Feature Engineering Report

Explanation of New Features Created

- **Study Time Per Credit**

Definition: This measures the average study time per credit for curricular units taken in both the first and second semesters, indicating student performance relative to their enrolled credits.

Formula: (Curricular units 1st sem (grade) + Curricular units 2nd sem (grade))/ (Curricular units 1st sem (enrolled) + Curricular units 2nd sem (enrolled))

Conclusion: This feature provides insights into student engagement and performance, serving as a valuable metric for academic analysis and decision-making. It can help identify trends, compare groups, and inform strategies to enhance student success.

- **Performance Consistency**

Definition: This quantifies the absolute difference between a student's grades in the first and second semesters, indicating the stability or variability of their academic performance.

Formula: abs (Curricular units 1st sem (grade) – (Curricular units 2nd sem (grade)

Conclusion: This feature helps assess how consistent a student's performance is across semesters. A lower value indicates more consistent performance, while a higher value suggests greater variability, providing insights into student stability and potential areas for intervention.

- **Performance ratio**

Definition: This measures the proportion of approved curricular units to the total enrolled units across the first and second semesters, indicating overall academic success.

Formula: (Curricular units 1st sem (approved) + (Curricular units 2nd sem (approved)/ (Curricular units 1st sem (enrolled) + (Curricular units 2nd sem (enrolled)

Conclusion: This feature provides insights into a student's effectiveness in completing their enrolled courses. A higher Performance ratio indicates better academic performance and course completion rates, which can inform academic advising and support strategies.

- **Study Engagement**

Definition: Study Engagement quantifies a student's overall engagement in their curricular units by summing the number of enrolled, credited, and evaluated units across both semesters.

Formula: (Curricular units 1st sem (enrolled)) + (Curricular units 1st sem (credited)) + (Curricular units 1st sem (evaluations)) + (Curricular units 2nd sem (enrolled)) + (Curricular units 2nd sem (credited)) + (Curricular units 2nd sem (evaluations))

Conclusion: This feature provides a comprehensive measure of student involvement in their academic activities. Higher values indicate greater engagement, which can be correlated with academic success and retention, offering insights for targeted support and interventions.

- **Dropout risk**

Definition: Dropout risk assesses the potential risk of student dropout based on their tuition fee status and debtor status, assigning a risk value that reflects financial obligations.

Formula: df['dropout_risk'] = np.where(

df['Tuition fees up to date'].isna(),

df['Debtor'] * (1 - df['Tuition fees up to date'])

Conclusion: This feature provides insights into financial factors that may influence student retention. A higher dropout risk value indicates greater financial pressure, which could correlate with higher dropout rates, aiding institutions in identifying at-risk students for intervention.

- **Age at enrollment * Admission grade (For Interaction Terms)**

Definition: Age at enrollment * Admission grade calculates the product of a student's age at enrollment and their admission grade, providing a composite metric that may reflect the impact of age and academic performance on student outcomes.

Formula: {Age at enrollment} *{Admission grade}

Conclusion: This feature can be useful for analyzing how the combination of age and admission performance influences academic success and retention. It may help identify trends related to age demographics and their relationship with academic achievement.

- **Polynomial features Transformation**

Description: This process involves creating polynomial features of degree 2 from selected numerical columns in the dataset. By applying this transformation, I enhance the dataset's ability to capture non-linear relationships among variables such as age, grades, and economic indicators. The original numerical columns are transformed into a new set of features that include both the original values and their interactions, allowing for a richer representation of the data.

Conclusion: These features are crucial for capturing non-linear relationships in data, allowing linear models to fit complex patterns. They improve model performance by enabling better fits and revealing interactions between features. By adjusting the polynomial degree, you can control model complexity, enhancing predictive capabilities while avoiding over fitting. Overall, they serve as a powerful tool for feature engineering and improving the interpretability of models.

Justification for the Transformations:

- **Log Transformation**: This is applied to normalize skewed data distributions, reducing skewness and bringing them closer to a normal distribution. This adjustment enhances data interpretability and improves the performance of statistical analyses and machine learning models by revealing underlying patterns.
- **Binning**: Binning group's continuous data into discrete intervals, simplifying analysis and visualization while reducing noise and highlighting trends. It transforms numerical features into categorical ones, improving model interpretability and handling outliers by placing extreme values within defined ranges. Overall, binning aids in data reduction and enhances the robustness of analyses.

- **Data Cleaning**: The code identifies and addresses NaN and infinite values by filling NaNs with the mean and replacing infinite values with the maximum finite value. After confirming the resolution of these issues, it standardizes numerical columns to have a mean of 0 and a standard deviation of 1, preparing the data for further analysis or modeling.

Analysis of Feature Importance and Selection Results.

- Recursive Feature Elimination (RFE): The analysis uses Logistic Regression to model a dataset after preprocessing and selecting features. Categorical variables are encoded, and the dataset is split into training and testing sets. It identifies the top 10 features, which are then used to fit the model. The model's accuracy is calculated, showcasing the effectiveness of the selected features in predicting the target variable.
- The feature importance was analyzed from a Random Forest model, highlighting that Target binary is the most significant predictor. Other key features include Performance, Curricular units 2nd sem (approved), and Curricular units 1st sem (grade), which also play vital roles. Several features contribute moderately, such as Study Time per Credit and Age at enrollment. Overall, the model emphasizes a few critical features while many others have minimal impact.
- The analysis uses Lasso regression for feature selection through stability selection, iterating 100 times with random subsamples of the dataset. Categorical variables are encoded, and features are standardized before applying Lasso with a predefined alpha value of 0.01. After counting how often each feature is selected across iterations, features that are chosen in more than 50% of the iterations are deemed stable. The final output displays the names of these robust features, indicating their significance in predicting the target variable. This method ensures that only the most reliable features are retained for further modeling, enhancing the robustness of the analysis.

Visualization of Dimensionality Reduction Results.

The t-SNE visualization displays a two-dimensional representation of a dataset, where the points are color-coded based on the target variable, which has three classes (0, 1, and 2). The clusters suggest that classes 0 and 1 are more distinct and separated from each other, while class 2 overlaps with both. This indicates that the model may differentiate between classes 0 and 1 effectively, but class 2 might be more challenging to distinguish. Overall, the visualization provides insights into the relationships and separability of the different classes in the dataset.

Thank you.