**FACTORS AFFECTING HIGH SCHOOL STUDENTS’ ACADEMIC PERFORMANCES.**

**BY AJIBOLA JAMES AKISANMI**

**ILLINOIS STATE UNIVERSITY**

**DATE: 12/09/2016**

**ABSTRACT**

The purpose of this study is to examine the factors that affect high school students’ academic performance in an attempt to identify the most important factors and to educate governing bodies in order to improve the quality of high school education. Presently, Portugal has one of the worst academic performances in Europe due to high failure and drop-out rates among its high school students. This abysmal situation is of huge concern to the Portuguese government as the well-being of the next generation depends heavily on the current youth of the country. In this work, an analysis is made on real-world data from the year 2006 based on two Portuguese high schools. Two different sources were used: mark reports and questionnaires. The mark reports contained scarce information (i.e. only the grades and number of absences were available), it was complemented with information from the questionnaires, which allowed the collection of several demographic, social and school-related attributes (e.g. student’s age, alcohol consumption, mother’s education etc.) (Cortes & Silva, 2008). The objective is to predict students final grade performance and if possible, to identify the most important variables that affect educational success or failure.

**SECTION ONE: INTRODUCTION**

Education is key for long-term success in terms of economic growth and the well-being of any country. However, In Portugal there is an overwhelming failure and dropout rate among high school students in the country, thus leaving their educational sector as one of the worst in Europe and a laughing stock among EU countries who perform dramatically better than Portugal. To be more precise, Portugal is the poorest and least educated country in Western Europe. There is also a debt crisis bearing down and the country must make massive reforms to fix its economy, and education is at the top of the list (World Street Journal, Charles Forelle). Put simply, Portugal must generate enough long-term economic growth to pay off its large debts, however, an unskilled workforce makes that hard. In addition, 28% of the Portuguese population between 25 and 64 has completed high school. Whereas the figure is 85% in Germany, 91% in the Czech Republic and 89% in the U.S (World Street Journal, Charles Forelle).

On a positive note, Data Mining promises a way to accurately identify the factors for the astonishing failure and drop-out rates in the country and based on the huge surge of databases due to the rise of Information Technology, it has become easier to access relevant data in order to improve decision making and optimize success. The new emerging field, called Educational Data Mining (EDM) is the application of Data Mining techniques on educational data. The objective of EDM is to analyze such data and to resolve educational research issues. Educational Data Mining researchers study a variety of areas, including the factors that are associated with student failure and the improvement of student models. Educational data mining uses many techniques such as Support Vector Machines, Linear Regression, Logistic Regression, Decision Trees, Neural Networks, and many others. Prediction and analysis of student performance is an important milestone in the educational environment.

Students’ academic performance is a crucial factor in building students’ future. Academic performance of students is not a result of only one deciding factor besides it heavily hinges on various factors like personal, socio-economic, psychological and other environmental variables. This paper identifies the factors associated in an attempt to improve the quality of education.

There are several interesting questions for this domain that could be answered using DM techniques (Luan 2002, Minaei-Bidgoli et al. 2003): Who are the students taking most credit hours? Who is likely to return for more classes? What type of courses can be offered to attract more students? What are the main reasons for student transfers? Is it possible to predict student performance? What are the factors that affect student achievement? This report will focus solely on the last two questions.

Therefore, the research questions proposed in this study are:

• Q1: What are the important factors used in predicting students’ performance?

• Q2: What are the most accurate prediction methods used for students’ performance?

Data mining methodologies utilized for this report are to study student performance at high school level and to identify the highly influencing predictive variables on the academic performance and to find the best prediction algorithm. This could be highly influential in order for school professionals to perform corrective measures for weak students (e.g. remedial classes). In summary, evaluation and prediction of students’ performance in high school will help to find important factors affecting students’ success in education and moreover they can have an important role in helping educational managers to improve the quality of schools.

**SECTION TWO: LITERATURE REVIEW**

Over recent years, several papers have attempted to predict student’s performances in different subjects and areas. Kotsiantis et al. (2004) applied several DM algorithms to predict the performance of computer science students from a university distance learning program. For each student, several demographic (e.g. sex, age, marital status) and performance attributes (e.g. mark in a given assignment) were used as inputs of a binary pass/fail classifier. The best solution was obtained by a Naive Bayes method with an accuracy of 74%. Ma et al. (2000) applied a DM approach based on Association Rules in order to select weak tertiary school students of Singapore for remedial classes. The input variables included demographic attributes (e.g. sex, region) and school performance over the past years and the proposed solution outperformed the traditional allocation procedure. Minaei-Bidgoli et al. (2003), online student grades from the Michigan State University were modeled using three classification approaches (i.e. binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). The database included 227 samples with online features (e.g. number of corrected answers or tries for homework) and the best results were obtained by a classifier ensemble (e.g. Decision Tree and Neural Network) with accuracy rates of 94% (binary), 72% (3-classes) and 62% (9-classes).

Pardos et al. (2006) collected data from an online tutoring system regarding USA 8th grade Math tests. The authors adopted a regression approach, where the aim was to predict the math test score based on individual skills. The authors used Bayesian Networks and the best result was a predictive error of 15%. V. Ramesh et al (2013) tried to identify the factors influencing the performance of students in final examination. The results from hypothesis testing reveal that type of school does not influence student performance but parent’s occupation plays a major role in predicting grades.

In conclusion, even though quite a number of researchers have studied this particular domain, more research needs to be done using more statistical techniques which happen to be missing in the current literature. Methods such as linear model selection and regularization could provide a new dimension which has not been explored yet. Finally, the results of the proposed research may add to existing literature and provide information which could guide new education driven procedures for countries such as Portugal.

**SECTION THREE: METHODOLOGY**

**Data Collection & Description**

The dataset was collected from kaggle’s website (www.kaggle.com) with the data provided to data scientists to continue research on students’ alcohol consumption and students’ academic performance in Portugal. According to the data collectors (Cortez & Silva, 2008), In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. There are several courses (e.g. Sciences and Technologies, Visual Arts) that share core subjects such as the Portuguese Language and Mathematics. Like several other countries (e.g. France or Venezuela), a 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. During the school year, students are evaluated in three periods and the last evaluation (G3 in the table below) corresponds to the final grade.

This study will consider data collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal. The database was built from two sources: school reports, based on paper sheets which included few attributes (i.e. the three-period grades and the number of school absences); and questionnaires were used to complement the previous information. Additional data was augmented with closed questions (i.e. with predefined options) related to several demographic (e.g. mother’s education, family income), social/emotional (e.g. alcohol consumption) (Pritchard and Wilson 2003) and school related (e.g. number of past class failures) variables that were expected to affect student performance.

Also, the questionnaire was reviewed by school professionals and tested on a small set of 15 students in order to get feedback. During the preprocessing stage, some features were discarded due to the lack of discriminative value. For instance, few respondents answered about their family income (probably due to privacy issues), while almost 100% of the students live with their parents and have a personal computer at home.

Finally, the data used for this study was based only on the Mathematics class (with 395 examples) and a total of 33 interesting predictors. Reason being that the failure in the core subject - Mathematics is extremely serious and it also provides fundamental knowledge for the success in the remaining school subjects (Cortez & Silva, 2008). The data was also randomly split into two in order to test for validity and accuracy of each statistical Data Mining technique used (i.e. 197 for training the data and 198 for testing the data).

Description of each variable used is as follows:

|  |  |
| --- | --- |
| **Attribute** | **Description(Domain)** |
| sex | Student’s sex (Binary: Female or male) |
| age | Student’s age (Numeric: from 15 to 22) |
| school | Student’s school (Binary: Gabriel Pereira or Mousinho da Silveira) |
| address | Student’s home address type (Binary: urban or rural) |
| Pstatus | Parent’s cohabitation status (Binary: living together or apart) |
| Medu | Mother’s education (Numeric: from 0 to 4) |
| Mjob | Mother’s job (nominal) |
| Fedu | Father’s education (Numeric: from 0 to 4) |
| Fjob | Father’s job(Nominal) |
| guardian | Student’s guardian (Nominal: Mother, father or other) |
| famsize | Family size (Binary: ≤3 or ≥ 3 |
| famrel | Quality of family relationship (numeric: from 1 – very bad to 5 – excellent) |
| reason | Reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| traveltime | Home to school travel time (Numeric: 1 - < 15min, 2 - 15 to 30 min, 3 - 30 min to 1 hour or 4 - > 1 hour) |
| studytime | Weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – >10 hours) |
| failures | Number of past class failures (numeric: n if 1 ≤ n < 3, else 4) |
| schoolsup | Extra educational school support (binary: yes or no) |
| famsup | Family educational support (binary: yes or no) |
| activities | Extra-curricular activities (binary: yes or no) |
| paidclass | Extra paid classes (binary: yes or no) |
| internet | Internet access at home (binary: yes or no) |
| nursery | Nursery attended nursery school (binary: yes or no) |
| higher | Wants to take higher education (binary: yes or no) |
| romantic | With a romantic relationship (binary: yes or no) |
| freetime | Free time after school (numeric: from 1 – very low to 5 – very high) |
| goout | Going out with friends (numeric: from 1 – very low to 5 – very high) |
| Walc | Weekend alcohol consumption (numeric: from 1 – very low to 5 –very high) |
| Dalc | Workday alcohol consumption (numeric: from 1 – very low to 5 –very high) |
| health | Current health status (numeric: from 1 – very bad to 5 – very good) |
| absences | Number of school absences (numeric: from 0 to 93) |
| G1 | First period grade (numeric: from 0 to 20) |
| G2 | Second period grade (numeric: from 0 to 20) |
| G3 | Final grade (numeric: from 0 to 20) – Response Variable |

To provide context, the response variable, G3, which is evaluated as the final grade for each student is usually divided into five different parts as illustrated in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country (Subject) | Excellent or Very good | Good | Satisfactory | Satisfactory | Fail |
| Portugal (Mathematics) | 16-20 | 14-15 | 12-13 | 10-11 | 0-9 |

**SECTION FOUR: ANALYSIS**

The statistical techniques used for analysis in this project are Linear Regression, Linear Model Selection and Regularization, Tree-Based methods and Neural Networks. All of which require supervised learning, where a model is adjusted to a dataset made up of k ɛ {1, ...,N} examples, each mapping an input vector (xk1 , . . . , xkI) to a given target yk (Cortez & Silva, 2008).

To go in detail, the Decision Tree (DT) is a branching structure that represents a set of rules, distinguishing values in a hierarchical form (Breiman et al. 1984). This representation can be translated into a set of IF-THEN rules, which are easy to understand by humans. The Random Forest (RF) (Breiman 2001) is an ensemble of T unpruned DT. Each tree is based on a random feature selection from bootstrap training samples and the RF predictions are built by averaging the outputs of the T trees. The RF is more difficult to interpret when compared with the single DT, although it is still possible to provide explanatory knowledge in terms of its input variable relevance. Whereas in the case of Boosting each tree is built sequentially i.e. each tree is grown based on the knowledge of the previously grown trees. Nonlinear functions, such as Neural Networks (NN) and Support Vector Machines (SVM), have also been proposed for DM tasks (Hastie et al. 2001), obtaining better results when a high nonlinearity is present. It should be noted that NN and SVM use model representations that are difficult to understand by humans.

Similarly, the Lasso is expected to outperform the Ridge Regression and OLS since it makes use of a tuning parameter that enables the model to produce sparse variables and as such, makes use of only the important variables while leaving the unimportant ones out of the model. In short, the coefficients are shrunken towards zero as compared to the least squares estimates. Not only is the prediction accuracy expected to be greater, the level of interpretability of the model should be highly enhanced. The subset selection, another interesting statistical technique, also promises to identify a key subset of predictors and considering the fact that it makes use of the easily understandable Least Squares, results of this technique might just be what answers all the research questions identified in the Introduction of this paper.

A complete list of all statistical methods considered are as follows: OLS, Subset Selection, Ridge Regression, the Lasso, Decision Trees, Random Forests, Bagging, Boosting, PCR, PLS and Neural Networks.

**SECTION FIVE: RESULTS & DISCUSSION**

In order to perform further investigation on the most important factors that affect students’ performances, each statistical method which was proposed in the Analysis section was recorded alongside their predictive performance which can be found in the table below. The list was sorted based on their predictive performances, with the best performance at the top and the worst performance at the bottom.

|  |  |
| --- | --- |
| **Statistical DM technique** | **Predictive performance (MSE)** |
| Boosting | 3.11 |
| Subset Selection | 3.12 |
| Lasso | 3.14 |
| Random Forest | 3.28 |
| Bagging | 3.60 |
| Ridge | 4.09 |
| Neural Network | 4.19 |
| OLS | 4.19 |
| PLS | 4.62 |
| PCR | 5.11 |
| Regression Tree | 6.06 |

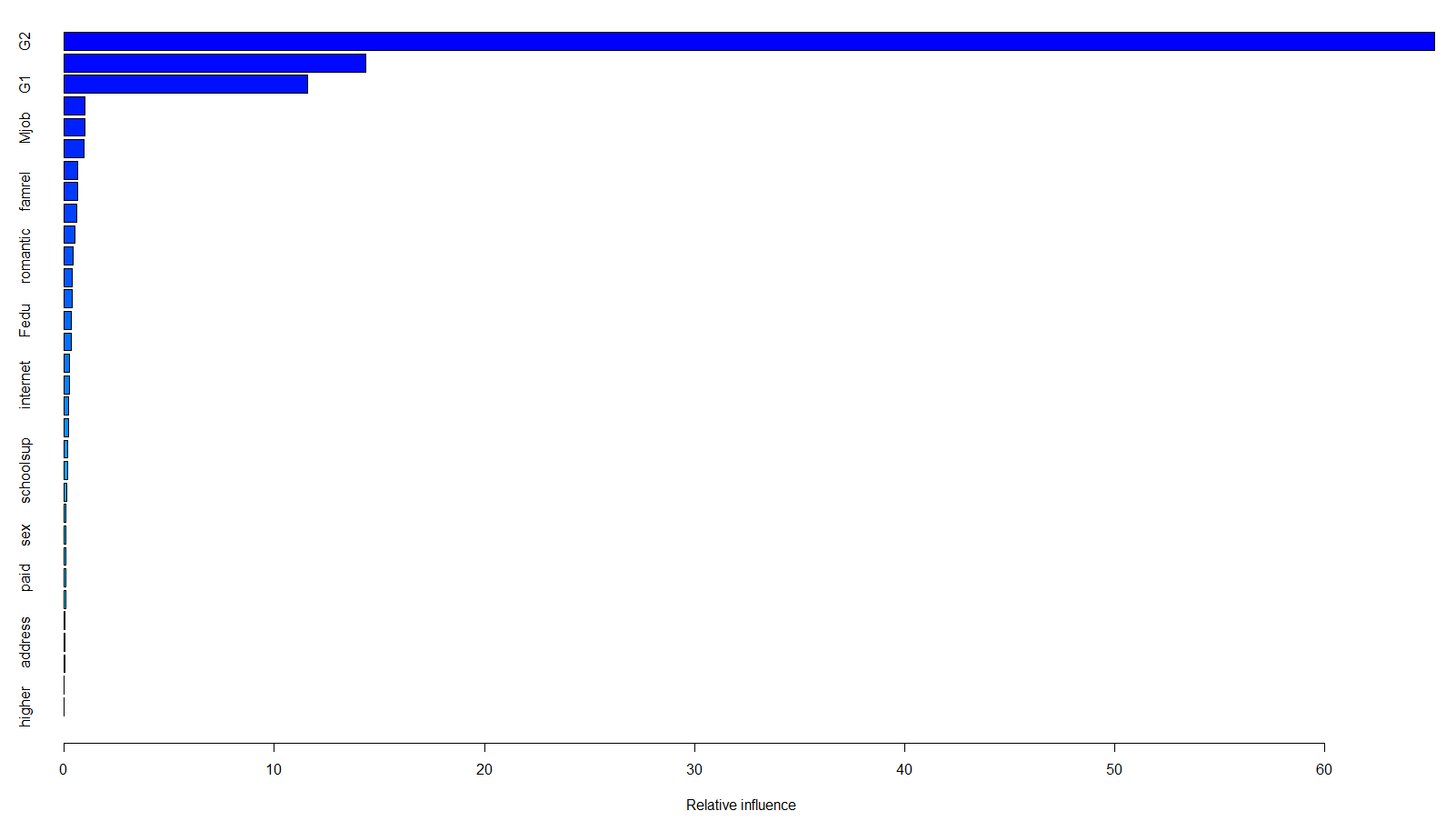
According to the results above, the best three models are the GBM (Generalized Boosted Model), the subset selection and the Lasso. It is important to note that all three models had performances relatively close to each other and thus must be considered when validating for the most important factors affecting High School students’ academic performances.

**Boosting**

A very powerful statistical method that improves on the predictive results of the decision tree and although it learns slowly, it makes up for it in terms of predictive capabilities. It doesn’t overfit and most importantly, it builds off trees that have been grown. That capability can only be found in Boosting and not even the more popular Random Forest or Regression Tree can boast of having such capabilities. It is no surprise that is was the best model in terms of predictive performance with an eye-opening MSE of 3.11.

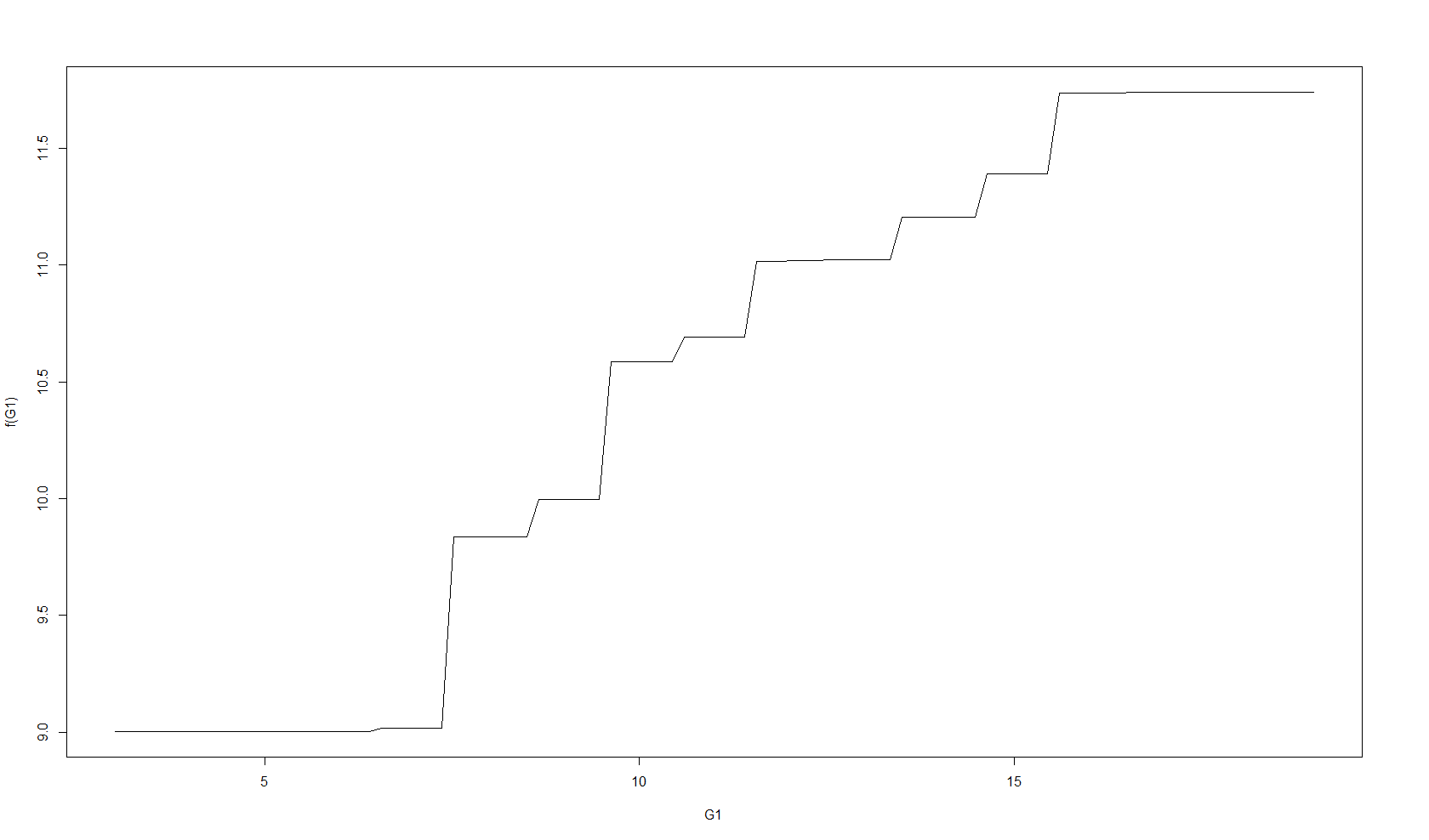
During the process of running the GBM code, there were three tuning parameters considered which were the total number of Trees to be used, the shrinkage parameter and the total number of splits. Initially, the default values of all three parameters were used and that gave a very good predictive performance. However, in order to fully make use of the potential of the GBM, further tunings of the parameters were made. The results of the boosted model are as follows:

|  |  |
| --- | --- |
| **Variable** | **Relative Influence of top 5 variables on G3 (%)** |
| G2 | 65.21% |
| Absences | 14.35% |
| G1 | 11.57% |
| Age | 1% |
| Mjob | 1% |



Graph of Relative Influence (GBM)

|  |  |
| --- | --- |
| G2’s influence on the y-predictor (G3) | Number of absences influence on the y-predictor (G3) |

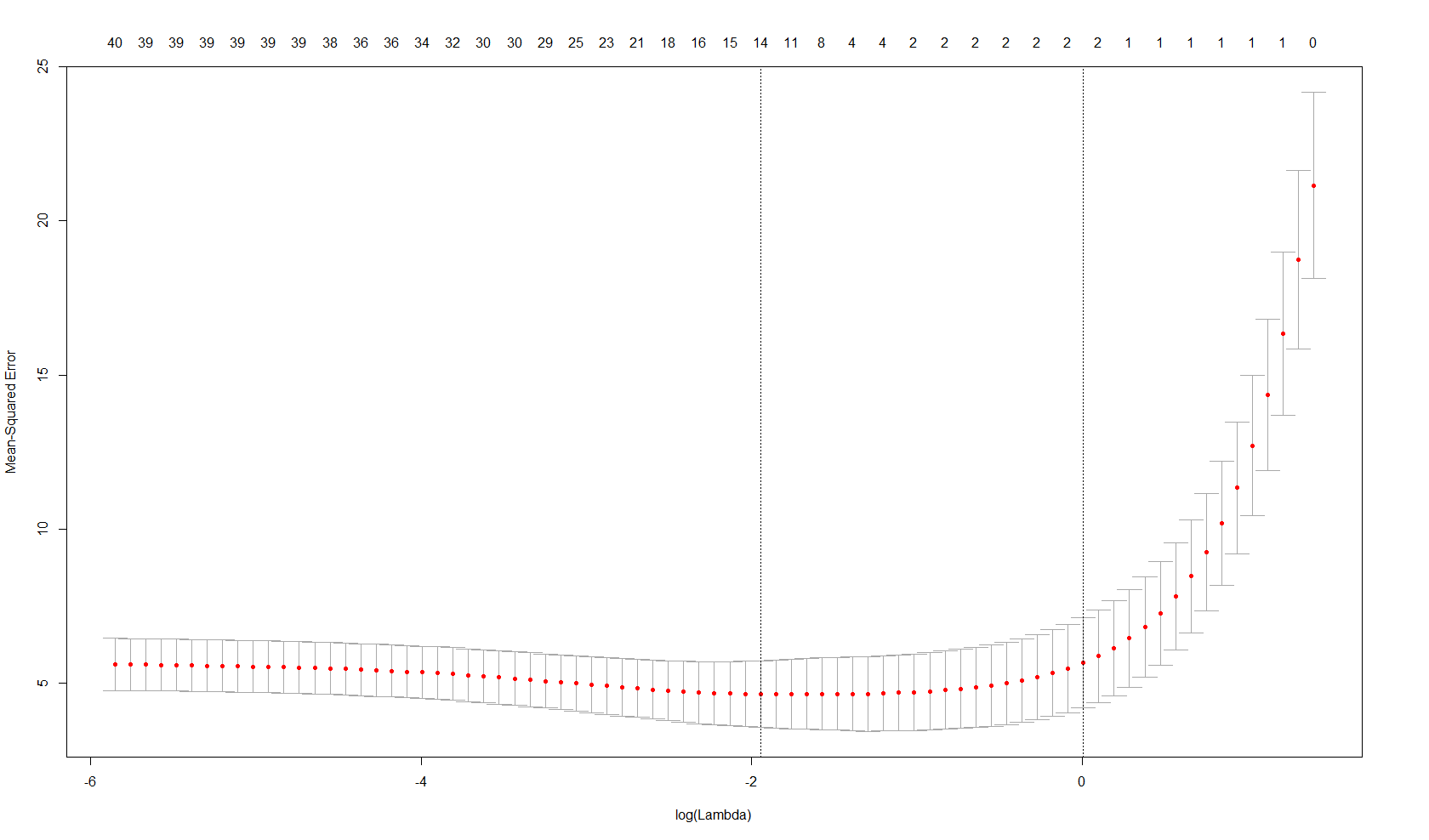


G1’s influence on the y-predictor (G3)

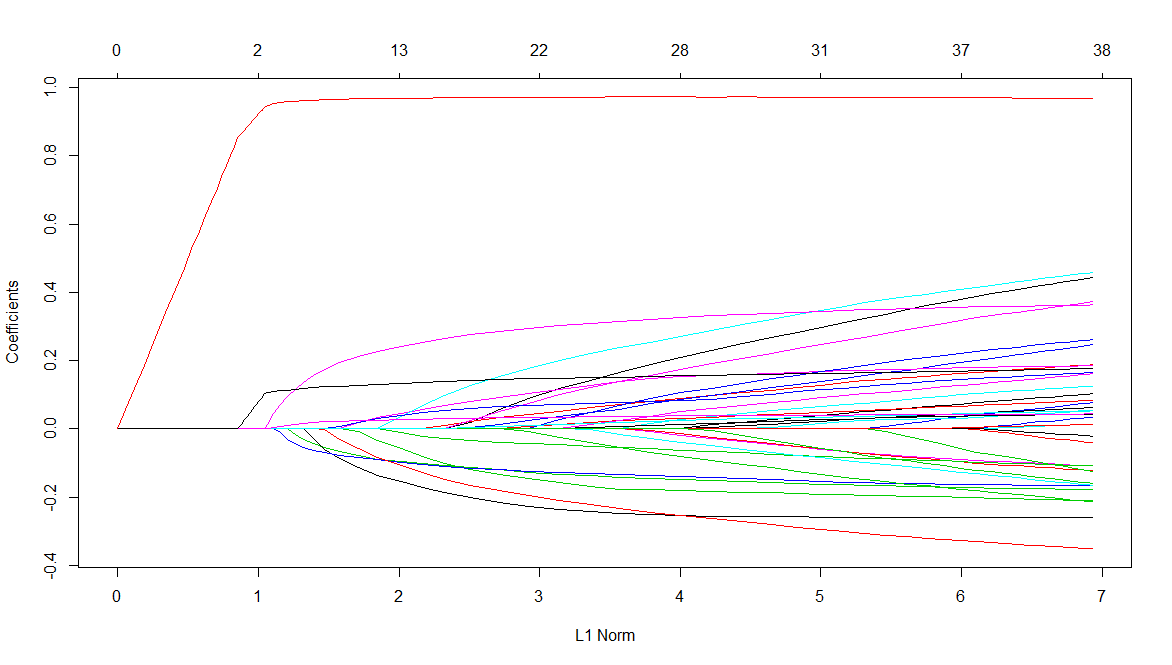
**The Lasso**

The lasso is one of the shrinkage methods that regularizes the coefficient estimates of its independent variables to a value of Zero. By so doing, it improves on the OLS by reducing the variance in the model. The ridge regression is very similar to the Lasso, however, it doesn’t produce sparse variables- an attribute which could help solve our research questions. Gladly, the Lasso will only include the important p predictors in its final model by making use of the l1 penalty that could force the coefficient estimates of some variables to be exactly zero.

This makes this approach very attractive and it tremendously answers both our research questions as it has one of the highest predictive performances for our data and it also boasts the sparse producing variable capability. Hence, making it very interpretable. The parameter considered for the Lasso was the lambda value, this is crucial in order to ensure that the tuning parameter is sufficiently large and thus predicted with an MSE of 3.14. This parameter was cross-validated using the training data and an optimal lambda value of 0.143 was obtained as shown below.



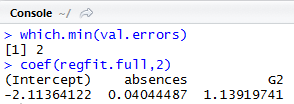
According to our Lasso results, the sparse variables produced include: Age, Fjobservices, reasonhome, guardian mother, study time, failures, school supervs, activties yes, famrel, walc, absences, G1 and G2 while the most important variables based on their respective coefficients are G2, G1, absences, famrel, walc and guardian mother as depicted in the figure below.



**Best Subset Selection**

The subset selection identifies a subset of the best p predictors that the model believes have a substantial effect on the response variable and then fits the least squares regression for every possible combination of the identified p predictors. In this project, all three subset selection methods were used (i.e. Best subset, forward stepwise and backward stepwise selections), all of which made very impressive predictions with MSEs of 3.12, which was extremely close to the best model(Boosting).

According to the Best Subset Selection, the best model identified contained only two variables that had significant impacts on the response variable, G3. This included the number of absences and second-period grade(G2) for each student. This is captured in the figure below.



**SECTION SIX: CONCLUSION**

In conclusion, the results of analysis produce three very impressive results that answer both research questions with regard to prediction accuracy of the model and estimation of the highly influencing variables. In addition, key variables such as the number of absences in a class and students’ second-period grade (G2) have remained high influencers on the response variable (G3) in the best three statistical methods used. How important can this information be? This information can be immensely valuable to high schools not only to the education sector in Portugal but also in other parts of the world who are suffering from a similar fate. Now, Managers, Principals and Heads of High Schools can use the results of this project to prioritize educational procedures and improve resource allocation by focusing their energies on reducing the total number of absences for each class as it has a sufficient amount of influence on the final grade as well as provide guidance and if possible, provide remedial classes or extra classes to ensure that each student has enough material and knowledge to prepare for the second period examination which has been determined to be a huge influence on the final grade of students.

The results have been really insightful by letting us know which predictors can be disregarded and which predictors must be of enormous priority. It is must also be noted, that one of the best models, the Lasso which produces sparse variables, indicated that other variables had influences on the final grade. Although they were not massive influences, they are still worth mentioning and putting light on; Interesting predictors such as the age of students, the Fathers job ( if Services), reason (if home), guardian (if mother), study time, past failures, extra educational school support, extra-curricular activities (if yes), quality of family relationship, weekend alcohol consumption were all used in the model. Therefore, if funds and resources are available these other predictors could possibly enhance a student’s final grade performance. Nevertheless, the main focus must be on reducing the total number of absences and improving the second-period grade(G2) of the students.

To digress a bit, education is a crucial element in our society. Data Mining (DM) techniques, which allow a high-level extraction of knowledge from raw data, offer interesting possibilities for the education domain. In this paper, the prediction of secondary student grades of one core class (Mathematics) by using past school grades (first and second periods), demographic, social and other school related data have all been addressed. Linear Regression, Shrinkage, Tree-based methods and the Neural Networks (NN) were all tested.

This study was also based on an off-line learning since the DM techniques were applied after the data was collected. However, there is a potential for an automatic online learning environment, by using a student prediction engine as part of a school management support system. This will allow the collection of additional features (e.g. grades from previous school years) and also to obtain a valuable feedback from the school professionals (Cortez & Silva, 2008). Furthermore, it could be enlarged to experiment more schools and school years, in order to enrich the student databases. Automatic feature selection methods (e.g. filtering or wrapper) (Witten and Frank 2005) can also be explored, since only a small portion of the input variables considered seem to be relevant. In particular, this is expected to benefit the nonlinear function methods (e.g. NN), which are more sensitive to irrelevant inputs.

Finally, EDM is in its infancy and it has a lot of potential for education. EDM opens promising and exciting opportunities for future research. Particularly, this project can be further improved by using more recent data as the data used for this report is from the year 2006, and the data should not be limited to just the Mathematics subject even though it was a good benchmark for other subjects. If possible, students’ overall CGPA should be investigated. To continue moving in the right direction, more predictors investigated as well more data from other schools could further improve the accuracy of all aforementioned statistical models.

**References**

Breiman,L. (2001). Random Forests. Machine Learning, 45(1), 5-32. //link.springer.com/article/10.1023/A:1010933404324

Breiman, L., Friedman, J., Ohlsen, R., & Stone, C. (1984). Classification and regression trees. Monterey, CA: Wadsworth.

Cortez, P., & Silva, A. (2008). Proceedings from 5th FUture BUsiness TEChnology Conference: *Using data mining to predict secondary school performance*. Porto, Portugal.

Forelle, C. (2011). A nation of dropout shakes Europe. European Business News. Retrieved from  
http://www.wsj.com/articles/SB10001424052748704076804576180522989644198

Hastie, T., Tibshirani, R & Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. NY: Springer-Verlag.

Jiawei Han Michelin Kamber(2011), “Data Mining-Concepts and Techniques”, Morgan Kaufmann Publishers.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500-508.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students’ performance in distance learning. Applied Artificial Intelligence, 18, 411-426.

DOI: 10.1080=08839510490442058

Luan, J. (2002). Data mining and its applications in higher education. New Directions for Institutional Research, 113, 17-36.

Ma, Y., Liu, B., Wong, C., Yu P., & Lee, S (2000). Proceedings from 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Targeting the right students using data mining. Boston, USA.

Minaei-Bidgoli, B., Kashy, D., Kortemeyer, G., & Punch, W., (2003). Proceedings from IEEE Frontiers in Education: Predicting student performance. Colorado, USA.

Pardos, Z., Heffernan, N., Anderson, B., & Heffernan, C., (2006). Proceedings from of 8th Int. Conf. on Intelligent Tutoring Systems: Using fine-grained skill models to fit student performance with Bayesian networks. Taiwan.

Pritchard, M., & Wilson, S. (2003). Using emotional and social factors to predict student success. Journal of College Student Development, 44(1), 18-28.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. International Journal of Computer Applications, 63(8), 35-39.

Shahiri, M. A., Husain, W., & Rashid, A. N. (2015). A review on predicting student’s performance using data mining techniques. Procedia Computer Science, 72, 414-422.

Witten, I., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques using Java implementations. San Francisco, CA: Morgan Kaufman.

Zahedifard, M., Attarzadeh, I., Pazhokhzadeh, H., & Malekzadeh, J. (2015). Prediction of students’ performance in high school by data mining classification techniques. International Academic Journal of Science and Engineering, 2(7), 25-33.