Team 26 Progress Report

Introduction

The education system and academics have been working to foster environments for success and curriculums to engage students. This comes with many challenges when working at scale with many different schools and districts comprised of a wide range of demographics. To make matters worse, there is no simple solution for how to ensure a student will succeed. By using the power of data tracking and non-traditional factors these complex challenges may be solved. Evaluation of student success and a school's ability to educate students is accomplished by tracking key metrics. These success metrics are typically focused on graduation rates, test scores and grade point averages. By tracking these various metrics, the education system has opened opportunities for analytics to find patterns in the data and even help improve operations in many school districts. One study has even shown that school sizes have an impact on student dropout ratings [1]. Optimizing school size may even reduce dropout prevention expenses for the school [1]. Analytics may also help educators and communities allocate funding to the appropriate programs that have been proven via data to improve student success. This analytical effort may be critical to schools struggling to improve student success.

Objective

The objective of this research is to identify key predictors that attribute to high school student success. These key predictors may then be used by districts or schools to optimize student success by improving upon the key predictive metrics we discover.

Problem Statement

The Texas Education Agency is responsible for collecting much of the education data in Texas and reporting on it publicly. This research will look at many different factors that may influence the outcome of graduation rates in Austin, Texas High Schools. These factors may include attendance which has been found to be a fundamental indicator of student engagement with the school, family median income by demographics, and district operating expenditures [2]. This analysis will determine the extent that these factors influence graduation rates and the divergent influences they have depending on demographic inputs.

Additional Heilmeier Questions

How is it solved today? How do you plan to approach the problem? The problem is solved with educator experience and community voting on fund allocation with no real data-oriented understanding of what the impactful factors are in student success.

How will your project and your team be organized? The project will be organized in a series of steps that need to occur to evaluate and select the final model. We have some team members specializing in the data engineering process while others are prioritizing model development. All team members are contributing an equal level of effort and support to the project.

How will intermediate results be generated? Results may be generated as our models start to develop. We may find significant predictors early on as evaluated by basic models and improve upon our understanding as we test future models with higher accuracy.

How will you measure progress? Progress will be measured as a binary completion of key tasks that need to occur before the results are ready. For instance, once the data collection process is finished, we can count that as complete and move on toward the next task in our model development pipeline.

What could it cost if one could implement your idea on a large scale? What financial impact would your project have if its results were implemented widely? The cost of implementing this idea on a large scale would be impacted by the data collection. The financial impact would outweigh the cost. With effective resource management and funding allocation, a school district could optimize funding which would reduce taxpayer dollars and enable higher student success ratings which leads to more educated professionals in the community.

Data Overview

Data collection first started by searching for key metrics like graduation rates, enrollment numbers, and teacher salaries. This led us into a pool of data on the Texas Education Agency (TEA) site. We decided to focus on Austin, Texas because were concerned with building datasets biased to rural communities and figured Austin was a great location with a variety of schools and districts available. The city of Austin also had additional data for graduation rates that we found interesting. To further extend our search for data we decided to include U.S. Census data to tie in a wide variety of demographics. Many of the data sets we found either had missing information, certain years or a lack of documentation resulting in some difficulty in connecting values to school campuses.

Raw data for this study were sourced from:

1. City of Austin, Texas open data portal
2. The Texas Education Agency
3. 2016 United States Census data summarized Code

Data Examples

1: Austin Open Data Portal:

- Travis County Four-Year High School Graduation Rates: Contains {Campus}, {District}, {Zip Code}, and {Graduation Percentages} for the years 2016 – 2020 by Ethnicity and Gender.
- Raw Datasets -
  Travis_County_4-Year_High_School_Graduation_Rates_by_Campus.csv

2: Texas Education Agency:

- Campus Type: Contains if the campus is Rural, Suburban, or City.
- State-Wide Staff Salary by District: Contains the {Total Base}, {FTE (Full-Time Employees) Count} (Full-Time Employee Count), and {Average Base Pay} by {District} for the years 2016-2020
- State-Wide Enrollments: Contains {Teacher_Count}, {Subject}, {Enrollments} by {District} for the years 2016-2020
- Total Operating Expenditures: Contains {Total Operating Expenditures},{Enrollment}, and {Operating Expenditures Per Student} by {District} for the years 2016-2020
- Raw Datasets -
  - operatingexpenditures.csv
  - traviscounty9-12TeacherSalaries.csv
  - District-Type1920.csv

3: United States Census Data:

- Demographic: Contains detailed racial demographics by {Zip Code}.
- Economic: Contains detailed economic data by {Zip Code}.
- Housing: Contains housing data by {Zip Code}.
- Raw Datasets -
  - demo.txt
  - econ.txt
  - housing.txt

- rural_urban.txt

The cleaned and joined data that comprises our raw data for analysis is comprised of the following types of data for 2016 through 2019. We have dropped the data for the year 2020 due to a significant amount of missing data for 2020. The missing data in 2020 is likely due to COVID-19 disrupting education and traditional testing.

Dependent variables: 8 different high school graduation rates for the schools in Austin. Campus graduation rate as well as graduation rates by race, gender and economic status.

Independent variables: 91 predictors at the start of the analysis which would be dwindled down to remove any unnecessary predictors. Variables included items like:

- School/District Info:
  - District Expenditures
  - Staffing
  - School type
  - Student distribution by socioeconomic classifications
- Community Data:
  - Population distributions by demographic
  - Housing information
  - Employment
  - Household distributions by demographic
- To view the full list of variables used see the data_dictionary.txt file.

Methodology

Data Collection/Cleaning: The data that we are collecting is described in more detail above. A considerable amount of effort is necessary to prepare the collected data for any form of modeling. We explored each dataset, finding discrepancies, missing data, and unnecessary information. Once we were able to understand the data, we joined the datasets to include the predictive factors we thought may show some form of significance. The following steps were taken to perform data cleaning using Python Pandas and create a final usable dataset for analysis:

1: Variable modification and Joins

- For educational data, joins were made on {District} and where possible, {Campus}.
- Census data was joined on {Zip Code}.
- Data sources from The Texas Education agency were in separate files by year. Year attributes were manufactured for each dataset and then concatenated into one set for each source prior to joining Austin Four-Year Graduation Rates.
- Campus ID is considered as the primary key for the final data set

2: Data imputation

- In the case of missing graduation rates for a subset, the overall graduation rate was substituted in its place.
- In the case of missing teacher salary information or missing operation expenditures, the row was dropped. This applied to 3 rows.
- In the case of missing 'mean income if on public assistance', the imputation value was the average of the column. This applied to 12 rows.

The final dataset contains 8 different dependent variables (the graduation rate for each school as well as graduation rates for key groups of students, e.g., Hispanic students, economically disadvantaged students, female students, etc.) and 91 potential independent variables. As the names for these variables can be quite long, aliases will be used to rename these variables. The data dictionary specified in the R code gives a list of the original variables with their alias variable names. As mentioned above, the dimensionality of the final dataset is still large. During analysis, the dimensionality will be reduced as we use feature engineering to find and manufacture suitable explanatory variables for graduation rates.

The final dataset “master_datafile.csv” and python code file (used for data cleaning purpose) is attached with the submission.

Exploratory Data Analytics: We ran exploratory and diagnostic plots on the final dataset to check linear regression assumptions as well as identify any discrepancies missed in the data cleaning process. The following steps were taken in this stage:

1: Checking the normal distribution of the dependent variables: Values of each dependent variable i.e., types of graduation rates were plotted individually to investigate if the dependent variable follows a normal distribution. This was done using the individual datasets created for each dependent variable. Here are the plots:

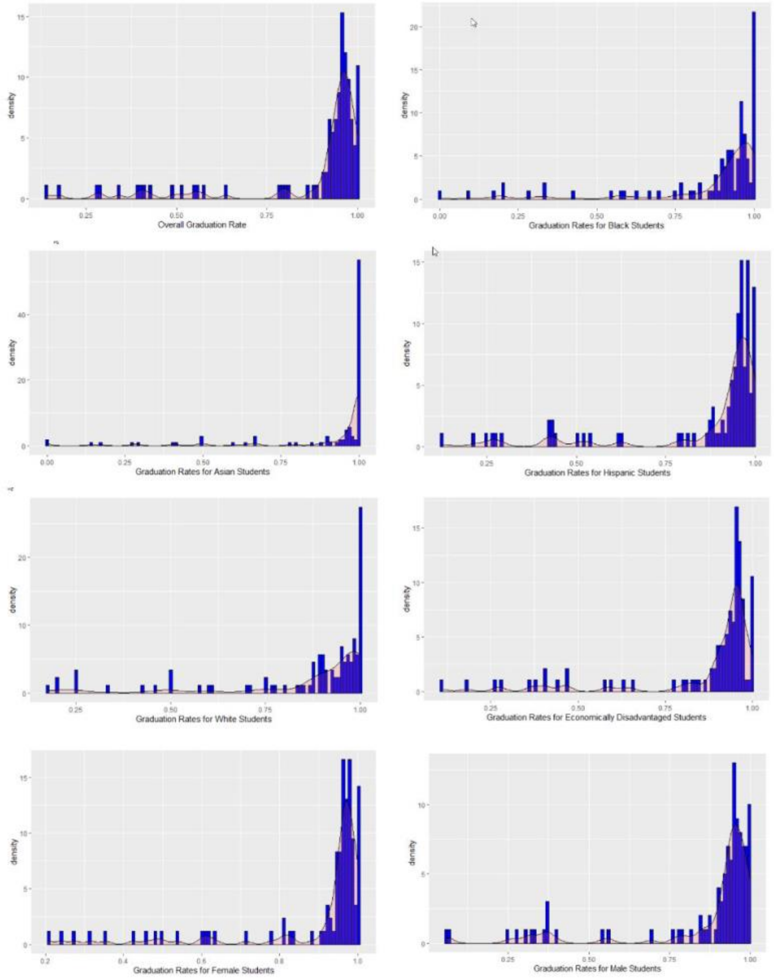

Figure 1: Distribution of Overall Graduation Rates, Graduation Rate for Black, Asian, Hispanic, White, Economically Disadvantaged, Female and Male Students (values of graduation rates have been adjusted to be in between 0 and 1)

Conclusion - As seen in figure 1 above, none of the dependent variables i.e., the overall graduation rates and graduation rate for Black, Asian, Hispanic, White, Economically Disadvantaged, Female and Male Students are not normally distributed. This could be due to most of the schools having graduation rates in the higher range i.e., close to 1, with a handful of schools with much lower graduation rates. This causes all dependent variables to be left skewed. This indicates a need for transformation

Transformation attempt - Traditional transformations considered for a left-skewed data set include a log transform, a square root transform, or a cube root transform. The dependent variable overall graduation rate is transformed using all three transformations and plots are created to test normal distribution. Below are the plots -

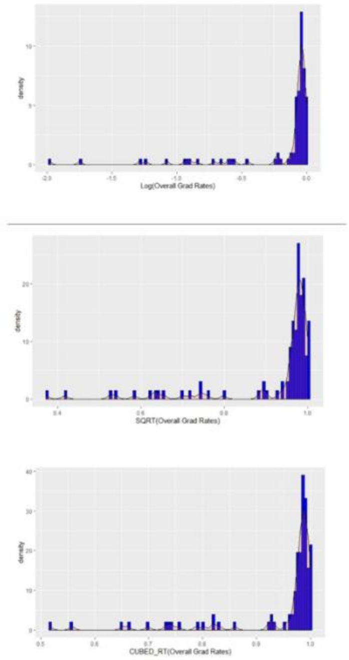

Figure 2: Distribution of log, sqrt and cubed transformed overall graduation rate

Conclusion - As seen in figure 2 above, the transformed overall graduation rates do not follow normal distribution. The left skewed distribution of the dependent variables needs to be kept in mind while analyzing further and deriving conclusions. An alternative would have been dropping the low schools or reducing the number of high performing schools in attempt to solve for the systemic bias.

2: Checking outliers: Boxplots were plotted to visualize the distribution of each numeric variable. If points lie beyond whiskers, then outlier values are present. However, the presence of an outlier does not automatically suggest a data point should be excluded from the overall data set. A few of the boxplots are shown below as examples. Due to space limitation, it was not possible to include boxplots for all the variables here. Please refer to the attached files for code and output for all the boxplots.

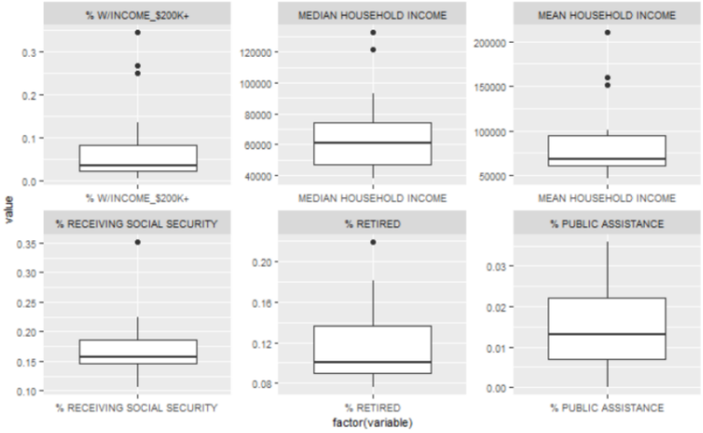

Figure 3: Examples of Boxplots of numeric variables

Conclusion – We noticed outliers in case of many variables. A domain expert knowledge would be needed to completely understand the outliers and remove/keep outlier values. We decided to leave the outliers in the final dataset so that we are not potentially removing real values.

3: Checking correlation: We created correlation matrix to check correlation among variables. Here is the list of variables grouped by correlation coefficient sorted in decreasing order. Please refer to the attached code file for reference on variable alias names.

x8/x9 : -1.00 (% Urban, % Rural)
x13/x14: -1.00 (% Homes Owned, % Homes Rented)
x27/x28: -1.00 (% Homes w/Mortgage, % Homes w/o Mortgage)
x66/x67: -1.00 (% Pop. Male, % Pop. Female)**

These inverse relationships are completely understandable as each variable pair is essentially the inverse of the other. The following variables are dropped as potential predictors: x8, x13, x27, x66.

x1/x3 : 0.999 (TOTAL_OP_EXPENDITURE, FTE_COUNT)
x1/x4 : 0.994 (TOTAL_OP_EXPENDITURE, TOTAL_SALARY_SPEND)
x3/x4 : 0.994 (FTE_COUNT, TOTAL_SALARY_SPEND)**

We see that the above 3 predictors track strongly to each other, which is a reasonable observance. If there isn't much variance in pay among employees, The total salary would effectively be the number of full-time employees (FTE) multiplied by the nominal salary. Similarly, if the operational expenditures budget is dominated by the amount spent on employee salaries, it would be understandable for these variables to also be strongly correlated. To simplify our model, the number of full-time employees variable will be kept while total_salary_spend and total_op_expenditures will be dropped.

x7/x38: 0.995 (TOTAL_POP,POP_16_YEAR_AND_OVER)

This is a reasonable observation if fraction of the total population who are adults are similar/identical across Austin.

**x41/x42: 0.989 (PERCENT_OF_LABOR_FEMALE_AND_16_AND_OVER,
PERCENT_EMPLOYED_FEMALE_AND_16_AND_OVER)**

Another reasonable observation. The number of adult women who are employed would reasonably track the number of adult women in the labor force. In this case, x41 will be dropped and x42 will be kept as 'employed' is a clearer descriptor than participating in the labor force, which has a number of caveats.

x55/x57: 0.978 (PERCENT_W_INCOME_200000_OR_MORE, MEAN_HOUSEHOLD_INCOME)

Austin is a relatively well-off city with a booming tech sector. These high income employees are likely skewing the mean household income.

x10/x38: 0.965 (TOTAL_HOUSING_AVAILABLE, POP_16_YEAR_AND_OVER)

The amount of housing available tracks population. This seems like an uncontroversial relationship.

x56/x57: 0.951 (MEDIAN_HOUSEHOLD_INCOME, MEAN_HOUSEHOLD_INCOME)

The relationship between median and mean is well-explained.

x7/x10 : 0.950 (TOTAL_POP, TOTAL_HOUSING_AVAILABLE)

This relationship is similar to the one between x10/x38 above.

x15/x71: 0.940 (AVERAGE_HOUSEHOLD_SIZE_OWNED, PERCENT_POP_15_TO_19)

A reasonable hypothesis is that the average household size of a homeowner is related to the number of teenage children who still live at home. Conversely, parents of young children (who are generally younger and early in their careers) may not be able to afford to own a home and still rent. This is an interesting observation as we are considering high school graduation rates, where students are generally aged 15-19 years of age. This possibly suggests that those students are generally coming from income-stable homes in Austin.

x24/x26: 0.944 (PERCENT_OF_HOMES_VALUED_500000_TO_999999, MEDIAN_HOME_VALUE)

As mentioned earlier, Austin does have a booming tech sector and thus has seen an influx of high-paid employees coming into the city. Able to afford nicer, more expensive homes, they likely are skewing the median home value upwards.

x78/x81: 0.947 (PERCENT_POP_65_TO_74, MEDIAN_POP_AGE)

Likely this indicates a significant elderly contingent in Austin, skewing the median population age upwards.

x63/x65: -0.926 (PERCENT_NO_HEALTH_INSURANCE, PERCENT_FAMILIES_W_CHILDREN_BELOW_POVERTY)

While it is not surprising that households that are below the poverty line are also, unable to afford health insurance for their adult members, it is notable that these variables do not also track with the percentage of uninsured children. Possibly children living in poverty are successful in being caught by state-wide safety nets?

x26/x55: 0.915 (MEDIAN_HOME_VALUE, PERCENT_W_INCOME_200000_OR_MORE)

People who earn more buy more expensive houses.

x12/x18: 0.907 (MOBILE_HOMES_PERCENTAGE_OF_HOUSING, PERCENT_OF_HOMES_VALUED_LESS_THAN_50000)

Homes in Austin are very expensive due to demand outstripping supply. It would appear that very cheap homes are largely of the mobile home variety.

x18/x64: 0.903 (PERCENT_OF_HOMES_VALUED_LESS_THAN_50000, PERCENT_CHILDREN_NO_HEALTH_INSURANCE)

The best hypothesis we have for this relationship is that within the working poor demographic, there is a population who makes too much money for social safety nets (and thus can afford the lowest tier of home ownership) but insufficient income to afford health insurance without assistance.

x30/x36: -0.900 (PERCENTAGE_OF_RENTERS_PAYING_500_TO_999, MEDIAN_RENT)

The percentage of renters paying 500 to 999 dollars a month is numerous enough to skew the median rent value.

x55/x56: 0.905 (PERCENT_W_INCOME_200000_OR_MORE, MEDIAN_HOUSEHOLD_INCOME)

The richest Austinites are numerous enough to skew the median household income.

x64/x88: 0.909 (PERCENT_CHILDREN_NO_HEALTH_INSURANCE, PERCENT_POP_HISPANIC)

Any attempt to explain this relationship is pure conjecture.

x76/x81: 0.910 (PERCENT_POP_55_TO_59, MEDIAN_POP_AGE)

This age bracket consists of the oldest Gen X-ers and the youngest of the Baby Boomers. Reasonably, this would be senior managers, etc. within the working population. Apparently, they are numerous enough to skew the overall median age in Austin.

In considering what variables drop due to collinearity, aggregates of multiple variables were favored over variables describing a sub-category (e.g., median age vs. percentage aged 65-74). In the end, the following variables were dropped as predictors: x1, x4, x10, x15, x18, x24, x30, x38, x41, x55, x63, x76, x78. Unfortunately, even after dropping these highly correlated variables does not allow us to compute VIF within R. As such, we turn to the 'alias' function to find which variables are linearly dependent. These were removed for the model analysis. Figure 4 shows the correlation matrix of the variables after dropping all the variables as mentioned above. Figure 5 shows the VIF values for each variable. As can be seen in the plot of VIF values, there is high multicollinearity in play. Thus, it is important to run principal component analysis as an option to build model. It is explained in the section on model building and evaluation.

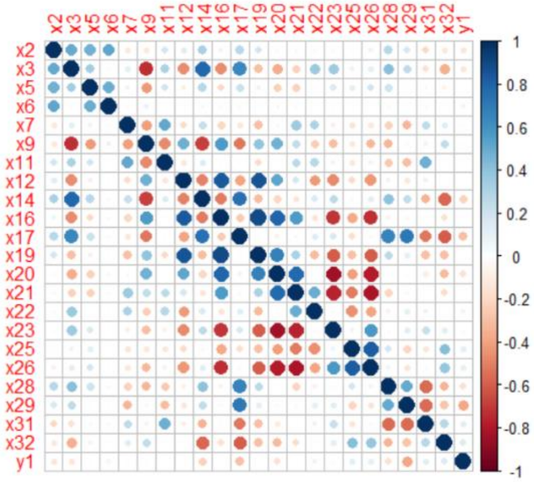

Figure 4: Correlation matrix after dropping the variables

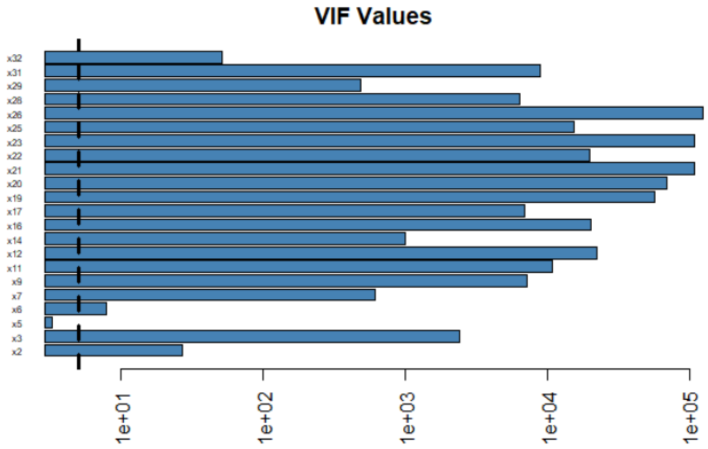

Figure 5: VIF values after dropping the variables

Going further, these predictors/independent variables for used for model building - x2, x3, x4, x5, x6, x7, x9, x11, x12, x14, x16, x17, x19, x20, x21, x22, x23, x25, x26, x28, x29, x31, and x32. Y1 ie. Campus graduation rate was selected as the primary dependent variable for our analysis since it is highly correlated to the other dependent variables and is not biased to any particular demographic.

Model Building & Evaluation: In the model building stage, we used 3 modeling techniques to build three different models. These approaches were used considering the higher than usual dimensional nature of the final dataset (although it is not qualified as a high dimensional dataset since n > p), high multicollinearity in the cleaned final dataset and a smaller number of rows overall. In addition to the above, the following tools were used during the model building and evaluation process -

- R programming language using R studio platform
- Within R, these libraries were used -
    - library(ggplot2): used for creating graphs
    - library(dplyr): used for data manipulation
    - library(reshape): used for data manipulation
    - library(tibble): used for data manipulation
    - library(tidyr): used for data manipulation

    - library(corrplot): used for performing correlation analysis
    - library(car): used as a supplement library for analysis
    - library(tidyverse)
    - library(pls): used for Principal component analysis regression
    - library(randomForest): used for creating data random forest for regression analysis
    - library(caret): used for variable selection using lasso regression

1: Principal Component Analysis: Principal component analysis was conducted since figure 5 shows signs of multicollinearity.
The following steps were taken to conduct the analysis -

- Few data modifications such as scaling were performed. Refer to the analysis file.
- The principal components were calculated using “prcomp” function of the “tidyverse” package
- Principal component scores were calculated i.e., how much of the variance of each predictor variable and total variance, each principal component explains.

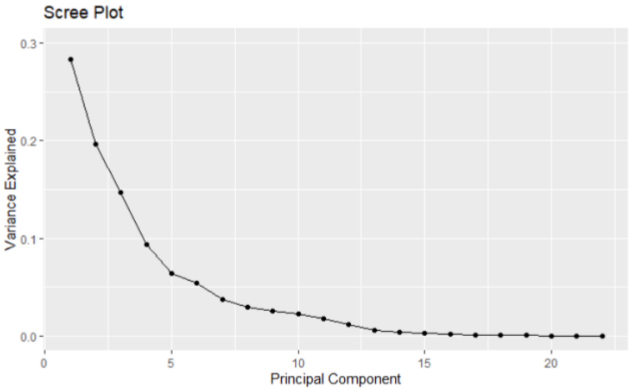

Figure 6: Scree plot showing total variance explained vs number of principal components

- Figure 6 shows a scree plot showing the drop in total variance explained with increase in the number of principal components. Around 5 principal components explain a substantial portion of the variance. The first five principal components explain 28.3%, 19.7%, 14.7%, 9.4%, and 6.4% of the total variance in the dataset, respectively.
- A linear regression model was built on all the 22 principal components calculated above using “pcr” function of the “pls” library with 10-fold cross validation. 10-fold cross validation helps with minimizing the error in the models and avoid overfitting of the models due to limited amount of data (105 rows).

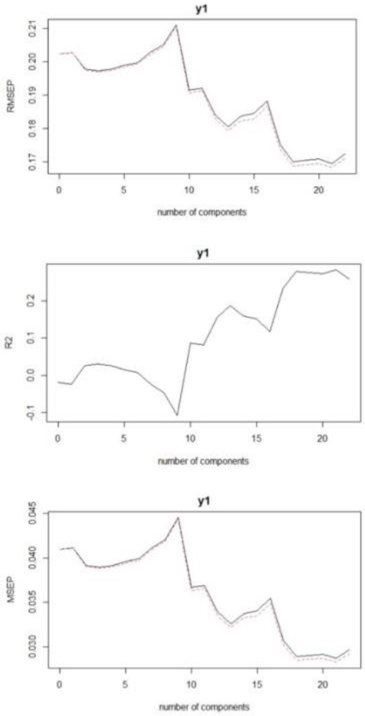

Figure 7: Plots showing variance in RMSEP, R2 and MSEP for predictions of overall graduation rate with number of principal components

- Figure 7 shows variance in RMSEP (root mean square error), R2 and MSEP (Mean Square Error) for models built using 1 principal component to 22 principal components.
- Based on results shown in figure 7, three cases using 2, 12, 16 principal components were tested to find out the best performing model. Data were standardized before running the regression analysis. Then, data was split into a 70% training set (which will also be used for validation) and a 30% test set. Again, a 10-fold cross validation is used with the regression analysis for each case.
- RMSE values were calculated for each case as 0.1882663, 0.2036243, and 0.2051205 respectively. Adjusted R^2 values for these cases were 0.0455, 0.197, and 0.1898 respectively.
- Conclusion – The PCA approach has some value in identifying significant features in the data and reducing multicollinearity. However, the predictive capability of this model is low as evident in the values of RMSE and adjusted R^2 above which may be explained by the dependent variables lacking a normal distribution.

2: Random Forest: Testing the random forest model with a lasso model for feature-selection was done in hopes to take advantage of its ability to work with many different variables and its ability to produce accurate predictions while being lightweight.

- Model setup – Random Forest analysis was run using the RF method in the 'caret' library, which also runs lasso regression technique for feature/variable selection. We also used 10-fold cross validation to train the model better.
- Model performance – There was only one data split done by the random forest algorithm as shown in the output attached. This could be due to several things such as a noisy dataset or insignificant variables from the lasso model. RMSE and adjusted R2 values were found to be 0.1952 and 0.4607 respectively.
- A variable importance chart was created using "varImp" function of the "caret" library with importance values between 0 and 100, 0 being the lowest importance and 100 being the highest importance. 20 most important variables are plotted as shown in figure 8

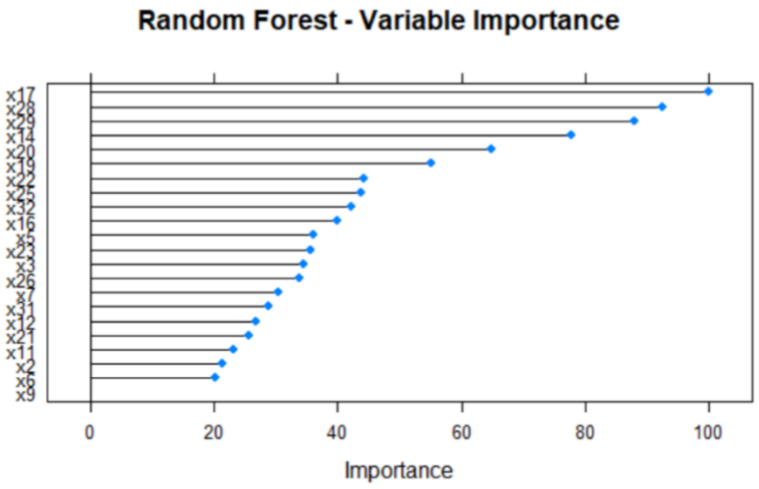

Figure 8 : Plot showing variable importance for 20 most important variables after random forest analysis

- Just as a sanity check we decided to run the random forest model without using lasso first for feature selection to see what the results would be.
    - In this case, the data only split one time thus no improvement over the previous case.
    - RSME and Adjusted R2 values for this case were 0.19014 and 0.45802 respectively.

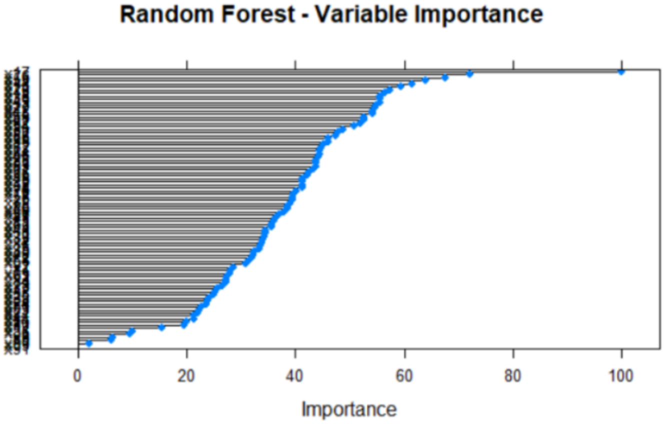

Figure 9 : Plot showing variable importance for 20 most important variables after random forest analysis using all the original predictor variables

- Conclusion – The random forest model only split once indicating that only one variable and value split were used to make the predictions. This is not what we were expecting and makes the model as useful as a coinflip.

3: Stepwise Regression: This modeling approach provides a way to simply remove variables by only including predictors that significantly influence the dependent variable. By doing this we remove the need for PCA and can better understand our results. Forward selection and bidirectional elimination were used to find the final set of predictor variables.

- Model setup -
    - Step function from the “stats” package was used to run stepwise regression. Direction “forward” and “both” were specified to build model using forward and bidirectional selection respectively.
    - The analysis using forward and both direction was run on the original dataset with all the possible predictor variables to allow this analysis to remove variables for us without any prejudice.
    - As the above models are created on the original dataset, the final model was created using simple linear regression and the final selected predictor variables found the step above and k-fold cross validation (to train the model and check

the accuracy of the model better). Figure 10 shows a correlation chart of all the final predictor variables.

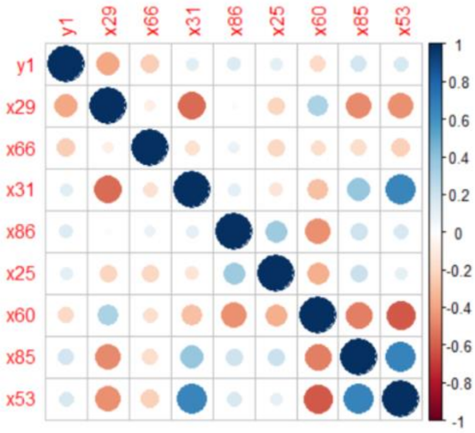

Figure 10 : Correlation chart of all the final predictor variables

- Model performance -
  - Based on stepwise regression using forward direction, x29, x66, x31, x86, x25, x60, x85, and x53 are the final predictor variables. Also, figure 11 shows the VIF values plotted for the final predictor variables. All these variables have VIF values less than 5.

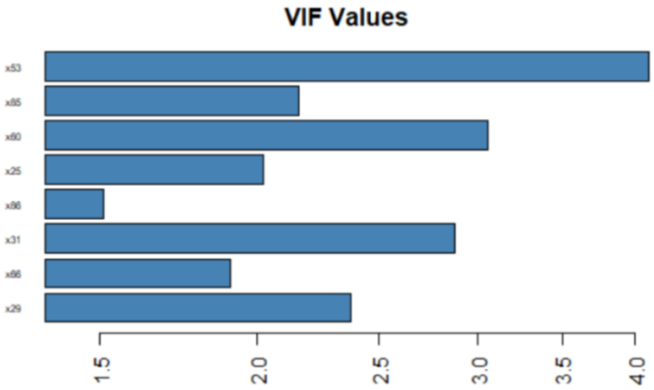

Figure 11: Plot showing VIF values for the final predictor variables based on forward direction stepwise regression

  - Based on stepwise regression using both directions, x29, x66, x31, x86, x25, x60, x85, and x53 are the final predictor variables.
  - The R2 and the adjusted R2 value of the final model were found to be 0.4285 and 0.3809 respectively.
  - Based on the results of the linear regression for the final model, x29, x66, x31, x86, x25, and x60 are statistically significant.
- Conclusion -
  - Stepwise regression analysis helped with coming up with a final set of variables which do not have multicollinearity as evident in figure 11.
  - The adjusted R2 value for the final model i.e., 0.38 is not great but an improvement on the principal component analysis approach.

Results

We explored three approaches to this analysis. Principal Component Analysis (PCA) with Principal Component Regression (PCR), Random Forest, and Stepwise Factor Selection with Regression. Our primary hurdle in all three approaches was removing collinearity in our potential explanatory variables. While PCA with PCR and Random Forest had results and error terms that in our view were not acceptable, the results we received from Stepwise Factor Selection reduced collinearity, yielded multiple statistically significant explanatory variables, and yielded an adjust R-squared of .38.

The highly statistically significant variables greater than 95% identified by stepwise factor selection and regression included:

- x25 = PERCENT_OF_HOMES_VALUED_1000000_OR_MORE
- x29 = PERCENTAGE_OF_RENTERS_PAYING_LESS_THAN_500
- x31 = PERCENTAGE_OF_RENTERS_PAYING_1000_TO_1499
- x60 = PERCENT_RECIEVING_PUBLIC_ASSISTANCE
- x66 = PERCENT_POP_MALE

The statistically insignificant variables less than or equal to 95% identified by stepwise factor selection and regression included:

- x53 = PERCENT_W_INCOME_100000_TO_149999
- x85 = PERCENT_POP_ASIAN
- x86 = PERCENT_POP_HAWAII_PAC_ISL

Interpretation

Over the course of this analysis, we used and sourced more data than we had initially expected. In broad terms, we collected data about the schools themselves and census data as the source of our potential explanatory variables. Intuitively when we began this analysis, we expected the school data, specifically expenditures and the number of teachers, to have more significance than we ultimately uncovered in our analysis.

While we were not convinced with the results of PCA with PCR and Random Forests, they provided useful context when considering the entirety of the analysis. Each model agreed on the class of independent variable that was identified as significant in explaining variance. This class included economic variables almost exclusively.

When we settled on stepwise and bidirectional factor selection many of the same types of variables where selected that appeared in the previous two models. Along with lower VIF values and multiple statistically significant variables, this provided additional confidence in our final model.

That final model yielded an adjusted R-squared of approximately .38. While we would have liked to yield a higher adjusted R-squared with our analysis, we are satisfied with the result we received. When considering the complexity of a problem such as education, we would not expect to find a high adjusted R-squared and would be suspicious of overfitting in the event we had.

Based on the model results we can say that economic factors are very significant in driving the graduation rate prediction. However, due to a high number of variables and randomness in the data our models may be skewed due to overfitting to the random effects. This could raise some uncertainty in the true significance. Seeing the pattern repeated over several models does provide some insight that is worth investigating further.

Next Steps

We believe that this analysis provides a basis for additional research and raises several additional questions. Primarily we believe the addition of subject matter expert in economics would provide valuable context into taking these initial findings and attempting to generalize the analysis outside of the confines of Austin, Texas. Would we find these factors to retain their significance on a national level?

References

1. Goenner, C. F., & Snaith, S. M. (2004). Predicting Graduation Rates: An Analysis of Student and Institutional Factors at Doctoral Universities. Journal of College Student Retention: Research, Theory & Practice, 5(4), 409–420. https://doi.org/10.2190/LKJX-CL3H-1AJ5-WVPE

2. Hintze, John & Silberglitt, Benjamin. (2005). A Longitudinal Examination of the Diagnostic Accuracy and Predictive Validity of R-CBM and High-Stakes Testing. School Psychology Review. 34. 10.1080/02796015.2005.12086292.

3. Johnson, K.A., Wilson, C.M., & Williams-Rossi, D. (2013). All Reading Tests Are Not Created Equal: A Comparison of the State of Texas Assessment ofAcademic Readiness (STAAR) and the Gray Oral Reading Test-4 (GORT-4).

4. McGowen, R. S. (2008). *The impact of school facilities on student achievement, attendance, behavior, completion rate and teacher turnover rate in selected Texas High Schools* (dissertation).

5. Ochs, Sarah & Keller-Margulis, Milena & Santi, Kristi & Jones, & H., J. (2019). Long-term validity and diagnostic accuracy of a reading computer-adaptive test. 10.1177/1534508418796.

6. Ritter, Barbara (2015). *Factors Influencing High School Graduation*. WSAC.

7. VanMeveren, Kalie & Hulac, David & Wollersheim Shervey, Sarah. (2018). Universal Screening Methods and Models: Diagnostic Accuracy of Reading Assessments. Assessment for Effective Intervention. 45. 153450841881979. 10.1177/1534508418819797.