**Analysis on NY State Education Department Dataset**

The Spearman rank coefficient was used to identify prominent features prior to machine learning. It must be noted that these are percent student population per school features and not individual student characteristic features. The correlations examines the effect of having higher percentages of any SES factor in a school.

The features that are most negatively correlated with the graduation rate in order are:

* Homelessness -0.58
* Economically disadvantaged -0.56
* Limited English proficiency -0.49
* Students with disabilities -0.46
* Percentage out of certification -0.39

Note that ethnicity also had correlations with graduation rate to a lesser degree:

* Black -0.39
* Hispanic -0.32

The female gender had a slightly positive correlation with graduation at 0.13

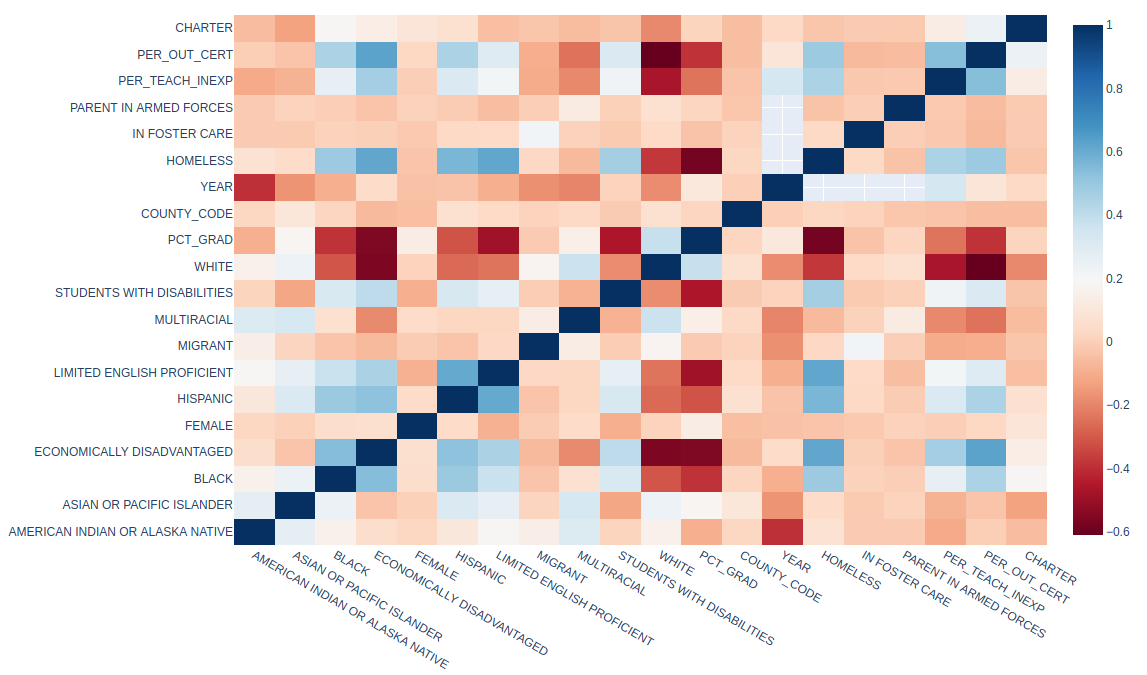


Figure 1. Spearman Rank Correlation matrix for socioeconomic factors, staff experience, and public/charter school

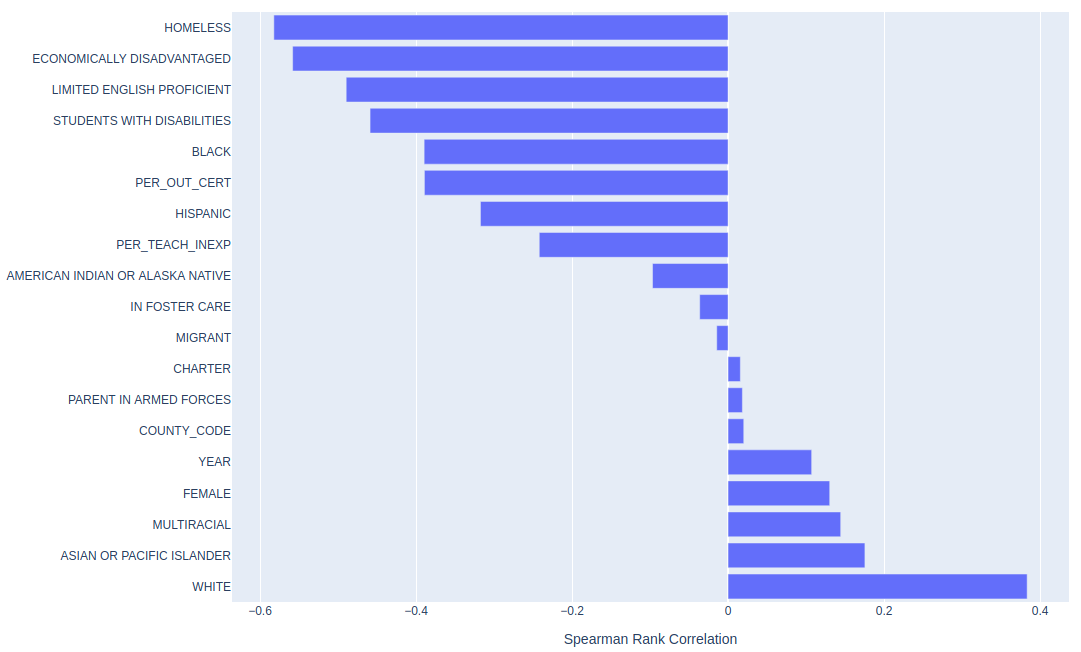


Figure 2. Spearman Rank Correlation for graduation percentage per school

Many of the factors that correlate with low graduation rate also correlate with Economically Disadvantage: Homeless, Limited English Proficient, Percentage Out of Certification, Percentage Teacher Inexperience, Students with Disabilities, and two ethnic subgroups. Additional analysis is required to separate whether the staff inexperience has a statistically significant effect when correcting for other factors.

Because Economically Disadvantaged was one of the largest common factors and the second largest negative factor for graduation rate, we examine factors that correlate positively with it.

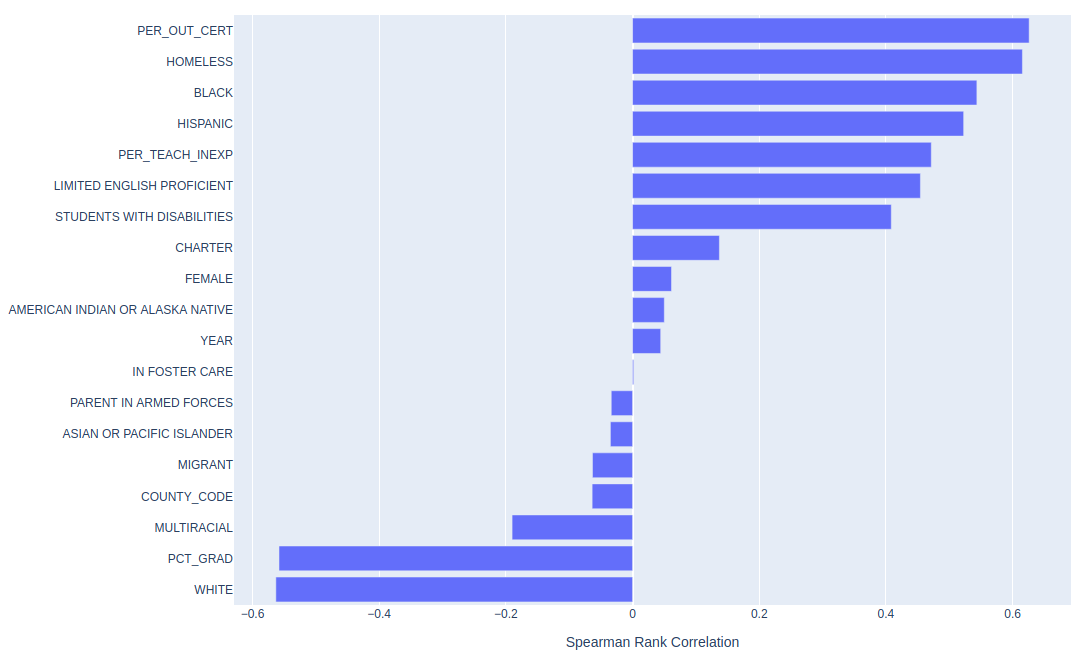


Figure 3. Spearman Rank Correlation for economically disadvantaged

A random forest regressor model is built to assess the feature importance based on random forest impurity and based on permutation importance.

The impurity-based feature importance produces a similar ranking of top features except that the Homeless feature is much lower and the Female feature is much higher. It is useful to see that some factors that seem to have high correlation with graduation are seen more clearly with the Spearman Rank Correlation than with the impurity feature importance.

A couple drawbacks to the impurity feature importance are a tendency to prefer features with high cardinality and can dismiss one feature over another if they are correlated. However, the impurity-based feature importance captured many of the same top features identified by the Spearman Rank Correlation.

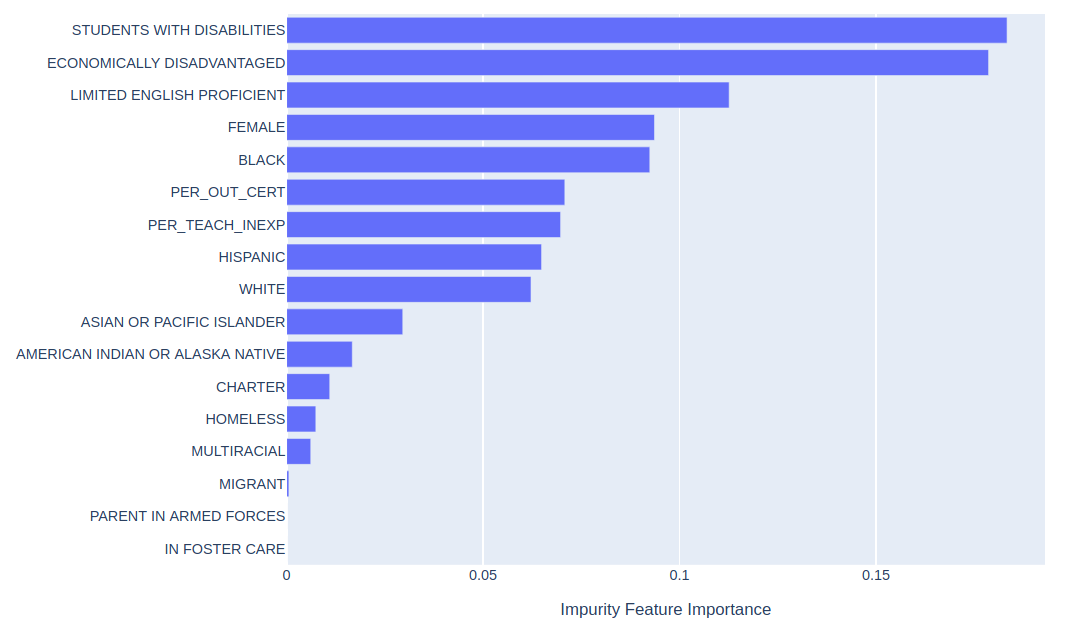


Figure 4. Impurity feature importance increases the Female feature and minimizes the Homeless feature compared to Spearman Rank Correlation.

The permutation importance produces similar results as the impurity-based feature importance with a few adjacent features changing places. The Female feature is again a higher factor, and the Homeless factor is minimized despite the known correlation with homelessness and graduation rate. The permutation importance could also have problems with highly correlated features though we do not see this issue here.

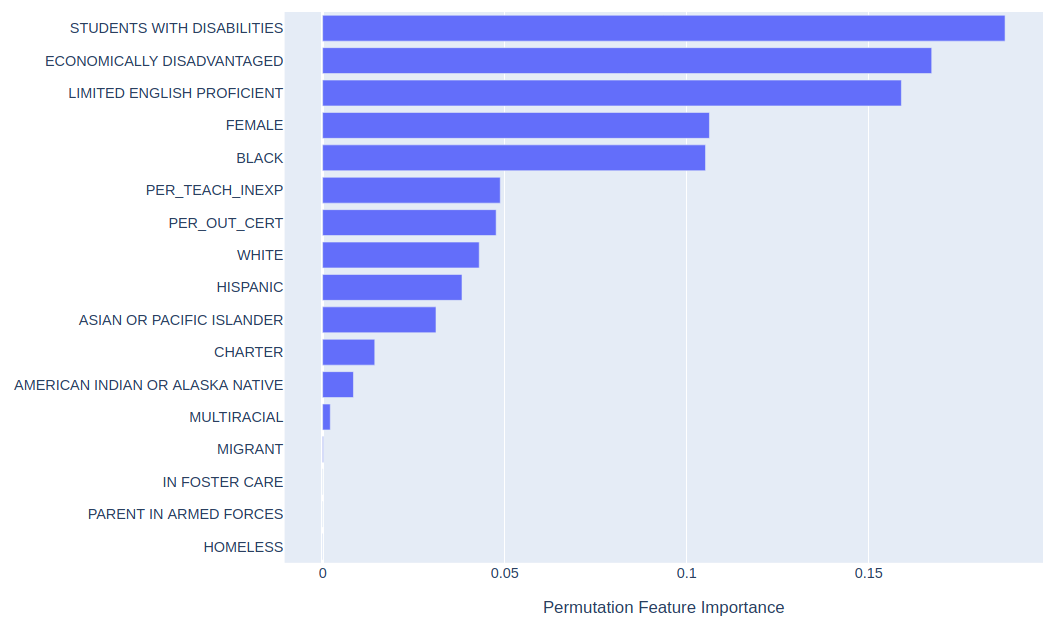


Figure 4. Permutation feature importance increases the Female feature and minimizes the Homeless feature compared to Spearman Rank Correlation.