READ ME

Model comparison:

There were two models that our group worked with. The first one being an OLS model. OLS models are ones that utilize linear regression to estimate the relationship between the dependent variable and one or more independent variables. In our case, we used OLS to estimate the relationship between the various parameters and the effect or lack thereof on graduation rates. The process of using this model included Lasso as seen by variables such as regions and student faculty ratios are set as zero as Lasso adds a “penalty” that forces less important coefficients to zero with the result being that Lasso selects for only the most relevant predictors and simplifying the model. This function additionally handles multicollinearity and prevents the problem of overfitting that is common with linear regression models, especially in the cases where the dataset in question has a large number of variables.

The resulting model found several values. The R squared score, an indicator of the percentage of the variance that can be explained by the model is about 76.5% which indicates that the model can explain a significant portion of the variability. The low RMSE value at 0.0817 indicates a low average difference between predicted and actual values indicating a reasonably low error for the model

The second model that our group worked with was PCA. PCA is typically used in cases when working with high-dimensional data sets in order to reduce the number of features in a dataset while reducing the variability into a dataset. This process ideally removes the noise by discarding variables that are of low variance and are thus of low relevance, speed up processing of the data, and most importantly serve as a base for our later work in linear regression. PCA performs feature extraction which allows our group to highlight the most important features of the dataset that we would then focus on in order to perform linear regression on.

Through PCA analysis, our group found that 95% of the data could be explained by 11 different components. From our PCA analysis, we found that the percent of students who take out loans, average ACT scores, and the school region are the best predictors of graduation rates.

In this case, our group would choose Linear regression with Lasso as the better choice. The reasoning behind this is multifaceted. Firstly, PCA works best when there is an extremely high degree of dimensionality, for example if there were hundreds of very granular regional indicators, and thus the need to consolidate or reduce the number of variables is crucial and if there is a high degree of multicollinearity. Though there is multicollinearity in this dataset, it appears that Lasso can handle this aspect as seen with it’s handling of the regional values and more importantly, Linear regression with Lasso fits more with the stated goals of our project.

The project goal aimed to find what variables would contribute to the perceived value of education and perhaps find a quantitative answer to the question “Is college a scam?” We aimed to do this by examining student outcomes such as graduation rates and income after graduation. In order to do this, we aimed to find the variables that would give the highest predictive value. Linear regression with Lasso is the best system to do so. The Linear regression model identifies specific variables that are correlated with graduation rates and provides coefficients that are straightforward to interpret as opposed to PCA. Lasso served as a means of dealing with the inherent limitations and pitfalls of linear regression by providing feature selection. PCA’s main downfall was the sacrifice of interpretability and the potential inclusion of variance to a target variable.

For these reasons, we feel that the OLS model was the best model to use.