Texas High School Dropout Rates

Emily Joseph
McNeil High School
15524 Belfin Drive, Austin, TX
+1 512-412-9381
emilyjesu@gmail.com

Collab Link:
https://colab.research.google.com/drive/1c8gwNpy2LL_al099cxO4EFmXPH1yHQYH?usp=sharing

Github Notebook Link:
https://colab.research.google.com/drive/1WIoq5qwSfLK9WCrAixJ6uTt166Z2p_bK?usp=sharing

ABSTRACT

Dropout rates in high schools throughout Texas have been steadily increasing, especially following the Covid-19 pandemic. Dropping out of high school has a permanent effect on the future of a person's life (as explained later), and as more students don't make it to graduation, the more serious this issue becomes. The only way to combat this problem is by identifying what factors or characteristics of a campus are causing students to drop out. This project aims to solve this problem through an artificial intelligence algorithm that can predict the dropout rate of a Texas high school campus based on specific characteristics of the school. This project uses a gradient boosting model on a dataset by the TEA (Texas Education Agency) with information about campuses throughout the state. The results of this model are also implemented into an app where people can input information about a school and get a dropout rate prediction which can help schools plan for the future.

1. INTRODUCTION

As Covid-19 hit the United States, it simultaneously affected the education systems throughout many states. Specifically in Texas, dropout rates increased dramatically. Not only have dropout rates increased, but an article in the Dallas News explains how many Texas public schools unrealistically inflate graduation rates due to cash incentives given to superintendents. Both of these issues help show how important it is to combat the ongoing dropout rate issue that has been ignored for many years. Studies also show that high school dropouts make around $200,000 less over their lifetime compared to someone who made it through high school. Additionally, highschool dropouts commit 75% of crimes in the US. This shows us how important it is to reduce the dropout rates in Texas so that we can improve the future and wellbeing of our students.

This project aims to combat the rising dropout rates in Texas through a machine learning model that is trained to predict dropout rates according to a number of inputed factors. This can help us understand which factors have the highest impact on dropout rates, and therefore help target the state funding on specific areas of groups within Texas. This model will help us identify possible issues and unfairness within our education system which can then be used by individual schools to figure out ways to improve graduation rates within their own campuses. This model can also be used to send supported recommendations to the TEA (Texas Education Agency) on how to improve the current education system within the state.

2. LITERATURE REVIEW

There are very few models that have been built to look at dropout rates within Texas. One study called Predicting K-12 Dropout by the Journal of Education for Students Placed at Risk (JESPAR), predicts factors that affect dropout rates in one specific ISD in Texas. They use their results to detect whether an individual student is at-risk for dropping out using specific factors about the student. A big limitation of most models that focus on this issue is that they only work within a specific ISD compared to the state of Texas as a whole. This model also focuses on helping individual students which has a lower range than giving recommendations to campuses as a whole. This specific model was also limited in the fact that it only predicted how likely a student is to drop out without giving insight or recommendations on how to support the student and help them graduate. There are other models that have looked at this issue but none of them have focused on Texas.

3. METHODS

3.1 MODELS

We tested our data on a variety of models so that we could compare the results and use the most accurate model to predict dropout rates, and therefore be able to find the most important factor causing high dropout rates. We tested our data on the Random Forest, Linear Regression, Decision Tree, and Gradient Boosting Models because we are looking at a regression problem, not a classification problem. Then to find the accuracy of the models, we used mean squared error and r squared.

A Decision Tree Model (Regressor) makes predictions based on how questions are answered. It resembles a tree with both decision nodes, as well as branches off of them that represent the different possibilities.
A Linear Regression Model makes predictions by establishing a line that best fits the data. This means that predictions are created solely based on other positioning and coordinates of the established line.
A Random Forest Model (Regressor) is basically many Decision Trees that work in conjunction to create the final prediction. For regression, this model takes the average of all outputs as the final output.
Gradient Boosting is similar to Random Forest because they both use decision trees. Gradient Boosting on the other hand starts off with a single model and improves that model slightly based on true values and what the previous model outputted as well as its weights. This continues to occur until the final model is highly accurate.

Mean squared error is the squared difference between the true value and the predicted value. Now because it is squared, the error or loss becomes more exaggerated as the distance between the true value and predicted value increases. This means that the MSE can never be negative and we want the mean squared error to be as close to 0 as possible.

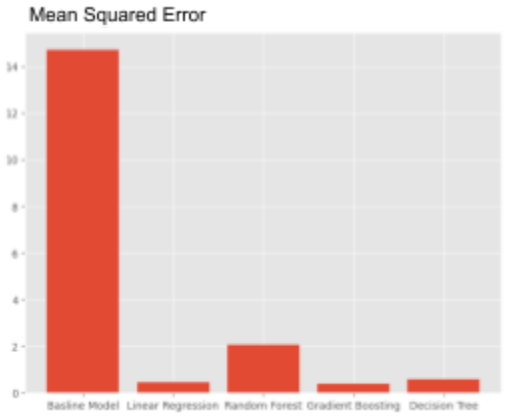
3.2 DATASET

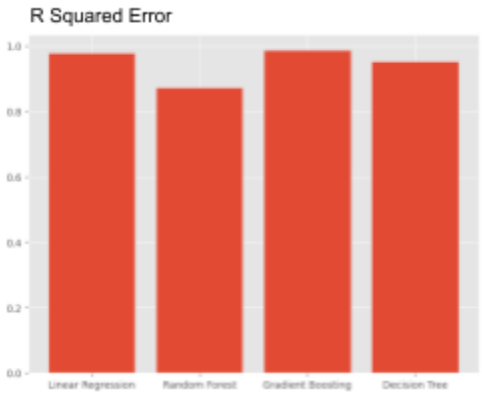
The dataset we used for this project was from the TEA (Texas Education Agency of Education). It consists of one excel file with the data, as well as a data dictionary. This dataset has a total 18,095 pieces of data. Each piece of data corresponds to a specific campus within Texas and has many different characteristics about the school. This includes factors like the annual dropout rate, the county name, the number of students, the demographic, the economic status, and more.

There were a few tweaks that needed to be made to the dataset so that it would work well for our purpose. First off, there were two rows for every school that were identical except for one row which calculated state accountability per Texas Education Code. I decided to eliminate every other row of the dataset to make the dataset more concise. Then there were also null values within the dataset. We looked at the distribution of null values and realized that all the null values were found in rows that calculated the annual dropout rate within a specific group of students (for example Asian students). We then realized that if there are no students in a campus within that group, then the denominator of the rate would be zero, making the rate invalid or null. This prompted us to fill all null values with 0. This dataset also had campuses with a grade span of only 7th to 8th grade. Since in this research project we are focusing on high school dropout rates, we also omitted any campus that ran from 7th to 8th grade. The last tweak that had to be made to our dataset was because there were some values that read "<0.1". These values were within rows that held numeric values and since <0.1 would be considered a string, we decided to replace all instances of <0.1 with 0.05 because it is between 0 and 0.1. This way all the values in these rows can be converted to a float or number easily.

4. RESULTS

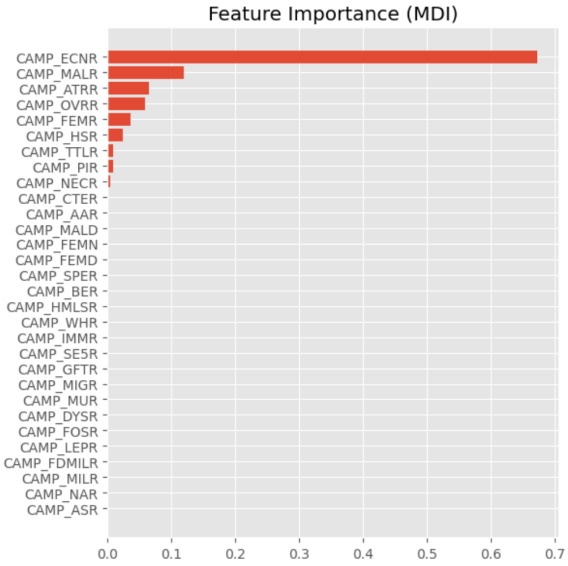
The accuracy for the different models was calculated in two ways; r squared and mean squared error. We then created a bar graph comparing the different accuracies of the models (one for mean squared error and one for r squared), which led us to the conclusion that the Gradient Boosting Model was the most accurate with an accuracy of 98.4%.





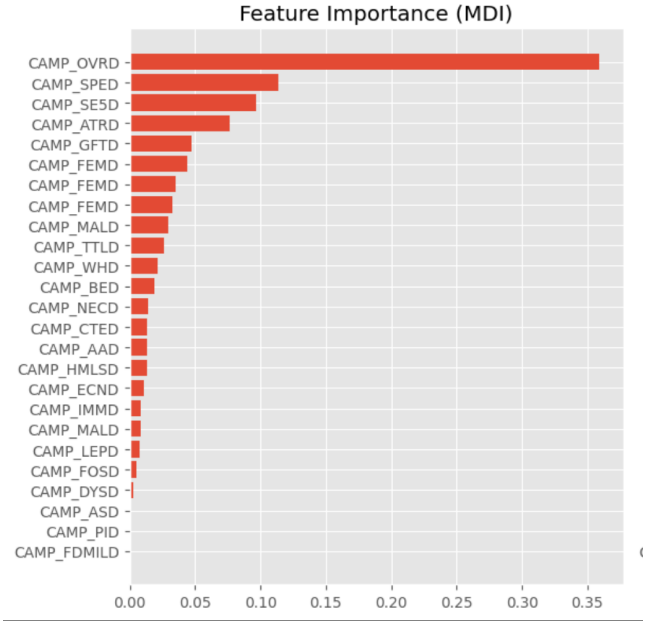
We then created bar graphs that show the feature importance and permutation importance within the Gradient Boosting model. This bar graph showed us that the majority of factors played little to no role in the dropout rate calculation. This led us to developing a

model that only took in the factors that had an effect on the dropout rate prediction(according to the bar graph).



This final model showed that the dropout rate within economically disadvantaged students was the biggest indicator of a high dropout rate. This factor was more than 4 times more important than succeeding factors (annual male dropout rate was the next most significant). This model shows that for Texas to decrease dropout rates throughout the state, we need to look into which areas have the highest economically disadvantaged dropout rates, why the rate is so high, and focus funding on these campuses to help give students the support necessary to graduate.

After creating this new highly accurate model, we realized that the input that would be going into the model is the actual dropout rates within each specific category (for example, the dropout rate within economically disadvantaged students). Our model would be more helpful if the input was basic characteristics about the school compared to individual dropout rates. Due to this, we decided to create a model that took in the actual numbers of total students in each column, not the dropout rate within each column. We created both a Linear Regression model as well as a Gradient Boosting Model with this altered x train and test. As expected, our accuracy for these models was much lower than when we used rates. For the Gradient Boosting Model, our r-squared error was around .42 and the mean squared error was around 7.42. The linear regression model performed slightly worse with an r-squared error of .123 (we want r-squared to be close to 1). Even though these models were significantly less accurate, they will still be very useful, especially compared to the expected human accuracy for predicting dropouts rates using this input data. We also used permutation importance to take out some factors that had close to no effect on the dropout rate predictions.



In this Gradient Boosting Model, the most predictive factor is the number of overage students in the campus. Overage students are students that are over the traditional age range for their grade and are either because they enrolled in school late or because they were retained/held back. This model shows that for Texas to decrease dropout rates throughout the state, we need to look into why students are being held back, specifically at the schools with high numbers of overage students. If this is because of an exam or grade requirement (including STAAR testing), the TEA should look into ways to help support schools and students pass these requirements.

5. CONCLUSION

In conclusion, our models can be used to predict dropout rates within Texas high schools and are able to show us the main indicators of high dropout rates. Our first model used the dropout rates within specific groups as the input and predicted the campus dropout rate with very high accuracy. This model also helped us conclude that the dropout rate within economically disadvantaged students is the largest indicator of the overall campus dropout rate. This shows that for Texas to help more students graduate high school, the government should focus funding and support to schools that have a larger number of economically disadvantaged students.

Our second model used the total number of students within specific groups as the input and predicted the campus dropout rate with significantly less accuracy. This model helped us conclude that the number of overage students within a campus is the largest indicator of a high dropout rate. This means that the TEA should look into the main reasons why students are being held back (not passing grade requirements, failing state issued tests like the STAAR test, attendance) and make sure to combat these specific issues.

Even though both our models are useful, you can see the trade-off between more realistic input data and accuracy. If we are focused on the prediction aspect of a school's dropout rate and we have the necessary information about the campus, the first model will be more useful. If we're less focused on the exact dropout rate

prediction, but more on improving the current campus for students, the results of the second model would be more useful.

6. LIMITATIONS OF THE STUDY

One of the limitations of our model is that it is not individually trained on each ISD, meaning that it focuses on more broad patterns throughout Texas compared to specifics of individual ISDs. Our model is also trained using a dataset from the school year of 2019 to 2020, meaning that the relevance of the results will decrease over time. We also have to consider the fact that there may be other factors that weren't included on this dataset that have a large effect on dropout rates.

7. REFERENCES

Critics scrutinize texas' unusual high school dropout rates. Dallas News. (2019, August 26). Retrieved October 10, 2022, from https://www.dallasnews.com/news/education/2015/08/30/critics-scrutinize-texas-unusual-high-school-dropout-rates/

Predicting K-12 Dropout. Digital Object Identifier System. (n.d.). Retrieved October 10, 2022, from https://doi.org/10.1080/10824669.2019.1670065

Texas Education Agency. (2021, August 13). *Annual Dropout Data, 2019-20*. Texas Education Agency. Retrieved October 10, 2022, from https://tea.texas.gov/reports-and-data/school-performance/accountability-research/completion-graduation-and-dropout/annual-dropout-data-2019-20