Socio-economic and Demographic Influence on Student Grades

Group 8
Dinesh Saraswat, Kyle McGonigle, Jeremy Manalo, Joel Saetern, Tia Cao

College of Business and Economics, California State University East Bay

Prof. Dr. Balaraman Rajan

December 17, 2023

Problem

Motivated by our own experiences as students, we aimed to investigate the association between several student socio-demographic factors and academic performance.

This study will not only deepen our understanding of learning but also explore the feasibility of interventions based on the findings to improve student grade outcomes.

Data

The data was obtained in a survey of students' Portuguese courses in secondary school. It contains a lot of interesting social, gender and study information about students. It is pulled from Kaggle data set source -https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

Sample size of 649 students, 32 variables, 1 output

- school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- 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)
- G3 - final grade (numeric: from 0 to 20, output target) -> G3_BIN

Objective

The goal of this analysis and model building analysis is to determine which socio-demographic properties of students influence the grades of the students and if we can do something about it to influence the grades of the students.

We would also like to provide cost effective solutions based on these observations so that we can propose solutions that would not only be theoretically feasible, but something that most communities could adopt. This narrows our scope to realistic solutions to the problem at hand.

Data Exploration

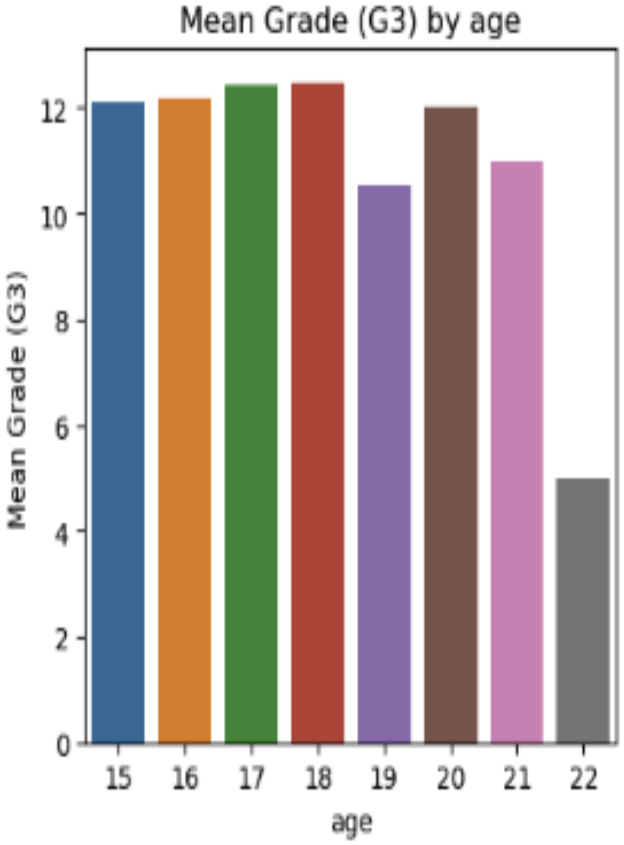
By observing the data, we found that GP students have twice as many MS students in this survey and the range of age is between 15 and 22 which mostly falls between 15 and 18 years old as secondary school; however, we still have a few students in our data who are between the age of 19 to 22. In addition, the majority of students are from urban areas and have large family sizes that are greater than three members who also stayed together with their parents. On the other hand, the number of absences is around 0 to 5, even though the dataset still shows a few students take 9 to15 days, compared to a full academic year. This is not a high number and most of the students still want to continue to come to school. Then, we found that the relationship between student’s free time after school and student’s go out with friends are similar (around 3 in the

range of 1 to 5), so students probably shared their free time with friends after school. Last but not least, we also noticed the Walc (Weekend Alcohol consumption) and Dalc (Workday Alcohol consumption) have big differences. On weekdays, the majority of students would have 1 drink which is very low; however, on weekends, most students would have more than 1 drink and some of those even have 5 drinks which are very high.

We will start with the data loading and start exploring the dataset and observe what is the type of the data and which model may suit the analysis and explore its performance along with the multiple models and compare its performance along with the naive model or mean models.

If the model works well and is able to identify more accurate results, it would be great to use it and make an impact on the life of the students.

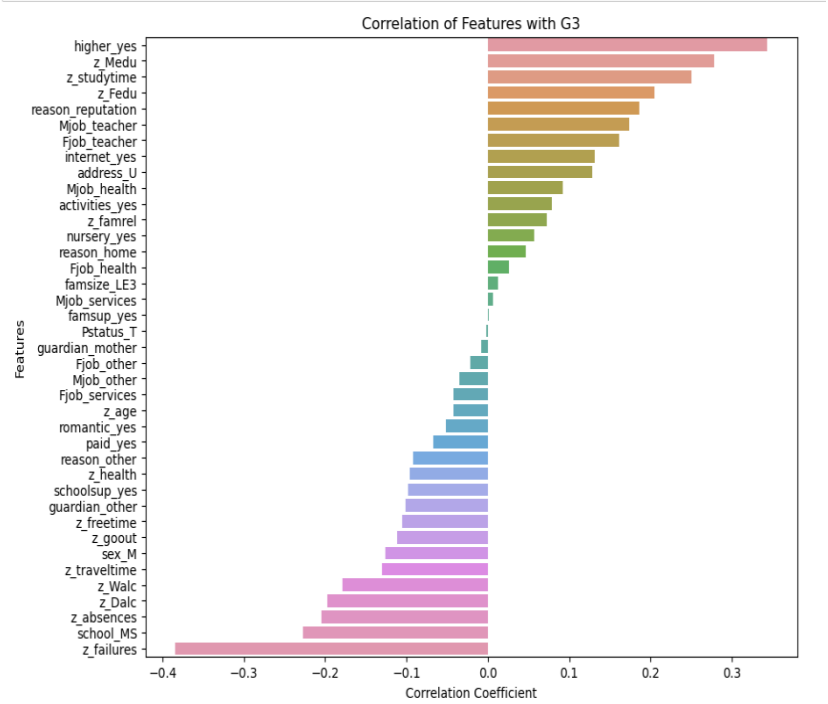
We initially took a look at the data in terms of just exploring the different variables visually. There were some interesting results, some of which we would end up incorporating into some of our data clean up and analysis. .



When we were looking at mean grade by age, we noticed a stark drop off for the group that was aged 22. We investigated why this might be the case and found that there were a multitude of factors such as this cohort going out more, having high alcohol consumption, higher absences, and a higher count of past failures. These people behaved more like outliers relative to the general population. Thus, we removed them from the dataset.

There were also some other interesting observations that we made when exploring the data via a correlation heat map and a correlation feature breakdown

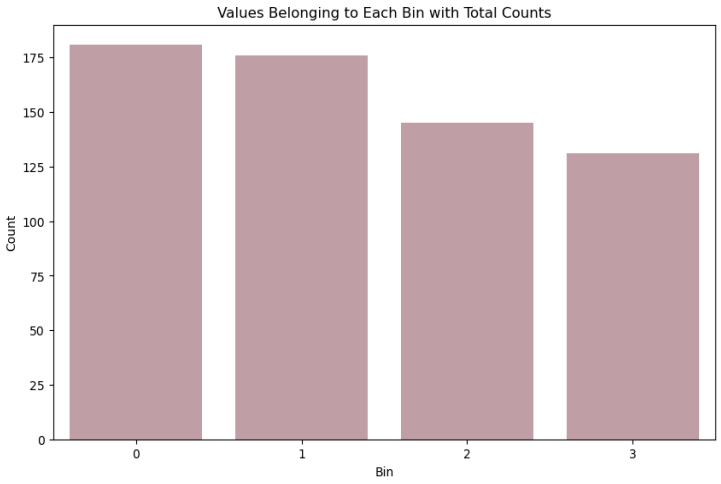
There was a strong correlation between G3- final grade and Father's education, Mother's education, study time, reason of reputation, parents' job as a teacher, having internet access, and urban address location has positive correlation.



On the other hand, there was a negative correlation with G3: past failures, absences, Workday and Weekend alcohol consumption, travel time, sex as male, and going out.

We should care about these features because there we can influence reinforcing the features that are positively correlated with the data as well as addressing issues that are negatively correlated with the features. Also, we might want to check for any highly correlated features for multicollinearity.

Data Clean Up



G3_bin	min	max	count
0	5	10	181
1	11	12	176
2	13	14	145
3	15	19	131

We started the data clean up by dropping G1- first period grade and G2- second period grade since they were already highly correlated with the model. We also dropped any outliers (like Age = 22) from the dataset.

We also decided to create Bins for G3- final grade. Our objective was to generally predict an outcome of a student's performance in terms of a success or failure, so binning seemed like a logical step.

With bins we were able to predict student approximate performance rather than actual grades. Therefore, we categorized grades in 4 bins of each around almost 25% of data and then normalized the continuous variables..

Then we decided to define the categorical variables (which we would keep) for model fitness.
We started with Multiple Regression and ran Backward Elimination to identify best factors associated with 'z_studytime', 'z_failures', 'z_famrel', 'z_health', 'z_absences', 'schoolsup_yes', 'higher_yes', 'school_MS', 'sex_M', 'Mjob_teacher', 'Fjob_health', 'Fjob_other', 'Fjob_services', 'reason_reputation'

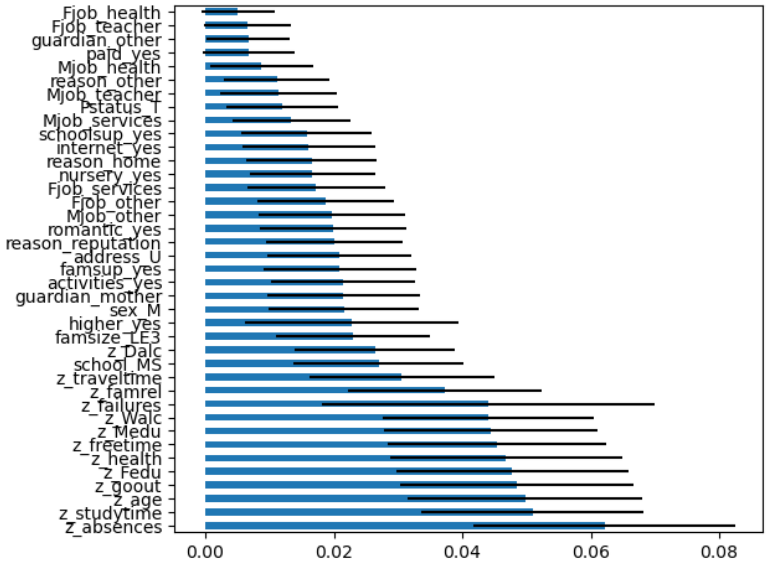
We identified that higher_yes" and "Fjob_Other" have strong VIF (identifies strong correlation with the dependent column (G3- final grade)).

	Variable	VIF
36	higher_yes	10.707842
23	Fjob_other	8.231278
17	Pstatus_T	8.011992
37	internet_yes	5.334340
35	nursery_yes	4.984757
24	Fjob_services	4.655083
29	guardian_mother	4.130585
15	address_U	4.044509
19	Mjob_other	3.181300
32	famsup_yes	2.830123
20	Mjob_services	2.633578
1	z_Medu	2.608084
21	Mjob_teacher	2.522472
14	sex_M	2.276130
34	activities_yes	2.198683
10	z_Walc	2.125139
2	z_Fedu	2.087526
13	school_MS	1.976517
25	Fjob_teacher	1.954323
18	Mjob_health	1.946154
9	z_Dalc	1.857122
28	reason_reputation	1.775952
38	romantic_yes	1.718174

We removed this and ran the multi regression again to yield results of RSME of .03 higher

Model Exploration

After our initial analysis using a multiple regression model, our team explored other methodologies for comparison. The **random forest model**, in particular, revealed new insights, identifying different influential variables.



Intrigued by this, we re-ran the multiple regression model, this time incorporating these variables identified by the random forest.

Interestingly, the results were similar to our initial multiple regression analysis. However, the updated variable list from the random forest model offers our team enhanced control over influencing factors. This finding suggests a potential avenue for more precise predictive modeling in our future analyses.

We decided to investigate further modeling with a few quick runs with the following results:

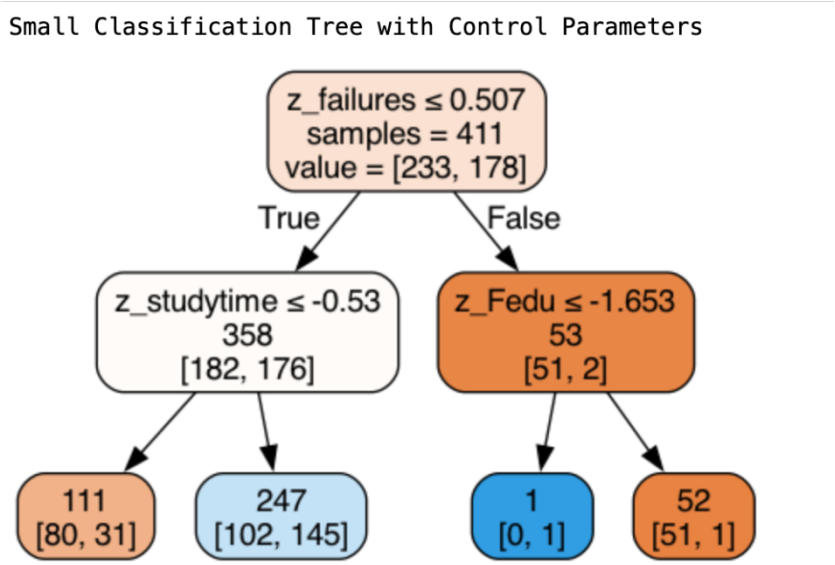
We tested the data using **kNN**, the accuracy turned out to be 34.23% with n = 9.
The accuracy of kNN was definitely not ideal for prediction (compared with 28.59% for Naive).
We then pursued further models.

We proceeded with running the data with the **classification tree** model, the accuracy returned was 36.94%. This, also, was not terribly predictive relative to the other variables. Overall, this classification model still gives you a better result than the naive model and kNN.

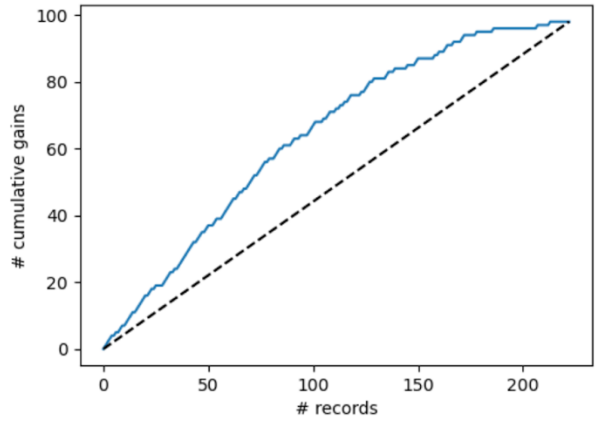
Logistic Regression, KNN, Classification Tree, Random Forest are evaluated as well for the model and have shown good insights.

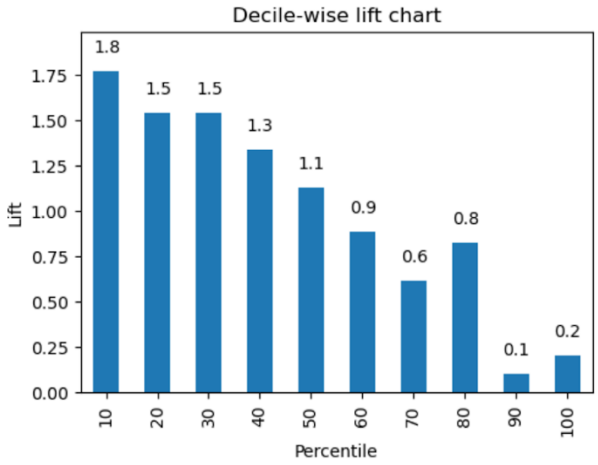
On the other hand, we also wanted to investigate students who have lower grades and would need more support or insights. First of all, we update the expected values as predicting Grade bins and we classify students who failed as having grade less than 10 and who did better as having above 10. Then, we used classification tree and logistic regression methods to determine:

We ran the classification tree with an accuracy of 60.81 % which is better than 28.59% of the naive model. Thus, this model is a good standard to understand if students would fail and do better. By having a classification tree, we can totally have the new rule by looking at the tree base and predict students who fail or do better. In addition, this model has good scope of further enhancement as we gather more data to understand the grade bins and see if the students will do better or need to intersect to help improve.



We also ran the logistic regression, we found that the z_studytime, z_Fedu (Father's education), z_Medu (Mother's education) and z_famrel (quality of family relationship) are associated with the failure result when the more study time, parents' education and family relation are positively related to the outcome. While z_absence, z_age, z_goout, z_health, z_freetime, z_Walc would bring negatively related to the success in students' performance. On the other hand, the Gains Chart did show that the trend is not really far off from the average line; however, this can be attributed to the facts that the data volume is really low and may be the model is not real fit.






```
intercept  -0.6399650857651512
       z_absences  z_studytime     z_age   z_goout    z_Fedu  z_health  \
coeff   -0.248603     0.460163  0.202708 -0.117768  0.204816 -0.263899

       z_freetime    z_Medu    z_Walc  z_failures  z_famrel
coeff   -0.003929  0.266664 -0.155486   -1.486059  0.151814

AIC 383.32830388406006
```

Final Model

Random Forest gave good insights on factors which have higher influence on the Grades, with decent accuracy of 37%. Almost all the other methods gave the accuracy from 33% -37%.
We decided to pursue **multiple regression** with the attributes identified using **Random Forest**, which gave a decent understanding of the coefficients and more control on the factors which could be influenced to better the grades.

Learnings and Insights

Predicting individual grades precisely was challenging. Therefore, we focused on predicting an approximate grade range using binning methods, addressing multicollinearity among variables.

In our recent project, we significantly improved our data cleaning skills by implementing a series of strategic steps. Our journey began with identifying and removing extraneous factors from our dataset. This process involved:

- Dropping unnecessary variables to concentrate on G3- final grade.
- Employing data cleaning techniques to prepare our dataset.
- Testing various strategies to identify the most effective predictive model.
- Choosing the multiple regression model for its simplicity and informative coefficients.
- Using the random forest model to identify the most influential variables.
- Assuming that final grades (G3) are influenced by Grade 1 and Grade 2, our analysis centered on G3.

- To reduce bias, we removed variables with high correlation but questionable predictive power, such as 'higher_education'.
- We conducted a grade range analysis to uncover insights. Factors like 'going out' and 'absences' emerged as controllable elements that could potentially improve grades.

This approach, focusing on influencing factors rather than just prediction accuracy, allowed us to identify actionable insights for academic performance improvement.

Our study led to a significant finding: student grades are influenced by a variety of socio-demographic factors and their interest in studies. To translate these insights into actionable recommendations, we relied on variables identified by the Random Forest (RF) tree. This approach not only yielded a similar Root Mean Square Error (RSME) to our backward elimination model but also highlighted factors we can actively influence. Our suggestions include:

- Minimizing student absences.
- Maximizing study time.
- Encouraging students to reduce social outings.
- Promoting healthy lifestyle choices.
- Reducing weekend alcohol consumption.
- Having more family support.

These measures proved effectively, offering better accuracy than a naive model and crucial insights into factors most influential in our model. We learned that an average predictive model, which allows more control over influential factors, is preferable to a slightly better model with less control.

Importantly, a lower prediction rate does not deter us from focusing on student well-being. Further, we plan to explore factor-to-factor correlations and their cumulative impact on students' grades, such as the interplay between family relationships and alcohol consumption.

However, our study has limitations due to the data's scope. The observations are confined to our sample dataset and may not be universally applicable. Future research with more extensive data will help isolate the most impactful factors and validate these trends across a broader population.

After school programs

We have identified key factors impacting student grades, such as study time, alcohol consumption, and health. To address these, we propose establishing an after-school program. This program would offer a secure environment for studying and socializing, provide necessary resources like internet access, and facilitate additional study sessions. It would particularly benefit students who are struggling or need to catch up, including older students.

While this initiative might require additional teacher resources, the cost would be offset by utilizing existing school facilities. We can also engage community volunteers for supervision. Providing healthy snacks could improve nutrition for students who otherwise might miss a meal. In this structured setting, we can promote healthy lifestyles and educational activities, thereby creating a foundation for their academic success

Internet Access

Internet access was one of the features that seemed to impact grades as well. This might be an indirect indicator of socioeconomic status of the student; however, this still does not negate the fact that having the internet is important to learning.

This seems to be a cost-effective way to subsidize student learning. This allows students to research topics on their own as well as provide a more level playing field for learning. Without the internet students are pretty limited. Students also might not have a computer so we might need to provide this as well.

Student Housing

The analysis indicates that factors such as travel time, absences, free time, study time, internet access, and alcohol consumption significantly impact student grades. An ambitious yet potentially impactful solution is the establishment of student housing. This approach, albeit more costly than other recommendations, addresses variables that are otherwise challenging to influence. For instance, on-site mentors, possibly teachers, could provide guidance to students lacking educational support at home.

The housing facility could enforce strict rules to curb negative behaviors like alcohol consumption and promote health, notwithstanding the risk of communicable diseases – a concern that exists in regular school settings as well. While this solution requires significant investment, its potential to create a controlled, supportive environment could yield substantial improvements in student outcomes. We could also explore phased implementation or alternative, less costly measures to achieve similar objectives.

Family factors

Family factors as parents' occupation also affect students' performance as parent's jobs are shown in our model. For example, if parents have flexible schedules or work from home, they will have more availability to provide support and assistance with schoolwork for their children. In addition, when parents work in education related fields, they will have better understanding of the educational system and also better connections to have better equipment or ways to support their children's learning so their children can improve their performance in school. On the other hand, having more family time will provide support and nurturing environments which absolutely help their children have academic success while students can have emotional support and communications for improving their learnings, difficulties and stress reduction.

Inspire through Opportunity

One key feature "reason" was indicated as an influential impact on final grades. If we show students why they should pursue education with successful members of the community in interesting positions, they might change their reason from things like "close" to something more meaningful. Instilling this understanding early on especially in secondary school is useful because they can still change their course in terms of underlying motivation before graduating.

Teaching the Teachers

An analysis of student grades (G3- final grade) reveals that Gabriel Pereia (GP) consistently outperforms Mousinho Da Silveira (MS). While other factors might influence these outcomes, a close examination of standardized test performance could provide insights into the effectiveness of GP's educational strategies. If a correlation exists, we could leverage this opportunity to enhance SV's performance.

Transforming the future of students doesn't necessarily require expensive solutions:

Rather, it demands well-considered strategies that directly impact key outcomes, such as academic performance or student wellbeing. For instance, simple changes like adjusting classroom layouts or implementing peer mentoring programs can have significant effects, demonstrating that thoughtful, targeted actions often outweigh mere financial expenditure. We then split the data into both training and validation datasets.