Predicting academic performance using demographic and behavioral Data

Zhengling Jiang, Colombe Tolokin, Franklin Aryee, Tien Nguyen

Table of contents

0.1 Summary

This project investigates whether a student's mathematics performance can be predicted using demographic and behavioral data, aiming to help educators support students and tailor educational strategies. Using a Ridge Regression model with optimized hyperparameters **(alpha = 10.0)**, we achieved strong predictive accuracy with a **cross-validation score of 16.67** and evaluation metrics on the test set including an **MSE of 17.407, RMSE of 4.172, and MAE of 3.272**. The Ridge model was particularly suitable for this task as it effectively handles multicollinearity among features while maintaining model interpretability. While the model demonstrates robust performance, future work could explore non-linear models to capture more complex relationships and provide confidence intervals for predictions, enhancing the model's interpretability and reliability. These improvements could further support educators in making data-informed decisions to optimize student outcomes.

0.2 Introduction

Math teaches us to think logically and it also provides us with analytical and problem-solving skills. These skills can be applied to various academic and professional fields. However, student performance in mathematics can be influenced by many factors, like individual factor, social factor, and family factor. Research has shown that attributes such as study habits, age, social behavior (e.g., alcohol consumption) and family background can significantly impact a student's academic success. Understanding these factors is crucial for improving educational outcomes. (Bitrus, Apagu, and Hamsatu (2016), Hjarnaa et al. (2023), Modi (2023))

In this study, we aim to address this question: **"Can we predict a student's math academic performance based on the demographic and behavioral data?"**. Answering this question is important because understanding the factors influencing student performance can help teachers support struggling students. Furthermore, the ability to predict academic performance could assist schools in developing educational strategies based on different backgrounds of students. The goal of this study is to develop a machine learning model capable of predicting student's math performance with high accuracy.

The dataset (Cortez (2008)) used in this study contains detailed records of student demographics and behaviors, such as age, study habits, social behaviors, and family background. The target variable, mathematics performance, is measured as a continuous score reflecting students' final grade. This dataset offers an excellent opportunity to explore meaningful relationships between features and academic outcomes.

0.3 Methods & Results

The objective here is to prepare the data for our classification analysis by exploring relevant features and summarizing key insights through data wrangling and visualization.

0.3.1 Dataset Description

The full dataset contains the following columns:

1. `school` - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. `sex` - student's sex (binary: 'F' - female or 'M' - male)
3. `age` - student's age (numeric: from 15 to 22)
4. `address` - student's home address type (binary: 'U' - urban or 'R' - rural)
5. `famsize` - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. `Pstatus` - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. `Medu` - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - " 5th to 9th grade, 3 - " secondary education or 4 - " higher education)
8. `Fedu` - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - " 5th to 9th grade, 3 - " secondary education or 4 - " higher education)

9. `Mjob` - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. `Fjob` - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. `reason` - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. `guardian` - student's guardian (nominal: 'mother', 'father' or 'other')
13. `traveltime` - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. `studytime` - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. `failures` - number of past class failures (numeric: n if 1<=n<3, else 4)
16. `schoolsup` - extra educational support (binary: yes or no)
17. famsup' - family educational support (binary: yes or no)
18. `paid` - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. `activities` - extra-curricular activities (binary: yes or no)
20. `nursery` - attended nursery school (binary: yes or no)
21. `higher` - wants to take higher education (binary: yes or no)
22. `internet` - Internet access at home (binary: yes or no)
23. `romantic` - with a romantic relationship (binary: yes or no)
24. `famrel` - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. `freetime` - free time after school (numeric: from 1 - very low to 5 - very high)
26. `goout` - going out with friends (numeric: from 1 - very low to 5 - very high)
27. `Dalc` - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. `Walc` - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. `health` - current health status (numeric: from 1 - very bad to 5 - very good)
30. `absences` - number of school absences (numeric: from 0 to 93)

These columns represent the grades:

- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

Attribution: The dataset variable description is copied as original from the UCI Machine Learning Repository.

0.3.2 Data Loading, Wrangling and Summary

Let's start by loading the data and reviewing the dataset's structure.

The file is a `.csv` file with `;` as delimiter. Let's use `pandas`to read it in.

This provides an overview of the dataset with 33 columns, each representing student attributes such as age, gender, study time, grades, and parental details.

Let's get some information on the dataset to better understand it.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetim
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   school      395 non-null    object
 1   sex         395 non-null    object
 2   age         395 non-null    int64 
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64 
 7   Fedu        395 non-null    int64 
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64 
 13  studytime   395 non-null    int64 
 14  failures    395 non-null    int64 
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64 
 24  freetime    395 non-null    int64 
```

```
 25  goout       395 non-null    int64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
```

The dataset contains 395 observations and 33 columns covering different aspects of student demographics, academic and behavioral traits.

We can see that there is no missing values. There is no need to handle NAs.

The dataset includes categorical (school, sex, Mjob) and numerical (age, G1, G2, G3) features.

There is a large range of features but not all of them are necessary for this analysis. Let's proceed and select only the necessary ones.

Let's selected the following key columns:

- Demographic attributes: sex, age
- Academic Attributes: studytime, failures, G3 (grades for three terms)
- Behavioral Attributes: goout (socializing), Dalc (weekday alcohol consumption), Walc (weekend alcohol consumption)

We will split the dataset into train and test set with a 80/20 ratio then set `random_state=123` for reproducibility.

0.3.2.1 Data Validation Checks

From heatmap shown in Figure 1, we observe no missing values, suggesting the dataset is entirely complete.

The histogram in Figure 2 visualizes the spread of the target variable. This distribution is critical to understanding how the target behaves and whether any transformations are needed to ensure better model performance.

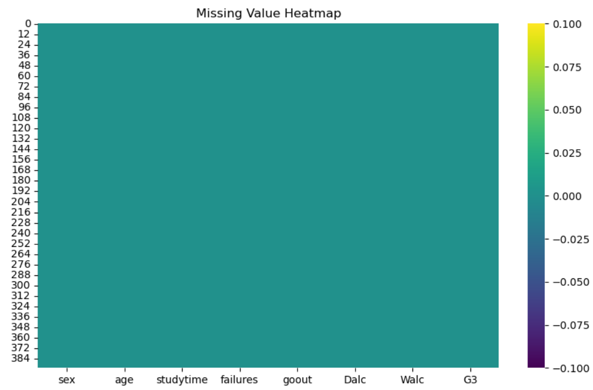

Figure 1: Missing Values Heatmap

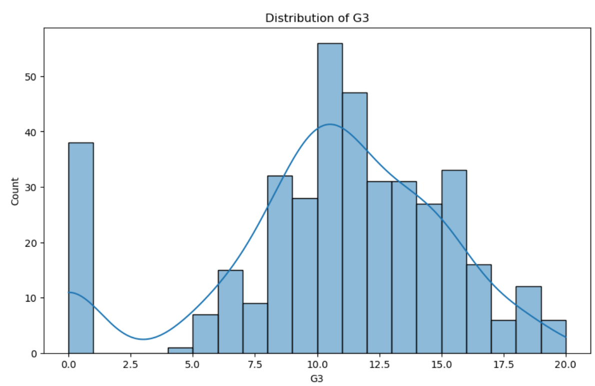

Figure 2: Distribution of the target variable

0.3.2.2 Checking for Outliers

There are few outliers in `failures`, `Dalc`, `age`, `studytime`, `G2`, and `G1`, as shown in Figure 3. Although these outliers are relatively few compared to the 395 entries, they could still influence model results. We will apply a `StandardScaler` transformation to the numeric variables, the effect of these outliers will be minimized. Therefore, we will not drop or modify these outliers at this step.

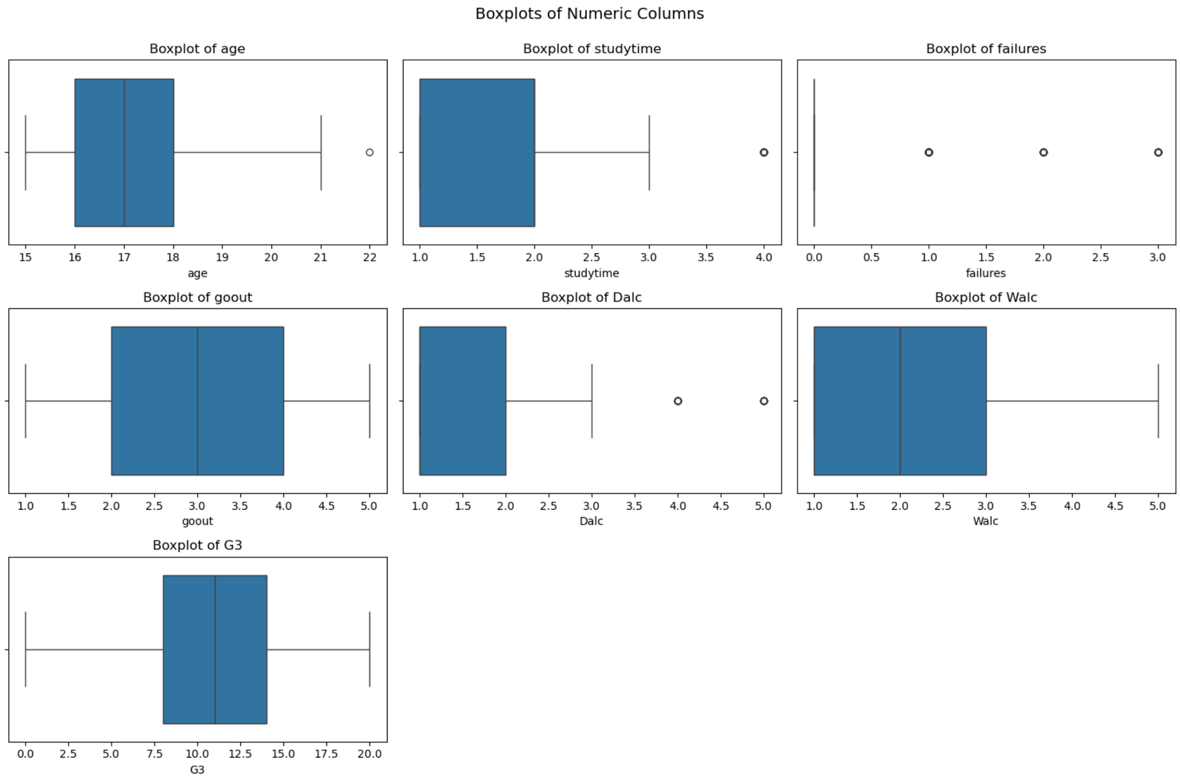

Figure 3: Visualization of Outliers

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 316 entries, 0 to 315
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   sex        316 non-null    object
 1   age        316 non-null    int64
 2   studytime  316 non-null    int64
 3   failures   316 non-null    int64
 4   goout      316 non-null    int64
```

```
 5   Dalc       316 non-null    int64
 6   Walc       316 non-null    int64
 7   G3         316 non-null    int64
dtypes: int64(7), object(1)
memory usage: 19.9+ KB
```

Let's get a summary of the training set we are going to use for the analysis.

Table 2: Summary statistics for columns

	age	studytime	failures	goout	Dalc	Walc	G3
count	316	316	316	316	316	316	316
mean	16.7563	2.05063	0.360759	3.0981	1.47152	2.30696	10.2627
std	1.29006	0.860398	0.770227	1.11833	0.855874	1.2589	4.52268
min	15	1	0	1	1	1	0
25%	16	1	0	2	1	1	8
50%	17	2	0	3	1	2	11
75%	18	2	0	4	2	3	13
max	22	4	3	5	5	5	20

Key takeaways from summary statistics from Table 2:

- The final grade `G3` ranges from `0` to `20`, with an average of around `10.26`.
- The average study time is about `2.05` hours.
- Most students have zero reported failures.
- Alcohol consumption (Dalc and Walc) and socializing habits (goout) appear to vary across the student population.

Let's create a visualization to explore the final grades `G3` distribution. We will use a histogram as it allows us to see the spread.

From Figure 4, The histogram shows that most students achieve grades between 8 and 15, with fewer students scoring very low or very high.

Some interesting observations from Figure 5 :

- The distribution of the grade `G3` is somewhat bell-shaped.
- Most student do not consume alcohol, or very minimally.
- Most students studied around 2-5 hours a week, and most of them also did not fail any previous classes.

Some interesting observations from Figure 6:

- Alcohol consumption is somewhat negatively correlated with grades
- Study time are somewhat positively correlated with grades/

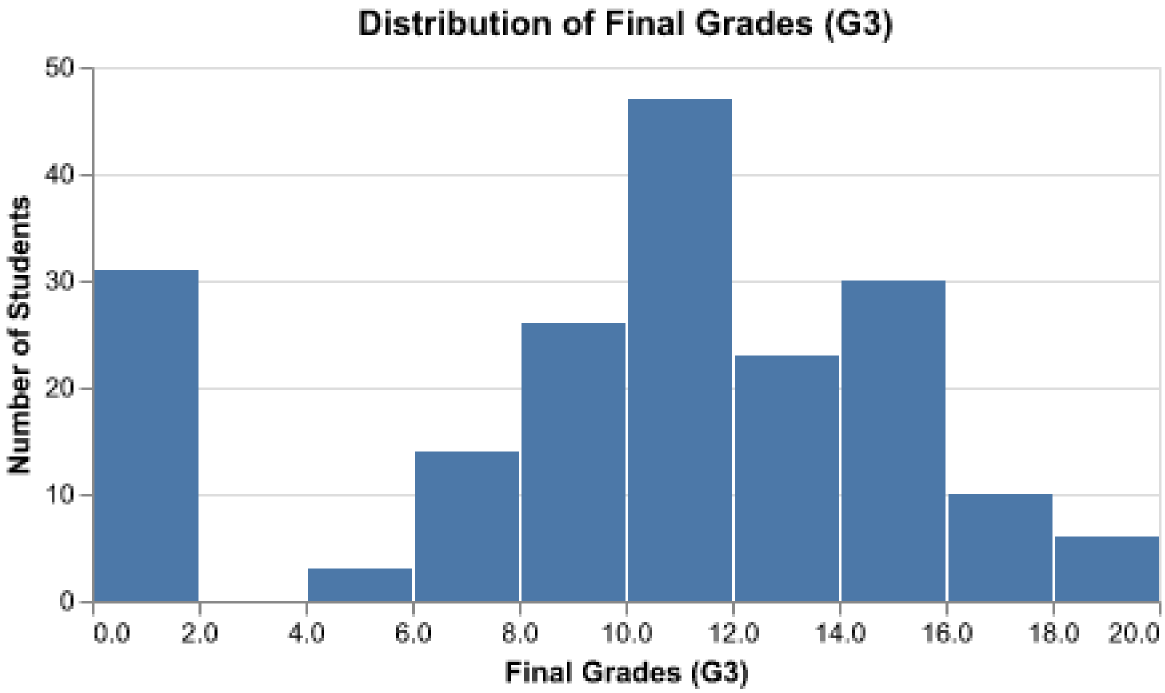

Figure 4: Distribution of Final Grades (G3)

0.3.3 Analysis

We begin our analysis by preparing the data, splitting it into features and target variables for both training and testing. To establish a baseline for comparison, we first fit a DummyRegressor and evaluate its performance, providing a benchmark against which to measure model improvements. Following this, we preprocess the data by distinguishing between categorical and numerical features, applying scaling to numeric features to standardize their range and one-hot encoding to categorical variables to make them interpretable by the model.

Next, we incorporate Ridge regression into a pipeline. Ridge regression is particularly well-suited for this task because it balances model simplicity and predictive performance by penalizing large coefficients. This helps to address potential multicollinearity in the features, ensuring that no single variable disproportionately influences the model while retaining interpretability. To further optimize performance, we fine-tune the Ridge model's hyperparameters using grid search with 5-fold cross-validation, a robust approach for mitigating overfitting and ensuring that the model generalizes well to unseen data.

Finally, we evaluate the Ridge model on the test set, analyzing the observed versus predicted values to assess its predictive accuracy. We also review the cross-validation results to gauge consistency and reliability across different subsets of the data.

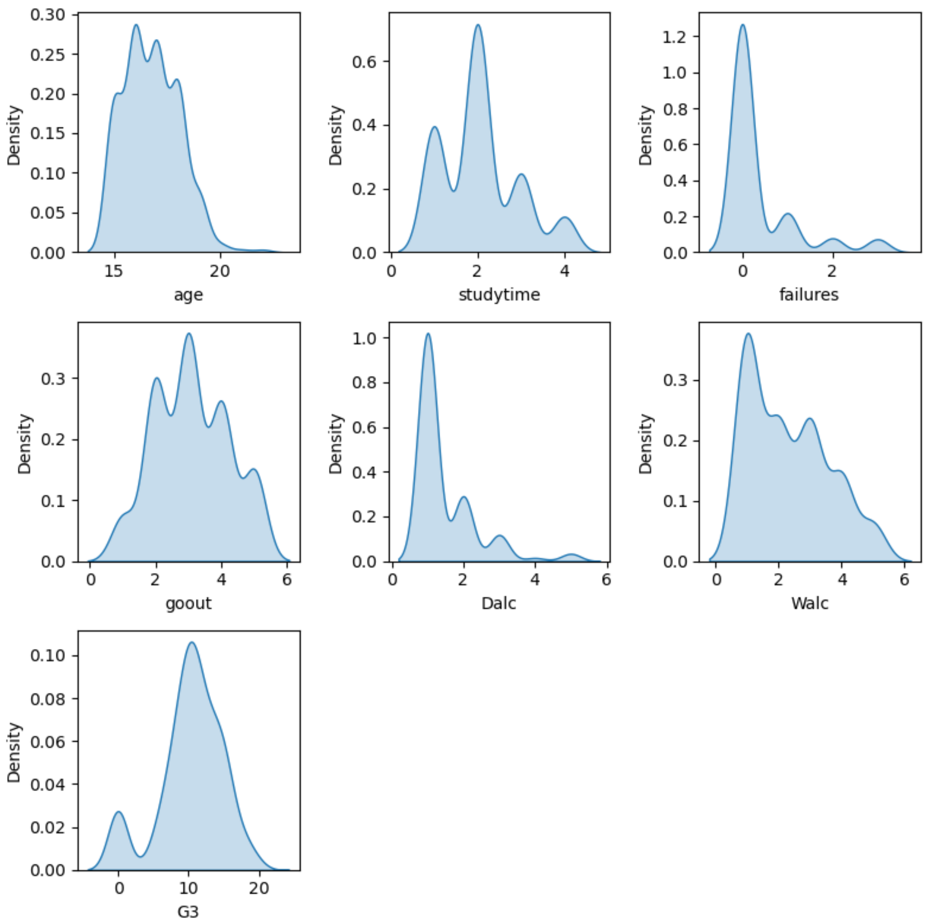

Figure 5: Density plot for each numeric columns

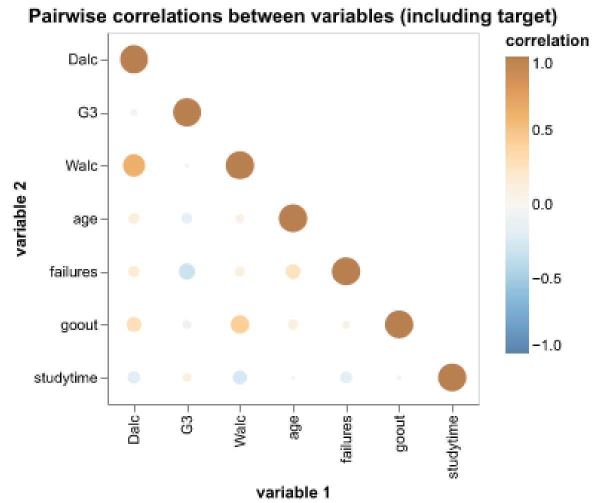

Figure 6: Correlation matrices for each numeric column

0.3.4 Model Evaluation

The Table 3 below summarizes the performance metrics of the model on the test dataset. The metrics used for evaluation are MSE, RMSE, and MAE.

- Mean Squared Error (MSE): The average of squared differences between predicted and actual values, giving more weight to larger errors.
- Root Mean Squared Error (RMSE): The square root of MSE, expressing errors in the same units as the data.
- Mean Absolute Error (MAE): The average absolute difference between predicted and actual values, showing overall prediction accuracy.

We use these metrics to evaluate model performance and understand how well predictions align with actual values, with each providing unique insights into error magnitude and distribution.

Table 3: Performance metrics on test data

Metric	Value
Mean Squared Error (MSE)	17.4068
Root Mean Squared Error (RMSE)	4.17215
Mean Absolute Error (MAE)	3.27234

Next, we analyze the coefficients of the Ridge regression model. The Table 4 shows the values of the coefficients, which indicate the importance of each feature in predicting the target variable.

Table 4: Coefficients of Ridge model

features	coefs
age	-0.199197
studytime	0.621031
failures	-1.16581
goout	-0.81515
Dalc	-0.0512919
Walc	0.254266
sex_M	0.85001

The following Figure 7 visualizes the coefficients of the Ridge regression model. Features with higher absolute coefficients have more impact on the model's predictions.

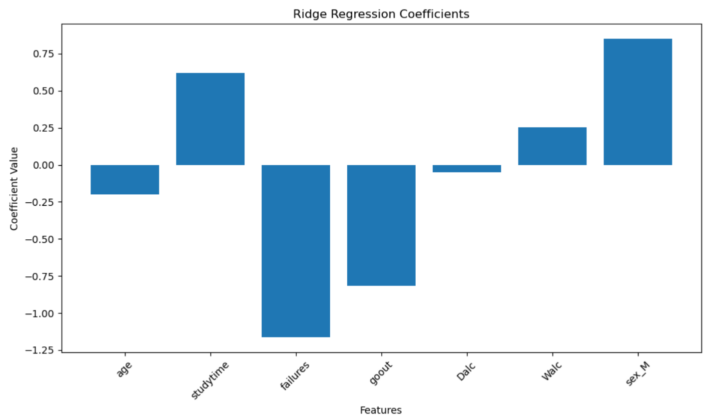

Figure 7: Ridge regression coefficients.

0.4 Results & Discussion

The Ridge Regression model, with tuned hyperparameters, demonstrated well predictive capabilities on student's math performance. The optimal hyperparameter for Ridge was found to be **alpha = 10.0**, and the **best cross-validation MSE** score is approximately **16.67**. This indicates a strong predictive accuracy during the model's validation phase.

Ridge Regression was chosen for the following reasons:

- The presence of correlated features made Ridge a suitable choice, as its L2 regularization shrinks coefficients to stabilize predictions.
- Ridge provides interpretable coefficients, making it easier to identify the most influential factors affecting student performance.

- Ridge Regression serves as a strong baseline for comparison with future models, such as ree-based algorithms or neural networks, which might better capture potential non-linear relationships and complex feature interactions.

Key Influencial Features

The Ridge regression coefficients Figure 7 provide insights into the relative impact of both academic and behavioral factors on student performance:

- `studytime`: The coefficient for study time is the most positive, highlighting that students who dedicate more time to studying tend to achieve higher grades. This aligns with expectations, as focused study enhances understanding and retention of material.
- `failures`: Prior academic failures have the most significant negative impact, indicating that repeated setbacks strongly hinder future performance. This result underscores the need for targeted academic support for struggling students.
- `age`: Age shows a slight negative influence, suggesting older students may face challenges such as balancing responsibilities or staying engaged with coursework.
- `Weekday Alcohol Consumption (Dalc)`: The negative coefficient for weekday alcohol consumption aligns with the idea that drinking during weekdays reduces study time and impairs cognitive performance, especially on critical school days.
- `Weekend Alcohol Consumption (Walc)`: Interestingly, weekend alcohol consumption shows a small positive effect. One hypothesis is that moderate weekend social drinking can act as a stress reliever, improving mental well-being and focus for the upcoming week.
- `Going Out (goout)`: The negative coefficient for socializing (goout) suggests that spending too much time on social activities takes time away from studying, which can hurt academic performance.
- `Gender`: The positive coefficient for "male" (sex_M) indicates a performance difference between genders in this dataset. This result should be interpreted carefully, as it may reflect underlying social, cultural, or educational factors not captured in the current model.

Model Performance

Based on the evaluation on the test set, the model achieved the following performance metrics:

- Mean Squared Error (MSE): 17.407
- Root Mean Squared Error (RMSE): 4.172
- Mean Absolute Error (MAE): 3.272

These evaluation metrics indicate that the model demonstrates reasonable accuracy in predicting students' final grades, with an RMSE of 4.172 suggesting that, on average, the model's predictions deviate from actual grades by about 4.172 points. The MAE of 3.272 further highlights that most errors are relatively small. However, there is still room for improvement since the model is not fully capturing the underlying patterns in the data.

Model Limitations

While Ridge Regression performed well, it has notable limitations that may affect its ability to capture certain relationships in the data:

- Linearity Assumption

Ridge Regression assumes a linear relationship between predictors and the target variable. However, some relationships in the dataset may be non-linear. For example, the impact of study time may exhibit diminishing returns; excessive study could lead to stress or fatigue, reducing its effectiveness.

- Multicollinearity

Ridge Regression helps reduce multicollinearity by shrinking the coefficients of correlated features (e.g., `Dalc` and `Walc`, or `goout` and `studytime`). This improves the model's stability and predictive accuracy. However, multicollinearity can still make it difficult to determine the exact contribution of each correlated feature, as their effects overlap.

- Feature Engineering

Ridge Regression does not automatically capture interactions between features. For example, the combined effect of socializing and alcohol consumption might impact performance in a way that the current model overlooks.

Model Improvement

To further enhance the model's robustness and interpretability, incorporating confidence intervals for predictions is a valuable next step. Confidence intervals would quantify the uncertainty around each prediction, helping stakeholders understand the range within which the true outcomes are likely to fall. This would improve trust in the model's reliability and support better decision-making, especially in real-world applications where uncertainty matters.

References

Bitrus, GA, KB Apagu, and PJ Hamsatu. 2016. "Marital Status and Age as Predictors of Academic Performance of Students of Colleges of Education in the Nort-Eastern Nigeria." *American Journal of Educational Research* 4 (12): 896–902.

Cortez, Paulo. 2008. "Student Performance." 2008. https://doi.org/10.24432/C5TG7T.

Hjarnaa, Louise, Sanne Pagh Møller, Alberte Brix Curtis, Ulrik Becker, Ove Andersen, Fartein Ask Torvik, and Janne Schurmann Tolstrup. 2023. "Alcohol Intake and Academic Performance and Dropout in High School: A Prospective Cohort Study in 65,233 Adolescents." *Journal of Adolescent Health* 73 (6): 1083–92.

Modi, Yushi Girishkumar. 2023. "The Impact of Stress on Academic Performance: Strategies for High School Students." *International Journal of Psychiatry* 8 (5): 150–52.