FEATURE ENGINEERING REPORT

By Blessing Ilesanmi

28th September, 2024

INTRODUCTION

In this analysis, I explored the impact of feature transformation and selection methods on the dataset aimed at predicting student dropout rate. No new features were created for this analysis. Instead, I focused on transforming skewed numerical features, identifying key variables using feature selection techniques, and visualizing the results through dimensionality reduction.

The first part of the analysis justified the need for log transformations on highly skewed features, which included variables like "Age at Enrollment," "Total Credits Earned," and several related to academic performance (e.g., "Curricular Units 1st Sem Enrolled"). This transformation reduced skewness, making the data more normally distributed and improving the accuracy and generalizability of the models.

Next, I employed five feature selection methods to identify the most significant predictors of student outcomes:

1. Filter Method (Correlation-based)
2. Chi-Square Test
3. Wrapper Method (Recursive Feature Elimination)
4. Embedded Method (Lasso Regression)
5. Stability Selection

Each method provided valuable insights, selecting features from different aspects of student performance, demographics, and socioeconomic conditions. Lastly, I performed a Principal Component Analysis (PCA) to visualize the impact of dimensionality reduction on the data.

JUSTIFICATION FOR FEATURE TRANSFORMATION (log transformation to skewed numerical features)

Skewness represents asymmetry in the distribution of data. When numerical features are heavily skewed, several issues arise:

- Bias in models: Many machine learning algorithms assume data is normally distributed. Skewed features can lead to inaccurate model learning and suboptimal generalization.
- Sensitivity to outliers: Skewed data often contains outliers, which can disproportionately affect model performance.
- Model performance degradation: Models built on skewed data may produce higher errors and reduced accuracy.

To mitigate these issues, transformations like the log function can be applied to make the data more normally distributed, thereby reducing the impact of outliers and improving overall model performance.

Identification of Skewed Numerical Features

Skewness was calculated for each numerical variable in the dataset to determine which features were highly skewed. A threshold of 1 was used to identify significant skew, with any variable whose absolute skewness value exceeded this threshold being flagged as highly skewed. This approach allowed for the identification of variables that would benefit most from transformation.

The following variables were identified as highly skewed:

- Application order
- Age at enrollment
- Curricular units 1st sem (credited)
- Curricular units 1st sem (enrolled)
- Curricular units 1st sem (grade)
- Curricular units 1st sem (without evaluations)
- Curricular units 2nd sem (credited)
- Curricular units 2nd sem (grade)
- Curricular units 2nd sem (without evaluations)
- Total Curricular Units 1st Semester
- Total Curricular Units 2nd Semester
- Total Credits Earned
- Total Units Enrolled
- Weighted Grade 1st Semester
- Weighted Grade 2nd Semester
- GPA

Why Use Log Transformation?

Log transformation is an effective method to reduce skewness in positively skewed data. It is particularly useful for data where large values are significantly more frequent than smaller values. The transformation compresses the range of data, bringing high values closer to the center and stretching out smaller values.

Adding a small constant to the data (e.g., 1) before applying the logarithm ensures that the transformation can be performed even for variables containing zeros. This method avoids undefined mathematical operations (logarithm of zero or negative values) and allows for smooth data transformation.

Impact of Log Transformation on Skewed Features

Log transformation was applied to the highly skewed variables listed above. The key effects of this transformation include:

1. Application order: Helps normalize the varying number of applications, reducing the influence of large application numbers on the overall data distribution.

2. Age at enrollment: Reduces skewness caused by non-traditional student ages, making the age data more comparable across different student groups.
3. Curricular units (credited, enrolled, grades, without evaluations): These variables, which represent student academic performance, are often skewed due to students with unusually high or low credit loads. Log transformation helps balance the data and reduce the impact of extreme values.
4. Total curricular units and credits: Variability in total units and credits earned can lead to skewness. Transforming these variables results in a more normal distribution, improving model performance.
5. Weighted grades and GPA: Student performance measures such as grades and GPA often exhibit skewness, especially in the presence of extreme high or low values. Log transformation helps compress these extremes and reduce the impact of outliers.

Benefits of Log Transformation

The benefits of applying log transformation to highly skewed variables include:

- Improved model accuracy: Machine learning algorithms perform better when the input data follows a more normal distribution.
- Reduced impact of outliers: Log transformation compresses the influence of extreme values, making models less sensitive to outliers.
- Enhanced interpretability: After transformation, variables become more comparable and easier to interpret, as their values fall within a more reasonable range.
- Increased stability of variance: Log transformation reduces heteroscedasticity (unequal variance), leading to more reliable models.

FEATURE SELECTION

I used the following feature selection methods:

1. Filter Method (Correlation-based)
2. Chi-Square Test (Statistical Test)
3. Wrapper Method (Recursive Feature Elimination)
4. Embedded Method (Lasso Regression)
5. Stability Selection (Lasso with Bootstrapping)

Each method has its strengths and focuses on different aspects of feature importance. Below is a detailed explanation of each technique and the selected features from each approach.

Filter Method (Correlation-based Feature Selection)

The filter method uses a correlation matrix to assess the relationship between independent features and the target variable. I selected features with a correlation coefficient greater than or equal to 0.47 in absolute value. This threshold was chosen based on the strength of correlation deemed significant.

Selected Features:

- Curricular units 1st sem (approved)
- Curricular units 1st sem (grade)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)

These selected features primarily relate to the student's performance in the first and second semesters, indicating that academic performance is a significant predictor of dropout rate. The correlation method is simple and computationally inexpensive, but it may overlook non-linear relationships between variables and the target.

Chi-Square Test (Statistical Test for Categorical Variables)

The Chi-Square test evaluates the dependence between categorical variables and the target variable. This method is useful when dealing with categorical features. The test was applied, and features with p-values less than the significance level of 0.05 were considered significant.

Significant Features (p-value < 0.05):

- Application mode
- Previous qualification
- Scholarship holder
- Debtor
- Curricular units 2nd sem (approved)
- Mother's qualification
- Curricular units 2nd sem (grade)

- Gender
- Father's occupation
- Tuition fees up to date

Chi-square testing revealed several significant features related to the demographic and financial background of students, such as application mode, scholarship status, debtor status, and parental occupation. This suggests that external factors, including financial stability and qualifications, may significantly influence student outcomes, in addition to academic performance.

Wrapper Method (Recursive Feature Elimination - RFE)

The Recursive Feature Elimination (RFE) method iteratively removes the least important features by training the model and ranking the features by importance. This method directly interacts with a specific model, in this case, Logistic Regression, to select the most useful features for prediction.

Selected Features:

- Curricular units 1st sem (approved)
- Curricular units 2nd sem (credited)
- Curricular units 2nd sem (approved)
- Total Credits Earned
- Total Units Enrolled

The RFE method selected features closely related to the student's academic achievements and workload. Features like Total Credits Earned and Total Units Enrolled imply that the volume of coursework successfully completed is critical for predicting outcomes. This method often aligns feature selection with the model's predictive performance, favoring variables that enhance classification accuracy.

Embedded Method (Lasso Regression)

Lasso (Least Absolute Shrinkage and Selection Operator) is a regression method that adds a penalty for the absolute size of coefficients to encourage simpler models. Lasso can shrink some coefficients to zero, effectively selecting only the most important features.

Selected Features from Lasso:

- Application mode
- Tuition fees up to date
- Scholarship holder
- Age at enrollment
- Curricular units 1st sem (evaluations)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)
- Total Curricular Units 2nd Semester

The Lasso method identified a mixture of academic, demographic, and financial-related features. It emphasizes both performance-based variables like Curricular units 2nd sem (approved) and Curricular units 1st sem (evaluations), as well as scholarship holder and tuition fees up to date, indicating that a student's financial situation is also a critical factor.

Stability Selection (Lasso with Bootstrapping)

Stability Selection combines bootstrapping with Lasso regression to identify stable features that consistently appear across multiple subsamples. This method mitigates the variability that can arise when fitting models on different subsets of the data.

Selected Features:

- Application mode
- Course
- Previous qualification
- Mother's occupation
- Admission grade
- Displaced
- Debtor
- Tuition fees up to date
- Gender
- Scholarship holder
- Age at enrollment
- International
- Curricular units 1st sem (credited)
- Curricular units 1st sem (evaluations)
- Curricular units 1st sem (approved)
- Curricular units 2nd sem (credited)
- Unemployment rate

Stability Selection identified a broader range of features compared to the previous methods. In addition to academic and financial features, it also highlighted unemployment rate and course, suggesting external economic factors and specific academic programs play a role in student success. This approach is highly robust and yields features that are stable across different resampling iterations.

Comparative Analysis of Methods

Overlapping Features:

Several features were selected across multiple methods:

- Curricular units 1st sem (approved)
- Curricular units 2nd sem (approved)
- Tuition fees up to date
- Scholarship holder

- Application mode

These features were identified by at least three methods, indicating their high importance in predicting student outcomes. Academic performance features (e.g., curricular units) and financial status indicators (e.g., tuition fees and scholarship) are consistently influential in multiple approaches.

Unique Features:

- The Chi-Square Test highlighted Mother's and Father's occupations**,** which did not appear in other methods.
- Lasso emphasized Age at enrollment, while Stability Selection identified Unemployment rate and Course as significant.

Each method uncovers different aspects of feature importance. Statistical methods like Chi-square prioritize demographic and categorical variables, while model-based methods (RFE, Lasso) tend to favor academic and workload-related features.

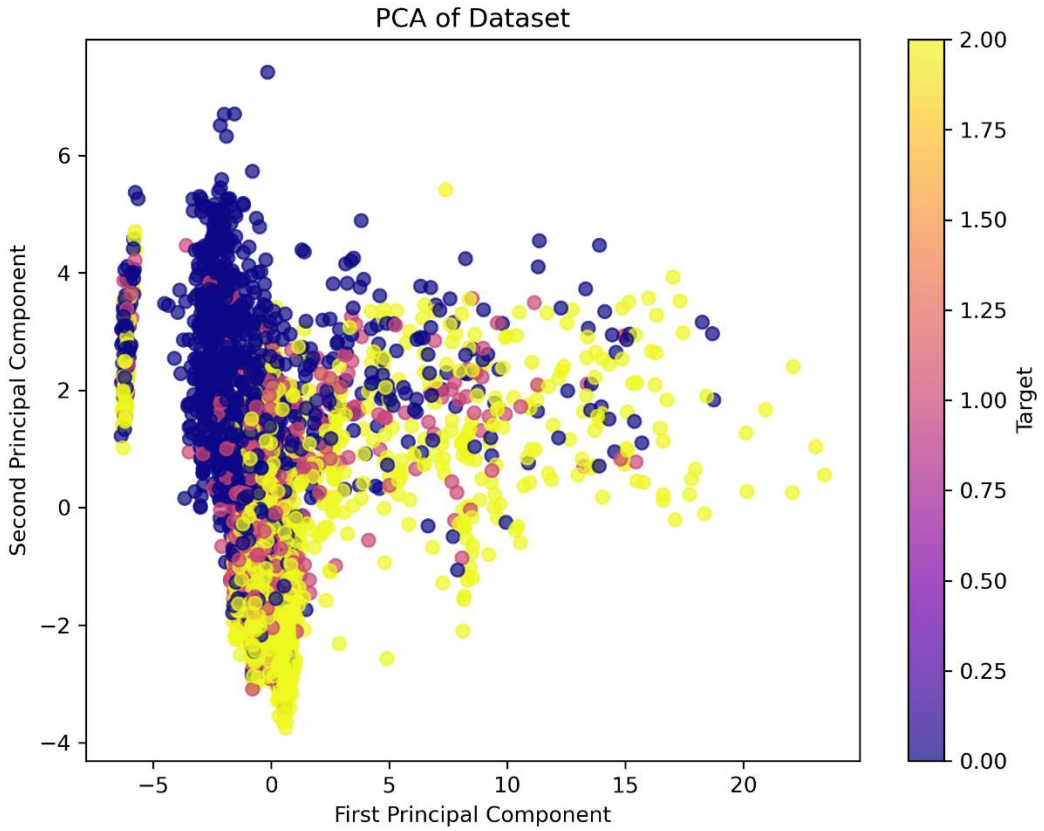
This feature selection process demonstrates that a combination of methods provides a well-rounded understanding of the most relevant features. Academic performance indicators, such as approved and graded curricular units, emerged as key predictors across all methods. Additionally, financial and demographic factors, including scholarship status, tuition fees, and application mode, were consistently significant.

Future work could involve testing the selected features using various machine learning models to evaluate their predictive power and refining the model further by investigating interactions between features. This analysis also underscores the importance of using multiple feature selection methods to ensure robustness in the selected features.

VISUALIZATION OF DIMENSIONALITY REDUCTION RESULTS

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms a high-dimensional dataset into a smaller set of uncorrelated variables called principal components. The PCA plot visualizes the distribution of data points projected onto the first two principal components. The color of each data point represents the corresponding "Target" variable (0, 1, or 2).



Key Observations

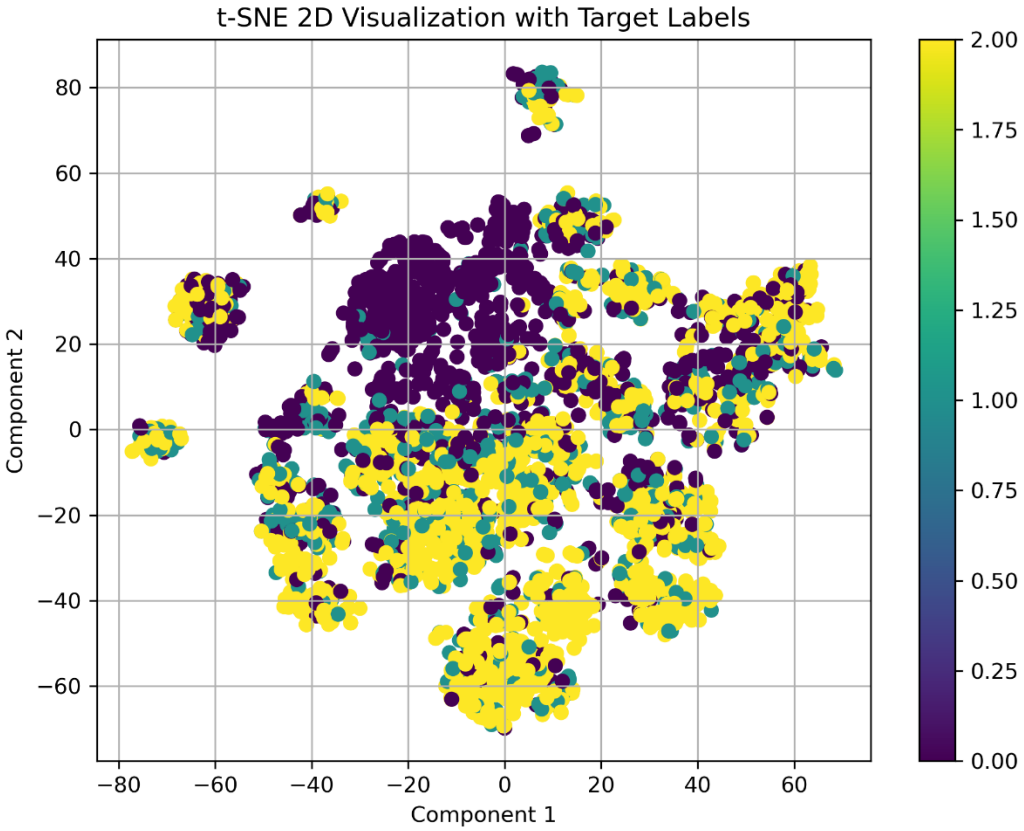
1. Clustering: There seem to be distinct clusters of data points based on the color (Target variable). This suggests that the first two principal components effectively capture the underlying structure related to the dropout, enrollment, and graduation outcomes.
2. Separation: The clusters appear to be somewhat separated, particularly between the "Dropout" (0) and "Graduate" (2) categories. This indicates that the principal components are able to differentiate these groups based on the combination of variables in the dataset.
3. Overlapping: There is some overlap between the "Dropout" (0) and "Enroll" (1) clusters. This suggests that there might be some similarity in the characteristics of students who drop out and those who enroll but don't graduate.

Interpretations Based on the Target Variable

- Dropout: Students who drop out are clustered in a specific region of the plot, potentially characterized by certain combinations of variables related to academic performance, personal factors, or socioeconomic conditions.
- Enrollment: Students who enroll but don't graduate fall in a region between the "Dropout" and "Graduate" clusters, indicating a combination of factors that influence both dropout and graduation.
- Graduation: Students who graduate are clustered in a distinct region, suggesting a combination of factors that contribute to successful completion of the program.

t-SNE (t-Distributed Stochastic Neighbor Embedding)

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that preserves local structure in high-dimensional data. It's particularly useful for visualizing complex relationships between data points. In the t-SNE visualization, the two axes represent the reduced dimensions (components 1 and 2). Each data point is colored based on its corresponding "Target" variable.



Key Observations

1. Clustering: There seem to be distinct clusters of data points based on color, suggesting that the t-SNE has effectively captured underlying groupings in the data related to the target variable.
2. Separation: Some clusters appear to be more separated than others, indicating that certain groups of data points are more distinct in terms of their characteristics.
3. Density: The density of data points within each cluster varies, suggesting different levels of concentration or variability within the groups.

Interpretations

- Distinct Groups: The well-defined clusters suggest that the data points can be naturally divided into distinct categories or groups based on their characteristics.
- Relationships: The proximity of data points within a cluster indicate similar characteristics or relationships between them.
- Outliers: Isolated data points or small clusters represent outliers or unusual observations that deviate from the main patterns.