Prediction Of Academic Performance With Social Factors Using Regression

- Aniruthan Sivakumar, John Huddleston

***Abstract*— Academic Performance of a student are influenced by multiple factors around them. The most dependent predictors had to be measured by putting all attributes in a Regression Model. Multiple regression models would have to be fitted to attain the most efficient model. The efficiency of the model is calculated through the RMSE (Root Mean Squared Error) which would give us the top dependent variables.**

I. Introduction

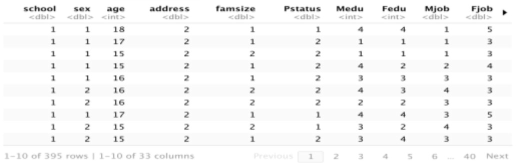
Quality of education is important for everyone's success with secondary school playing a significant role about future educational pursuits. In this Project, we quantify multiple social and demographic attributes that impact the academic performance of a student. Students try spending immense time in studying, they pay for extra tuition, they cut their social time to avoid distraction. The choice of students distancing themselves from these factors, are thought to be related towards a higher grade. Our analysis is here to state that, affecting the non-educational aspects of student life, doesn't positively increase the grade. The students don't feel the importance of their social-economic factors and put in a lot of effort towards their grade, and most parents support this for their child to be educationally bright [4]. The community around has normalized the fact that, the students could give up anything for education, which at end, doesn't benefit both the grade and the individual. The reason for this perspective of education could be investigated with context of policy optimization, which would help create socio-economic polices, helping education. The study cultivates important insights which can potentially improve development and policies refinement to relatively improve students' academic performance. The approach we used here for the research was interdisciplinary.

We were able to evaluate student performance from students in two specific schools to attain a diverse population rather than studying a group of students from the same education system. Some social factors that were surprisingly present in the evaluation were "Alcohol" and "Quality of Family relationships". Beyond these attributes, romantic relationships of the student, and their participation in extracurricular activities were surprisingly significant predictors, looked from the outline. Looking at the dataset, it was also easy to tell that the increased participation in extracurricular activities had a positive correlation with the final grade. So, by not jumping into deep analysis, we were able to view the fact that social features support education, which made us to select the dataset. We tried Linear Regression, Ridge Regression and Support vector regression as our models, with the final grade being our response variable.

The main reason we wanted to research on the topic was to understand the significance of student behavior outside classroom, and the impact has on student grades. We learnt about how social and mental factors that could potentially affect the student's mind to influence their final grades. We knew that the intuition is present behind the theory, but we were curious to gain data-driven insights to support the hypothesis, and we were able to analyze and get them.

II. Dataset

We chose the student dataset from Kaggle. The dataset was obtained from students taking a math course over two secondary schools. The set had 33 attributes with close to 400 observations which is not a lot. Most of these attributes were categorical, and the biggest challenge was to convert and standardize those values to maintain uniformity over the dataset [6]. Mainly, the attributes focused on student's behavior outside of school, such as their study time, screen time, free time, family situation, extra-curricular activities, etc. These examples show an idea of on what the dataset provides information about. We also carefully analyzed each categorical and binary attribute and assigned them the appropriate standardized numerical values for each category to prevent biasing. The response variable was "G3" which was the final grade in terms of 0-20. We also discounted the unordered the data with respect to its handling. The final structure of the dataset would show how the dataset was modified.



school <dbl>	sex <dbl>	age <int>	address <dbl>	famsize <dbl>	Pstatus <dbl>	Medu <int>	Fedu <int>	Mjob <dbl>	Fjob <dbl>
1	1	18	2	1	1	4	4	1	5
1	1	17	2	1	2	1	1	1	3
1	1	15	2	2	2	1	1	1	3
1	1	15	2	1	2	4	2	2	4
1	1	16	2	1	2	3	3	3	3
1	2	16	2	2	2	4	3	4	3
1	2	16	2	2	2	2	2	3	3
1	1	17	2	1	1	4	4	3	5
1	2	15	2	2	1	3	2	4	3
1	2	15	2	1	2	3	4	3	3

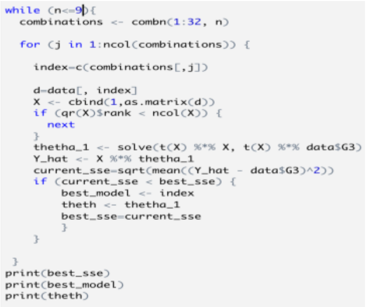
1–10 of 395 rows | 1–10 of 33 columns

After the structure was modified, we were able to get a myriad of attributes, which detailed student's life outside the education system. Post-modification, the dataset was ready to proceed with regression, trying to get the best subset of predictors by training the model with the 32 attributes towards the response variable. Luckily, the dataset didn't have 'N/A' values which gave us a non-sparse dataset. If there were 'N/A' values, then the predictions would have been more misleading, questioning the efficiency of the model. So, the only drawback of the dataset, is that there are less observations, other than that the dataset provides the accurate details.

III. Methods And Measures Used

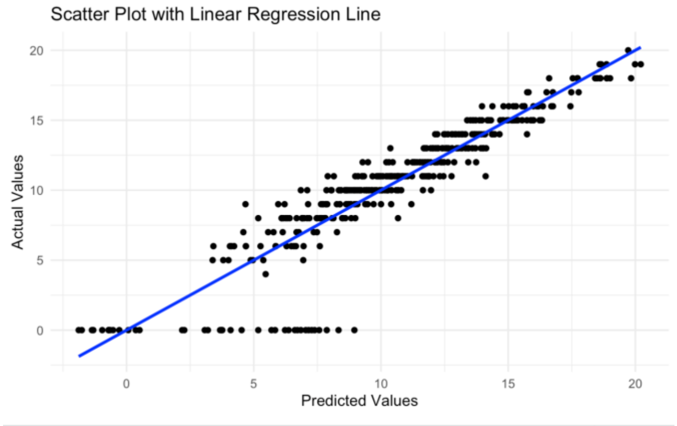
The whole project was completed using R and Python. We used Python primarily on the Support Vector Regression model. For the Support Vector Regression model, 75% of the dataset was assigned as the training data while the 25% was separated as testing data for the model evaluation. Several values were tested for the margin and regularization (epsilon and C), until the best value was seen. After continuous testing, we were able to attain a regularization parameter of 1 and a margin of 0.2.

R was used for linear regression. The Linear regression model was trained and tested on the same data, while it also went an 80-20 training-test split, which gave a rough 0.5 increase in the Root Mean Squared Error (RMSE). The formula used for the regression was the High dimension linear regression formula used in class, which was Y=X*θ where X is a column-bind matrix. Subsets of variables were sent to the linear model, and the linear model that acted with the lowest RMSE was selected.



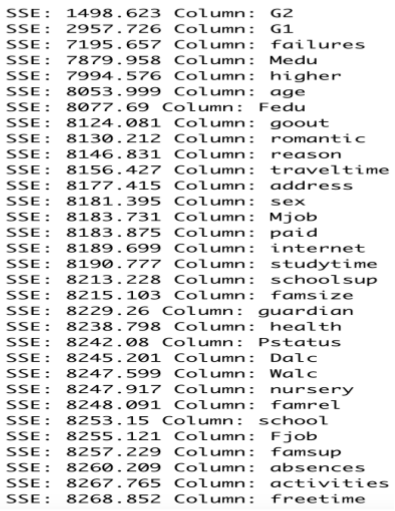
```
while (n<=9){
  combinations <- combn(1:32, n)

  for (j in 1:ncol(combinations)) {

    index=c(combinations[,j])

    d=data[, index]
    X <- cbind(1,as.matrix(d))
    if (qr(X)$rank < ncol(X)) {
      next
    }
    thetha_1 <- solve(t(X) %*% X, t(X) %*% data$G3)
    Y_hat <- X %*% thetha_1
    current_sse=sqrt(mean((Y_hat - data$G3)^2))
    if (current_sse < best_sse) {
        best_model <- index
        theth <- thetha_1
        best_sse=current_sse
        }
    }

 }
print(best_sse)
print(best_model)
print(theth)
```

Before implementing high dimension regression, single variable was regression was implemented to find the top dependent variables. This was also measured with respect to the RMSE. After proceeding with the best regression model (High dimension regression model), we graphed the plot to look at the fit between the actual values and the predicted values (Linear fit). The fit looked almost consistent with not very many outliers. This reassured that the model was a very good fit.



IV. Regression Analysis

We separated the data into 75% training data and the remainder testing data. This was done so that we had sufficient training data. The Support Vector Regression using all variables after hyperparameter tuning yielded a root mean squared error of 2.1. Using only the social factors, i.e. excluding the G1 and G2 variables, the root mean squared error was 3.1. The multiple linear regression using the 8 most independently predictive variables yielded mean squared error 2.3. The midterm grade variables G1 and G2 [6] were by far the most independently predictive, which may explain why most variables contributed very little to improvement of the multiple linear regression model. This may also explain why the multiple linear regression compares well to the support vector regression as these grades are on the same scale and highly predictive of each other. Insights can also be gained from the contrast in accuracy between the support vector regression including or excluding past grades as predictive. The significant decrease in predictive accuracy when excluding midterm grades indicates that social factors likely have limited predictivity for academic performance.



```
SSE:  1498.623 Column:  G2
SSE:  2957.726 Column:  G1
SSE:  7195.657 Column:  failures
SSE:  7879.958 Column:  Medu
SSE:  7994.576 Column:  higher
SSE:  8053.999 Column:  age
SSE:  8077.69 Column:  Fedu
SSE:  8124.081 Column:  goout
SSE:  8130.212 Column:  romantic
SSE:  8146.831 Column:  reason
SSE:  8156.427 Column:  traveltime
SSE:  8177.415 Column:  address
SSE:  8181.395 Column:  sex
SSE:  8183.731 Column:  Mjob
SSE:  8183.875 Column:  paid
SSE:  8189.699 Column:  internet
SSE:  8190.777 Column:  studytime
SSE:  8213.228 Column:  schoolsup
SSE:  8215.103 Column:  famsize
SSE:  8229.26 Column:  guardian
SSE:  8238.798 Column:  health
SSE:  8242.08 Column:  Pstatus
SSE:  8245.201 Column:  Dalc
SSE:  8247.599 Column:  Walc
SSE:  8247.917 Column:  nursery
SSE:  8248.091 Column:  famrel
SSE:  8253.15 Column:  school
SSE:  8255.121 Column:  Fjob
SSE:  8257.229 Column:  famsup
SSE:  8260.209 Column:  absences
SSE:  8267.765 Column:  activities
SSE:  8268.852 Column:  freetime
```

Those social factors that were strongest as independent predictors were generally related to socioeconomic status such as educational attainment of the student's parents. Also, of note some of the most predictive social factors were frequency with which students interacted with friends and romantic partners. That this prosocial behavior predicted greater student success indicates that student social and academic well-being may be related [3].

Further into the linear model, the formula that was derived from the Linear model was "Final Grade = -1.323716 – 0.214581 * Student Age + 0.34237*Student's level with extra-curricular activities + 0.32*Romantic Relationships (1 or 0) + 0.372*Healthiness level of the family relationships + 0.123*Weekend Alcohol Consumption rate + 0.04* Level of Absences in Attendance +0.18*Midterm 1's Grade + 0.9622* Midterm 2's Grade. The slopes mentioned above were taken from the θ vector and is supposed to be part of one of the efficient models. Each slope shows its dependent correlation towards the Final grade. We can also see if the equation supports the intuitiveness behind the uniqueness of the dataset. Age relatively has a steep negative correlation with the Final grade variable. So, for this model, a person of a higher age is highly likely to score a lower grade than a younger student. This is understandable, as in secondary school, the coursework gets harder as the grade increases, and is going to result in a lower grade average, so this slope looks intuitively final. Moving towards the slopes of social and physical, all of them have a positive correlation (Romantic Relationships, Family Relations, Extra-curricular activities), which is further justified in the conclusions So social health in general has a positive correlation with the final grade. Then the next slope we see is of the weekend alcohol consumption rate, which surprisingly has a positive correlation [2] against the final grade. So, does that mean higher weekend alcohol consumption results in a higher grade? Definitely not![2]

When we looked at the reasons for its positive correlation, we saw that the dataset unusually had students with higher rate of alcohol consumption. One way to look at it would be that the dataset is taken from Portugal, which had a higher alcohol consumption percentage [1]. So, this was a potential reason for the positive correlation, which in terms can be seen as a non-significant predictor or dataset-specific predictor. Another potential reason could be that alcohol could be dependent on another variable, which could influence the behavior of its slope (Alcohol could be related to social gatherings, and not representing the very effect of it). If we really want to assume Alcohol consumption boosts our grade, we can see how badly they perform against the Final grade in the single variable regression, when they are used as predictors. Then we could look at the slopes of the class-related predictors which are class absence and grades of midterm 1 and midterm 2. The slope of class absence is positive, but is very minute, so a statement can't be constructed around that slope, as it is not very meaningful. But when we look at the grades of midterm 1 and 2, we can see both have a significant positive correlation against the final grade. But comparatively, the relation of midterm 2 and final grade is very strong, as the slope is 0.96 (almost 1). So, the interesting analogy here would be that if a student has a wonderful midterm 2, then they are likely to have a similar final grade. Also, if we have two students, student 1 and student 2, where student 1 has a 95% in their midterm 1, and 70% in midterm 2, while student 2 has a 65% in midterm 1 and a 90% in midterm 2. The intuition would be to predict student 1 to have a higher final grade, but our data driven insights in this dataset support student 2 to have a better final grade than student 1, which is another interesting insight.

V. Conclusion

Our insights between performance provided by the academics and different social factors. First, we can state that the final grade has a strong positive correlation with the previous exams. We can conclude that a strong base of understanding is needed to have the high probability of getting a good grade. Even though we can interpret that social factors affect student lives, all these factors cannot conclude and determine the final grade, as the slight causation [2] should be seen as a combination of all factors, which directly or indirectly implies social, physical or mental health [3]. After the consideration of certain variables, the three most social factors that showed a positive correlation between were "Parent's Education", "Student's Romantic Relationship", and the frequency of going out with friends. As we stated earlier, all these factors could be seen as a sense of social health [3]. Now with additional analysis, we can state that a student is highly likely to get a better grade if they understand the material and has a good social health. So, a good human interaction cannot just be a distraction but can at times, positively help the mind to study.

We also learnt that just because a variable is present in the most efficient model to predict the output, it cannot be related to the response variable, as the very variable could be affected by the dependence of other variables, or even the response variable could be heavily affected by another group of variables, and one such variable in our case was the Weekend Alcohol Consumption variable.

VI. Author Contributons

Aniruthan Sivakumar was responsible for doing background research on our topic of interest as well as locating and processing the dataset that was used for the project. he also implemented and tested the single and multiple linear regression models in r, as well as making the part of the presentation detailing these models. Aniruthan wrote the parts of this paper concerned with detailing the problem background, dataset and measures and methods. John Huddleston implemented and tested the support vector regression model in python, created the part of the presentation concerned with its details and wrote the results and interpretations section of this paper.

References

[1] C. Bello, “Europeans are the world’s heaviest drinkers: How do countries compare?,” euronews, https://www.euronews.com/next/2023/06/30/so-long-dry-january-which-country-drinks-the-most-alcohol-in-europe#:~:text=Daily%20alcohol%20consumption%20in%20the%20EU%20by%20age&text=In%20the%20EU%2C%20drinking%20every,cent%20in%20Latvia%20and%20Lithuania. (accessed Dec. 13, 2023).

[2] R. Mart&iacute;nez, “If correlation does not imply causation, then what does?,” Medium, https://medium.com/gradiant-talks/if-correlation-does-not-imply-causation-then-what-does-8fa462943b84 (accessed Dec. 13, 2023).

[3] The Lancet Public Health, “Education: A neglected social determinant of health,”The Lancet Public health, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7326385/ (accessed Dec. 13, 2023).

[4] M. Shareef, “Why most of the parents do insists their children to choose medical and engineering fields?,” Medium, https://medium.com/age-of-awareness/why-most-of-the-parents-do-insists-their-children-to-choose-medical-and-engineering-fields-155e129d4295 (accessed Dec. 13, 2023).

[5] K. Nordhausen, Robust linear regression for high-dimensional ... - wiley onlinelibrary,https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.1524 (accessed Dec. 14, 2023).

[6] U. M. Learning, “Student Alcohol Consumption,” Kaggle, https://www.kaggle.com/datasets/uciml/student-alcohol-consumption/data (accessed Dec. 13, 2023).