A regression study on the impact of Demographic and Poverty features on student success.

The regression model seeks to determine if a relationship exists between a student's success and demographic and poverty factors. These factors include Race, Gender, English language learners, Total cohort, Total regents, Advanced Regents #, Regent’s w/o Advanced #, Total enrollment, still enrolled, #poverty, and Economic Need Index.

We use a supervised learning method for regression as our target is a continuous variable. We tried different machine learning algorithms, such as Random Forest Regressor (RFR), Decision Tree Regressor (DTR), Support Vector Regression (SVR), and Multiple Linear Regression (MLR). We evaluated these algorithms by r2\_score, Mean Absolute Error (MLE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Of all these algorithms, Multiple Linear Regression performed the best. So, we decided to go with Multiple Linear Regression.

Chart, bar chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

We built a Multiple Linear Regression Model on NYC School Data. The model will predict the student success measured on the predicted outcome, i.e., 'Dropped out #' based on demographic and poverty features.

Here we have a quick overview of creating a model:

1. we have imported all the libraries; We are using sklearn, the Scikit-learn machine learning library for Python. It has many modules, including one for Multiple linear regression, which we have used here.

2. We have defined x and y for the model. x and y are input and output features of the dataset. So, taking x features as independent input values, our model will predict the outcome, which is y that are dependent.

3. we have split our dataset into two parts in which 80% of the dataset will go into the training set, and 20% of the dataset will go to the testing set.

4. Create an instance of the model with model = LinearRegression().

5. Train the model with the dataset with the model. fit(X\_train,y\_train).

6.Create predictions with y\_pred = model.predict(X\_test).

We have performed multiple linear regression analyses using statsmodels. Statsmodels is a Python module that provides classes and functions for the estimation of different statistical models, as well as different statistical tests.

Analysis and results:

A screenshot of a computer

Description automatically generated with low confidence

With the rsq 99%, we can say that the data best fit the regression model.

The adjusted R-squared penalizes the R-squared formula based on the number of variables; therefore, a higher adjusted score means most of the variables are contributing to our model's R-squared properly.

After looking at the p values, some of the demographic features (Asian, black, Hispanic native Americans, white), total cohort, regents’ data, and the number of English language learners are significant as they are less than 0.05. so, we can reject the null hypothesis that these features make a difference in the student dropout rate.

Poverty features (poverty number, Economic Need Index) are not significant as they are higher than 0.05. so, we can conclude that these features make a difference in the student dropout rate.