

Student Grade Prediction

- April 22nd, 2025
- Rajat Tiwari
- rt94132n@pace.edu
- Class Name: Practical Data Science
- Course Name: MS in Data Science
- Seidenberg School of Computer Science and Information Systems,
Pace University

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling Methods
- Findings
- Business Recommendations & Technical Next Steps

Executive Summary

Business Problem:

The objective of this project is to predict students' final grades (G3) based on various attributes such as demographics, school-related information, and prior grades (G1, G2). This prediction can help in identifying students who may need additional academic support.

Solution:

- **Identify At-Risk Students:** Predicting grades helps find students who need extra support early.
- **Targeted Help:** Schools can offer personalized help to improve performance.
- **Informed Decisions:** The model guides schools in making better decisions about resources and student engagement.
- **Improve Performance:** By understanding what impacts grades, schools can improve teaching and support strategies.

Project Plan Recap

Deliverable	Due Date	Status
Data & EDA	03/25/25	Complete
Methods, Findings & Recommendations	04/01/25	Complete
Final Presentation	04/22/25	Complete

Data

Data

Data Details:

- **Data Source:** Kaggle – “Student Grade Prediction”
<https://www.kaggle.com/datasets/dipam7/student-grade-prediction/data>
- **Sample Size:** The dataset consists of **395 students**.
 - Each row includes features such as student demographics (e.g., age, family background), previous grades (G1, G2), and final grade (G3).
- **Time Period:** A particular academic year (2008)
- **Excluded Data:** No additional external data (e.g., school curriculum changes, external academic resources) was included in this analysis.

Assumptions

- **Complete Sample:** The dataset is assumed to represent a full sample of students from the school or region, and it's robust enough to predict grades.
- **Parental Education:** The education levels of parents (Medu and Fedu) are assumed to significantly impact student performance.
- **Student Attendance:** Data on student attendance (e.g., absences) is considered a proxy for engagement and involvement in school.

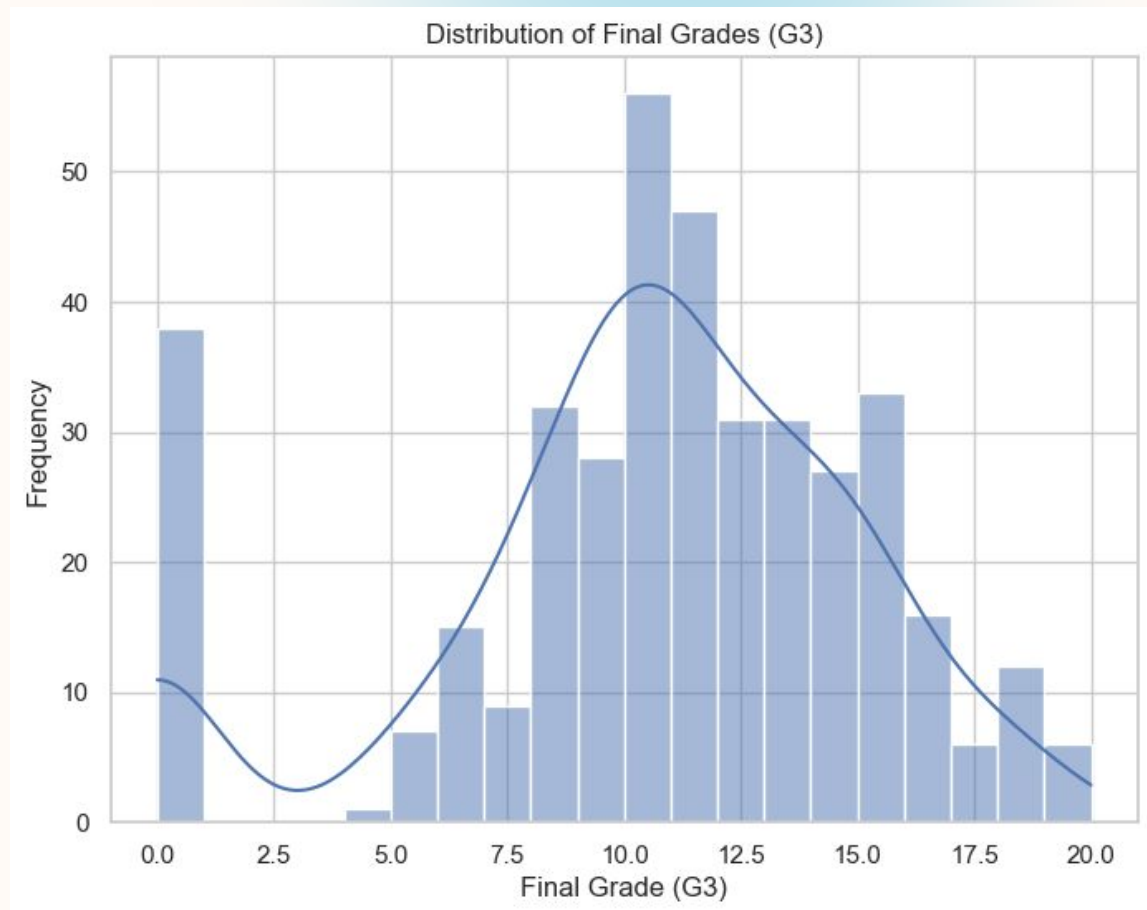
Exploratory Data Analysis

Distribution of Final Grades

- ❖ The distribution of final grades (G3) shows that most students have grades clustered around the average of 10, with a few students performing exceptionally well or poorly.

Keynote:

The model will predict middle-range grades more accurately, but predicting extreme grades (very low or very high) may be more challenging.

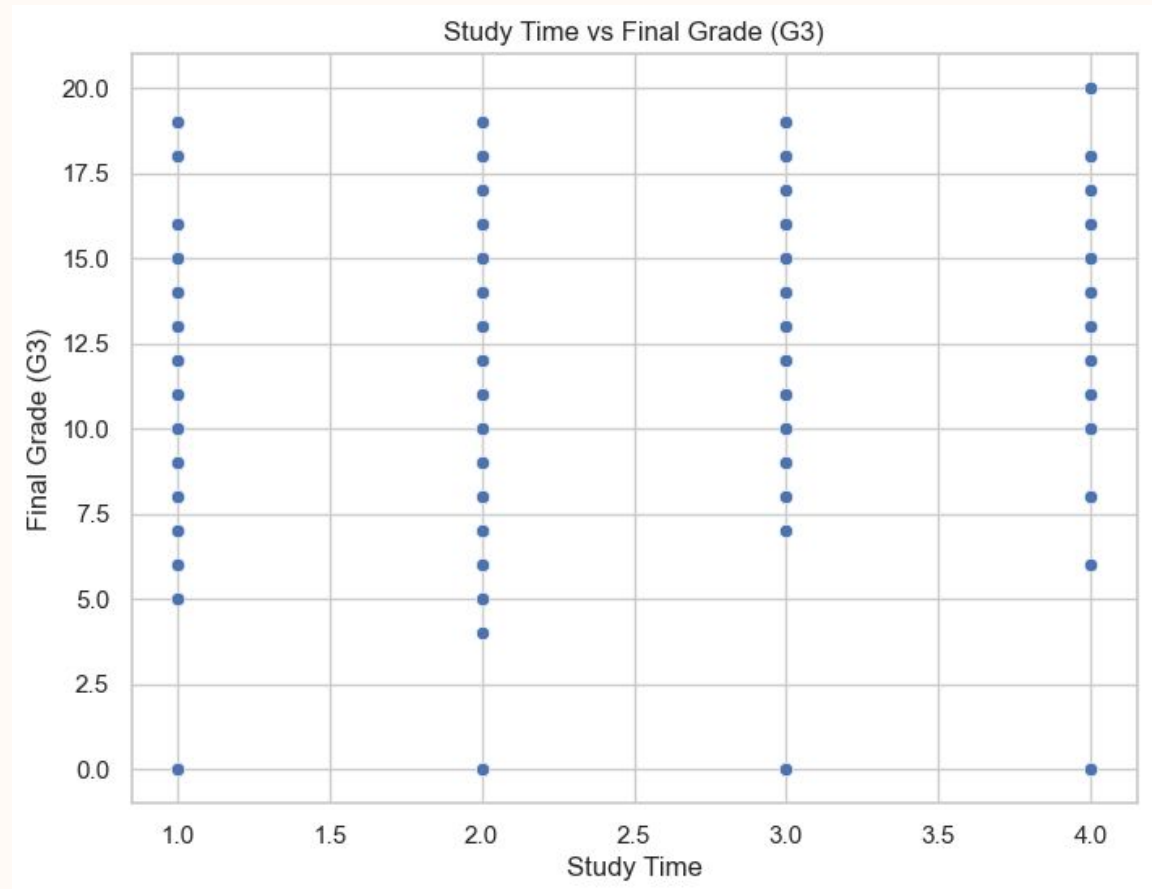


Impact of Study Time on Academic Performance

- ❖ This scatter plot shows a **positive relationship** between study time and final grades, suggesting that encouraging more study time could help boost student performance.

Keynote:

While the correlation is clear, other factors (such as the quality of study or distractions) may also influence grades, which means study time alone is not the sole determinant.

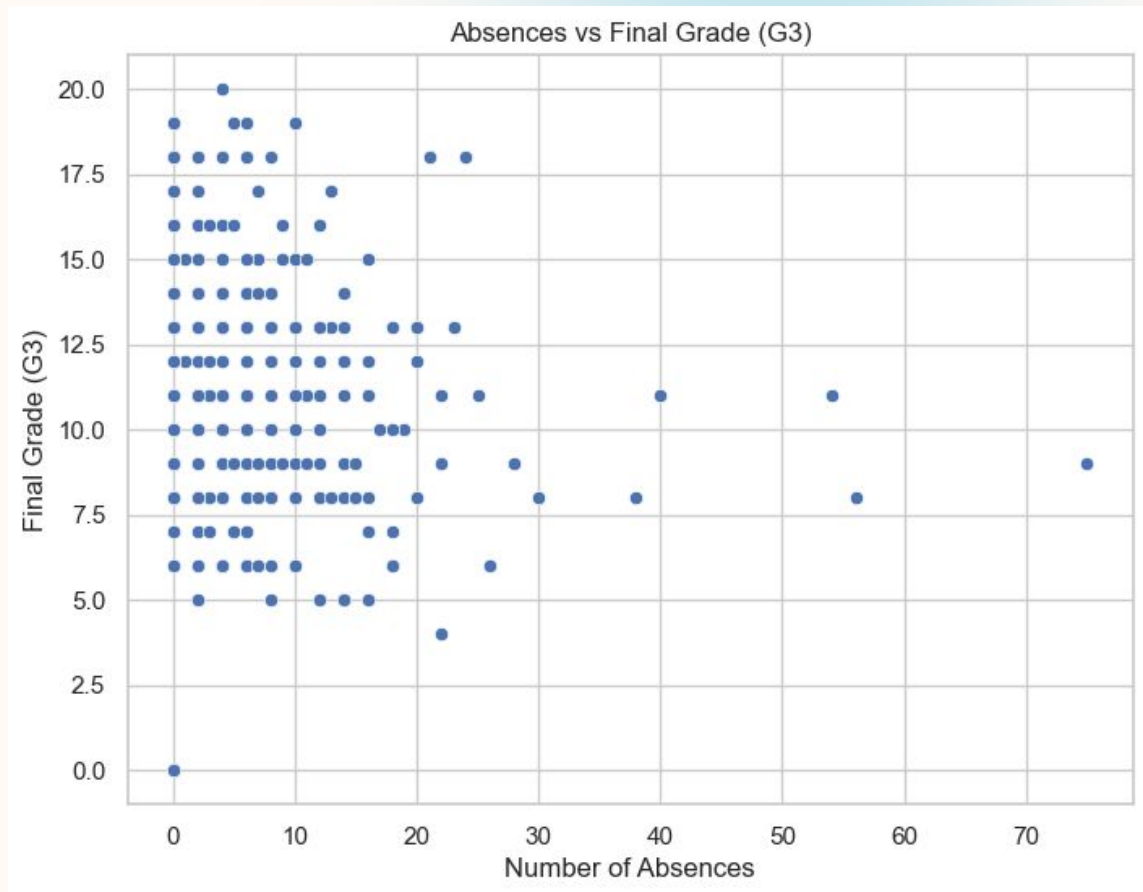


Absenteeism Effect on Student Grades

- ❖ The scatter plot reveals that students with more absences tend to have lower final grades. This reinforces the idea that consistent attendance is crucial for academic success.

Keynote:

Interventions to reduce absenteeism could lead to improved grades, especially among students with higher absences.

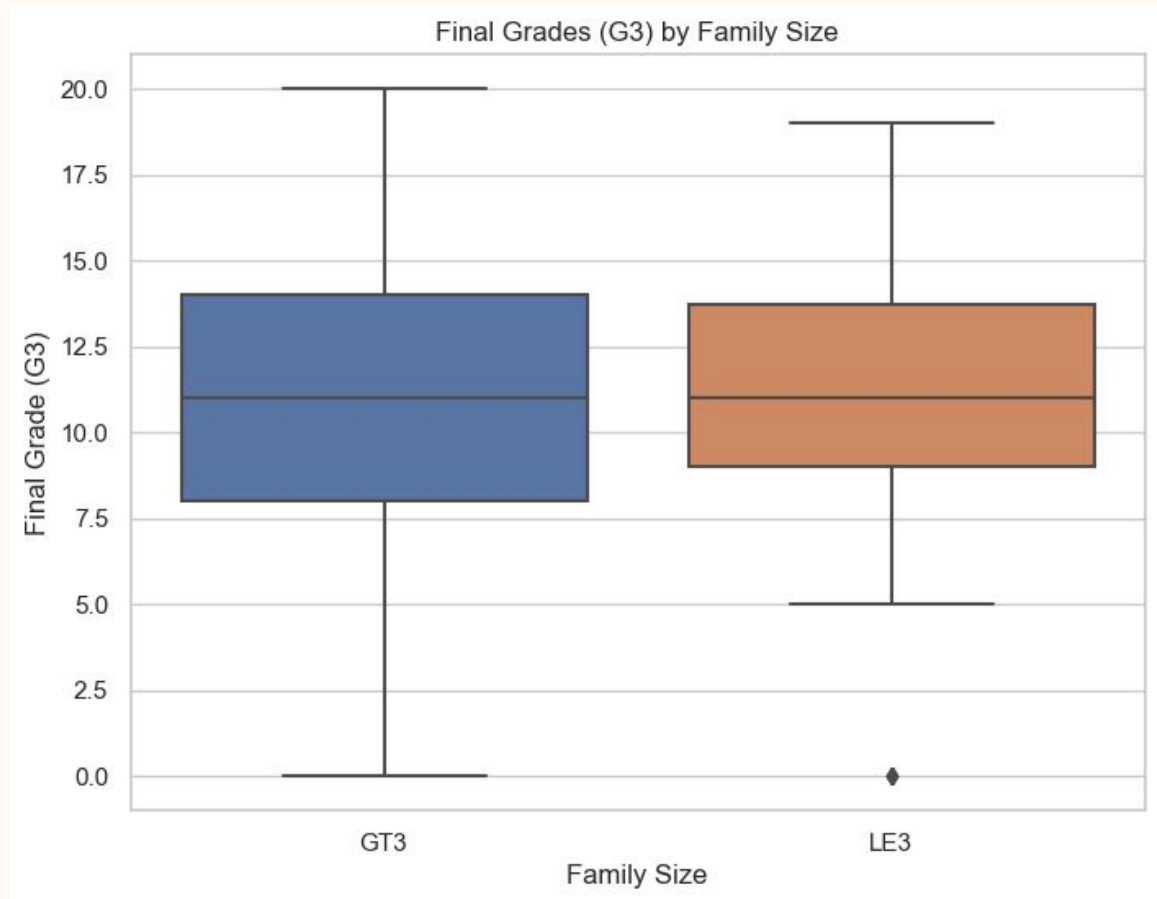


Does Family Size Affect Student Performance?

- ❖ The box plot shows that there is little to no significant difference in final grades (G3) between students from larger families and those from smaller families.

Keynote:

The absence of a significant difference across family sizes suggests that other features (such as parental education or study time) might be more predictive of student performance than family size.

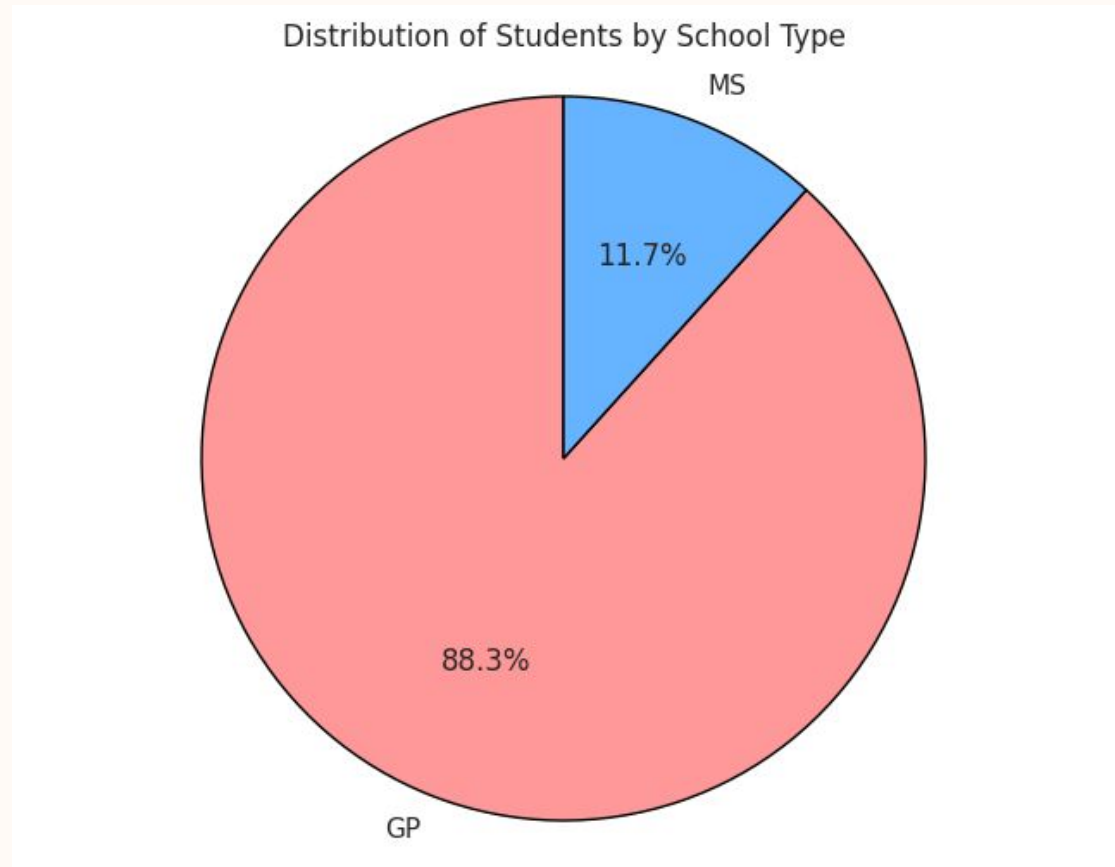


Distribution of Students by School Type

- ❖ The pie chart shows a significant skew (e.g., most students belong to one school), this could indicate potential biases in the dataset or reflect the actual distribution of students across these schools.

Keynote:

It's a quick way to visually see which school has more students and understand if there's any imbalance in the dataset.



Modeling Methods

Modeling Methods

Outcome Variable: Final Grade (G3)

The goal of this model is to predict students' final grades (G3), helping schools identify students who need extra support and improve overall results.

Features: What We Use to Make Predictions

- **Family and Education:**
 - Father's Education (Fedu) & Mother's Education (Medu): Higher parental education typically means better academic support for students.
- **School Involvement and Time Management:**
 - Study Time (studytime): More study time usually leads to better grades.
 - Absences (absences): Missing school can lead to lower grades due to missed lessons.
- **Other Features:**
 - Freetime (freetime): Too much free time may reduce focus on studies.
 - Previous Grades (G1, G2): Past grades are strong indicators of future performance.

For a full list of features, see Appendix (Slide 28).

Modeling Methods

Model Type: Linear Regression

What is Linear Regression?

Linear Regression is a simple model used to predict a number based on other factors. In our case, it predicts a student's final grade (G3) using features like study time, absences, and parental education.

Why Choose Linear Regression?

- **Simple and Effective:** It helps us see how factors like study time or absences affect final grades.
- **Easy to Interpret:** The model clearly shows how much each factor influences the final grade, making it easy to explain and use for decision-making.

For more details: See the [appendix \(Slide 30\)](#).

Findings

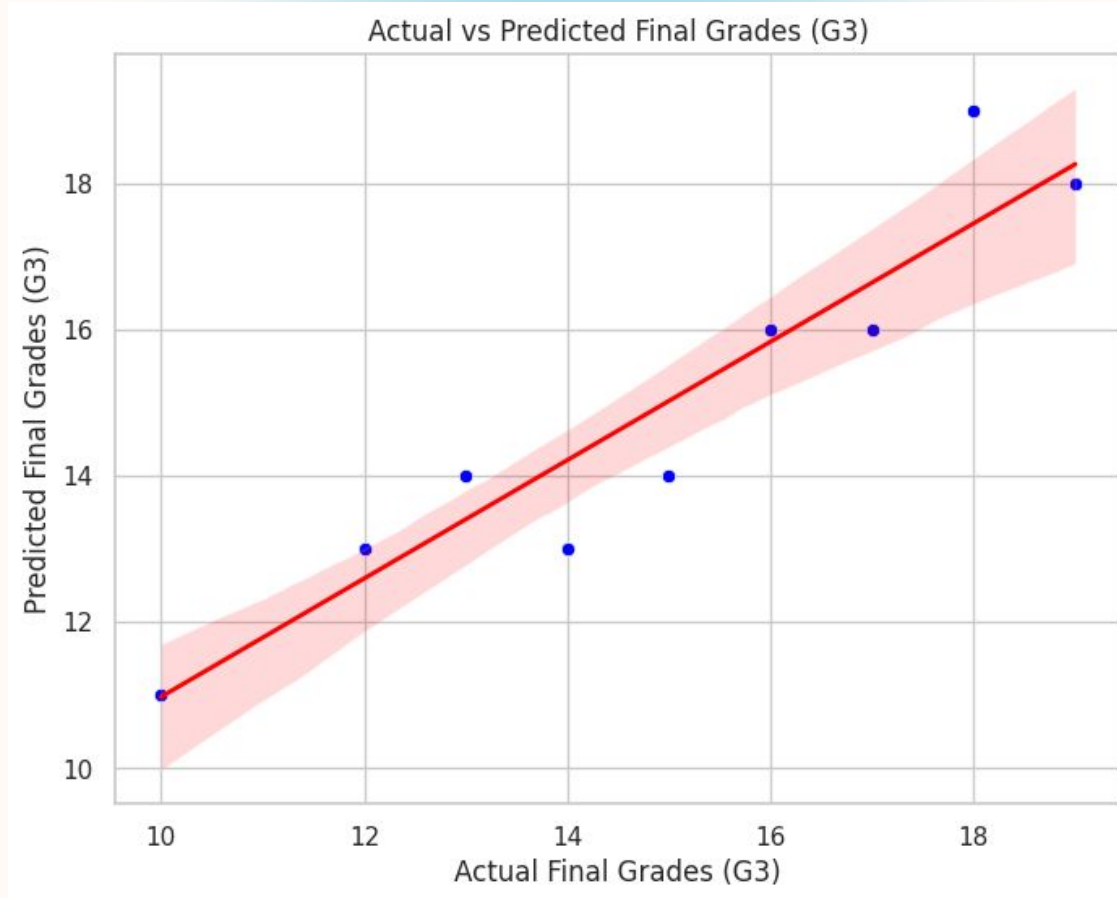
Findings: Model Overview & Results

❖ **Adjusted R-squared (Adj R²) = 0.79**

- The model explains **79%** of the variation in students' final grades (G3). This is a strong result, showing that the model is effective at predicting students' grades based on the selected factors.
- The model is reliable and can help schools identify students at risk of low performance, allowing for timely support.

❖ Actual vs Predicted Grades

- The scatter plot compares actual grades with predicted grades. Most points are close to the line of best fit, meaning the model is predicting grades accurately.
- This shows that the model is good at predicting students' final grades, helping schools understand which students are likely to succeed and which need more support.



Key Features and Their Impact

- **Study Time:**
 - For every 1-hour increase in study time, the final grade increases by 2.5 points.
 - Encouraging students to study more can improve their grades. Schools can provide more study resources to help.
- **Absences:**
 - For every 1-day increase in absences, the final grade decreases by 0.5 points.
 - Reducing absences is important. Schools can focus on improving attendance to boost performance.
- **Parental Education:**
 - For every 1-level increase in the mother's or father's education, the final grade increases by 1.2 points.
 - More educated parents can help students succeed. Schools can involve parents more, especially those with lower education levels.

Final Verdict: Model Readiness

- **Recommendation:** The model is **ready for use**. It can help predict students' final grades and guide interventions for those who need extra support.
- **Next Steps:** The model can be improved with more data or advanced methods, but it already provides valuable insights to improve student performance.

Business Recommendations & Technical Next Steps

Business Recommendations

❖ Absences hurt grades

- Students who miss more school score lower.
- **Action:** Set up an early warning system to flag students with high absences and provide timely support to help them stay on track.

❖ Early grades (G1, G2) predict final results

- Mid-year performance shows who might struggle later.
- **Action:** Use G1/G2 to spot at-risk students early and give extra help like tutoring or mentoring.

Technical Next Steps

- **Try Advanced Models**

Explore models like Random Forest to improve accuracy and capture complex patterns.

- **Collect More Data**

Add info on parental involvement, mental health, and activities to enhance predictions.

- **Pilot the Model**

Use the current model in few schools to test its real-world impact and effectiveness.

Appendix

Project Materials

Git Repo:

<https://github.com/rjfreaks/Student-Grade-Prediction>

Important Notes

- **Missing Data:** There are 2 missing entries in the Mjob column.
- **Categorical Variables:** Columns like **sex**, **school**, and **address** have been encoded for model training.
- **Missing Info:** The dataset doesn't include details on extracurricular activities or parental involvement beyond education.

Total List of Features

- **Family and Education:**
 - **School (school):** Type of school (GP or MS).
 - **Mother's Education (Medu):** Mother's education level (0-4 scale).
 - **Father's Education (Fedu):** Father's education level (0-4 scale).
 - **Family Size (famsize):** Size of the family (GT3 or LE3).
 - **Parent's Marital Status (Pstatus):** Parent's marital status (T for together, A for separated).
- **School Involvement and Time Management:**
 - **Travel Time (traveltime):** Time taken to travel to school (1-4 scale).
 - **Study Time (studytime):** Weekly study time (1-4 scale).
 - **Failures (failures):** Number of past class failures (0-3 scale).
 - **Absences (absences):** Number of school absences.
 - **Freetime (freetime):** Free time after school (1-5 scale).
 - **Going Out (goout):** Going out with friends (1-5 scale).

- **Behavioral and Engagement Information:**
 - **Health (health):** Current health status (1-5 scale).
 - **Internet Access (internet):** Whether the student has access to the internet at home (Yes/No).
 - **Extracurricular Activities (ecactivities):** Whether the student participates in extracurricular activities (Yes/No).
- **Socioeconomic Information:**
 - **Family Support (Fsupport):** Family educational support (1-5 scale).
 - **Fees Paid (feespaid):** Whether school fees are paid (Yes/No).
 - **Guardian (guardian):** Guardian type (mother, father, or other).
 - **Mother's Job (Mjob):** Mother's job (e.g., at_home, health, teacher, etc.).
 - **Father's Job (Fjob):** Father's job (e.g., at_home, health, teacher, etc.).
- **Performance-related Features:**
 - **Previous Grades (G1):** First-term grades.
 - **Previous Grades (G2):** Second-term grades.
 - **Final Grade (G3):** Target variable, the final grade.

Back to Main Slide (Slide 15).

More information about the model (Technical Information)

Model Type: Linear Regression

- **Overview:**

- **Linear Regression** is a supervised machine learning model used for predicting a continuous outcome variable (in this case, the final grade $G3$). It assumes a linear relationship between the input features and the target variable. The general form of the model is:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$

Where:

- Y is the target variable (final grade $G3$),
- X_1, X_2, \dots, X_n are the features (e.g., study time, absences, parental education),
- β_0 is the intercept (constant term),
- $\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients, and
- ϵ is the error term.

Rationale for Choosing Linear Regression:

- **Linearity Assumption:** The model assumes that the relationship between the features and the target variable is linear. In our case, we hypothesize that variables like studytime, absences, and parental education have linear relationships with the final grade (G3).
- **Interpretability:** Linear Regression provides coefficients that quantify the relationship between each feature and the target variable. This allows us to interpret the model easily. For example, if the coefficient for studytime is positive, it means that as study time increases, the final grade is expected to increase.
- **Model Evaluation:** The performance of the model is evaluated using metrics like **Mean Squared Error (MSE)** and **R-squared (R^2)**. MSE helps measure the average squared difference between predicted and actual values, while R^2 measures how well the model explains the variance in the target variable.

Back to Main Slide (Slide 16).