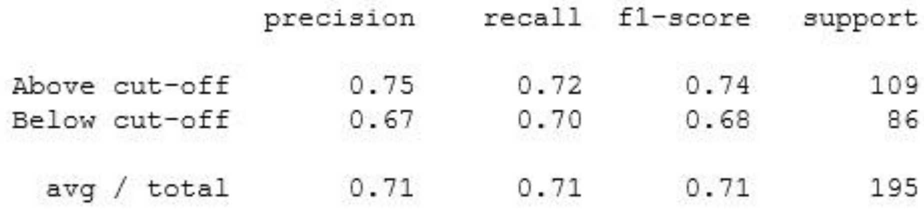
```
              precision    recall  f1-score   support

Above cut-off       0.73      0.79      0.76       109
Below cut-off       0.70      0.63      0.66        86

  avg / total       0.72      0.72      0.72       195
```

RF has performed poorly than Logistic regression with selected features as the recall value is 63% for Below cut-off category and but has done well in Above cut-off as its recall is 79%. Let’s tune the parameters such that recall value for ‘Below cut-off ‘category.

RESULTS AFTER TUNING



```
              precision    recall  f1-score   support

Above cut-off       0.75      0.72      0.74       109
Below cut-off       0.67      0.70      0.68        86

  avg / total       0.71      0.71      0.71       195
```

In comparison to un-tuned RF model, the model has much better results. In comparison to logistic regression, recall value for Below cut-off has increased by 3% and recall value of Above cut-off has come down but our priority was to capture Below cut off students and hence this model is slightly better than logistic.

ANALSYIS OF STUDENTS PERFORMANCE IN MATHEMATICS

SPECIFIC AIM 1: DATA EXPLORATION AND VISUALIZATION

Dataset Summary:

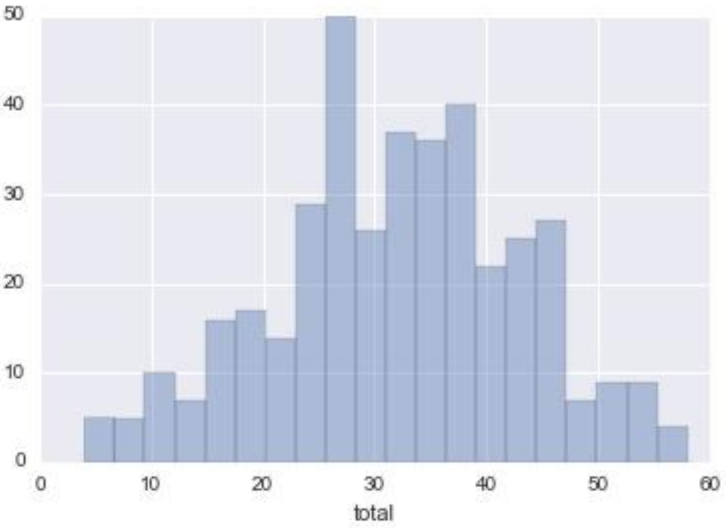
No. of Observations:395

No. of Continuous Variables: 5 – Age, Absences, G1, G2, G3 (grades at different periods)

No. of Categorical Variables: 29 categorical variables based on s

Feature Creation:

1. Total Score (total)
   Variable total is defined as the sum of grades in 1st term, 2nd term and final term.
   total=G1+G2+G3



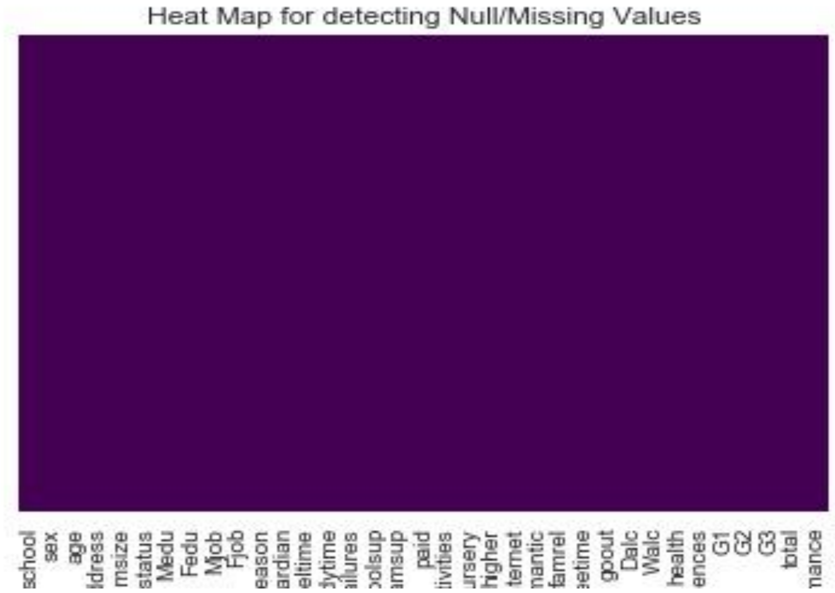
Obs	P40	P41	P45	P46	P50	P54	P55	P56
1	29	29	30	31	32	33	34	34

Above table shows different values corresponding to different percentiles for total scores. For example, 55th percentile is a score of 34.

2. Performance: Performance variable divides students into two groups based on their total score percentile. From the analysis of total score, 55th percentile seems to be a good dividing point as there is a distinct divide between 55th and 56th percentile. We choose cut off value of 33.5 so than we can divide the data in suitable manner.
   Performance levels are defined as below:
   Level 1: Below cut-off (total<33.5)
   Level 2: Above cut-off (total>33.5)

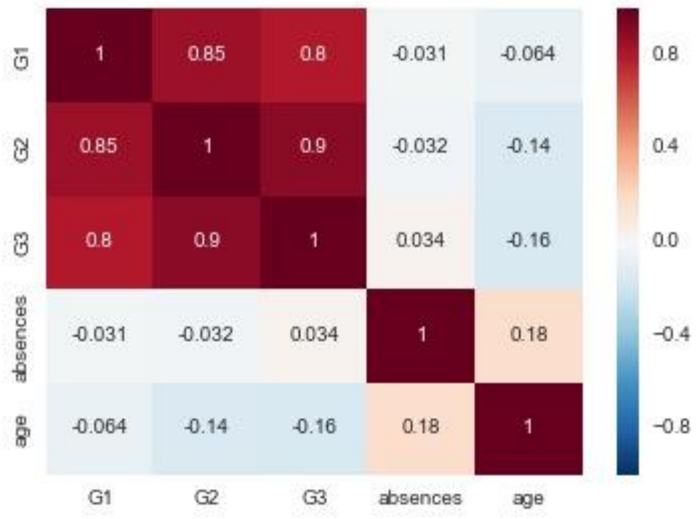
Data Visualization:

Detecting Null or Missing Values:



In order to detect any missing or null values visually, heat map has been used. Since the map colour is consistent, it is concluded that there are no missing values for any of the variables.

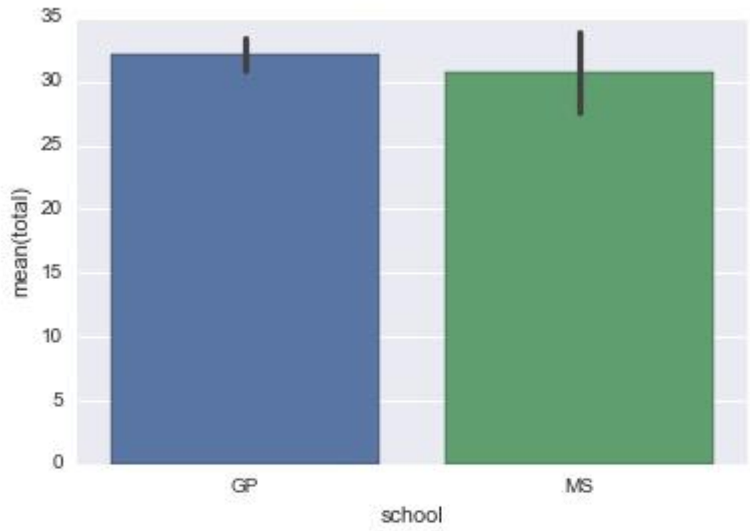
CORRELATION



We can observe that G1, G2 and G3 are strongly correlated. It was expected as G1, G2, G3 represent grades in first, second and third period. So performance of students has generally been consistent. Number of absences have a small negative correlation with all of G1 and G2 and small positive correlation with G3, apparently, absences don't really matter much. Age has a negative correlation with the grades, it is understandable as young students learn especially languages really quickly and hence that reflects in their grades as well. Absences and Age have small positive correlation, which implies that as older (in terms of age) students tend to miss classes.

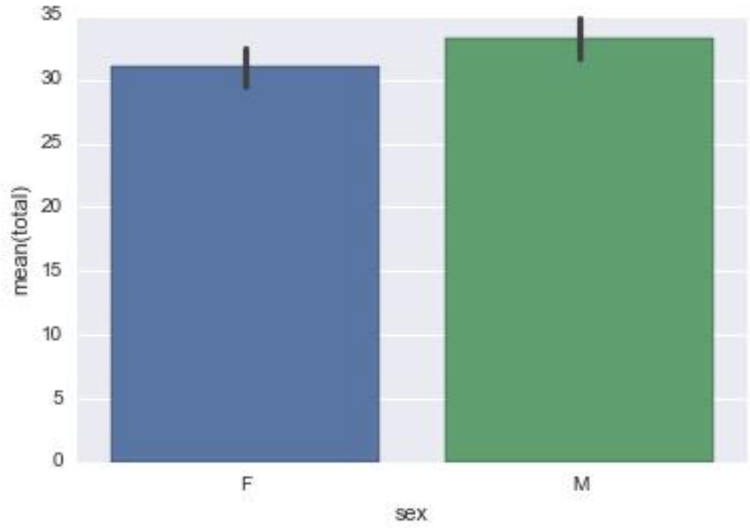
Total grade Vs Categorical variables

SCHOOL



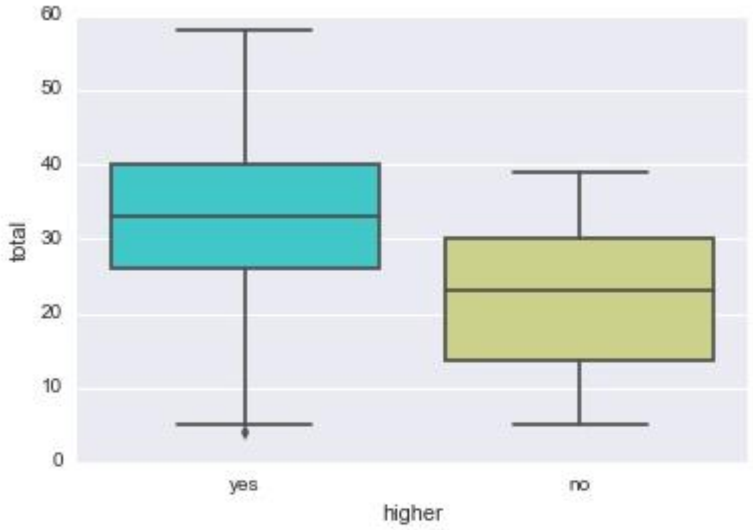
Average total score for GP school is greater than MS school but the difference is not that much. So it may not be a significant predictor.

SEX



Males have done better in Mathematics. Average total for males is greater than females. It may be a significant factor.

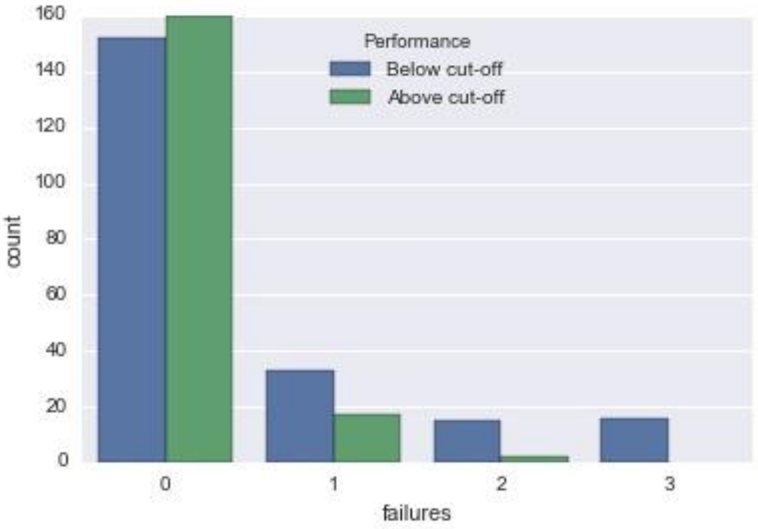
HIGHER EDUCATION



It can be observed from the box plot of total score with respect to categorical variable- higher that students who are willing to go for higher education have mean score better than the students who are not willing to go for higher education. Maximum total score among student who said ‘no’ is close to 40 and for students who choose ‘yes’ is 56. So students who are willing to go for higher education have outperformed the other category.

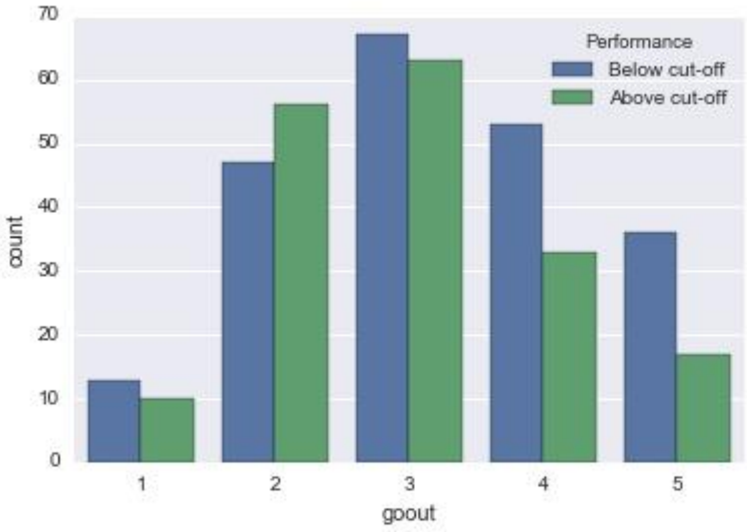
PERFORMANCE

Number of Past course failures



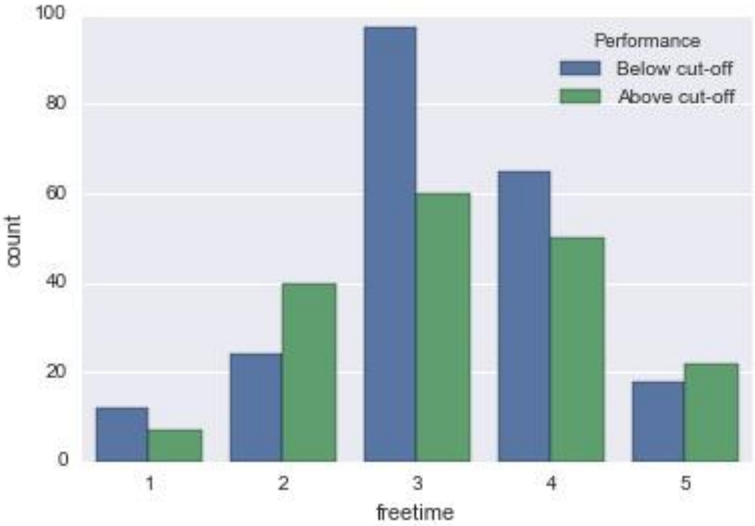
In the Performance count plot with respect to past failure in courses, we can observe that most of the students haven’t failed in any course before. There are very few students who have failed a course before and we can observe that there is not a single observation for ‘Above cut-off’ performance in category 3.

Going out intensity



Graph representing distribution of number of students in from 2 different performance categories with respect to different levels of going out intensities. Most of the students fall have medium intensity of going out (3rd category). If going out intensity is high and very high (4th and 5th), difference between two performance levels seems significant.

FREE TIME



Graph representing distribution of number of students in from 2 different performance categories with respect to different levels of free time intensity. Most of the students fall have medium intensity of free time (3rd category) and in that category, there seems to be a significant difference between the two performance levels.

SPECIFIC AIM 2: SIGNIFICANT FACTORS FOR PERFORMANCE

Objective: To find the significant factors impacting the performance of students. Performance has two levels: ’Below cut-off’, ‘Above cut-off’. Analysis will give insight into how changes in significant factors impact the Performance level.

Model used: Logistic Regression

Logistic Regression model is fit with event as performance as ‘Above cut-off’ as we are more interested in what factors are significant in prediction of ‘Above cut-off’ performance of students.

Feature Selection: Stepwise Selection

We fit a model with all the variables and use stepwise selection to come up with the significant factors.

Summary of Stepwise Selection

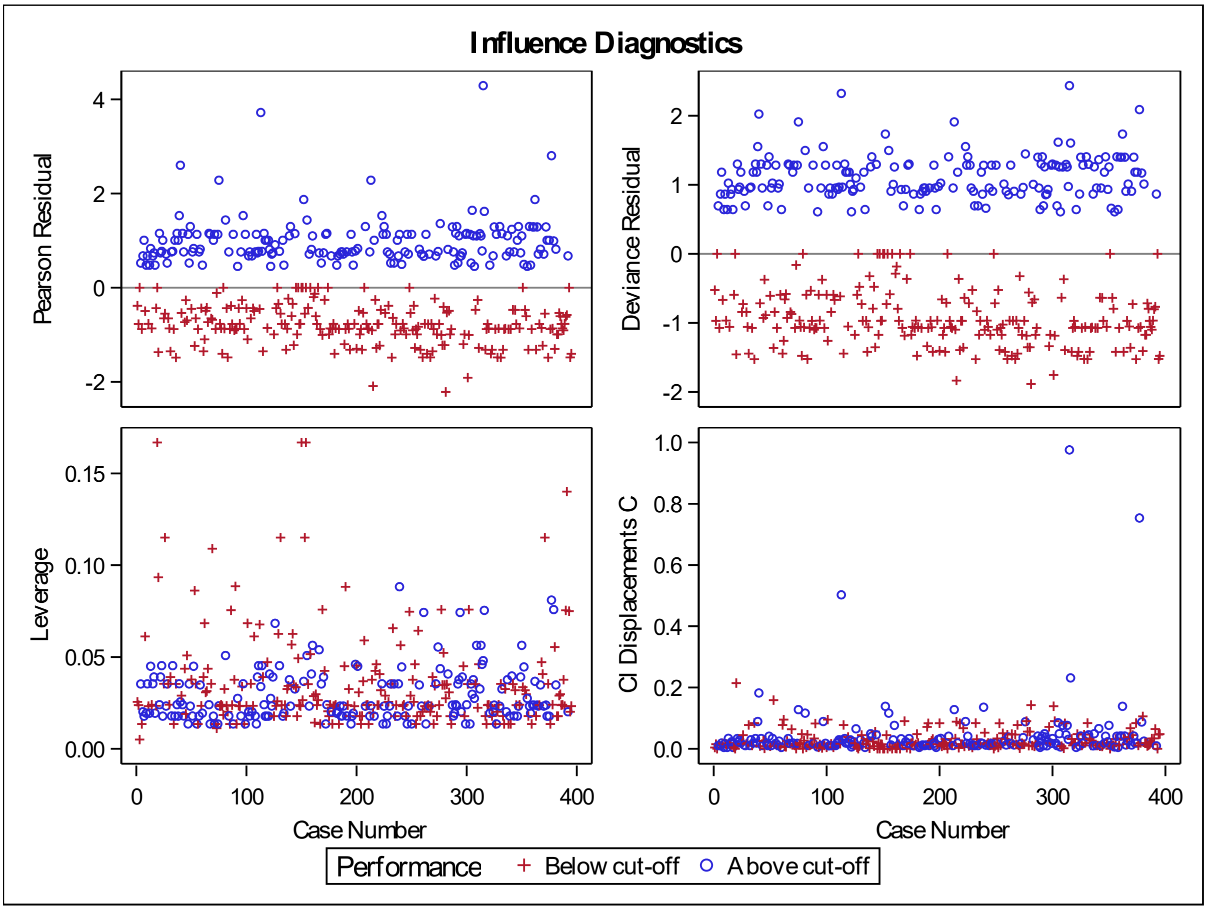
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	failures		3	1	28.0466		<.0001
2	schoolsup		1	2	14.3556		0.0002
3	Mjob		4	3	16.9323		0.0020
4	freetime		4	4	11.8693		0.0184
5	goout		4	5	9.6488		0.0468
6		Gout	4	4		9.4046	0.0517

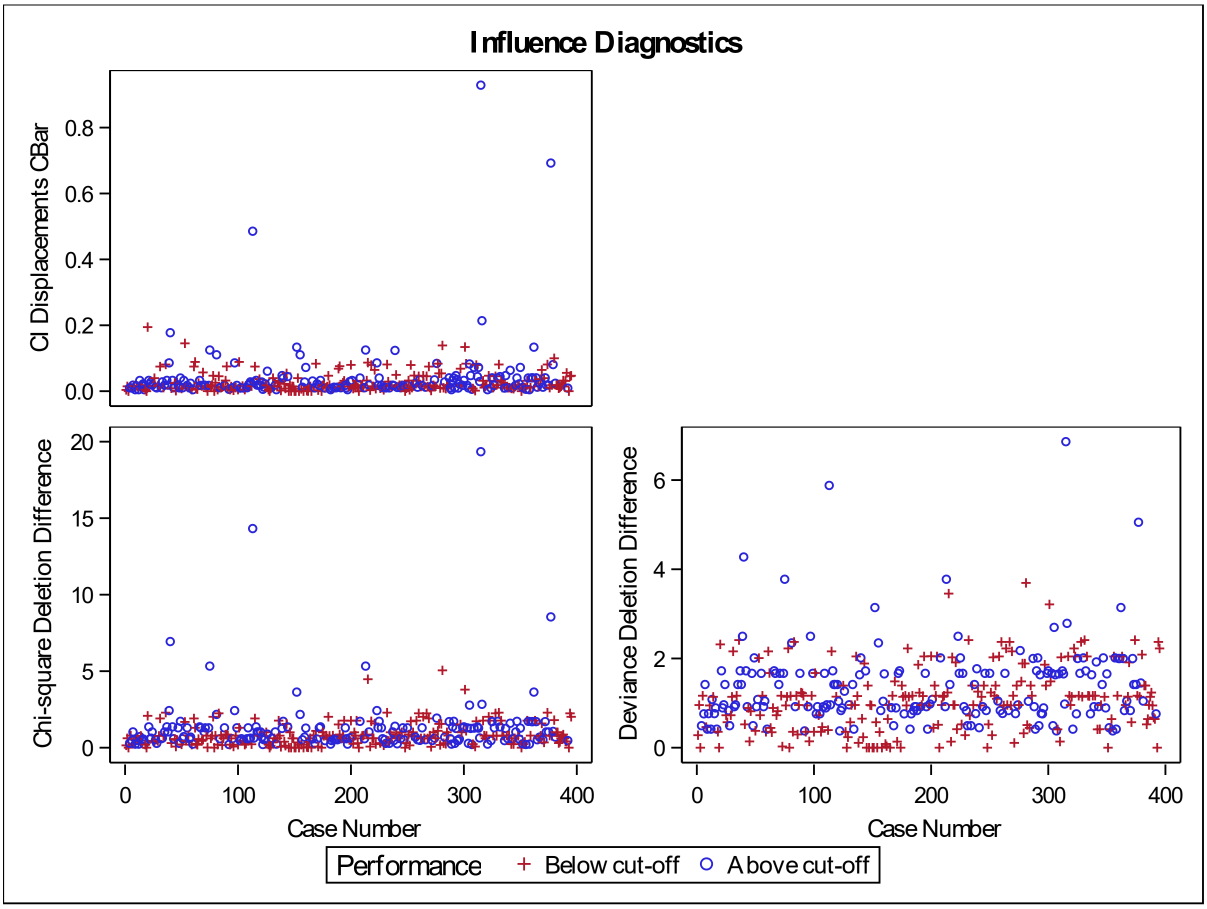
Failures, School Support, Mother’s Job and free time are most significant factors. Since inclusion of failure caused issues in Portuguese data. Let’s check the odds ratio first.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Mjob at_home vs teacher	1.012	0.450	2.273
Mjob health vs teacher	2.440	0.960	6.199
Mjob other vs teacher	1.314	0.682	2.529
Mjob services vs teacher	2.923	1.427	5.991
failures 0 vs 3	>999.999	<0.001	>999.999
failures 1 vs 3	>999.999	<0.001	>999.999
failures 2 vs 3	>999.999	<0.001	>999.999
schoolsup no vs yes	4.065	1.909	8.657
freetime 1 vs 5	0.460	0.130	1.625
freetime 2 vs 5	0.893	0.360	2.216

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
freetime 3 vs 5	0.354	0.157	0.798
freetime 4 vs 5	0.448	0.196	1.021





As we can observe that confidence interval for odds ratio for failures categories in comparison with 3rd category vary from less than 0.001 to greater than 999.99. The reason for this behaviour is as we observed in our data exploration analysis, for failure category other than 0, data points for 'Above cut-off' data points are very few.

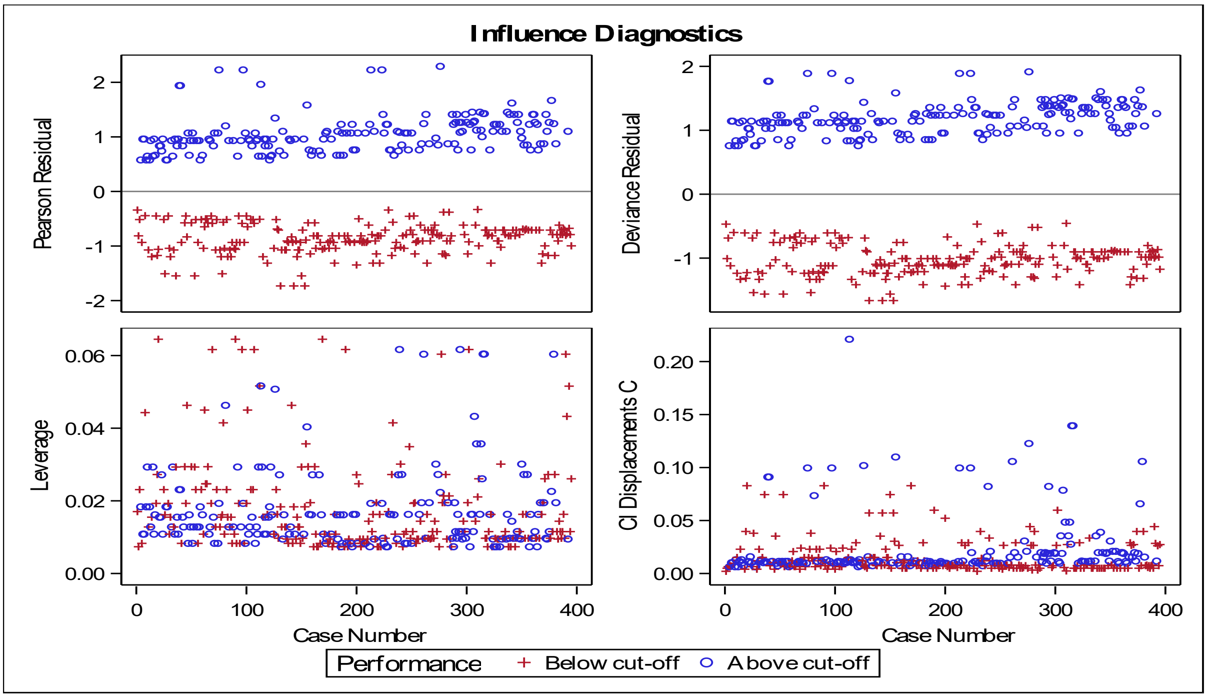
Refit model after removing 'failure' predictor

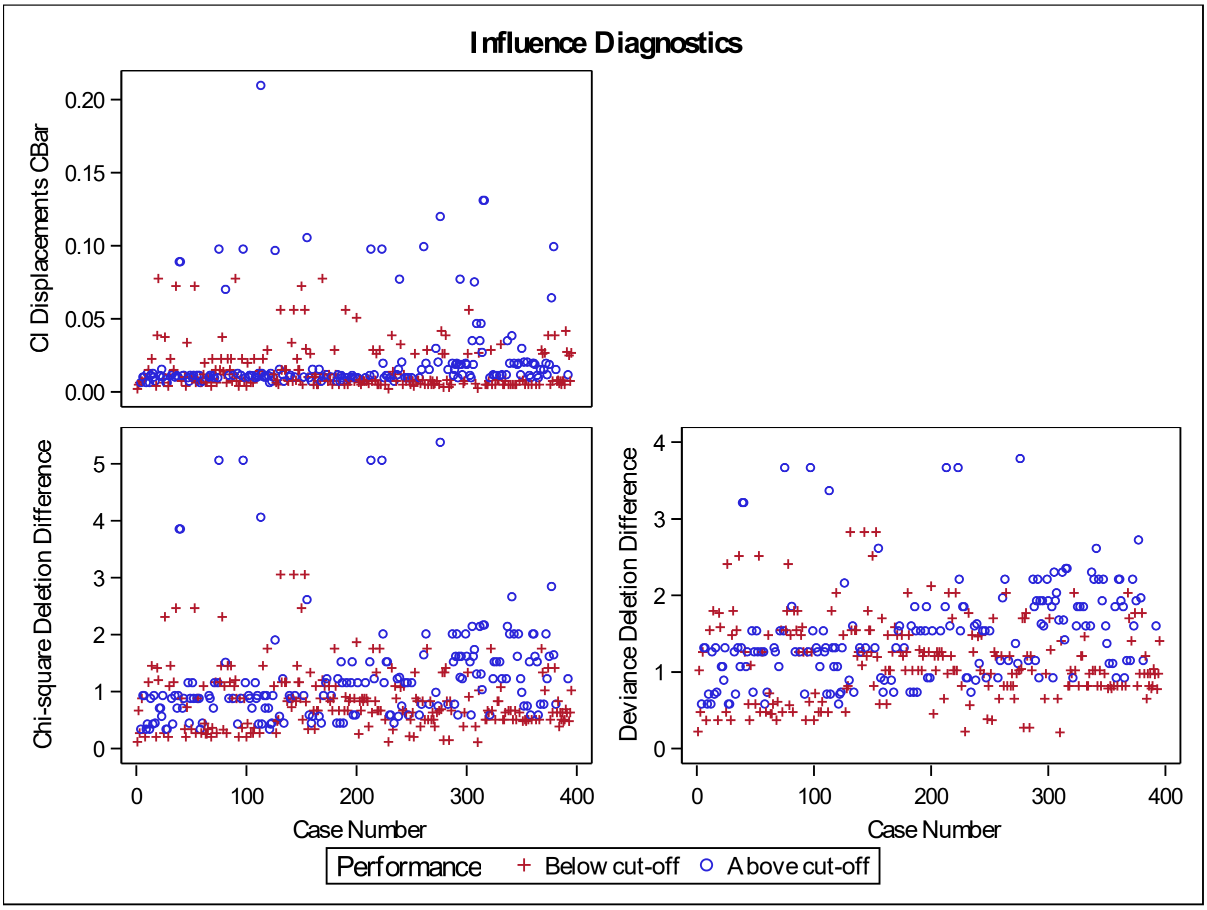
Feature Selection for New Model:

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	schoolsup		1	1	13.3274		0.0003
2	age		1	2	9.4882		0.0021
3	freetime		4	3	11.4918		0.0216

We observe that School Support, Father's education, and Mother's job are significant predictors and since all of them seem plausible, we continue with the model check for influential points.

INFLUENCE DIAGNOSTICS





Since there is high Cbar value and maximum value of Cbar is around 0.20, there is no need for removing observations. Though the point is relatively away but it’s value is not that big. So we continue with the model.

MODEL STATISTICS

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	546.115	522.682
SC	550.094	550.535
-2 Log L	544.115	508.682

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.4329	6	<.0001
Score	33.6319	6	<.0001
Wald	30.6711	6	<.0001

AIC for the model is relatively close to Intercept only model. SC value for model is actually greater than SC value for intercept only model. Hence this model is not an exceptional model. And we may not get too much insight by fitting this model. Let's check goodness of fit test.

GOODNESS OF FIT

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
1.9376	9	0.9924

The Hosmer-Lemeshow test, is highly insignificant indicating no problem with lack of fit in the model. Since model fits well, we continue with the model.

Parameters Estimates

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	3.9826	1.4541	7.5017	0.0062
schoolsup	**no**	1	0.7325	0.1894	14.9558	0.0001
freetime	**1**	1	-0.1792	0.4075	0.1933	0.6602
freetime	**2**	1	0.5257	0.2479	4.4993	0.0339
freetime	**3**	1	-0.4329	0.1923	5.0689	0.0244
freetime	**4**	1	-0.2140	0.2055	1.0846	0.2977
age		1	-0.2762	0.0881	9.8255	0.0017

We observe that school sup (no), level 2,3 for free time and age are significant. We will explain the impact of each on Performance using odds ratio.

ODDS Ratio

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
schoolsup no vs yes	4.328	2.060	9.094
freetime 1 vs 5	0.619	0.193	1.983
freetime 2 vs 5	1.253	0.546	2.876
freetime 3 vs 5	0.480	0.231	0.998
freetime 4 vs 5	0.598	0.281	1.270
age	0.759	0.638	0.902

In comparison to having school support, if students don’t have school support, they are more likely to be in ‘Above cut-off’ performance category.

In comparison to free time intensity 5 (very high):

a) Students who have very less free time (level 1), they are less likely to be in ‘Above cut-off’ performance category.
b) Students who have less free time (level 2), they are more likely to be in Above cut-off’
c) Students who have medium free or high free time (level 3 and 4), they are much less likely to be in Above cut-off’

Younger people are more likely to be in ‘Above cut-off’

SPECIFIC AIM 3: CLASSIFICATION POWER

Priority: Correct classification of ‘Above cut-off’

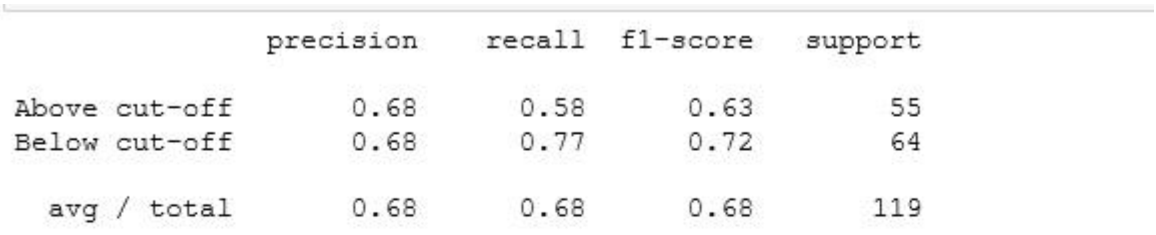
We now, predict the class of performance based on the given predictors. To increase the prediction accuracy, we use all the predictors and will split the data into training and testing data.

Data is randomly split into 70:30 as training and testing. We would train our models on same training set and test it onto the test data. The model selection would be done on the basis of test data.

MODEL 1: LOGISTIC REGRESSION

Logistic Regression with all the features gives better results in comparison to logistic regression with only selected variables. The classification report is as follows:

RESULTS:



```
               precision    recall  f1-score   support

Above cut-off       0.68      0.58      0.63        55
Below cut-off       0.68      0.77      0.72        64

  avg / total       0.68      0.68      0.68       119
```

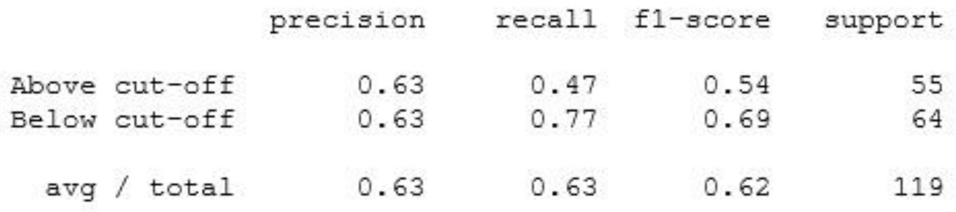
COMMENTS:

Above cut-off category has recall value of 58% which means 58% of total student whose performance is in ‘Above cut-off’ category have been captured. The recall value of ‘Below cut-off is much better and is 77% indicating that Logistic regression is doing much better job capturing ‘Below cut-off ‘category students.

MODEL 2: RANDOM FOREST

Fit a Random Forest classification model with all the features.

RESULTS:



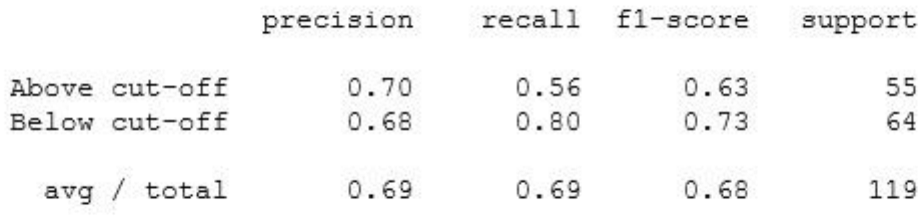
```
               precision    recall  f1-score   support

Above cut-off       0.63      0.47      0.54        55
Below cut-off       0.63      0.77      0.69        64

  avg / total       0.63      0.63      0.62       119
```

We can observe that recall value for Above cut-off is 47% whereas recall value for Below cut-off value is 77%. Precision for both the categories is 63%. We will tune the parameters in such a way that recall value of Above cut-off increases.

TUNING RESULTS:



```
               precision    recall  f1-score   support

Above cut-off       0.70      0.56      0.63        55
Below cut-off       0.68      0.80      0.73        64

  avg / total       0.69      0.69      0.68       119
```

After tuning the model and finding the appropriate parameters, we refit the model which gives 80% recall value for Below cut-off and 56% recall value for Above cut-off which is a considerable movement. Since there is increase in percentage points of recall values of both categories, precision has increased as well.

CONCLUSIONS

1. For Portuguese data set, with cut-off value of 33.5, students were divided into two performance based groups ‘Below cut-off’ and ‘Above cut-off’.
2. The most significant features that for prediction of performance were: **school, preference for higher education, number of absences, sex, school support, Daily Alcohol Consumption intensity and Mother’s education.**
3. For Portuguese data set, priority was to predict ‘Below cut-off’ category, so model selection was based on recall value of ‘Below cut-off’. Logistic Regression with selected features produced **recall value of 67%.** Support Vector Machine gave recall value of **69%** and Random Forest gave the recall value of **70%.** Random forest had 68% of precision as well. Hence Random Forest is selected as best model for Prediction of Below cut-off for Portuguese data set.
4. For Mathematics data set, with cut-off value of 33.5, students were divided into two performance based groups ‘Below cut-off’ and ‘Above cut-off’.
5. The most significant features that for prediction of performance were: **school support, free time intensity and going out intensity.**
6. Logistic model with covariates wasn’t such a great model as it’s AIC value was pretty close to the one with intercept. But the goodness of fit test was highly insignificant.
7. For Mathematics data set, priority was to predict ‘Above cut off’ category, so model selection was based on recall value of ‘Above cut-off’. Logistic Regression with all the produced **recall value of 58%.** Random Forest gave the recall value of **56%.** Logistic regression had 68% of precision as well. Hence Logistic regression model is selected as best model for Prediction of Above cut-off for Mathematics data set.
8. Random Forest works well if training data set is huge but in our case number of observations were far less and hence though it outperformed logistic regression for Portuguese data set, the difference wasn’t much significant. Random Forest is considered as one of the strongest classification algorithms.
9. Analysis and Code for this project can be replicated and applied various other student performance analysis data sets.