Predicting Kentucky's Failing Schools

Elizabeth Miller
April 8, 2021

Introduction

Kentucky ranks 45th in the nation in uneducated population-- only 5 states have a worse education system. Therefore, our state is one of the least educated in the country. Additionally, our average dropout rate is 6.6%, which is above the national average of 6%. As an educator for the last decade, this is hugely concerning. While Kentucky does post this data on their Department of Education website, no studies have been done on dropout rates in Kentucky. The purpose of this research is to provide the state of Kentucky with some guidelines moving forward in order to reduce dropout rates. First by creating a model to predict dropout rates then by using these to determine what features help or hinder the most.

Data Cleaning

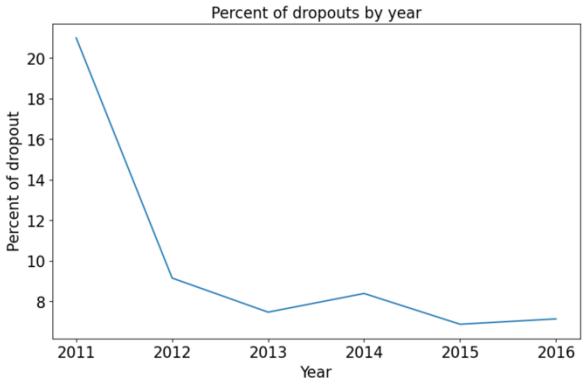
The data collected from this project comprised of 33 separate csv files containing state testing data, learning environment data (qualifications of teachers, free and reduced lunch, technology available, etc.), and school safety data (behavior referrals) for each school in Kentucky. In addition to that, I added county level demographic data on child poverty, low birth weight, born to smoking mothers, unemployment and median household income from Kids Count – a nonprofit organization which gathers any data relating to children in the US. The data spanned from 2011 to 2017.

Combined this was around ~18 million data points. I began by compressing the data into each row being a single school, initially there were about 80 rows per school with a very sparse format. Then I went further and make each row into a single high school in Kentucky (1 entry for each of the 6 years studied) -- I changed middle and elementary data into columns, then I dropped any schools that do not participate in state testing. In the end I had 1,355 rows and 52 columns, about ~70,460 data points total. This meant the data was reduced by 99.6% overall.

The data was very dirty overall. Many reports did not use a typical system for reporting NaN values. There were S's, *'s, **'s etc. and I had to email the Department of Education to get clarification on the meaning of each in order to decide how to clean it. Some years did not report either. Additionally, a lot of schools were reported with different names in different years so tracking them all down was tough. I also had to add the county to each school so that the demographic data could be mapped onto it as a key.

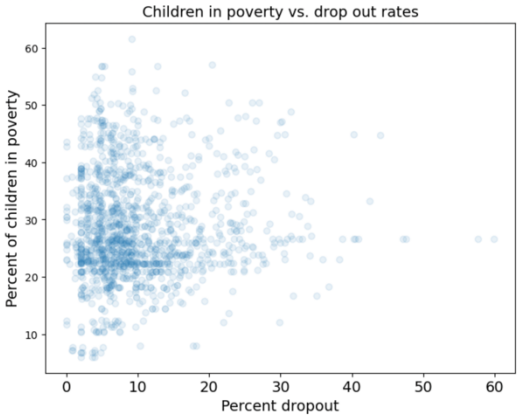
EDA

First, it was clear that drop out rates have reduced since 2011. The law changed in 2012 requiring that students be 18 before they could drop out. Previously the age was 16, so 2011 to 2012 there was a large drop in dropouts due to this.



Predicting Kentucky's Failing Schools

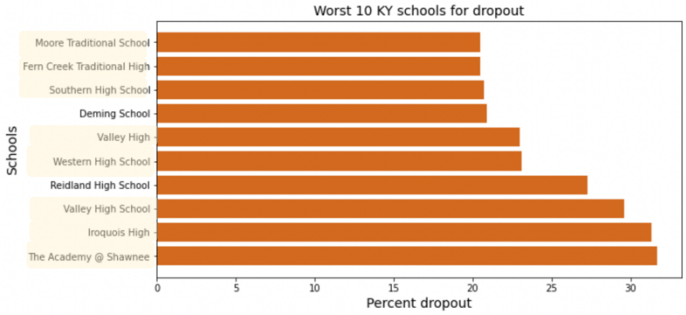
Elizabeth Miller
April 8, 2021



I also saw that poverty was not a big indicator for dropout which was shocking for me. The state of Kentucky often blames poor education results on the high poverty rates within. It's interesting that this isn't actually the case.

Additionally, it was clear that the average dropout rate from 2011 to 2017 was 10%, not the 6.6% reported. I have looked at the dropout rates for 2018 and 2019 separately form this report and they are both above 6.6% as well so I'm not sure where Kentucky is getting these numbers.

I also found that Louisville, KY is an educational battleground. The majority of the best schools in the state (in terms of dropout rates and test scores) are from Louisville and the majority of the worst schools in the state are also from there. In this graph, the highlighted schools are Louisville ones.



Pre-processing

In the pre-processing phase the clean dataframe that was already train/test split was scaled. But I also wanted to look at a variety of data sets so I created a data set with correlated features removed and then I created a dataset comprised of only features that schools can actually change (dropped the demographic data).

First, I mapped the correlated features in a heat matrix and dropped any features with > 60% correlation with each other. After that I looked at the multicollinearity remaining with the VIF score but decided to continue forwards regardless because the strength of the model was more important given the time frame. I did the same for the school actionable dataset. I then scaled both.

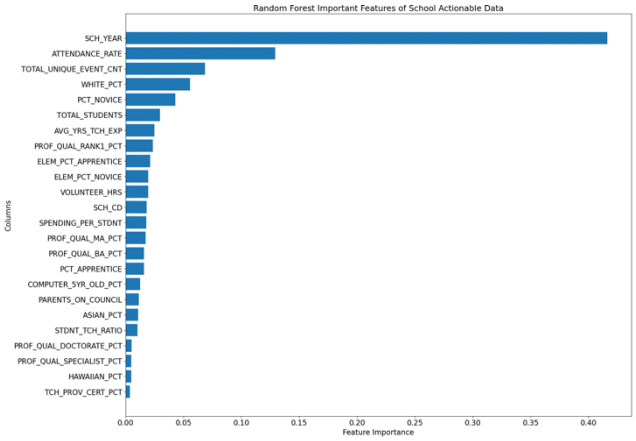
Lastly, I decided to use the following regression models to predict dropout rates: linear regression, KNN regressor, decision tree regressor, and random forest regressor. Although KNN's don't give the feedback needed on feature importance, the nonlinear decision boundaries seemed like they might be useful. To determine what features, lead to greater dropouts, I decided to use feature permutation importance on my decision tree and random forest models since they don't have the coefficient function that linear regressions do.

Modeling

For the prediction models, I got the following results seen in this table. To evaluate which model was best I used both rsquared value and root mean squared error. In the end the random forest was the best predictor model with 72% rsquared and 4.4 RMSE.

Algorithm Used	Best R2	Best RMSE
Linear Regression	50%	5.9
KNearest Neighbor Regressor	53%	5.8
Decision Tree Regressor	65%	5.3
Random Forest Regressor	72%	4.5

The coefficient function of the linear regression model showed me that a few demographic issues did effect dropouts… low birth weight (an indicator of mothers who did drugs while pregnant), and percent of missing child support, but no direct poverty feature such as free and reduced lunch, or the actual percent of children living in poverty feature. Additionally, the percent of students graduating college within 4 additional years was import. In terms of what schools could actually control, attendance rate, minority makeup of the school, and poorly qualified teachers negatively impacted dropout rates, but good elementary and middle school test scores and highly qualified teachers positively affected dropout rates.



The decision tree and random forest permutation importance showed similar findings, but also the number of behavior incidences (suspensions) in schools were important as well.

Conclusions

The Kentucky Department of Education should focus on Jefferson County Public School District as their top priority in the coming years. This district is in the major city of Louisville and contains most of the top performing schools and also the lowest performing schools in the state. The vast inequity in that district needs to be resolved.

Additionally, they need to accurately report the state's average dropout rate, because based on all of the data available on their website, the correct dropout average is 10%.

For districts I found the following goals:

1. Increasing teacher quality: We saw that schools with better dropout rates had teachers with higher degrees and more years of teaching experience. Districts should provide more incentives to attract more qualified teaching candidates
2. Improving attendance rate: Schools and districts should work on increasing daily attendance. Perhaps by rewarding high attendance, students might attend more.
3. Reducing behavior incidences: The latter models showed that having more suspensions in schools resulted in more dropouts. Perhaps a school wide behavior policy changes should be looked into.
4. Reducing novices on state testing: Districts focus on test scores every year; this should continue to be a major focus.
5. Improving diversity among all schools in the district: There is some political issues with this one, but schools should have roughly equal diversity amongst the race and household incomes of their students.
6. Focus less on technology: Kentucky has been pushing for increased technology lately, but none of the models showed that technology effects dropout rates. This isn't to say it doesn't affect student outcomes of achievement… that wasn't the focus of this research.

These findings greatly answer my initial goal of the project (find which features lead to dropouts). Next steps are to give this report to Kentucky's Department of Education via the commissioner, and to my old school district to use. Future steps would be to apply a neural network to see if prediction could be increased.