Dimensionality Reduction

Juanita Ramirez V.
STA6704 Data Mining II
UCF
Orlando, FL.

Abstract:

Keywords—PCA, Manifold, FA, clustering

I. DATA - *"PREDICT STUDENTS DROPOUT AND ACADEMIC SUCCESS"*

The dataset is extracted from a higher academic institution, each row corresponds to a student and the columns are related to academic, demographic and socio-economical features. The original dataset shape has 37 columns and 4427 rows, with 1 Ordinal feature, 18 Numerical features and 18 Categorical features (8 of them Binary).

A. *Data Preprocessing*

To approach the Categorical features with high number of categories such as: Previous qualification, Mother's occupation, Father's occupations, Course, and Application mode. OneHotEncoder was fit to transform the data.

Other Categorical features were transformed to binary class merging several classes to focus in more general characteristics such as whether or not the parents pursued higher education instead of what type of education, and the column International was dropped because it was the same as Nationality.

B. *Data Exploration*

- The distribution of the Target variable is unbalanced as observed in Fig. 1.
- Mother's and Father's occupation are positively correlated, likewise Mother's and Father's qualification are positively correlated.
- The Chi Squared Test of Independence 'chi2_contingency', rejected the null hypothesis for the variables Mother's occupation, Father's occupation and course when compared to the Target variable, therefore there is a significant association between the variables and the Target.

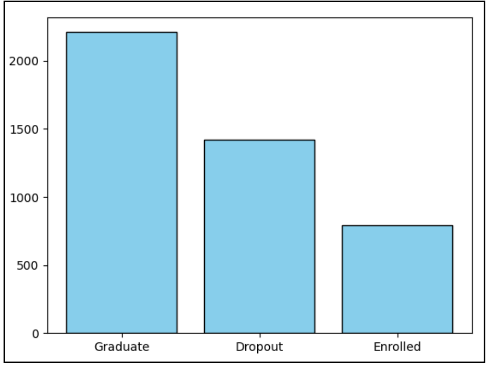

Fig. 1. Target Distribution.

II. DIMENSIONALITY REDUCTION

A. *Principal Component Analysis (PCA)*

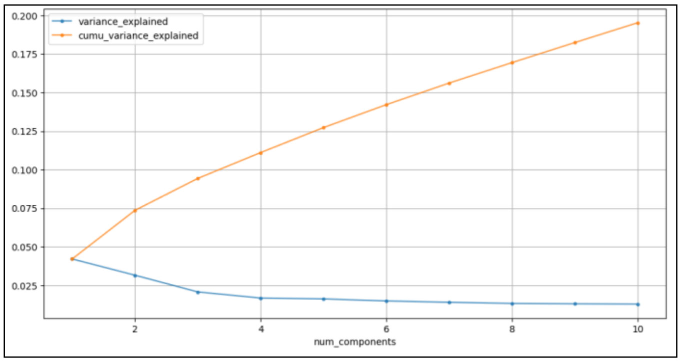

Fig. 2. PCA with all the features of the encoded data.

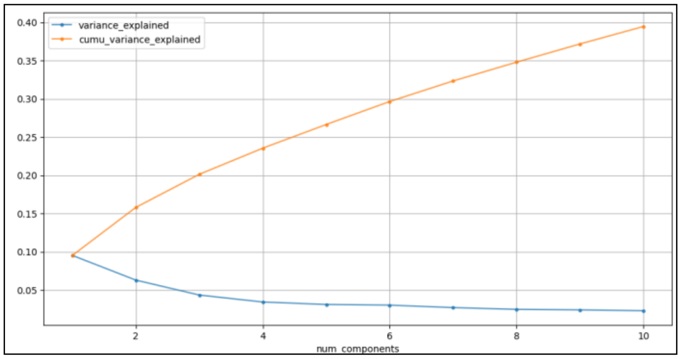

Fig. 3. PCA dropping small categories of encoded data.

Using OneHotEncoder can introduce multicollinearity, to avoid that we can drop some columns, however in this data when the parameter drop = 'first' is used, the results are similar to drop=None (in this case the encoded data has 161 columns), and we can't observe an improvement in the variance explained. The cumulative variance explained in the first 10 principal components is close to 0.2 in both cases (Fig. 2).

When dropping categories with sizes close to the average category size, for that encoded feature, we notice an increase in the cumulative variance explained, approaching nearly 0.4 (Fig. 3), after dropping the category columns with an specific size threshold the data has 69 columns.

Examining the first two principal components in Fig. 4, when place on the x-axis, suggest that positive values generally indicate a graduate, negative values suggest a dropout, and values close to zero at both ends could indicate an enrolled student. However, these distinctions are not well-defined and seem to overlap.

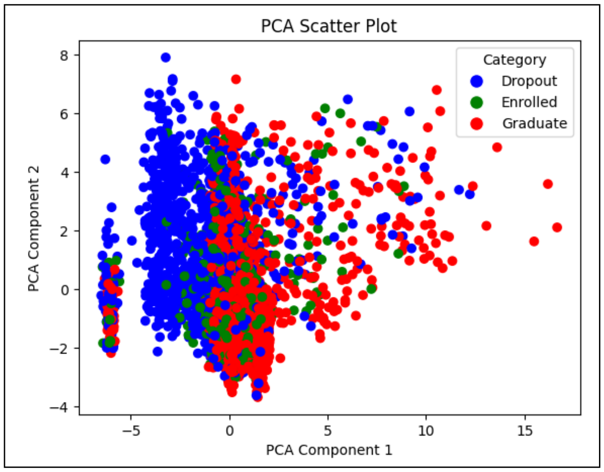

Fig. 4. PCA Scatter Plot with the First 2 Principal Components

B. *Factor Analysis*

The Factor Analysis (FA) decomposition performed slightly worse compared to PCA. When the data has 161 columns, the cumulative explained variance for the 10th component is 0.16. In contrast, when the data is reduced to 69 columns, the cumulative explained variance for the 10th component increases to 0.32 (Fig. 5.)

.

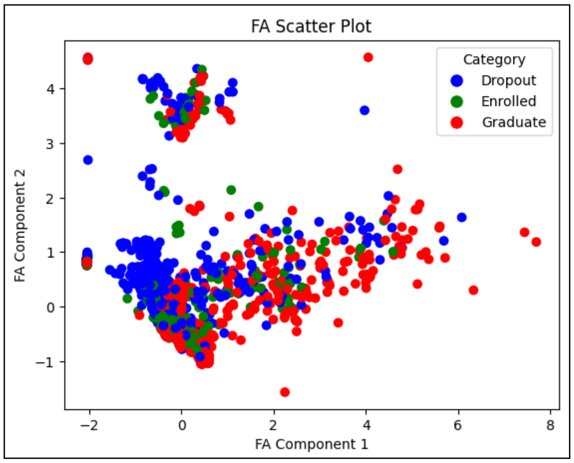

Fig. 5. Factor Analysis Scatter Plot with the First 2 Components

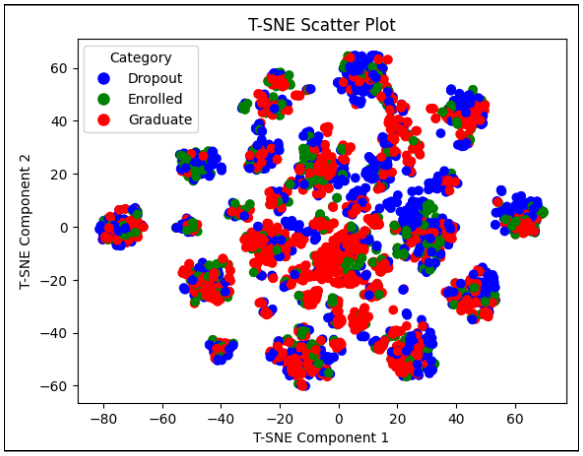

Fig. 6. T-SNE Scatter Plot with the Fist 2 Components

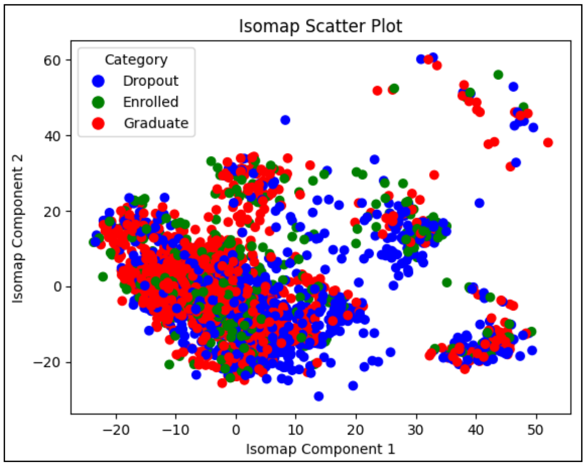

Fig. 7. Isomap Scatter Plot with the First 2 Components

C. *T-SNE and Isomap*

Observing Fig. 7 and Fig. 6, there is more overlapping between the categories compared to PCA and FA. This suggests that these methods might be less effective in distinguishing between categories.

III. Models

A. *K-means:*

	PCA – 10 Components	Reduced Data with 69 Features
Rand Score	0.58	0.50
Adjusted Rand Score	0.17	0.04

TABLE 1: K-means performs better after PCA decomposition, however the adjusted rand score is close to zero which means that the model is slightly better than random.

B. *XGBoost*

	PCA – 10 Comp.	PCA – 20 Comp.	Data 69 Features	Data 161 Features
Accuracy	0.7	0.7	0.76	0.76

TABLE 2: Accuracy in XGBoost model, fitted to different datasets.

In TABLE 2 we can conclude that:

- XGBoost perform the same for 10 Principal Components and 20 Principal Components, indicated that the variance explained after 10 Components is small, and is better to choose the first 10.
- XGBoost performs the same for the reduced data and for the encoded data, which means that reducing the number of features did not lead to loss of important information and the columns dropped were redundant data.
- We can observe a 6% of decrease in accuracy when comparing PCA – 10 comp. data with reduced or encoded data, which indicates that PCA reduced the features from 69 or more to 10 while preserving a significant amount of variability.

IV. References

[1] V. Realinho, M. Vieira Martinis, J. Machado y L. Baptista, « Predict Students' Dropout and Academic Success.,» *UCI Machine Learning Repository,* 2021.

[2] M. V. Martins, D. Tolledo, J. Machado, L. Baptista y V. Realinho, «Early prediction of student’s performance in higher education: a case study,» *Trends and Applications in Information Systems and Technologies,* vol. 1, 2021.