Regression on Student Performance Dataset

Ioustina Harasim
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
iharas01@ucy.ac.cy

Leonidas Ioannou
Department of Mathematics and Statistics
University of Cyprus
Limassol, Cyprus
lioann09@ucy.ac.cy

Fedor Turchenko
Department of Business
University of Cyprus
Saint Petersburg, Russia
fturch01@ucy.ac.cy

I. Introduction

A. Problem statement

The Student Performance dataset contains information on student achievement in secondary education in Portugal. It includes various attributes such as demographics, family background, academic and behavioral factors, as well as the student's final grade in the course.

Student's academic performance is determined by all these above mentioned factors as they all influence ones life. The school's directors of studies cannot make accurate guess of how well student is going to pass the exams due to variety of these factors and their combinations. In addition, people are commonly influenced by unconscious bias, making such decisions based on prior experience, thinking patterns, self-determined assumptions, interpretations. It makes directors' guesses biased and in inaccurate.

However, one may rise the following issue: why is it actually important to predict final exam score of a student? The better average score of students is, the more prestigious university they can apply to after graduation. This will promote the school, raise governmental investment to this exact educational unit, which will help raising salaries, improving infrastructure, etc. By being able to accurately forecast the score, directors of studies may help certain number of students by, for example, offering them a session with school's psychologist in case their poor academic performance is driven by certain personal matters.

B. Goal

The goal of this research is to develop a predictive model that can forecast a student's final grade in the course based on their demographics, family background, academic and behavioral factors. By understanding which factors contribute to student success or failure, educators and policymakers can develop targeted interventions to improve student outcomes and promote academic success.

II. Methodology

A. Features and short description

The dataset contains 382 rows of observations and 33 columns of variables related to different students. The dataset contains information on the academic performance of students, including socio-economic and demographic attributes such as family size, parents' education, and student's health status.

- school: student's school
- sex: student's sex
- age: student's age
- address: student's home address type
- famsize: family size
- Pstatus: parent's cohabitation status
- Medu: mother's education
- Fedu: father's education
- Mjob: mother's job
- Fjob: father's job
- reason: reason to choose this school
- guardian: student's guardian
- traveltime: home to school travel time
- studytime: weekly study time
- failures: number of past class failures
- schoolsup: extra educational support
- famsup: family educational support
- paid: extra paid classes within the course subject
- activities: extra-curricular activities
- nursery: attended nursery school
- higher: wants to take higher education
- internet: Internet access at home
- romantic: with a romantic relationship
- famrel: quality of family relationships
- freetime: free time after school
- goout: going out with friends
- Dalc: workday alcohol consumption
- Walc: weekend alcohol
- G1: first period grade
- G2: second period grade

Target variable

- G3: final period grade

B. Data source and brief description

The dataset[1] used for this project is provided by the UCI Machine Learning Repository and contains data on the academic performance of students in two Portuguese schools. The primary objective of this regression problem is to predict student performance, which is indicated by the target variable "G3". The target variable represents the final grade received by students in each course, and is a continuous numeric value.

C. Handling Missing Data and Outliers

The dataset does not contain any missing values, but we did identify a percentage of 3.14% duplicates. After applying the interquartile range (IQR) method to detect outliers, we identified several features with outlier values. The feature "failures" had a particularly high outlier value of 17.28%. Other features such as "age", "absences", and "G2" also had some outliers. It is important to note that we did not remove the outliers, as these observations may contain valuable information and represent extreme values and by removing them may lead to a loss of information.

Even though there were no missing values detected, target variable, final score, contained 39 zero values. Since it accounts for approximately 10% of observations, simply dropping these rows was not the option. These values could be scores of those students who cheated during exam. Sometimes it is quite subjective whether a student cheated or not. One can argue if the decision to force student out of exam was fair or not. Thus, the imputing strategy was adopted in order to replace these zeros. Specifically, these were substituted with medians of corresponding age and gender. For instance, if the student with zero score was a 16 years old male, then his grade was replaced with median score of boys at this age.

D. Preliminary Feature Selection

The variety of variables available in the dataset aroused a potential collinearity issue. Certain factors potentially could be correlated, bringing the same information to the model. This is a violation of the predictors independence assumption for both Linear and Poisson regression models, which were used in this study. This issue was checked and tackled during the exploratory data analysis.

At first, the correlation matrix was created to explore the relationship between the features. It is important to note that most of variables are categorical, which makes building Pearson correlation matrix not fully correct and accurate from the standpoint of methodology. However, it still provides the useful overview of the relationship between variables and helps identifying potential collinearity prior to fitting models.

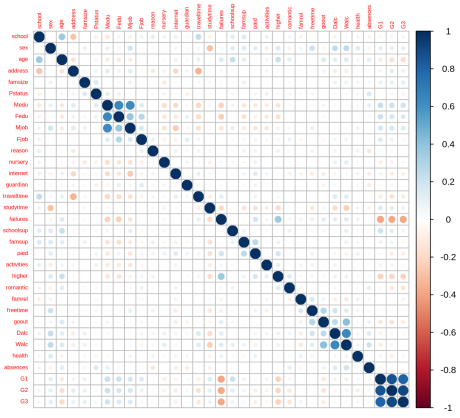

Fig. 1. Correlation Matrix

The obvious outcome, which is easily observable at the first glance is the extremely high correlation between G1, G2 and G3, i.e., grades for the previous periods and the last period. One may, undoubtedly, oversimplify the task and make use of this observation, stating that the final grade is obviously determined mostly by previous grades, therefore it could be possible that, since the correlation is positive, to assume that for every next trial the student is going to achieve at least a little bit higher score than for the previous one even without running any regression models.

This assumption, though, has two main downsides. Firstly, it is the generalization of the picture for all of the students, meaning that if one relies on this statement, then they would expect every single student to improve scores, while it obviously might not be the case. Secondly, it is based on the fact that the exams would be taken at least once before the final period. The school's policy might shift in this sense or the final period exams may start differing much from the previous ones, so that the relationship could change.

Taking this into consideration, it was decided not to keep the G1 and G2 in the model in order to make it more robust and free from the obvious collinearity issue with these variables.

In addition, there was quite a considerable correlation between student's either rural or urban address and home to school travel time. Both variables represent quite similar concept of time that it takes to go to school, since usually it takes longer to get to school from rural rather urban area. The null hypothesis was formulated as follows: there is no dependency between type of area (rural/urban) and time it takes to go to school. Fisher's exact test was performed in order to test this hypothesis.

TABLE I
CONTINGENCY TABLE FOR FISHER'S EXACT TEST OF TRAVEL TIME BY ADDRESS TYPE

	Rural Area	Urban Area
$<$ 15 minutes	31	219
15-30 minutes	32	71
30-60 minutes	13	8
$>$ 60 minutes	5	3

Since the p-value is lower than 0.01 ($p = 1.789\text{e}{-9}$), there is enough evidence to reject the null hypothesis at 99% confidence level. Therefore, there is a relationship between address and travel time, and thus these variables might be collinear. Only travel time variable was kept in the data, since it captures the concept of time it takes to get to school more accurately.

Next pair of variables to test for potential dependence was the number of past class failures and willingness to take higher education. One obvious assumption here might be that those students who want to take higher education should have either none or at maximum one class failure. As previously, the same approach with Fisher's exact test was followed, which started by building a contingency table.

The vast majority of students (more than 80%) want to take higher education and never failed a class, so, in fact,

TABLE II
CONTINGENCY TABLE FOR FISHER'S EXACT TEST OF NUMBER OF FAILS BY HIGHER EDUCATION DESIRE

	No	Yes
0 failures	61	310
1 failure	4	34
2 failures	1	10
3 failures	7	10

running a test was not a necessity here. It was decided that most probably, both of these variables will not provide much information, and so both could be dropped.

Another two variables between which there was quite a considerable correlation captured were level of frequency of going out with friends and level of alcohol consumption. Originally, alcohol consumption was divided into a weekday and weekend variables but they were combined into a single one, since it is in general alcohol consumption. Combination of initial variables was performed by summing up values of 1-5 scales which resulted in 1-10 scale, which shows the aggregated figure of how much alcohol student consumes both on weekdays and weekend. Considering its relationship with going out, one may assume that there might be a dependency in a sense that the more frequently students go out, the more alcohol they consume. This statement could be supported by such argumentation that students how often go out are worse controlled by their parents in terms of how they spend their free time, and so these students may start consuming alcohol. The following contingency table was obtained for these variables.

TABLE III
CONTINGENCY TABLE FOR FISHER'S EXACT TEST OF ALCOHOL CONSUMPTION LEVEL (1-10) BY FREQUENCY OF GOING OUT WITH FRIENDS

	Very Rarely	Rarely	On Average	Frequently	Very Frequently
2	16	49	45	25	9
3	3	22	29	8	6
4	2	13	26	10	7
5	2	7	13	16	4
6	0	4	3	14	11
7	1	2	4	3	7
8	0	0	1	5	3
9	0	1	0	0	1
10	0	1	2	0	6

Starting from the level 5 of alcohol consumption, it is quite evident from the contingency table that at such levels of drinking alcohol there are commonly more observations among students who go out with friends on average level, frequently, very frequently. Nevertheless, the Fisher's exact test was performed here as well. Null hypothesis was stated as: there is no dependency between level of alcohol consumption and frequency of going out with friends. Since the p-value obtained was lower than 0.01 ($p = 0.0004998$), there was enough evidence to reject the null hypothesis at 99% confidence level. Therefore, there is a relationship between these two variables and they might be collinear. The statement of the more frequently students hang out with friends, the more alcohol they consume, thereby could be proved. Since going out is a broader concept, which also involves other activities than drinking alcohol, it was decided to drop alcohol consumption variable to reduce potential collinearity.

E. Feature Engineering

The data contains years of father's and mother's education. The distinction between the effect of each parent's years of education was examined through the following plots.

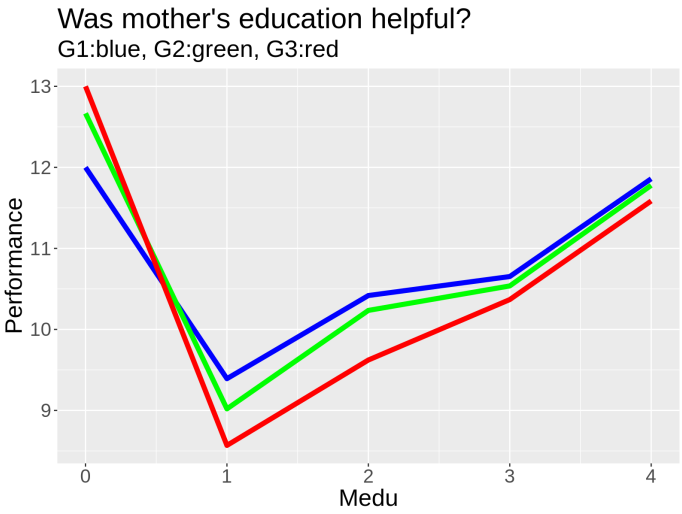

Fig. 2. Effect of Years of Mother's Education on Average Exam Scores

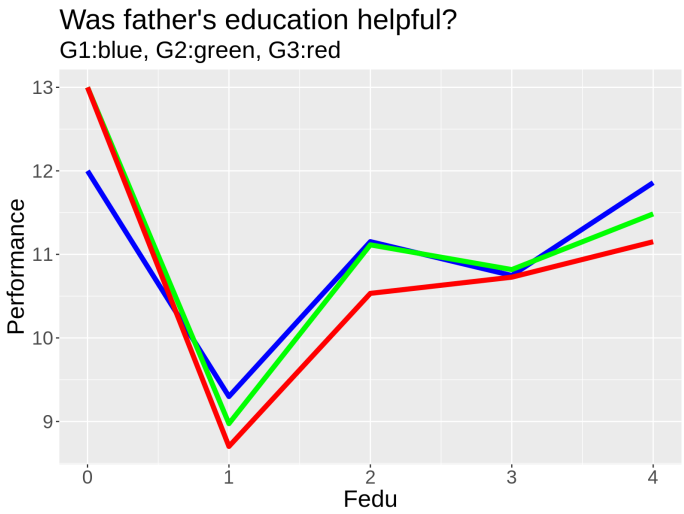

Fig. 3. Effect of Years of Father's Education on Average Exam Scores

The relationship patterns between years of education and average score are highly similar for both father's and mother's years of education. Therefore, in order to avoid collinearity, these variables were combined into a single feature, which represented the aggregate of parents' education.

In addition to parents' education, dataset contained their occupation as well, classified as "teacher", "healthcare", "services" (administrative or police), "at home", and "other". The same kind of plots were considered as for education.

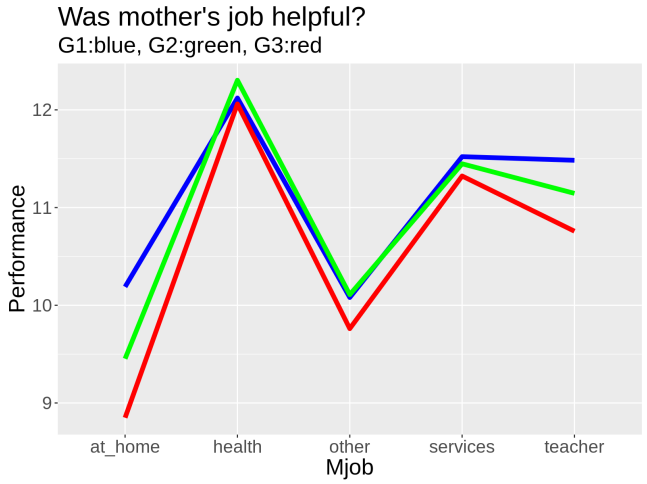

Fig. 4. Effect of Mother's Job on Average Exam Scores

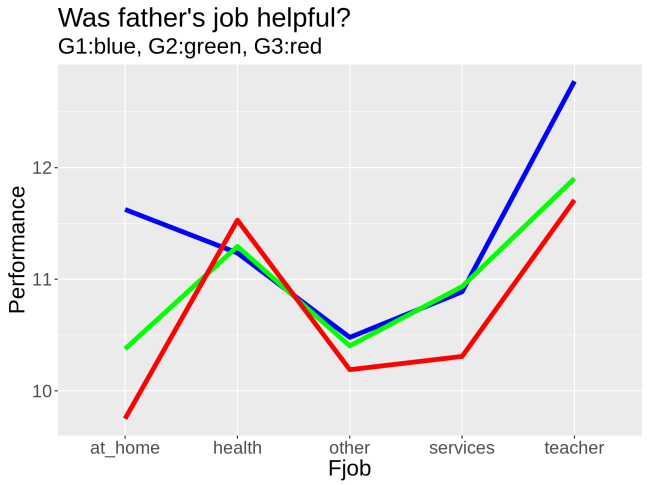

Fig. 5. Effect of Father's Job on Average Exam Scores

The highest average scores were achieved for students whose mother is either a healthcare worker or father was either a teacher. It is important to note the average grade of students, whose mother is a teacher, is close to the second best mother's occupation – public services. In terms of class frequencies, mother's and father's jobs are in majority represented by 'other' class, which does not provide much evidence in terms of explaining which exact parents' job could force students to study harder. However, one may assume that if one of the parents works in a healthcare, then they spent more years when getting their education, so they can positively influence child's attitude to studying. Moreover, if one of the parents works as a teacher, then they better know how to motivate their child to study harder, explain some material in a more detail, and so the grades will be better. Since these jobs of parents considerably influence the grade and there is a certain assumption behind such relationship, a new variable was created. It was a dummy variable, indicating if at least one of parents either works in healthcare or as a teacher.

Finally, the testing for potential interaction terms was performed. Specifically, for variables representing gender and number of absences. There was a goal to investigate if there is a case when for representatives of one gender the number of absences has different relationship with the final grade than for another. In order to test this assumption, the scatter plot of 'absences' and 'sex' was made and the linear model was fitted to visualize relationship. For a more accurate and representative plotting, there was used an approach of Poisson regression. Therefore, we would like to plot $\log(\lambda_i)$ of G3 by number of absences. λ_i is unknown, so it was approximated by taking average value of final score at every value of number of absences and respective gender.

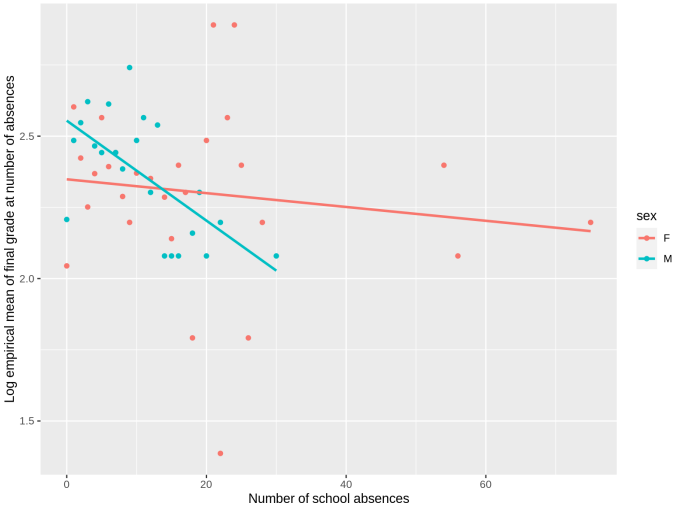

Fig. 6. Relationship between Final Score and Number of Absences by Gender

For male students there is quite considerable negative linear relationship between number of school absences and final score. However, it is not that obvious for girls, which means that girls can actually achieve quite high score even at higher number of absences. This might be caused by the fact that girls may skip classes for a good reason, while boys tend to intentionally do it. Therefore, it was considered to be useful to add an interaction $sex \times absences$ to the model.

F. Models

The data has been fitted to multiple models to compare their performance and select the best one. The preferred algorithms for this problem are Linear Regression and Poisson Regression, and the sub-models of the algorithms are Lasso and Ridge. In order to compare the performance of these models, various evaluation metrics such as MSE, RMSE, and MAE have been used.

- Linear Regression

Initially, a Linear Regression model was fitted to the entire subset of variables. The model summary revealed that several predictor variables, including sex, age, study time, schoolsup, and MplusFedu, were statistically significant. This indicated a significant impact on the students' final grade. The assumptions of the model were examined to determine whether the model assumptions of linearity, normality, and homoscedasticity were met.

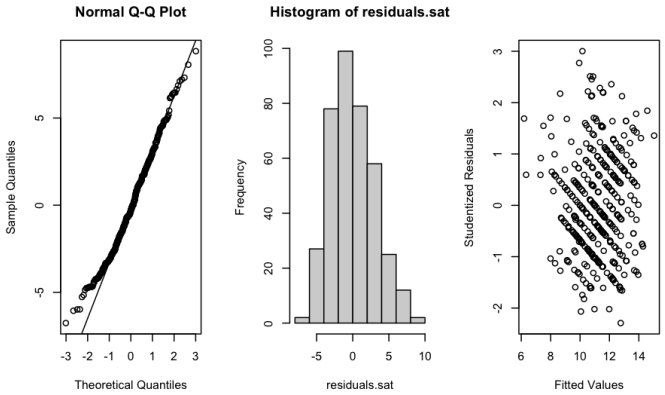

Fig. 7. Residuals

The q-q plot indicates that the residuals in the regression model are a bit heavy tailed. Also, this is confirmed by the histogram plot. As a result, the quantiles are larger than the theoretical.

Next, feature selection was performed using three approaches: best subset selection, forward selection, and backward elimination. The top eleven variables identified by each approach were similar, and had identical coefficients as they all were nested within each other.

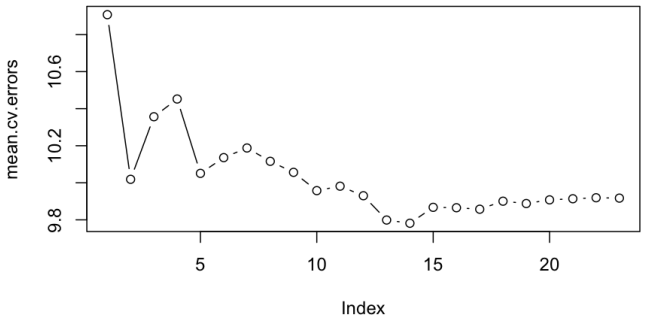

Fig. 8. Mean Cross-Validation Errors

The validation set and cross-validation approaches were used to select the best model, from a set of models with varying sizes. To implement the validation set approach, we first divided the observations into a training set and a test set, and subsequently performed best subset selection. As shown in Figure 8, the cross-validation approach selected a 14-variable model, with the lowest mean cross-validation error. After fitting a Linear Regression model using the features selected through cross-validation, we found that the variables sex, age, studytime, schoolsup, famsup, and absences were statistically significant.

We used the anova function to compare the two models. Model 1 represents the linear model that includes all predictors, while Model 2 includes only the selected features from cross-validation. The anova function performed a hypothesis test to compare these models. The null hypothesis was that the models are equally good and on the other hand, the alternative hypothesis was that the reduced model (with fewer predictors) is worse than the full model. Based on the output and the null hypothesis, we rejected the alternative hypothesis since the p-value of 0.81 is greater than the significance level of 0.05.

An alternative approach to generalizing Linear Regression is to implement Lasso, which applies an L1 penalty to the model. Depending on the chosen tuning parameter, some coefficients may be set to exactly zero, resulting in a sparse model. The sparsity of the model is a substantial advantage of Lasso over Ridge Regression. We used cross-validation to fit a Lasso regression model and select the optimal regularization parameter, lambda. We evaluated the performance of the model on a test set by calculating the mean squared error using two different values of lambda: the one that minimizes the cross-validation error (bestlam) and the one that is one standard error away from the minimum (bestlam1se). The results showed that the mean squared error was 8.32 for bestlam and 8.62 for bestlam1se. These findings suggests that using the value of bestlam, which was equal to $\lambda = 0.12359772$, may result in slightly better predictions on new data. We have observed that eight coefficient estimates are precisely zero (reason, guardian, freetime, school, parent_health_teacher, paid and famrel) in this case. Therefore, the Lasso model includes only 15 variables and by using only the features selected by Lasso and the optimal value of bestlam, we observed an improvement in the mean squared error compared to using all the features.

The presence of numerous categorical predictors with more than two levels in the dataset used for this study can lead to a large number of coefficients. The ordinary least squares method may lead to over-fitting, where the model fits the noise rather than the true underlying relationship between the predictors and the response variable. To address this issue, the L2 penalty (Ridge) was employed as an alternative approach to linear regression. By shrinking the coefficients toward zero, the Ridge approach helps to reduce the model's variance. The optimal value of the regularization parameter lambda was determined using cross-validation, which yielded a selected lambda value of $\lambda = 1.711687$.

The following table represents the performance of all linear regression models.

TABLE IV
METRICS FOR LINEAR REGRESSION MODELS

Models	Metrics		
	MSE	***RMSE***	***MAE***
All predictors	8.61	2.93	2.37
Best 11 predictors from Feature Selection	8.05	2.84	2.26
Best 14 predictors from CV	8.44	2.91	2.34
Lasso	8.32	2.89	2.32
Ridge	8.34	2.89	2.33

Based on the table of metrics for linear regression models, it can be viewed that the model with the best 11 predictors selected from feature selection has the lowest MSE, RMSE, and MAE values, indicating a better performance compared to the other models. However, the difference in performance between the models is relatively small. The Lasso model also performed well and had similar performance to the best 14 predictors selected from cross-validation.

- Poisson Regression

Since the final score is a discrete variable, it was worth trying to implement a model which can favor such type of data. The Poisson regression model was considered in this essence. Undoubtedly, grades essentially is not really a variable, which naturally follows Poisson distribution, since they are not events which occur with a certain rate over time. Still, the assumption of equality between mean and variance of Poisson random variable was tested in order to check if there is some potential and evidence to implement this model.

Firstly, the relationship between G3 and age was considered. The following table summarises mean and variance of final score at each age.

TABLE V
AGGREGATED SUMMARY OF FINAL SCORE BY AGE

Age	Mean of G3	Variance of G3	Students Count
15	11.90	12.04	81
16	11.67	10.68	107
17	10.88	9.62	100
18	10.57	10.47	81
19	8.82	1.96	11
20	18.00	NA	1
22	8.00	NA	1

While the size of the group remains considerably big, the mean of final score is approximately equal to variance. In addition, the table shows that the responses at each level of X become more varied with increasing means. These observations favor Poisson regression assumption of equality of mean and variance. It could be also depicted graphically for a better perception.

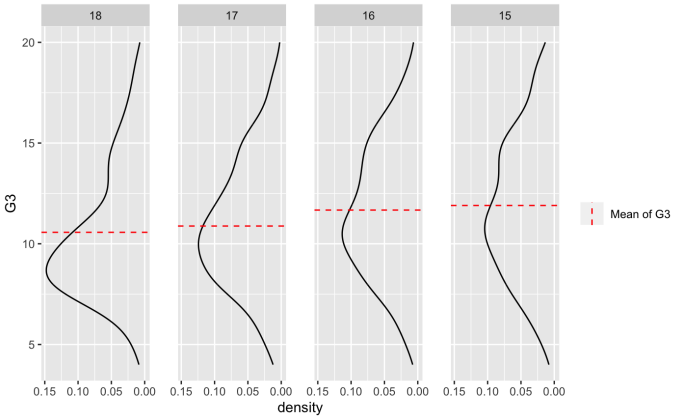

Fig. 9. Distributions of Final Score by Age

The same procedure was performed for the study time variable. One may logically assume that the more students study, the better is the final grade. However, in case of Poisson random variable one should not only expect higher mean of final grade but the variance as well, meaning that this mean is not influenced by higher number of observations, concentrated around it. This will assure that for different students studying more does not necessarily lead to a certain "guaranteed" score.

TABLE VI
AGGREGATED SUMMARY OF FINAL SCORE BY WEEKLY STUDY TIME

Weekly Study Time	Mean of G3	Variance of G3	Students Count
$<$ 2 hours	11.11	11.45	103
2-5 hours	10.82	9.71	190
5-10 hours	12.15	10.88	62
$>$ 10 hours	12.15	13.67	27

In this case there is quite a considerable difference between mean and variance of study time of 5-10 hours per week. Still, for other levels it is close. The graphical representation is as follows.

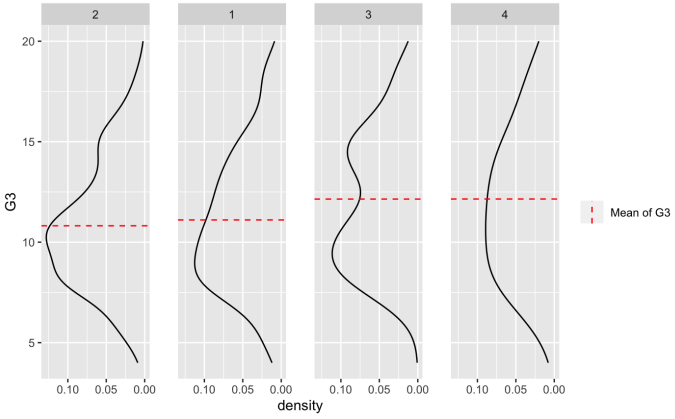

Fig. 10. Distributions of Final Score by Weekly Study Time

Even though for this variable the assumption of equality between mean and variance is not perfectly held, still there is quite sufficient evidence to try Poisson regression for the task of this study.

At first, the Poisson regression was fitted on the whole subset of variables. It resulted in only 6 statistically significant coefficients. Such observation allowed to assume that feature selection algorithms might help identify the smaller subset of variables which will still be explanatory powerful. Feature selection was performed through forward selection and backward elimination approaches, while the target metric to minimize was Akaike's Information Criterion (AIC). Forward selection did not help reduce number of predictors considerably, while backward elimination decreased it from 23 to 11. In addition, it reduced AIC from 1958.9 of model with all variables to 1937.2.

Alternative approach to Poisson regression generalization was its implementation with L2 penalty (Ridge). The dataset used in this study contains a lot of categorical predictors, ones with more than 2 levels. These predictors introduce a lot of coefficients to the model, which may result in over-fitting. Ridge regression shrinks coefficients towards zero, so that the model's variance is reduced. The best lambda parameter for ridge regression was selected using 10-fold cross-validation and was set to $\lambda = 8.111308$.

An alternative approach to generalizing Poisson regression is to use L1 penalty (Lasso) instead of L2 penalty (Ridge). In datasets with many categorical predictors with more than 2 levels, Lasso can be particularly useful because it can perform variable selection by shrinking some coefficients all the way to zero, effectively eliminating those predictors from the model. This can prevent over-fitting and improve the model's interpretability by reducing the number of predictors.

To select the optimal lambda parameter for Lasso, we use cross-validation and choose the value that minimizes the mean squared error. For example, after performing cross-validation on our dataset, we found that the best lambda parameter for Lasso was $\lambda = 0.102613$. Using Lasso regression with this lambda value can help us build a more accurate and parsimonious model.

The following table represents the performance of all Poisson regressions fitted.

TABLE VII
METRICS FOR POISSON REGRESSION MODELS

Models	Metrics		
	MSE	***RMSE***	***MAE***
All predictors	11.59	3.41	2.69
Best 11 predictors from Feature Selection	11.03	3.32	2.61
Lasso	8.35	2.89	2.32
Ridge	10.94	3.31	2.67

Since the target variable is discrete, it does not make much sense to give weight to the points far from the mean, since they are in fact in general not so far from it. In this sense, using MAE might be a more accurate approach to assess models performance. Therefore, the model which was fitted with Lasso is slightly better than the rest.

G. Proposed best model

The proposed best model for the given data and problem statement is Linear Regression with Feature Selection. This model was selected after performing feature selection with best subset selection and evaluating multiple models. The model uses the best eleven variables identified through feature selection to predict the numerical output variable. The model performance was evaluated using three metrics:

TABLE VIII
LINEAR REGRESSION WITH FEATURE SELECTION

MSE	RMSE	MAE
8.05	2.84	2.26

Overall, the Linear Regression with Feature Selection model is a suitable choice for the given data and can be used to make accurate predictions on new data. Linear Regression has several advantages as a predictive modeling technique. One of them is that, it allows for the interpretation of the relationship between the input and output variables. Moreover, it is a computationally efficient algorithm, making it suitable for large datasets. These are some advantages that make Linear Regression with Feature Selection a reliable choice for predicting numerical output variables in a variety of applications.

III. DISCUSSION

A. Findings of primary significance

One of the key findings from our data analysis is that feature selection can improve the performance of the model. Through feature selection, we identified a subset of input variables that were most relevant to the prediction of the output variable. This finding highlights the importance of feature selection in improving the interpretability, efficiency, and accuracy of the model. Feature selection can help to identify the most important input variables and reduce the complexity of the model.

Our data analysis has revealed that Linear Regression is a better fit for our analysis than Poisson Regression. This finding suggests that the relationship between the input and output variables is better captured by a linear relationship than a logarithmic relationship. This finding may be due to the fact that our output variable is not a count variable and its distribution is better approximated by a normal distribution. Furthermore, the input variables may have a linear relationship with the output variable rather than a logarithmic relationship.

B. Findings of secondary significance

The analysis revealed several significant findings related to student performance and the factors that influence it. This provided valuable insights that can contribute to academic success.

One significant finding is that students whose mothers' are healthcare or services workers, or whose fathers' are healthcare workers or teachers, tend to achieve higher average scores. Interestingly, students whose mothers' are teachers also perform well, with their average grades being comparable to those of students whose mothers' work in public services, which is the second-best performing category.

It was also revealed that for male students, there is a significant negative linear relationship between the number of school absences and their final score. In contrast, for female students, this relationship is not as clear, and they can still achieve a high score even with a higher number of absences.

Another finding was the significant correlation between the type of area (rural/urban) and the time it takes to go to school. This observation aligns with the expectation that students who live in rural areas may have to travel longer distances to reach school than those who live in urban areas.

Additionally, a significant dependence was found between the number of past class failures and the willingness to pursue higher education. The majority of students (over 80%) who expressed a desire to pursue higher education had no record of past class failures. This observation aligns with the expectation that students who are motivated to pursue further education tend to have a stronger academic record.

IV. Conclusions and Future work

In conclusion, the analysis of the student performance dataset has provided valuable insights into the factors that influence academic success. Through the exploratory data analysis and predictive modeling techniques, several variables have been identified which are significant predictors of student performance, including traveltime, sex, mother's and father's education levels, and absences. The linear model with 11 selected predictors from feature selection has provided promising results in predicting students' final grades. The model achieved a relatively low mean squared error (MSE) of 8.05, a root mean squared error (RMSE) of 2.84 and a mean absolute error (MAE) of 2.26. This suggests that the model can provide reasonably accurate predictions of student performance.

For future work, a possible way to increase the performance of the model for predicting students' final grade is to explore the potential impact of other variables such as extracurricular activities and family income. Additionally, further data collection from a more diverse population could ensure that our findings are more generalized to a wider range of students. Overall, further research in this area can help to better understand the complex factors that influence student performance and ultimately lead to improved educational outcomes.

References

[1] 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.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.