**Research on Which Factor Decides**

**Student Performance**

**BBangya  
Hansol\_Kim / Youtaek\_Jeon / Youngha\_Lee / Hyoshin\_Soh**

**Abstract** We used multiple regression at first, but before long we found it was not appropriate to analyze our data. Therefore, we used logistic regression and clustering analysis instead. Logistic regression is for the inference and prediction of the student performance. At the same time, clustering analysis is for assigning each observation to the nearest centroid.

We tried to figure out which factor affected the student performance more, an internal one or an external one. With logistic regression, we can see many factors decided students' grades in Portuguese and Math. In Portuguese, personal factors affected student performance such as "failures" and "higher". "Failures" means the number of past class failures, and "higher" means willingness to take higher education. The less failure they had, the better grade they got. Also students who wanted more education had higher scores. In Math, male students were better than female students. It means a biological factor affected grades. Also, "failure" and "paid" decided them too. "Paid" is an extra educational support from home. Students with less failure and more educational support were better. Therefore, math grade was influenced by not only personal factors but also a home background.

As to clustering analysis, we can see following results. We grouped variables together according to their features and selected five notable variables to show the result. The less class students skipped and the less failure or romantic relationship they had, they got better grades. On the other hand, more school supports make students' grades better.

**Introduction** Our team conducted a research on student performance since so many Koreans are so enthusiastic about education and student performance problem. In Korea, education is a huge issue because 12-year studies in childhood can decide children's future lives. In addition, Korean people believe there is a strong connection between environmental factors and student performance. There is a reason why people hope to live in Kangnam, the best place to educate their children in Korea, which is notorious for a high rent.

We used data on Portuguese student performance. We investigated which factor affects students' studies the best with this data. We have a limitation because there could be a cultural difference between Portuguese education environment and Korean education environment. However, we think we can predict Korean student performance with the result of this report.

**Methods**

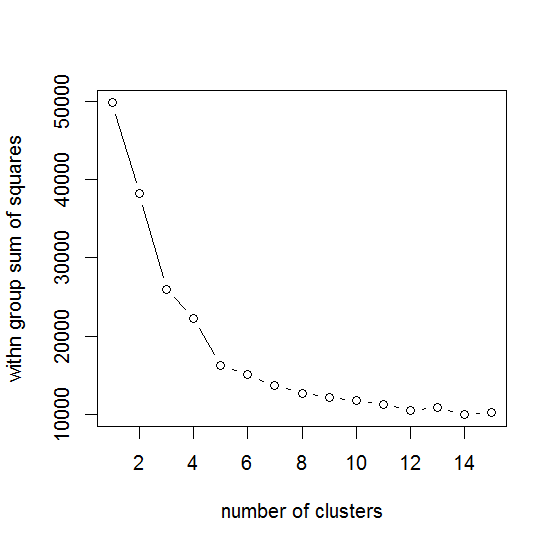
**Logistic Regression** Logistic regression is used to infer and predict student performance. Before analysis, we determined student performance rating. In math and Portuguese final grades(G3), we divided them into two groups. The bottom 25% is “Bad” and the rest 75% is “Good”. And we made them binomial response like “Good (Y=1)” and “Bad(Y=0)”. “ISLR” package in R was used, and following steps should be done.

1. Using new response variable, we initially model full model and remove not significant variables.
2. Choose the final reduced model based on AIC(Akaike Information Criteria).

**Clustering Analysis (K-means Method)** Clustering Analysis is used to Assign each observation to the nearest centroid. Following Steps are taken.

1. Decide a number of clusters (K).
2. Take initial values which would be centroid value.
3. Assign observations to the centroid value.
4. Calculate the new centroid based on the result, and reassign the other observations.
5. Repeat the steps until each cluster seldom change.

We used "Cluster" package in R, and there is a way to decide the number of clusters (K). We have to find the number of clusters which minimizes within sum of square (WSS). WSS means the minimal first differentiation sum of distance square.



the minimal first difference point

**Results**

**Logistic Regression** In Portuguese data, the following results were shown.

|  |
| --- |
| Log[π(x)/1- π(x)] = 0.8388 -1.0694 \* failures + 1.4774 \* higher |
|  |

H0 : Coefficients in model are not significant.  
Under [Null deviance – Residual deviance = 557.79 – 467.38]   
 ~ Chi-squared distribution (degree of freedom = 648-646)  
Under the null, p-value is 2.33174E-20. So we reject null hypothesis and choose model.

Students who replied “yes” to the question about "failures" have 0.343214 times bigger odds ratio of student performance than those who replied “no”. Students who replied “yes” to the question on "higher" have 4.381539 times bigger odds ratio of student performance than those who replied “no”.

These results mean that student performance in Portuguese is influenced by experience of failures and volition to higher education. Two factors are students' personal aspect.

In math data, the following was shown.

|  |
| --- |
| Log[π(x)/1- π(x)] = 1.216 + 0.7854 \* sex – 0.7691 \* failures + 0.8985 \* paid |
|  |

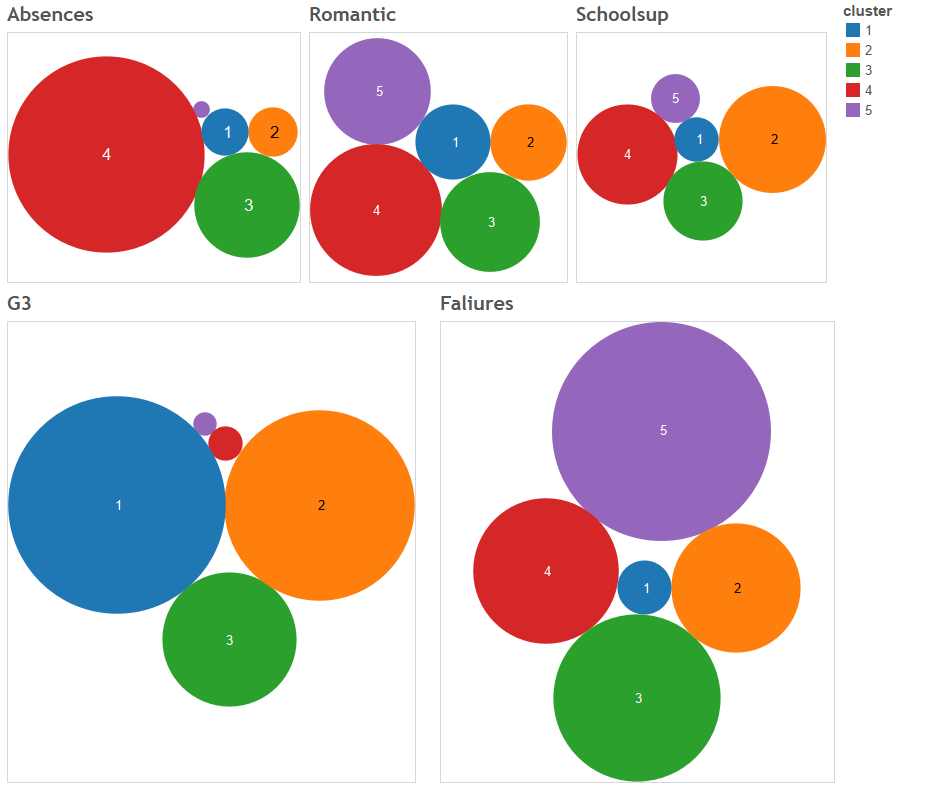
H0 : Coefficients in model are not significant.  
Under [Null deviance – Residual deviance = 369.04 – 326.42]   
 ~ Chi-squared distribution (degree of freedom = 394 - 391)  
Under the null, p-value is 1.24075E-08. So we reject null hypothesis and choose model.

Students who are male have 2.193284 times bigger Odds ratio of student performance than those who are female. Students who replied “yes” to the question about "failures" have 0.46343 times smaller odds ratio of student performance than those who replied “no”. Students who replied “yes” to the question on "paid" have 2.455916 times bigger odds ratio of student performance than those who replied “no”.

These results mean that student performance in Math is influenced by various factors. "Sex" is a biological factor. Male students are better in math than female ones. "Failure" is similar to above. It affects final scores negatively. "Paid" is an extra educational support from home. It means that math grade is influenced by not only personal aspects but also a home background.

**Clustering Analysis (K-means Method)** The reason why we used clustering analysis is that it can be a good method to try various approaches according to students' traits. With this, we are going to group some observations together into several clusters.

Close observations can be put together by clustering analysis, and then they can be classified according to their properties. We separated all the variables of this data into five clusters according to K=5. By using data visualization of five distinctive variables, we can show you how observations are bound together by their properties.



The picture above shows an average amount of each cluster's variables. Sizes of a circle shows sizes of averages, and colors and letters show which clusters they are. For example, students in Cluster 1 have less absence and romantic relationship. They have the best grade, though they have less school support. On the other hand, students in Cluster 4 have the most absences, romantic relationship, and school supports. However, they often get failures, and their grades are the second lowest. With this kind of clustering analysis, we can see some features in common of the observed values.

We can try various instructions according to students' traits for them to get better grades with these results. For instance, students in Cluster 4 often skip the class and have had many failures. Therefore, we should make them to participate in the class more. In case of Cluster 5, we need some education systems to encourage school supports.

**Discussion**

**Can We Predict Korean Student Performance with This Data?** As we told you before, we admit our limitation that a cultural environment in Portugal and Korea could be different. However, education-wise, it seems that many educational aspects are in common in Portugal and Korea. Many teachers and parents in Korea tell children not to go out with someone because it can disturb their studies. Surprisingly, the data collected in Portugal shows the same result, romantic relationship discourages one's studies. Likewise, boys are usually better in Math than girls are in Korea, too. There are so many similarities, and we can sure human beings are not that different as to education.

**Conclusion and Future Directions** As you can see the results above, many different kinds of variables decide the student performance. It's rather complex, not simple. However, many Korean parents tend to only care how many extracurricular lessons their kids take, how long a day they study, etc. They usually care quantity-related factors and ignore other things such as their children's own properties. Yet, this report shows we must not overlook other factors, and we should instruct children in different ways according to their individual features. Some might be influenced extremely by atmosphere of their neighborhood or school. Then they had better move to a better town like Kangnam. On the other hand, some might be affected not by it, by family support. Then they don't need to bother to move. All they need can be parents' care. The thing is we should consider student's own traits and use different methods. Unfortunately, it is not common in Korea. Therefore, we have to research more on relationship between one's properties and grades to develop education means. Then, we can raise our children smart enough not living in Kangnam.

**Refernces** P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7