Predicting Students' Performance in Math

Wei-Chen Wang[*]

May 13, 2020

Abstract

Understanding the factors of students' performance is the key to improve our education system. This paper uses different machine learning methods to train models to predict students' performance in math. Models are evaluated by their mean squared error (MSE), and our results show that a linear regression model yields the lowest MSE. The best model includes four predictors, in which the first and second midterm grades have positive influences of 0.119 and 0.870 respectively, while an additional absence from class and going out with friends have negative influences of 0.011 and 0.078 respectively, on the final grade. All models show significant importance of the two midterm grades in the prediction, while most other predictors have weak correlation with the final grade. This paper demonstrates the difficulty for teachers to design personalized education, as it is nearly impossible to effectively predict students' final performance solely based on their background information. In addition, recommendations are provided on how this topic can be further studied and improved.

[*] University of Illinois at Urbana-Champaign, Email: `wcwang2@illinois.edu`

Contents

1 Introduction

Education inequality has become a serious issue in recent years. Race, ethnicity, wealth, sex and many other factors play significant roles in the inequalities of the education system, in which privileged students tend to perform better academically. Students with less education resources receive less help, and have to allocate more time outside of school, such as having a part-time job. The difficulty of predicting students' performances makes it harder for teachers to identify and assist students who are falling behind.

This project explores potential factors that can impact students' academic performance in math. The dataset is obtained from the machine learning repository at the University of California, Irvine [1]. The data are collected using school reports and questionnaires from two Portuguese secondary schools. Attributes include grades, demographic, social and other school related attributes. In this report, we introduce multiple machine learning models, where most of the descriptions and objective functions of each model are referenced from the class lecture slides (Khazra & Farhoodi, 2019). We evaluate each model by comparing their mean squared errors. In particular, we run the following models:

- Ordinary least squares (OLS) linear regression
- Least absolute shrinkage and selection operator (LASSO) regression
- Ridge regression
- Decision tree
- Bootstrap aggregating (Bagging)
- Random Forest
- Boosting
- Extreme Gradient Boosting (XGboost)
- Neural Networks
- K-nearest neighbors (KNN)

[1] Student Performance Data Set: `https://archive.ics.uci.edu/ml/datasets/Student+Performance`

2 Literature Review

Making accurate predictions of students' final grades is essential for bridging education inequality, as teachers can identify students who are falling behind in advance. In recent years, researchers use machine learning techniques to predict future performances, with logistic regressions, support vector machine and decision tree among the most common machine learning architecture (Dalipi, Imran, Kastrati). Other researchers also propose the idea of using hybrid selection algorithms with multiple machine learning classifiers to perform feature selection (Turabieh, 2019). However, the field faces the difficulties in which students come from diverse backgrounds, and courses are not equally informative for predictions (Xu, Moon & Schaar, 2017). Data collection of this study is also complicated, as some factors which have profound impact on learning cannot be easily recorded and quantified. For example, some believe that the student's progress in solving a single exercise can also provide valuable information, which makes it challenging to capture his knowledge state (Wang, Sy, Liu & Piech, 2017).

Recent studies depict a wide range of potential factors that can impact student outcome, including exploration strategies (Kaser, Hallinen & Schwartz, 2017) and parents' occupation (Ramesh, Parkavi & Ramar, 2013). Their influences are expected to change in the future as an increasing number of classes are moving online. In particular, researchers show that the negative impacts of online courses are more common for students with disadvantaged backgrounds or low level of preparation (Figlio, Rush & Yin, 2013). This paper uses similar predictors and machine learning algorithms to examine how the variables contribute to predicting students' academic performances.

3 Data

The response variable in this study is the student final grade *G3* (numeric: from 4 to 20). Records with final grades of 0 are removed because there may be potential non-academic factor, as there exists a gap between the lowest non-zero final grade and a grade of zero. The remaining final grade records roughly follow a normal distribution, with a mean of 10 (Appendix A). There are multiple potential predictor variables, but only the essential ones were listed below. A full list of predictors can be found at Appendix B.

Variables	**Type**	**Description**	**Possible Values**
activities	binary	extracurricular activities	*yes* or *no*
absences	numeric	number of school absences	from 0 to 93
age		student's age	15 to 22
failures		past class failures	n for $1 < n \leq 3$, 3 for $n \geq 3$
Fedu/Medu		father/mother's education	0 for none 1 for primary education 2 for 5th to 9th grade 3 for secondary education 4 for higher education
G1/G2		first/second period grade	from 0 to 20
goout		going out with friends	from 1 (very low) to 5 (very high)
studytime		weekly study time	1 for less than 2 hour 2 for 2 to 5 hours 3 for 5 to 10 hours 4 for more than 10 hours

Most of the variables have a low level of correlation with the final grade, except for the two midterm grades (Appendix C).

4 Models

4.1 Linear Regression

A linear regression model assumes a linear relationship between the response and the predictor(s). The model seeks to minimize the least squares objective function:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij})^2$$

Intuitively, students' grades in their first and second period grade have significant impact on their final grade. If a student performs well in the beginning and middle of the semester, he is likely to receive a good final grade. First, we run the following model:

$$Y = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \epsilon$$

where Y is the final grade in math, G_1 and G_2 are the first and second period grades, ϵ is the error term, and β_i are the coefficients. Even though the model has a high adjusted R^2 (0.934), the result is expected because the two midterms directly count towards part of the final grade. However, the project is more interested in factors that do not directly derive from midterm grades of the same class. Using a full model with all the predictors without the two midterm grades, we fit the model and used stepwise selection to obtain another linear regression model:

$$Y = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 FS + \beta_4 ME + \beta_5 T + \beta_6 F + \beta_7 H + \beta_8 G + \epsilon$$

where S is the dummy variable for sex (0 for female, 1 for male), A is age, FS is the dummy variable for family size (0 for family size more than 3, 1 otherwise), ME is mother's education, T is the travel time from home to school, F is the number of past failed courses, H is the dummy variable for wanting to take higher education (0 for no, 1 for yes), G is the frequency of going out with friends. Compared to the first model, even though the model includes more predictors and is more complicated, the adjusted R^2 is significantly lower (0.277). We therefore believe that it is impossible to make accurate predictions without previous midterm grades, both of which are included in all future models.

We build another model which initially includes all predictors, and use the stepwise selection to filter valuable predictors to keep. We get the best linear model so far as follows:

$$Y = \beta_0 + \beta_1 A + \beta_2 G + \beta_3 G1 + \beta_4 G2$$

where A is the number of absences, G is the frequency of going out with friends, and $G1, G2$ are the two midterm grades. We get the coefficients

$$\beta_0 = 0.613, \beta_1 = -0.011, \beta_2 = -0.078, \beta_3 = 0.119, \beta_4 = 0.87$$

This indicates that for every additional absence, the student is expected to perform 0.011 points worse. For every additional unit of going out with friends, the students is expected to perform 0.078 points worse. For every additional point scored in the first and second midterm, the student is expected to perform 0.119 and 0.87 points better, respectively. These results should not be surprising because absences and going out with friends are expected to have negative impacts on the final grade, while getting a better grade in midterms are likely going to increase the final grade. The best linear model has a MSE of 0.663.

4.2 Lasso

Compared to linear models, LASSO models penalize coefficients of predictors in norm forms, as it seeks to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ controls the shrinkage penalty. The model performs feature selection, in which irrelevant predictors' coefficients are shrunk to zero as λ increases. Notice that when $\lambda = 0$, the model reverts to a linear OLS regression.

Data are first partitioned into a training (80%) and testing (20%) dataset, which are used to evaluate this and all other models which require cross-validation. The training uses all predictors, and the best λ is 0.069. At the chosen λ, most coefficients have shrunk to zero, except for the followings:

Variables	Coefficients
Mjobother	-0.039
famrel	0.074
goout	-0.028
health	-0.011
absences	-0.003
G1	0.099
G2	0.872

Compared to the linear regression model, the LASSO includes more predictors, even though they have very minimum coefficients. Notice that the coefficients for all predictors except for that of $G2$ shrunk significantly, which features the feature selection functionality of the model. None of the coefficient signs should be surprising, except for *health*, which has a negative coefficient. Nevertheless, the predictor does not have huge influence on the final grade, as the scale for the variable only ranges from 1 to 5. The LASSO model has a MSE of 0.735.

4.3 Ridge

A Ridge model is similar to a Lasso model such that both models penalize the magnitude of coefficients. The objective of the model is to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

Similar to LASSO, λ controls the shrinkage penalty, and the model reverts to a linear OLS model when $\lambda = 0$. The biggest difference is how the coefficients are penalized, in which the model does not perform feature selection. As λ increases, the coefficients approach but will not equal to zero.

Using a similar algorithm, we are able to obtain a Ridge model with the best λ of 0.313. At the best λ, none of the coefficients have shrunk to zero. Here, we will only list the coefficients which are greater than 0.2 in magnitude:

Variables	Coefficients
PstatusT	-0.203
Mjobhealth	0.252
guardianother	-0.216
schoolsupyes	-0.270
G1	0.275
G2	0.634

Compared to the previous two models, most coefficients shrunk, while a small number of coefficients increases in magnitude. Looking at the coefficients, it may be surprising that having additional school supply has negative impact on final grade, even though the impact is minimum. The Ridge model has a MSE of 0.753.

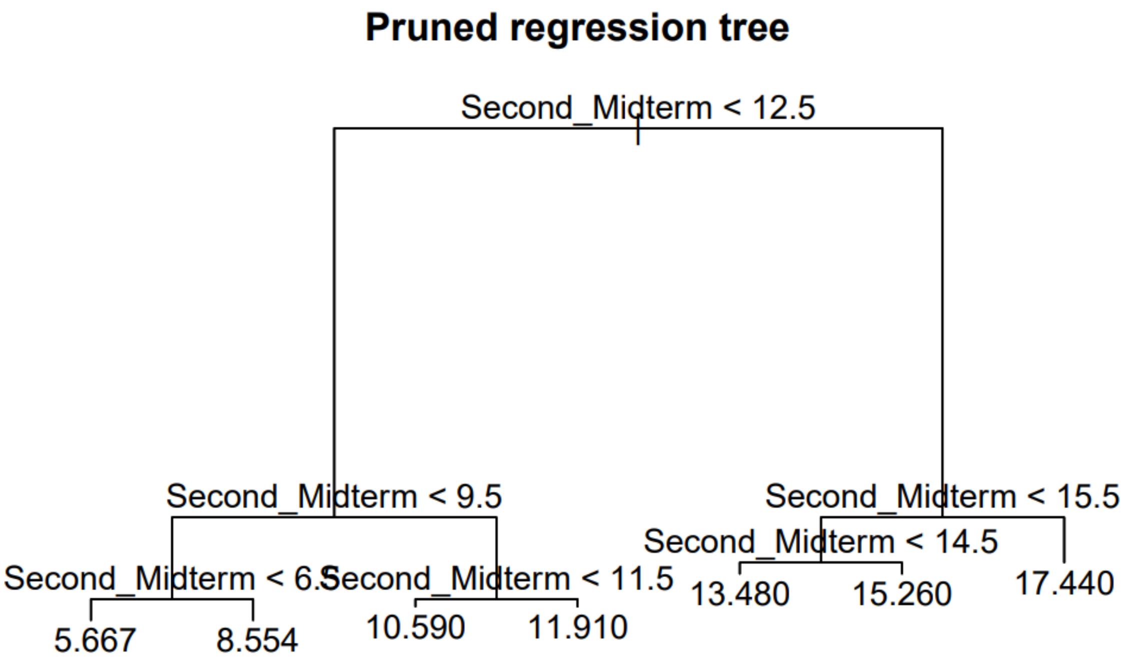
4.4 Decision Tree

A decision tree is one of the easiest models to interpret, as it is similar to how humans make decisions. They can be displayed graphically, and can handle qualitative variables without creating dummy variables. Starting from the root of the tree, the tree finds a best split that minimizes the RSS. The cutoff is selected to minimize the sum of RSS of the two regions. For instance, suppose the partition divides the dataset into two boxes $R1$ and $R2$, in which $R_1 = X \mid X_j < s$ and $R_2 = X \mid X_j \geq s$, then s is selected to minimize

$$\sum_{i:x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2} (y_i - \hat{y}_{R_2})^2$$

Each of sub-tree will find the best split recursively, until it reaches the stopping criterion, such as each region has no more than 5 observations, or the maximum tree depth is reached.

We first fit the training dataset to a big tree, and prune the tree using cross-validation. The tree size vs. deviance plot shows that the deviance is minimized at 7 nodes. Therefore, we prune the tree and obtain the following decision tree:



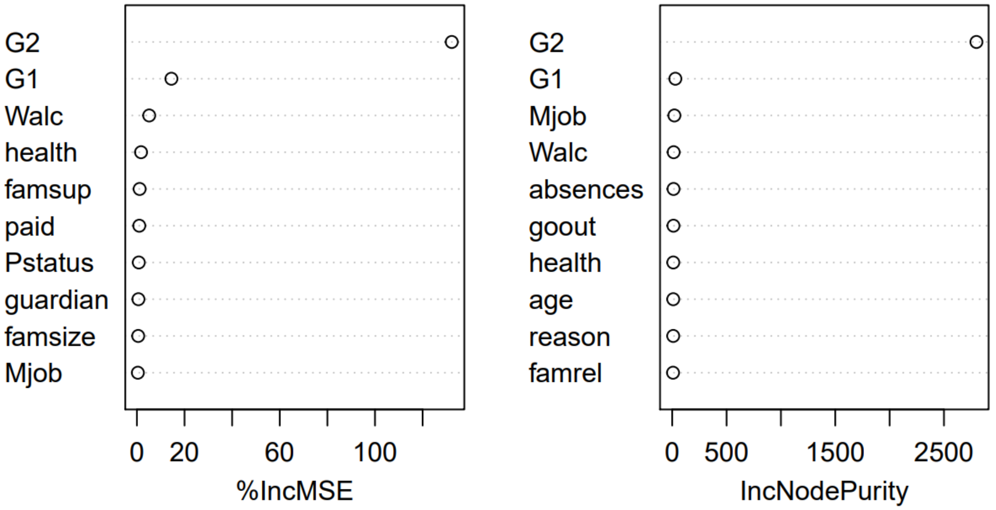
Interestingly, all splits are based on G2, which indicates that G2 has a profound impact on the final grade, which is also an expected result. The decision tree model has a MSE of 1.102.

4.5 Bagging

Bootstrap aggregating (Bagging) models take advantage of bootstrapping to make the model robust. Compared to decision trees, Bagging models are not sensitive to how the training dataset is partitioned, and is the average of multiple trees generated from bootstrap samples. The function is as follows:

$$\hat{f}_{bag} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(X)$$

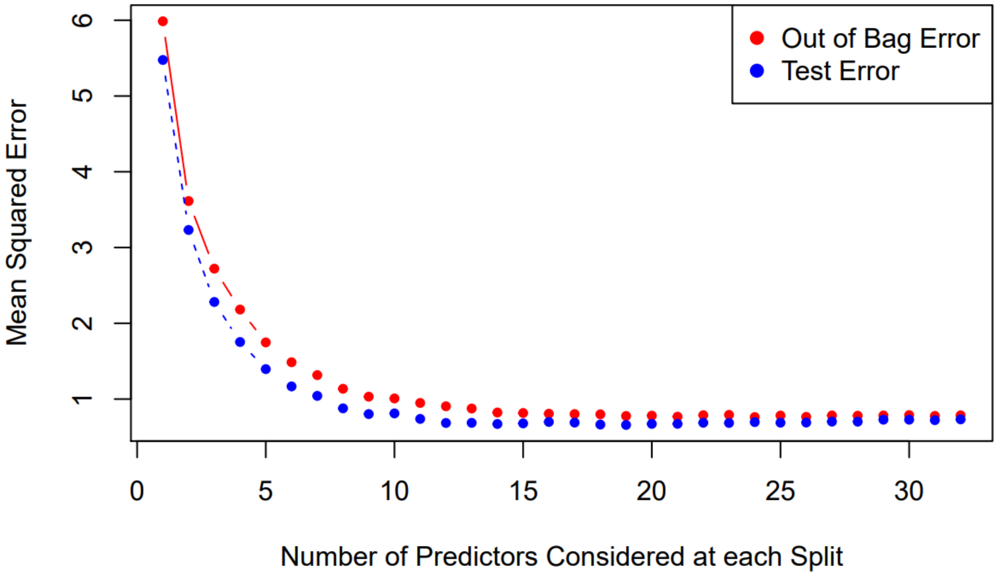
A predictor is considered more important if its split results in a greater overall reduction in RSS on average. Using the same training dataset, we get a model with 400 trees, which shows the relative importance of the top variables as follows. We can see that the second midterm grade still has a dominate effect on the final grade, while other predictors are significantly less important. The Bagging model has a MSE of 0.727.



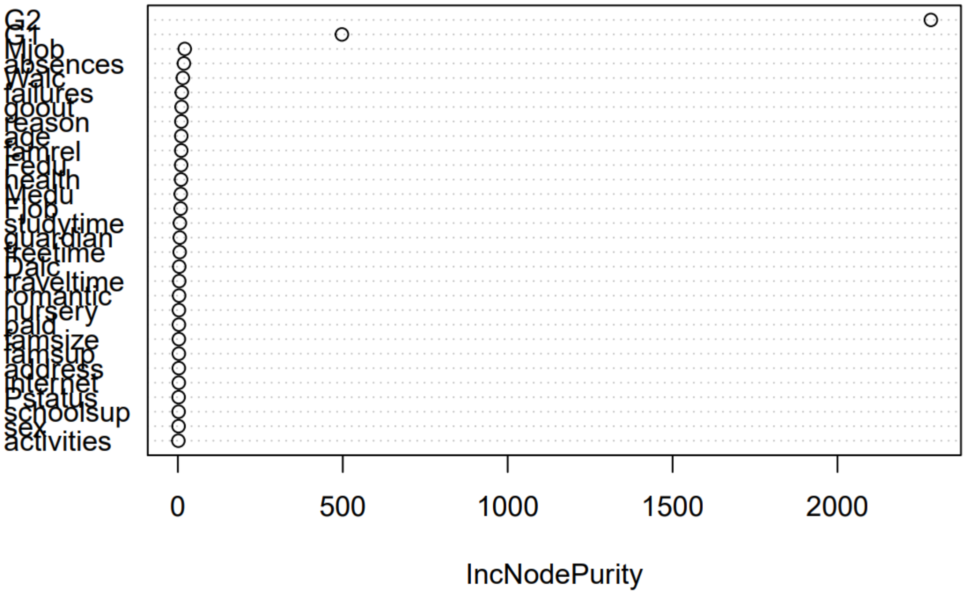
4.6 Random Forest

Random forest models are similar to Bagging but with an important difference: random forests de-correlate trees or force each split to use a subset of predictors. This solves the problem of bagging, in which if there exists a strong predictor, each tree considers it on top of its splits, and all generated tree will look similar to each other. This is exactly the case we are seeing right now in the bagging model, in which the second midterm grade is a strong predictor along with other moderate predictors.

We train models with numbers of predictors considered at each split ranging from 1 to 30. We then compare their out-of-bag error, and choose the parameter value which minimizes the MSE. The plot is as follows:



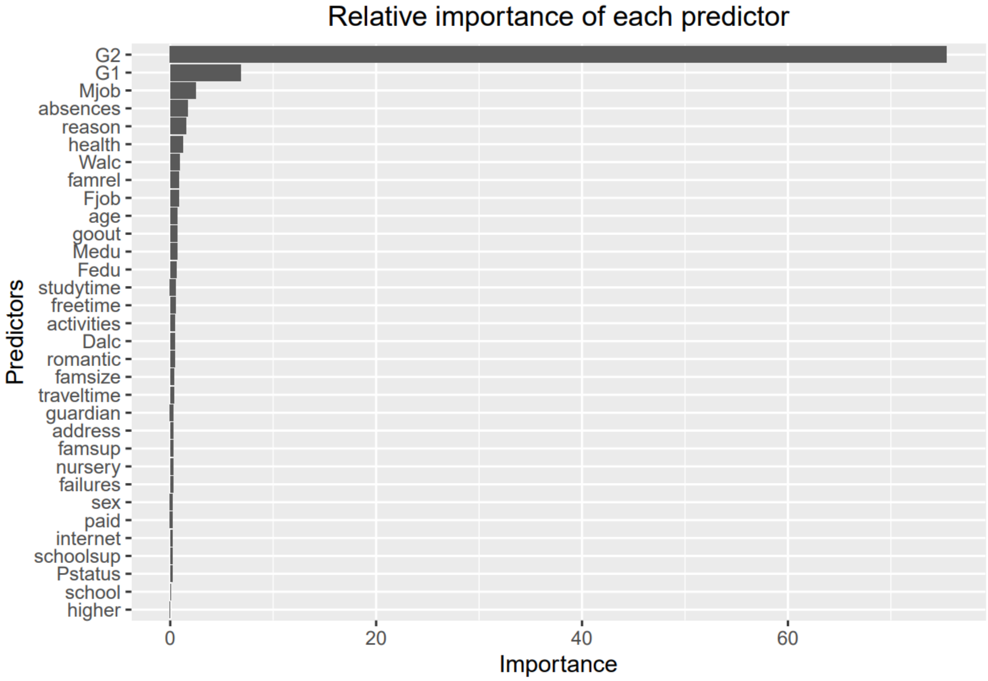
The test error and out-of-bag error decreases at first when the number of predictors considered at each split decrease at first, and remains stable when predictors exceed 15. Notice that they follow a similar trend, while the out-of-bag error is always slightly higher than the test error. The MSE for out-of-bag error is minimized when the number of predictors is 24, which will be used to train the model. We are able to get the following graph of importance: Even though $G2$ is still the most important variable in the model, we can see other variables, such as $G1$, have an increasing importance compared to the bagging model. The random forest model has a MSE of 0.684.



4.7 Boosting

Boosting models are common in tree based statistical learning. Unlike Bagging and random forest algorithms, Boosting trees are grown sequentially instead of separately, and does not perform bootstrap sampling, so the model can overfit with large number of trees. The algorithm enables the model to learn slowly, and in each period the estimation is slightly improved. Therefore, the construction of the tree is strongly dependent on the previous tree.

We first use the same parameters to obtain a relative importance of predictors, then tune the parameters using a grid search to minimize the MSE. The relative importance of each predictor is as follows:



As shown above, $G2$ has a high relative importance, and the model has a MSE of 0.787. We then tune the shrinkage parameter, total number of trees to fit, and the interaction depth. The grid is as follows:

$$\text{shrinkage parameter } (\lambda) \in \{0.001, 0.002, ..., 0.010\}$$

$$\text{total numbers of trees to fit } \in \{1000, 2000, 3000, 4000, 5000\}$$

$$\text{interaction depth } \in \{1, 2, 3, 4, 5\}$$

After the grid search, the model achieved the lowest MSE of 0.695 when λ is 0.005, number of trees is 1000, interaction depth is 5. This is an substantial reduce of MSE from 0.787.

4.8 XGBoost

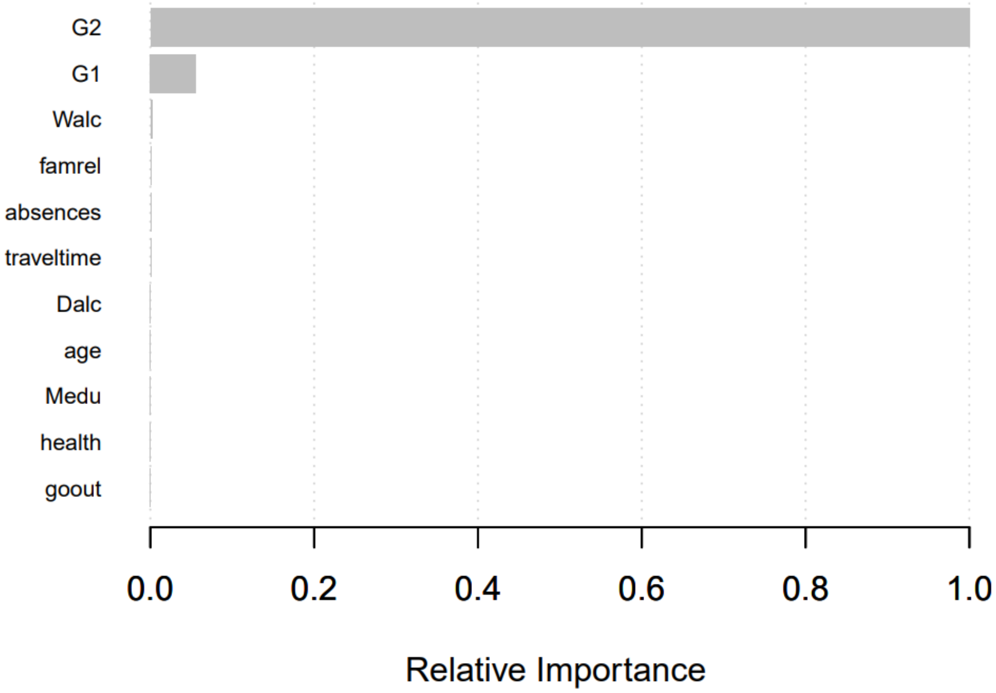
Extreme gradient Boosting (XGBoost) is an enhancement of the traditional boosting models, which has the following objective function:

$$\sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_t)) + \Omega(f_t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_t)) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$$

with the following parameters:

- number of trees
- η: same as learning rate in the boosting model, defaulted at 0.3
- max depth: similar to number of splits d in the boosting model, defaulted at 6
- γ: T is the number of leaves, defaulted at 0
- λ: $L2$ regularization term on weights, defaulted at 1. ω is the vector of scores on leaves

We train the model and obtain the following graph of variable importance:



As shown above, the two midterm grades have very dominant effects on the final grade, while other predictors have very little relative importance. The XGBoost model has a MSE of 0.742.

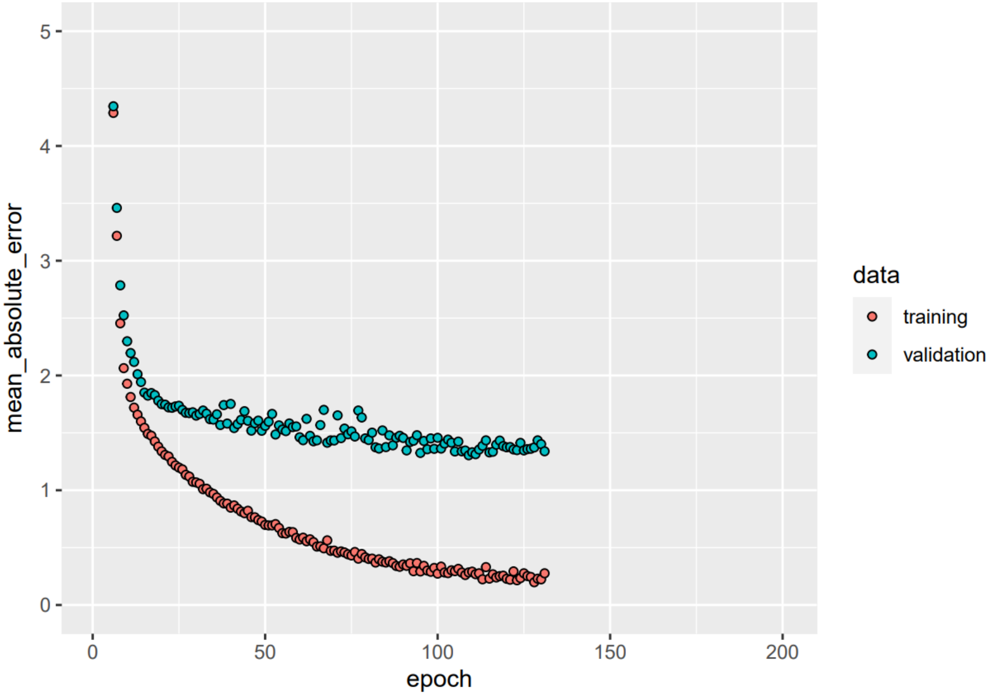
4.9 Neural Network

Neural network enables data to go through a network layer by layer until an output is produced. Each neutron produces an output or activation, and we fit the model by finding the right weights to connect the neurons. Each layer can be a linear or non-linear functions.

Common layers include:

- softmax: $g_k(\beta_0 k + \sum_{m=1}^{M} \beta_{mk} Z_m) = \beta_{0k} + \sum_{m=1}^{M} \beta_{mk} Z_m$ (regression)
- sigmoid: $\sigma(v) = \frac{1}{1+e^{-v}}$
- rectified linear units (ReLU): $\sigma(v) = max(0, v)$

We train the model using two rectified linear units (ReLU) functions, and invoke an early stop callback if we set the epoch too high. We obtain the epoch versus mean absolute error graph for training and validation datasets. As shown below, mean absolute error stabilizes after epochs reach 100, and the early stop callback is invoked. Even though the mean absolute error decreases as epoch increases, the error is still significantly higher than all previous models. The MSE for this model is 2.817.

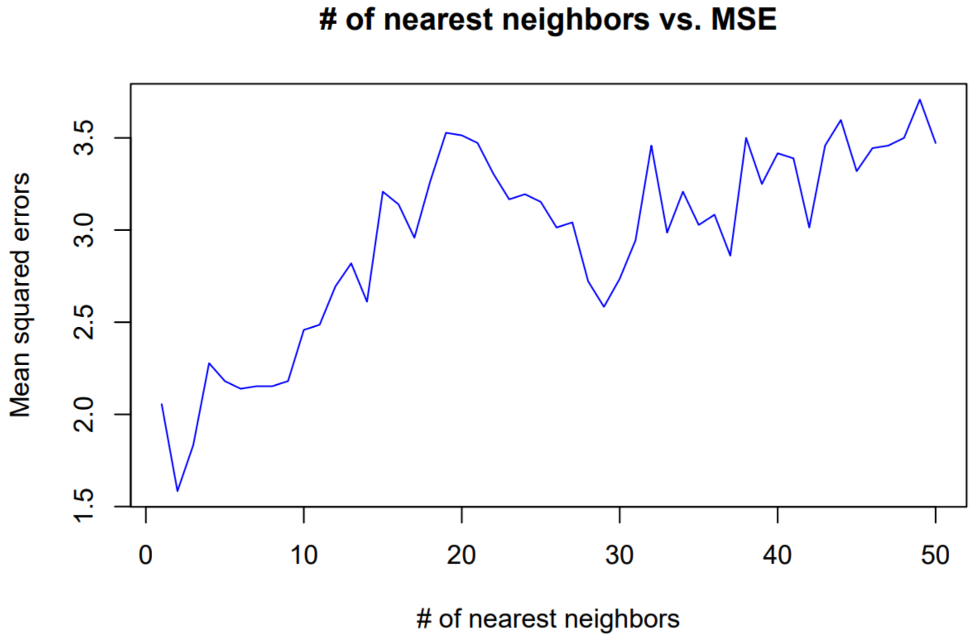


4.10 K-Nearest Neighbors

K-Nearest neighbors (KNN) algorithm finds K neighbors of each observation, and overage the outcome in each group as follows:

$$\hat{y}(x_0) = \hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

There is not a strict rule of how to find the best K, but if K is too small, the model is not smooth and fits to the noise of the data, which results in overfitting. If K is too large, the model is too smooth, and results in underfitting. Therefore, we will use a grid search which ranges from 1 to 50 to obtain the best K. The following graph shows the number of nearest neighbor versus MSE:



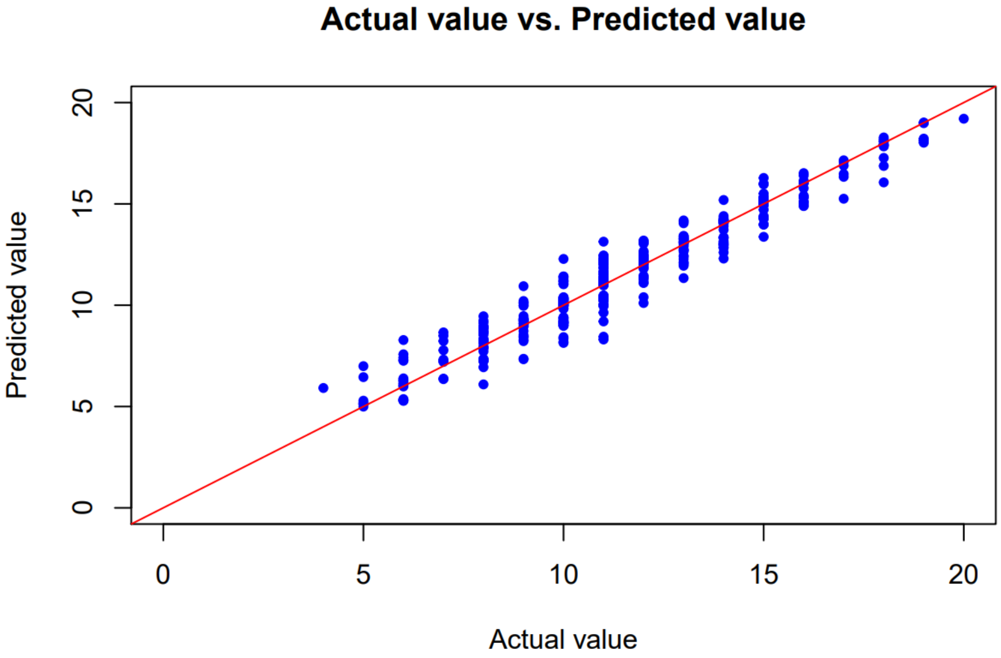
As shown above, MSE is minimized at 1.583 when $k = 2$.

4.11 Comparing all Models

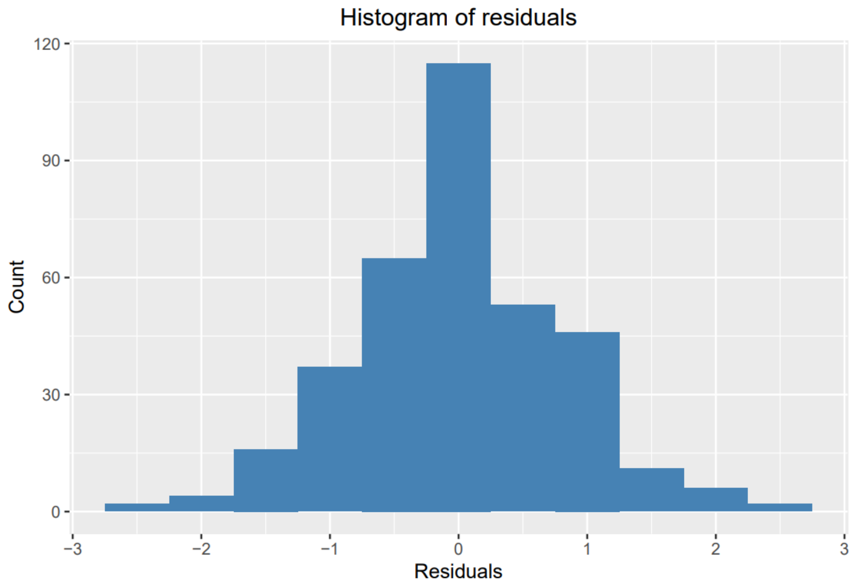
In this report we will compare each model by their mean squared errors (MSE) to assess their accuracy rates. Here, we list all the models, ranked by their MSE from low to high. Note that all models (if needed) are fit on the same training dataset, and tested on the same validation dataset to ensure that each MSE would not be biased against the partition of training and validation datasets.

Model	MSE
Linear regression	0.663
Random Forest	0.684
Boosting (tuned)	0.695
Bagging	0.727
LASSO	0.735
XGBoost	0.742
Ridge	0.753
Boosting	0.787
Tree	1.102
KNN	1.583
Neural network	2.817

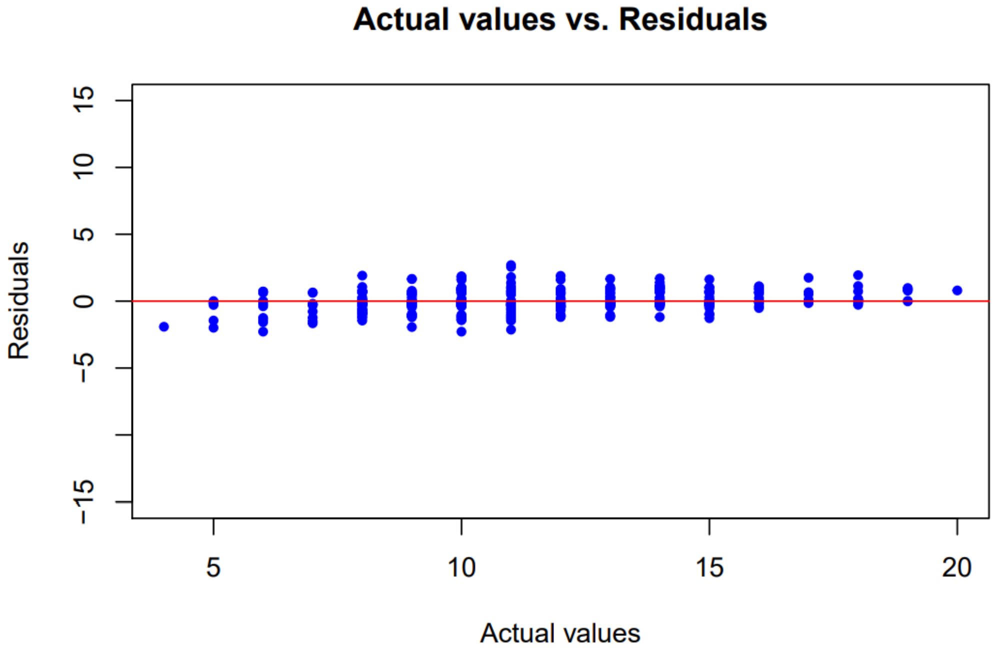
It may be surprising that the linear regression model, which is one of the simplest and easiest models to interpret, turns out to be the model with the least MSE. As a result, we plot the predicted values versus the actual values to see how the best model we have performs.



As shown above, the model performs well in predicting the values. In addition, we examine the distribution of residuals and plot them against the actual value to see if the model violates any linear regression assumptions.



As shown above, the model satisfies the normality assumption because its histogram roughly follows a normal distribution.



As shown above, there does not seem to be heteroscedasticity as the residuals mostly remain constant with varying actual values. We do see that the model slightly underestimates the value when the actual values are low, and slightly overestimates the value when the actual values are high. Overall, we can conclude that the linear regression is the best model, and performs fairly well in predictions.

5 Conclusion

This paper faces a tradeoff of whether the two midterm grades should be used as predictors. If they are not used, the model becomes more meaningful because teachers can predict students' performance on the final grade even before the semester starts. Students with educational disadvantages can be identified in advance for teachers to make personalized adjustments, and help them succeed in the class. However, the project quickly identifies that it is infeasible to make a relatively accurate prediction of final grade without taking into account of midterm grades. Adding these two predictors limits teachers' ability to project students' performance early, and their decisions have to rely on past midterm grades.

This is perhaps an unsurprising result, because student performances are volatile, and do not heavily depend on students' background and past experience at a middle school level. It is not rare to see a middle school student perform well in one year, and bad in another. Data are one of the most challenging part of the research, as they are mostly collected from questionnaires answered by middle school students. In addition, many questions are subjective, as students can have wide range definitions of "going out a lot with friends". The paper still provides meaningful feedback for teachers, parents and students because they can use a simple linear regression model to estimate their final grades. In particular, for students who are aiming for a particular letter grade, or those who simply want to pass the course, the model gives them a sense of whether their goals are achievable.

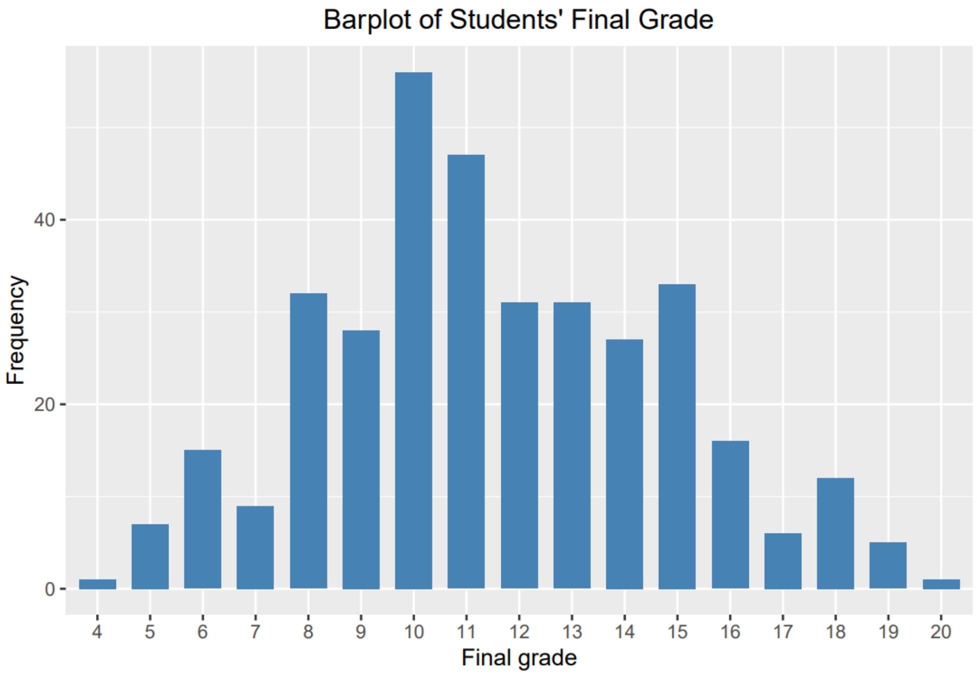
To summarize the best model, the first two midterm grades unsurprisingly have dominant impacts as they directly contribute to the final grade. Being absent from classes and spending more time hanging out with friends also tend to decrease the final grade. Even though the linear regression yields the lowest MSE, other models have their own advantages, such as being easier to interpret, or performing feature selections. Furthermore, all other models depict that the first two midterm grades have tremendous impact on the final grade. The paper also shows that MSE can be significantly reduced by tuning the parameters.

References

1. Dalipi, F. Imran, A. Kastrati, Z. (2018). MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges. *IEEE Global Engineering Education Conference.* Retrieved from `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8363340`. Accessed May 11, 2020.
2. Kharza, N. Farhoodi, A. (2019). ECON 490: Applied Machine Learning in Economics. [Powerpoint slides]. *University of Illinois at Urbana-Champaign.* Retrieved from `https://compass2g.illinois.edu`. Accessed May 14, 2020.
3. Ramesh, V. Parkavi, P. Ramar, K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *International Journal of Computer Applications. Vol. 63, No. 8.* Retrieved from `https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/4170`. Accessed May 11, 2020.
4. University of California Irvine Machine Learning Repository. (2014). Student Performance Data Set. *Center for Machine Learning and Intelligent Systems.* Retrieved from `https://archive.ics.uci.edu/ml/datasets/Student+Performance`. Accessed May 11, 2020.
5. Turabieh, H. (2019). Hybrid Machine Learning Classifiers to Predict Student Performance. *Second International Conference on new Trends in Computing Sciences (ICTCS).* Retrieved from `https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8923093`. Accessed May 11, 2020.
6. Wang, L. Sy, A. Liu, L. Piech, C. (2017). Deep Knowledge Tracing On Programming Exercises. *L@S.* Retrieved from `https://dl.acm.org/doi/pdf/10.1145/3051457.3053985`. Accessed May 11, 2020.
7. Xu, J. Moon, K, Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal of Selected Topics in Signal Processing, Vol. 11, No. 5.* Retrieved from `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7894238`. Accessed May 11, 2020.
8. Figlio, D., Rush, M. Yin, Lu. (2013). Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. *The University of Chicago Press Journals, Vol. 31, No. 4.* Retrieved from `https://www.journals.uchicago.edu/doi/pdfplus/10.1086/669930`. Accessed May 11, 2020.

Appendix

A Distribution of students' final grade



B Full list of Predictors

Variables	Type	Description	Possible Values
activities	binary	extracurricular activities	*yes* or *no*
absences	numeric	number of school absences	from 0 to 93
age	numeric	student's age	15 to 22
address	binary	student's home address type	U (urban), R (rural)
Dalc	numeric	workday alcohol consumption	from 1 (very low) to 5 (very high)
failures	numeric	number of past class failures	n for $1 < n \leq 3$, 3 for $n > 3$
famsize	binary	family size	$LE3$ (less than or equal to 3) $GT3$ (for greater than 3)
famsup	binary	family educational support	*yes* or *no*

Variables	Type	Description	Possible Values
Fedu / Medu	numeric	father's / mother's education	0 (none) 1 (primary education) 2 (5th to 9th grade) 3 (secondary education) 4 (higher education)
fmrel	numeric	quality of family relationships	from 1 (very bad) to 5 (excellent)
Fjob / Mjob	nominal	father's / mother's job	teacher, health, services at_home, others
freetime	numeric	free time after school	from 1 (very low) to 5 (very high)
G1 / G2	numeric	first / second period grade	from 0 to 20
guardian	nominal	student's guardian	mother, father or other
goout	numeric	going out with friends	from 1 (very low) to 5 (very high)
health	numeric	current health status	from 1 (very bad) to 5 (very good)
higher	binary	wants to take higher educ.	*yes* or *no*
internet	binary	internet access at home	*yes* or *no*
nursery	binary	attended nursery school	*yes* or *no*
paid	binary	extra paid classes within the course subject	*yes* or *no*
Pstatus	binary	parents' cohabitation status	T (living together), A (apart)
reason	nominal	trason to choose this school	home, reputation, course, other
romantic	binary	in a romantic relationship	*yes* or *no*
school	binary	student's school	GP (Gabriel Pereira) MS (Mousinho da Silveria)
schoolsup	binary	extra educational support	*yes* or *no*
sex	binary	students' sex	F (female), M (male)
studytime	numeric	weekly study time	1 (less than 2 hours) 2 (2 to 5 hours) 3 (5 to 10 hours) 4 (more than 10 hours)
traveltime	numeric	home to school travel time	1 (less than 15 min) 2 (15 to 30 min) 3 (30 min to 1 hour) 4 (more than 1 hour)
Walc	numeric	weekend alcohol consumption	from 1 (very low) to 5 (very high)

C Correlation between major predictors and final grade

Variables	Correlation
activities	0.059
age	-0.140
failures	-0.294
Fedu	-0.159
Medu	0.190
G1	0.892
G2	0.966
goout	-0.177
studytime	0.127