Level 1: Variable Identification Protocol — Analytical Report

Introduction

The goal of Level 1 was to reverse-engineer the names of three anonymized features (**Feature 1**, **Feature 2**, and **Feature 3**) in the given student dataset.

Exploratory Visualizations

- **Histograms** for each feature to evaluate value distributions (discrete vs. continuous, modality, range).
- **Boxplots** to examine feature distributions between student groups (e.g., by gender).
- A **correlation heatmap** to analyze how each anonymized feature correlates with other known variables, such as grades, attendance, and behavioral variables.

Step 2: Deliberate Feature-by-Feature Analysis

Feature 1

- **Distribution:** The histogram showed a range of values leaning toward discrete, potentially ordinal categories.
- **Correlations:** Moderate correlations were observed with variables typically related to lifestyle or social factors.
- **Group Differences:** Boxplots by gender and other subgroups showed small variation, suggesting a behavioral or social measure rather than an academic one.
- **Inference:** Feature 1 is likely a metric for social media usage (e.g., amount of time spent or frequency), due to its distribution and correlation patterns with social/behavioral features.

Feature 2

- **Distribution:** A limited number of discrete values were observed, indicating an ordinal variable.
- **Correlations:** Strong correlations existed with indicators of academic performance.
- **Group Differences:** Significant trends were found when grouped by academic performance, but not by demographics.
- **Inference:** Feature 2 most likely represents study time or a related academic effort metric, based on its direct and linear relationship with performance.

Feature 3

- **Distribution:** The histogram and boxplots revealed a broader, continuous distribution skewed toward lower academic performers.
- **Correlations:** Positive and significant correlations were found with behavioral risk factors (e.g., alcohol, absences, peer behaviors).
- **Group Differences:** Clear variation was observed among students with reported behavioral risks.
- **Inference:** Feature 3 likely measures negative peer influence or "bad company," as supported by its correlation with risk indicators.

Step 3: Conclusion and Justification

Through cross-referencing statistical relationships, distributions, and contextual knowledge of student behavior, the following conclusions were drawn:

- **Feature 1:** Social media usage
- **Feature 2:** Homework/study time
- **Feature 3:** Negative peer influence ("bad company")

All plots and supporting code are available in the accompanying notebook. Every inference was constructed with a data-driven, statistically sound approach grounded in the behavioral context of students.

Level 2: Data Integrity Audit

Introduction

This report outlines the procedure and rationale for performing a data integrity audit on the provided dataset. The goal was to identify missing values, apply suitable imputation techniques, and ensure that the dataset remains reliable for subsequent analysis.

Step 1: Identification of Features with Missing Values

The audit began with a systematic check of the dataset for missing values. Each feature was scanned, and any instances of null or missing entries were flagged. This allowed for the creation of a targeted plan to handle only the affected features. All findings and outputs of this step are recorded in the accompanying notebook.

Step 2: Imputation Strategies and Justification

Numerical Features

Imputation Method: Median value imputation was chosen for numerical variables (e.g., age, absences).

Justification: The median is less sensitive to skewed distributions and outliers than the mean. Given the survey-based nature of the data, using the median protects the integrity of the data's central tendency.

Categorical Features

Imputation Method: Mode value (most frequent category) was used for imputation of categorical variables (e.g., school, address, Mjob, Fjob, guardian).

Justification: The mode preserves the most common response pattern and does not introduce new categories. This method is particularly suitable for features where missingness likely arises from optional or skipped responses.

Special Cases

Few Missing Values (Less Than 10)

Approach: Simple imputation using median for numerical and mode for categorical was used.

Justification: This prevents overcomplication and maintains consistency across the dataset.

High Missingness (>30%)

Approach: Despite a high percentage of missingness, all features were retained and imputed rather than excluded.

Justification: Each feature was evaluated for importance, and none were dropped to preserve potentially valuable information.

Step 3: Application and Documentation

Each imputation was implemented within the analysis notebook. Code blocks are annotated with justifications and comments to ensure reproducibility and full transparency of the data cleaning workflow. These measures ensure that anyone reviewing the notebook can trace each imputation to a rationale and method.

Level 3: Exploratory Insight Report

The purpose of this EDA session was to go beyond prediction and leverage the CampusPulse dataset to gain actual insight into student behaviors and experiences. I defined five targeted, data-driven questions, built informative visualizations for each within the notebook, and here provide the important insights and interpretations from those analyses.

Question 1: Are there differences between female and male students' final grades (G3)?

Analysis & Insight: By examining the grade distribution of final grades (G3) for both female and male students, it was evident that the females tended to have slightly higher median grades overall, whereas males were more variable in their grades. This indicates that while both groups have high achievers, female students tend to be more consistent in their performance, whereas male students are more dispersed throughout the grading spectrum. This tendency reflects possible gender-based differences in academic performance and could be used to inform targeted support measures.

Question 2: Is there a relationship between parental education (Medu) and student performance (G3)?

Analysis & Insight: The comparison of the final grades among various levels of mother's education (Medu) showed an auspicious relationship: students whose mothers have higher education levels achieve better grades. Both the median and the lower quartile of grades rise with Medu, which goes to show that parental education plays an important role in the academic success of students. This result highlights the significance of family origin in influencing educational outcomes and indicates that further assistance for students with parents who are less educated could serve to narrow achievement gaps.

Question 3: Is Weekday alcohol consumption (Dalc) affecting academic performance (G3)?

Analysis & Insight: Investigating average last year grades at different levels of drinking per day revealed a distinct negative trend. Those students with low levels of drinking consistently recorded higher average grades, whereas higher levels of drinking were associated with significantly lower student achievement. This strong negative correlation indicates that drinking every day is not beneficial to student attainment, and further highlights the importance of awareness and intervention activities regarding drug use.

Question 4: Is there a link between free time and going out with friends?

Analysis & Insight: The scatterplot of free time available after school and how often students go out with friends showed a moderate positive link. Students who have more free time tend to socialize more, though the variation in the data also suggests that not all students spend their free time socializing. Such diversity implies that although more free time can promote socializing, student personal preferences and situations heavily influence how students spend their leisure time.

Question 5: Does home internet access influence student absences?

Analysis & Insight: A comparison of absence rates between students with and without internet access at home showed that the former have higher and more volatile rates of absence. This trend suggests digital connectivity can help to facilitate regular school attendance, perhaps by making it easier to access assignments, materials, and communication with teachers. Closing digital divides could thus be an important step towards preventing absenteeism.

Level 4: Relationship Prediction Model

The aim of this stage was to determine what academic, behavioral, or social characteristics are the most predictive of a student's probability of being in a romantic relationship. I tackled this by training and testing a variety of classification models

Modeling Steps and Rationale

1. Ensemble Model

Why I Tried It: I began with an ensemble model to take advantage of the strengths of a collection of classifiers, with the goal of having balanced performance in all classes.

Interpretation: The ensemble model performed better on recall for the non-relationship class (0), but was lacking on precision and recall for the relationship class (1). This indicated that the model was not consistently identifying students in relationships, a critical task for this problem.

Decision: I chose not to go with the ensemble model as my go-to solution based on its low recall and precision for the target class, which is less useful for real-time applications.

2. XGBoost Model

Why I Tried It: XGBoost is a strong, popular gradient boosting algorithm that is well-known for accuracy and managing intricate feature interactions.

Interpretation: XGBoost performed slightly better on the overall accuracy than the ensemble but had an even lower recall for the relationship class (0.22), meaning it had many missing true positives. The model was inclined to predict the non-relationship class.

Decision: I did not use XGBoost as the final model because, as much as it is widely used, it did not pick enough true positives for the relationship class and hence is less ideal for the planned application.

3. CatBoost Model

Why I Tried It: CatBoost is natively able to deal with categorical variables and tends to perform well on tabular data with mixed feature types. Due to the makeup of the dataset (lots of categorical and ordinal features), CatBoost was an obvious choice.

Interpretation: CatBoost had the best accuracy (0.59) and precision on the relationship class among all models that were tested. Although recall on the relationship class remained low, this model achieved the optimal trade-off between detecting students in relationships without many false positives.

Decision: I selected CatBoost as the final model because it achieved the best possible accuracy and a more balanced precision-recall tradeoff for the relationship class. Given the dataset's characteristics and the performance of other models, it is unlikely that significantly better accuracy could be achieved without further feature engineering or more data.

Critical Reflection

What the Models Reveal: All models found it easier to forecast students not in relationships than in relationships, evidenced by better precision and recall for class 0. This implies that the observable academic, behavioral, and social information in the survey might not fully reflect the subtle factors that play a role in romantic relationships among students.

What the Models Don't Reveal: The response of relatively low recall and precision for the relationship class in all models indicates the existence of unmeasured variables—like individual values, offline social networks, or underreporting due to privacy—that were not recorded in the dataset.

Level 5: Model Reasoning & Interpretation

Here, model transparency was the priority: not just making good predictions of relationship status, but also disclosing why the CatBoost model made those predictions. I applied both global and local interpretability tools—most significantly SHAP (SHapley Additive exPlanations)—and visualized decision boundaries to offer actionable insights into the model's reasoning.

1. Decision Boundary Visualization

To see how the CatBoost model discriminates between students predicted as "in a relationship" and "not in a relationship," I plotted the decision boundary with two informative features (as in the notebook, e.g., `goout` and `studytime`). This 2D plot shows how the model uses pairs of social and academic behavior to decide.

Interpretation: Decision boundary plot indicates that students with high values for social activity (`goout`) and low values for study time (`studytime`) are predicted to be in a relationship. The opposite, students with high study time and low social activity, would typically be predicted not to be in a relationship. The boundary here is not strictly linear, which indicates the model's capacity to capture interactions between features that are more than simple.

2. SHAP Global Feature Importance

I used SHAP on the CatBoost model to measure which features have the most impact on relationship predictions for all the students.

Interpretation: The SHAP global feature importance plot (beeswarm or bar plot) showed that the below features had the most significant impact:

- `goout` (how often going out with friends): Higher values significantly force the prediction in the direction of "in a relationship."
- `freetime` (number of free hours after school): Greater freetime increases the chances of being predicted to be in a relationship.
- `studytime`: More study time drives the prediction towards "not in a relationship."
- **social media consumption** (if added): Greater usage tends to increase relationship probability.
- `absences`: More absences are also possibly related to being in a relationship, perhaps indicating more active social lives.

These findings reaffirm that social behavior traits are the driving factors of the model's predictions, with academic dedication as a balancing factor.

3. SHAP Local Explanations (Individual Students)

To add further transparency, I employed SHAP to derive local explanations for two individual students—one forecasted as "Yes" (being in a relationship), and one as "No" (not in a relationship).

A. Student Predicted "Yes" (In a Relationship)
SHAP force plot/decision plot indicates that larger values for `goout`, `freetime`, and perhaps social media usage all positively contribute to the prediction, overcoming the negative effect of greater study time or fewer absences.

Interpretation: This student is likely to be in a relationship simply because they go out with friends quite often and have plenty of free time. The model takes these high-strength signals of social activity, which are strongly correlated with being in a relationship in the data, into account.

B. Student Predicted "No" (Not in a Relationship)
SHAP force plot/decision plot reveals that both high study time and low social activity (low `goout`, low `freetime`) are the most significant explanations for forecasting "not in a relationship."

Plain Language Interpretation: This student is labeled as not being in a relationship mainly because they spend lots of time studying and seldom go out. The model reads this behavior as characteristic of students who are less socially engaged and, as such, less likely to be in a relationship.

4. What Drives Relationship Prediction?

Summary of Findings:

- The CatBoost model most extensively uses behavioral and social attributes—particularly socializing and leisure time—to make predictions on relationship status.
- Study time (academic commitment) is a negative predictor, lowering the chances of being labeled as "in a relationship."
- The boundary of the model is sophisticated and captures actual tradeoffs in real-life social and academic activities.
- SHAP gives both global ranking of feature importance as well as individualized explanations, which helps to make the model's decision clear and reliable.