School Dropouts

Anthony Mansur

December 19, 2021

Contents

The code to reproduce this report is available on Github.

1 Executive Summary

Problem. In 2019, there were about 2 million status dropouts, or people between the ages of 16 through 24 who are not enrolled in school and have not earned a high school credential. Although this number isn't alarming, and has been decreasing over the years, the rates aren't equal among U.S. counties. There are several reasons why students decide to drop out. These may include not feeling supported, challenged, or motivated; not being able to keep up with school work; having to take care of a single parent or younger siblings; just to name a few.

In this report we take a look at the highschool dropout rates in 2007 of several U.S. counties, along with statistics on teacher compensation and level of education, and county-level data to see if we can find out what are the most important factors that impact whether or not a student may drop out.

[https://nces.ed.gov/fastfacts/display.asp?id=16]

Data. The dataset is pulled data from two sources. The first source comes from the National Center for Education Statistics (NCES). From the NCES, there was two datasets that were pulled, one containing the dropout rates for 2007 of over 18,000 Local Education Agencies (LEAs), or public authorities within a state that maintains administrative control of public schools. The other dataset is a teacher compensation survey from 2006-2007 from 17 participating states that includes the mean base salary segregated by level of experience and the percentage of teacehrs within each bracket of teaching experience. The second source is from the National Historical Geographic Information System (NHGIS) with a dataset from the 2009 American Community Survey. This dataset includes county-level statistics over a five year period (2005-2009) and includes both economic data (i.e., gini-index, per-capita income, percentage of pop. on welfare) as well as demographic data (i.e., percetanges of pop. with high school degree or higher, that are married, and that are single mothers). The statistics on the LEAs were averaged and mapped to the corresponding counties that are located in.

Analysis. This report contains several predictive models that attempt to see which are the best factors to use to predict the high school dropout rate of a county. The report includes six different cross-validated models: ordinary least squares, ridge regression, LASSO regression, elastic net regression, random forest, and boosting. Of the regression models, ordinary least squares had the least test error, while both random forest and boosting had approximately the same test error (and the lowest overall). All models had test errors lower than the intercept-only model.

Conclusions. Based on all the models that were analyzed, the two strongest predictors of a county's dropout rate are the domestic setting under which the student is in and the teachers they have. All models used the fact that counties with the higher percentages of single mothers tended to have higher school dropout rate while, on the contrary, those with a higher married population tended to have lower dropout rates. Furthermore, counties with more experienced teacher, not just necessarily those that were paid more, tended to have lower dropout rates. Although keeping the limitations of this analysis in mind, the data supports the claim that if a county wants to lower their dropout rates, it should focus on the quality of its teachers and further analyze how domestic affairs impact students and their ability to stay in school.

2 Introduction

Background. TODO

Analysis goals. TODO

Significance. TODO

3 Data

3.1 Data sources

The first source of the data comes from the Common Core of Data (CCD) nonfiscal surveys submitted annualy to the national center for Education Statistics (NCES) by state education agencies (SEAs). The first dataset from this source contains the number of students who received a regular diploma or alternative credential, the Average Freshmen Graduation Rate (AFGR), the enrollment base for the AFGR, and the number of dropouts for LEAs from the NCES Common Core of Data LEA Universe. This dataset can be obtained via this link: https://nces.ed.gov/ccd/drpagency.asp

The second dataset from this source is a Teacher Compensation Survey taken at the LEA level. It includes data from 17 participating states in 2006-2007: Arizona, Arkansas, Colorado, Florida, Idaho, Iowa, Kansas, Kentucky, Louisiana, Maine, Minnesota, Mississippi, Missouri, Nebraska, Oklahoma, South Carolina, and Texas. The data items aggregated at the LEA level includes: the percentage distribution of teachers by education background and by teaching experience grouping, the mean number of years of teaching experience, the mean base salary, and the mean base salary by teaching experience groupings. For more information, see https://nces.ed.gov/ccd/tcssurv.asp

The third dataset from this source is a geographical reference file that is used to map each LEA to a U.S. county. For more information, see https://nces.ed.gov/ccd/data_grf.asp

The second source of the data comes from the NHGIS via their data extraction tool https://data2.nhgis.org/main. The dataset extracted is the 2009 American Community over the span of 5 years (2005-2009). See https://www.census.gov/data/developers/data-sets/acs-5year/2009.html for more information. Only a select few variables were selected from this dataset, including economic and demographic factors at a county level. See below for more information.

3.2 Data cleaning

To clean the data, the two county files were joined into a table by the unique identifier, and only the final selected variables of interest were extracted from this file, per county. Then, the variables were modified to be with respect to population percentage. For instance, the proportion of married households is equal to the number of married households divided by the total number of households recorded.

After the county-level data was compiled into one table, the teacher compensation dataset was joined with the dropout dataset, with the LEA being the unique identifier. Only the dropout rate was extracted from the dropout dataset, while all the variables from the teacher compensation dataset were used. Note, the teacher compensation dataset had some missing values, so for complete analysis of an LEA, we dropped all the observations that had an incomplete variable.

Once all the teacher compensation statistics and dropout rate were joined into a table at the LEA level, the LEA and County mapping table was used to map and aggregate each LEA statistics to the county that LEA belonged to.

Finally, the county-level data was joined with the aggregated teacher and dropout dataset to get the final table. See below for the final variables that were used for analysis.

3.3 Data description

3.3.1 Observations

The dataset has a total of 1238 observations, corresponding to U.S. Counties

3.3.2 Response Variable

Our response variable is the dropout rate of that county, measured by taking the mean of all the measured LEAs corresponding to that county.

3.3.3 Features

We had a total of 17 features used for analysis.

- `married`: Number of married-couple family household over the total number of households
- `single_mom`: Number of Female family households with no husband present over the total number of households
- `gini_index`: the gini index of income inequality
- `per_capita_income`: the log of Per Capita income in the past 12 months (in 2009 Inflation-Adjusted Dollars)
- `welfare`: Number of households receiving public assitance over the total number of households
- `educated`: Number of people with a high school degree or higher over the total population 25 years and over.
- `advanced_educ`: Percentage of full-time teachers with a masters or doctorate degree
- `exp_1_5`: Percentage of full-time teachers with 1-5 years of teaching experience
- `exp_6_10`: Percentage of full-time teachers with 6-10 years of teaching experience
- `exp_11_20`: Percentage of full-time teachers with 11-20 years of teaching experience
- `exp_21_30`: Percentage of full-time teachers with 21-30 years of teaching experience
- `exp_gt_30`: Percentage of full-time teachers with over 30 years of teaching experience
- `mbsal_1_5`: Mean base salary of teachers with 1-5 years of teaching experience
- `mbsal_6_10`: Mean base salary of teachers with 6-10 years of teaching experience
- `mbsal_11_20`: Mean base salary of teachers with 11_20 years of teaching experience
- `mbsal_21_30`: Mean base salary of teachers with 21_30 years of teaching experience
- `mbsal_gt_30`: Mean base salary of teachers with over 30 years of teaching experience

3.4 Data allocation

The data was randomly split into training and test groups, with 80% of the observations being used for training and the remaining 20% for testing. A set seed was used so that the results were reproducible. Note, any N/As were removed from the data prior.

3.5 Data exploration

3.5.1 Response

The response variable's distribution is shown in Figure 1. As seen in the box plot We first sought to understand the response variable's distribution. As seen in the boxplot of the dropout rate variable, the data appears right-skewed with several outliers to the far right. The median dropout rate is 2.9 percent. The sorted data (Table 1) shows the counties in this dataset with the highest dropout rates. From the 17 states being analyzed, the top ten counties came from Lousiana, Colorado, and Minnesota.

3.5.2 Features

Histograms of each explanatory variable is shown in Figure 2

The correlation of the features are broken down into three correlation tables. Two for demographic and economic variables (see Figure 3), and one for the total correlations (see Figure 4).

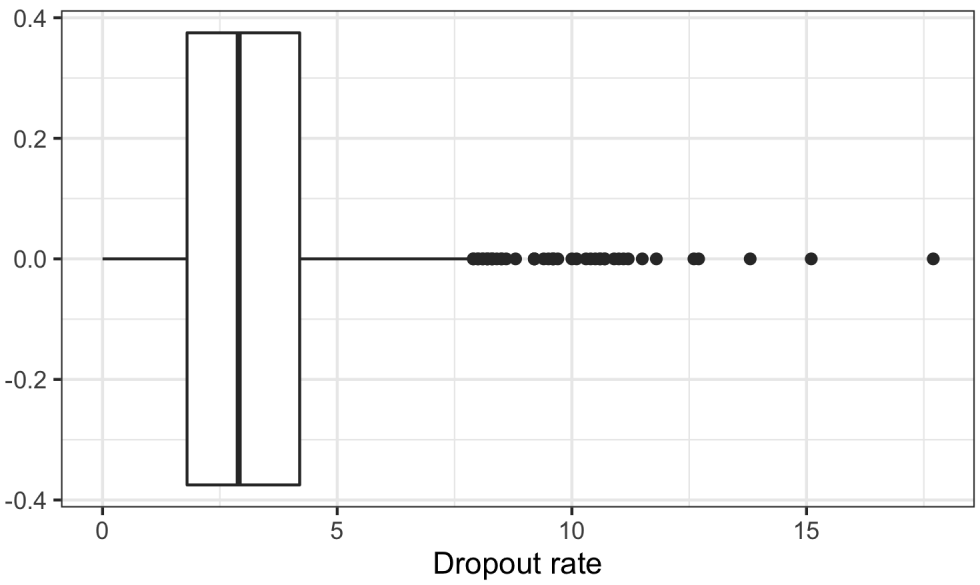

Figure 1: Distribution of dropout rate for the counties observed.

Table 1: Top ten counties by dropout rate (expressed as a percentage).

County	State	Dropout rate
Red River Parish	Louisiana	17.7
Morehouse Parish	Louisiana	15.1
Denver County	Colorado	13.8
Huerfano County	Colorado	12.7
Madison Parish	Louisiana	12.6
Mahnomen County	Minnesota	11.8
Lake County	Colorado	11.5
Iberville Parish	Louisiana	11.2
Union Parish	Louisiana	11.1
Beltrami County	Minnesota	11.0

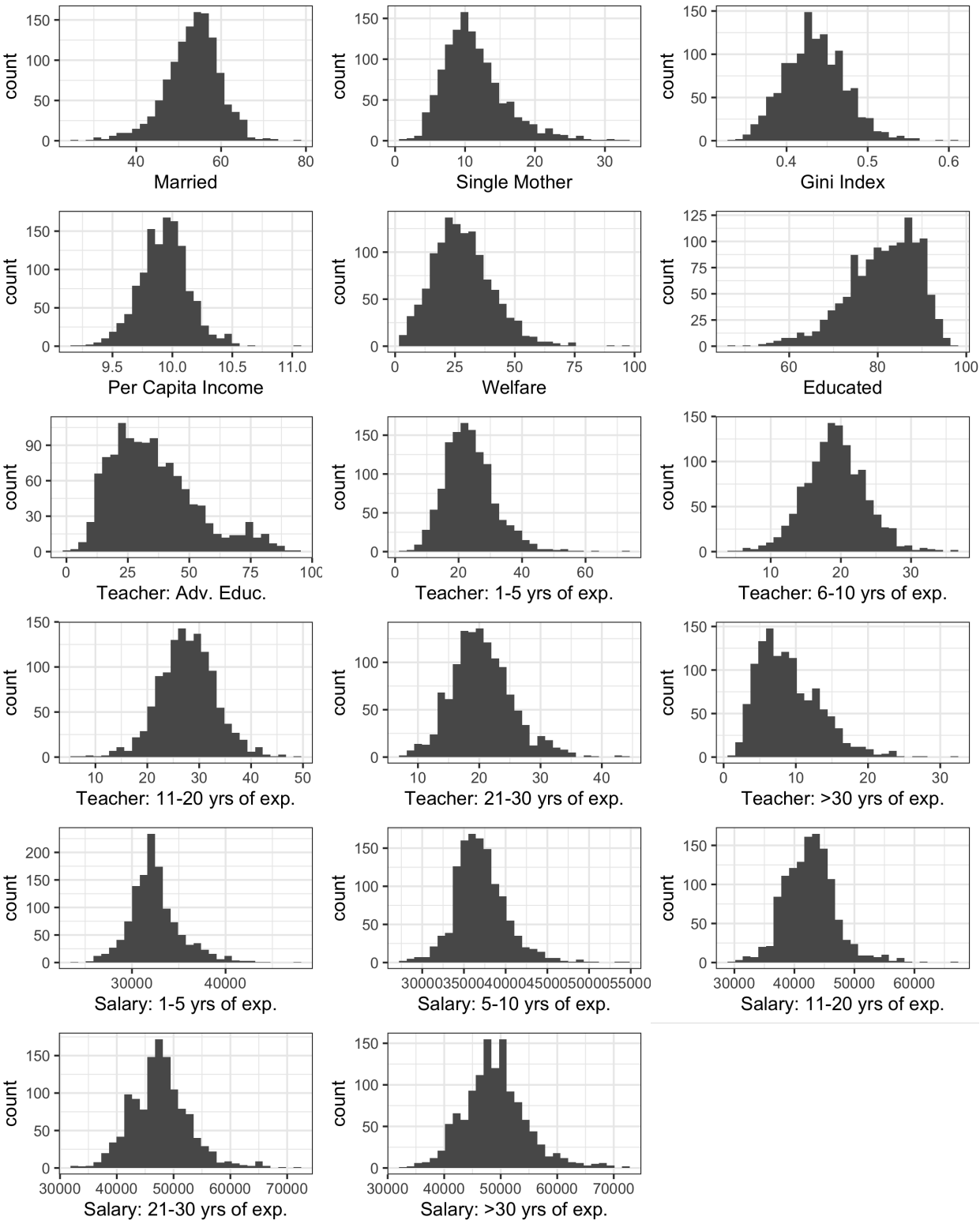

Figure 2: Distribution of explanatory variables.

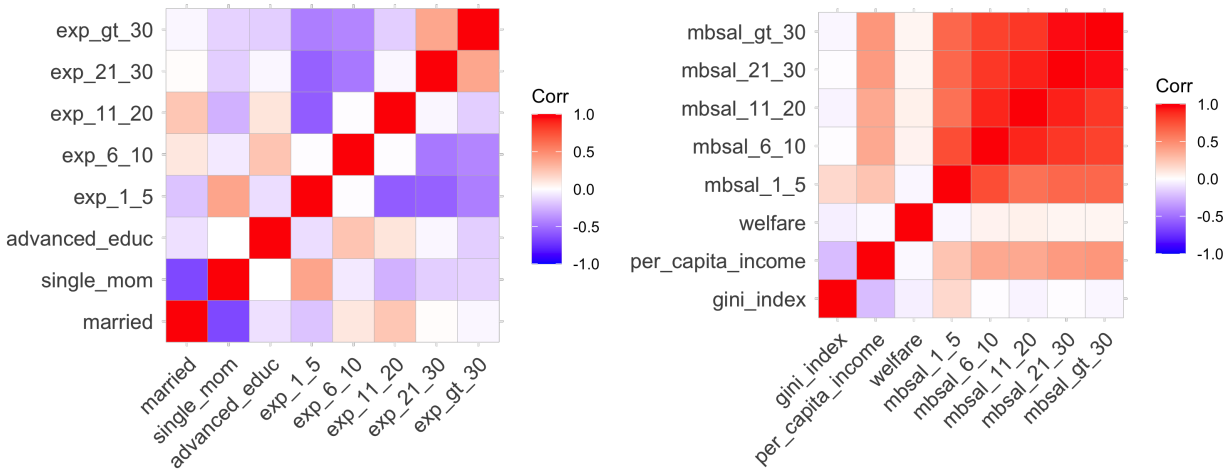

Figure 3: Distribution of explanatory variables.

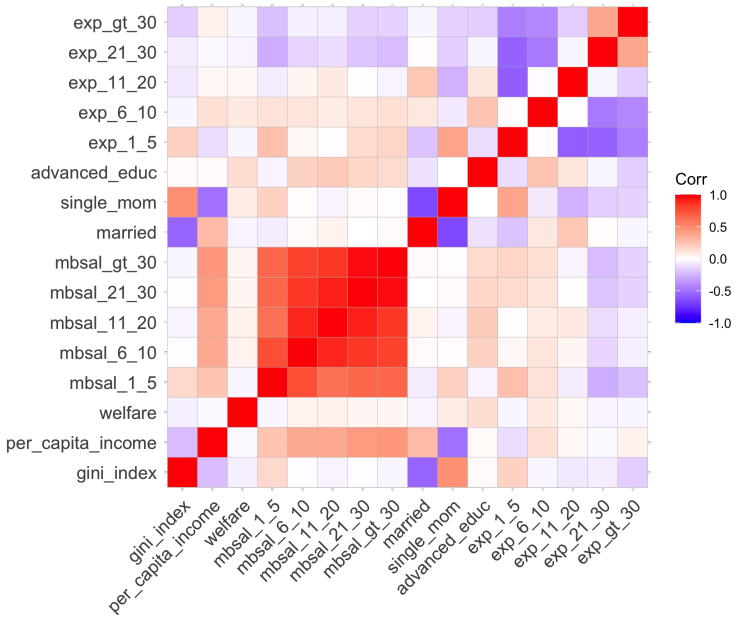

Figure 4: Distribution of explanatory variables.

Plotting some explanatory variables (see Figure 5) with the response gives some indications to the following claims: counties with higher proportions of married couples, with better experienced teachers, lower inequality, higher income, and less welfare needed tend to have less dropouts than those with higher single mothers, less experienced teachers, higher inequality, more welfare needed, and more unequal counties.

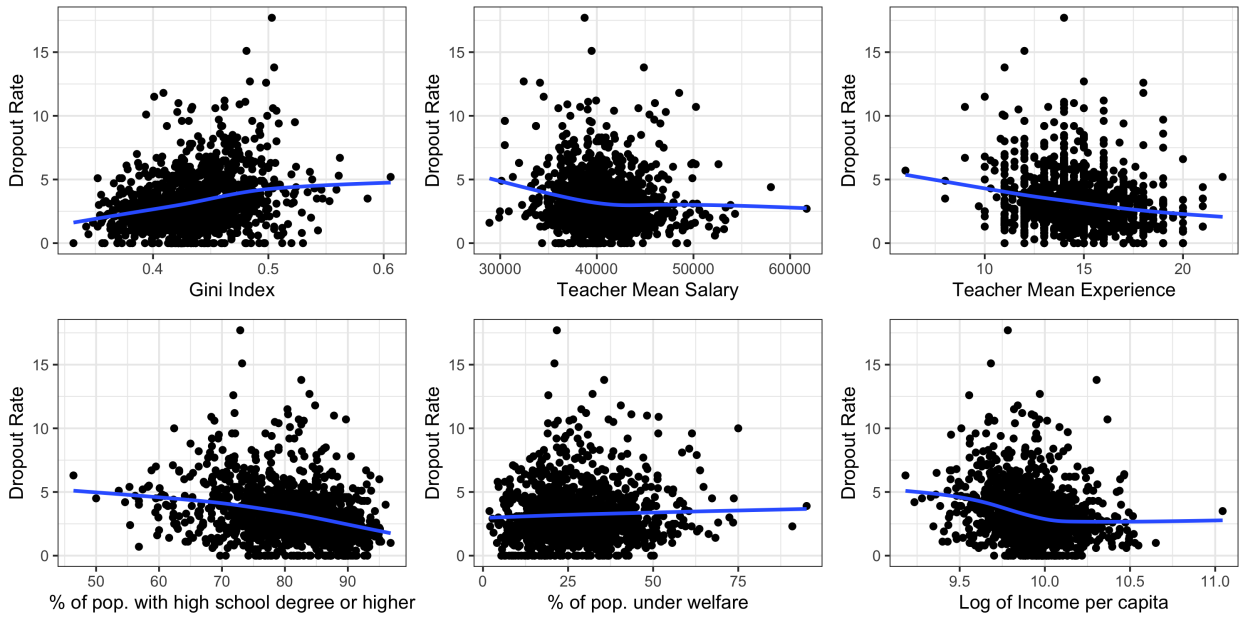

Figure 5: Initial findings with response and explanatory variables

4 Modeling

4.1 Regression-based methods

4.1.1 Ordinary least squares

The analysis for begins with an ordinary least squares regression of dropout rates on all 17 explanatory variables.

The summary plot is shown in Table (2)). The multiple R-squared is 0.247, indicating that these features explains 24.7% of the variation in the response.

4.1.2 Ridge Regression, LASSO Regression, and Elastic Net Regression

When fitting linear models with many features, they incur lots of variation and they tend to overfit the training data. Hence why other penalizations methods are performed to combat this and provide more insights to the feature variables that are most important.

Figures 6, 7, and 8 shows the trace plots for ridge regression, LASSO regression, and elastic net regression respectively.

The variables selected by LASSO are shown in Table 3. The top 5 featurs selected by elastic net regression are show in in Table 4.

Table 2: Ordinary least squares summary table on explanatory variables

...1	term	estimate	std.error	statistic	p.value
1	(Intercept)	91.5528	81.6907	1.121	0.2627
2	married	-0.0737	0.0152	-4.864	0.0000
3	single_mom	0.0992	0.0231	4.291	0.0000
4	gini_index	-1.2907	2.2510	-0.573	0.5665
5	per_capita_income	0.8333	0.5684	1.466	0.1429
6	welfare	0.0068	0.0050	1.359	0.1744
7	educated	-0.0338	0.0145	-2.328	0.0201
8	advanced_educ	-0.0054	0.0040	-1.357	0.1751
9	exp_1_5	-0.8826	0.8148	-1.083	0.2790
10	exp_6_10	-0.8804	0.8160	-1.079	0.2808
11	exp_11_20	-0.9122	0.8149	-1.119	0.2632
12	exp_21_30	-0.9031	0.8158	-1.107	0.2686
13	exp_gt_30	-0.9275	0.8158	-1.137	0.2559
14	mbsal_1_5	0.0001	0.0000	3.146	0.0017
15	mbsal_6_10	-0.0001	0.0001	-1.315	0.1890
16	mbsal_11_20	0.0001	0.0001	1.614	0.1068
17	mbsal_21_30	-0.0002	0.0000	-4.079	0.0000
18	mbsal_gt_30	0.0001	0.0000	2.383	0.0174

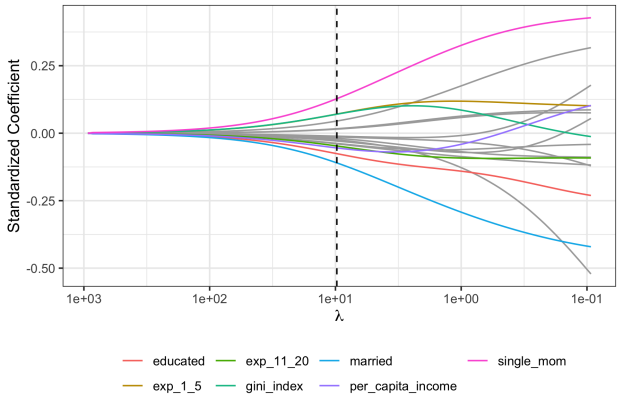

Figure 6: Trace plot for ridge regression

Table 3: Explanatory variables selected by LASSO regression

feature	coefficient
single_mom	0.464
married	-0.037

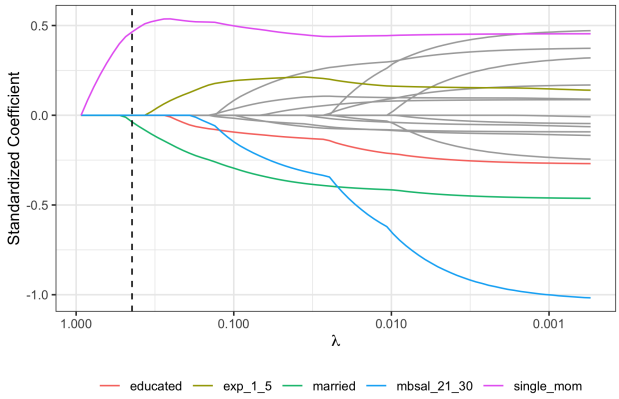

Figure 7: Trace plot for lasso regression

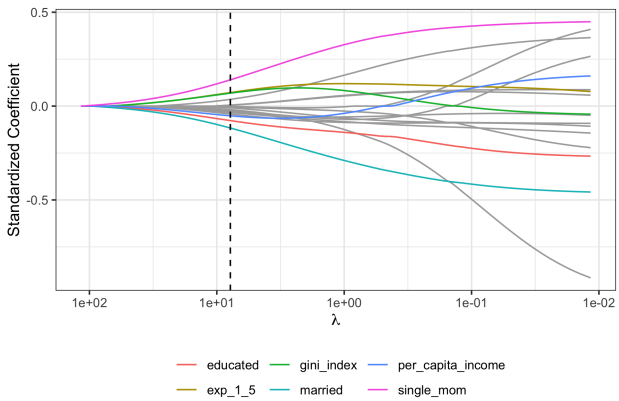

Figure 8: Trace plot for elastic net regression

Table 4: Explanatory variables selected by elastic net regression

feature	coefficient
single_mom	0.140
married	-0.117
educated	-0.077
exp_1_5	0.072
gini_index	0.069

4.2 Tree-based methods

4.2.1 Random forest

For random forest, we first optimize the number of trees to consider. The out of bag error is shown below. As we can see, the error decreases and remains relatively constant after around 200 trees. See Figure 9

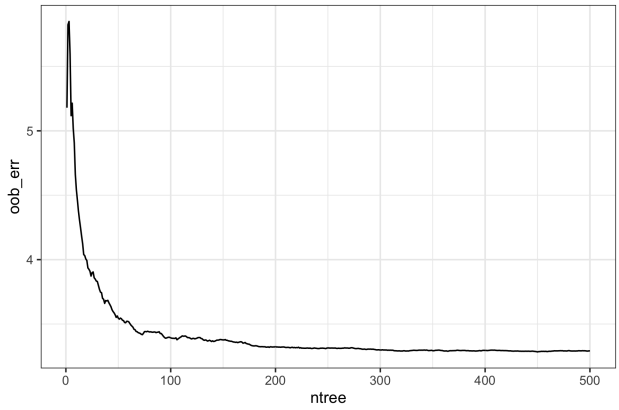

Figure 9: Out of bag error for number of feature tuning for random forest

For random forest, we first optimize the number of features m to consider at each split by trainig the model on different values of m, ranging from 1 to 17. The out of bag error for each value of m can be observed in 10. The out of bag error is minimized at a value of $m = 5$.

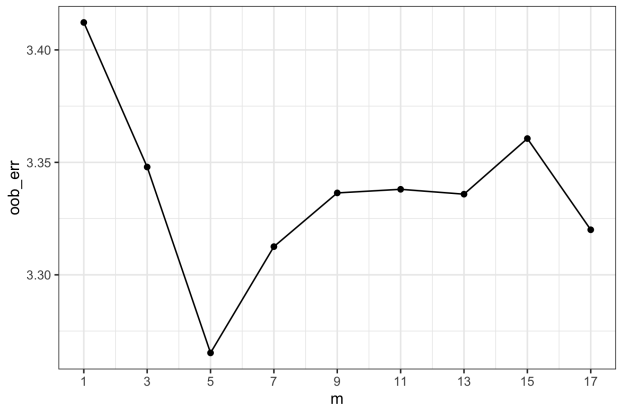

Figure 10: Out of bag error for number of feature tuning for random forest

With this fitted model, we can take a look at the out of bag variable importance (how much prediction accuracy lowers with a given feature out of bag) and purity-based importance (how much node purity improves as a result from splitting that feature). Refer to Figure 11

4.2.2 Boosting

The last model is boosting, utilizing the outputs of multiple decision trees to achieve better prediction performance.

First, find the optimal number of trees. As can be shown in Figure 12, the optimal number of trees is 1223.

Next, iterating over different iteration depths, as seen in Figure 13, the optimal number of iterations is 3.

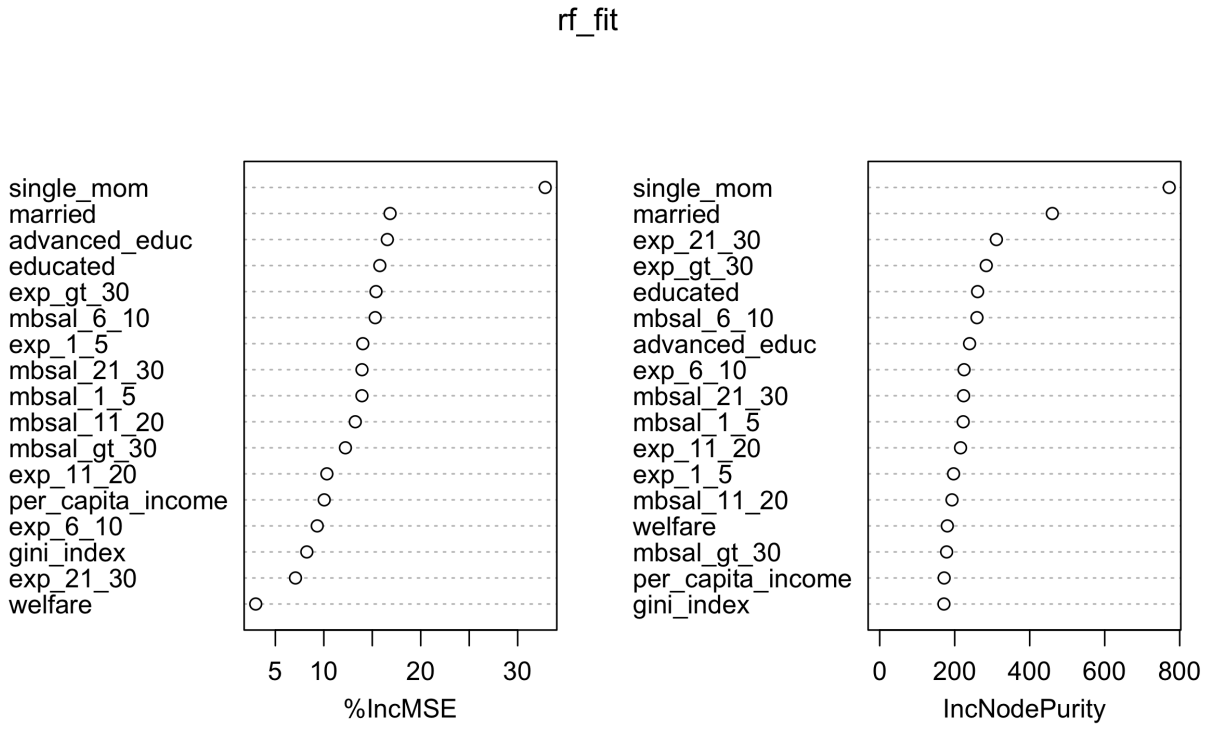

Figure 11: variable importance and node purity of out of bag

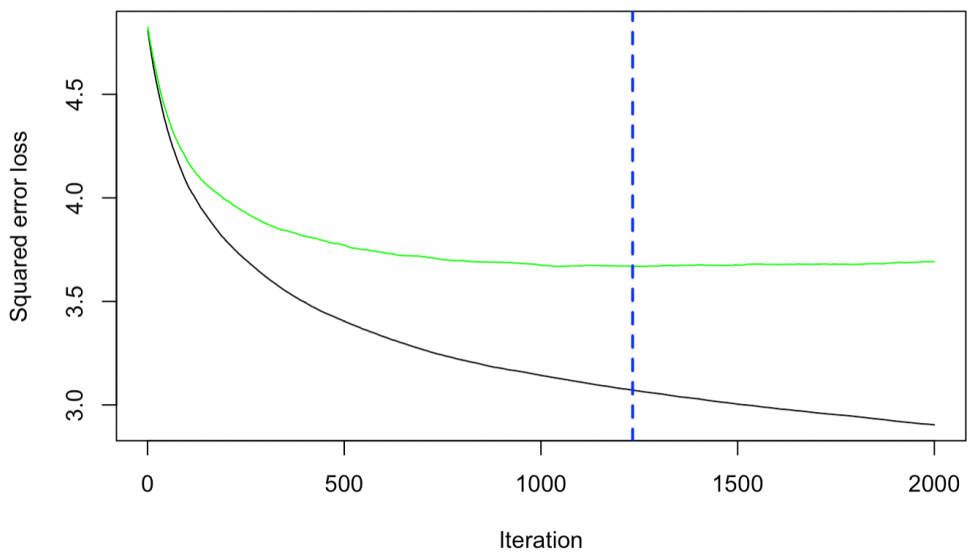

Figure 12: CV error over number of trees for boosting

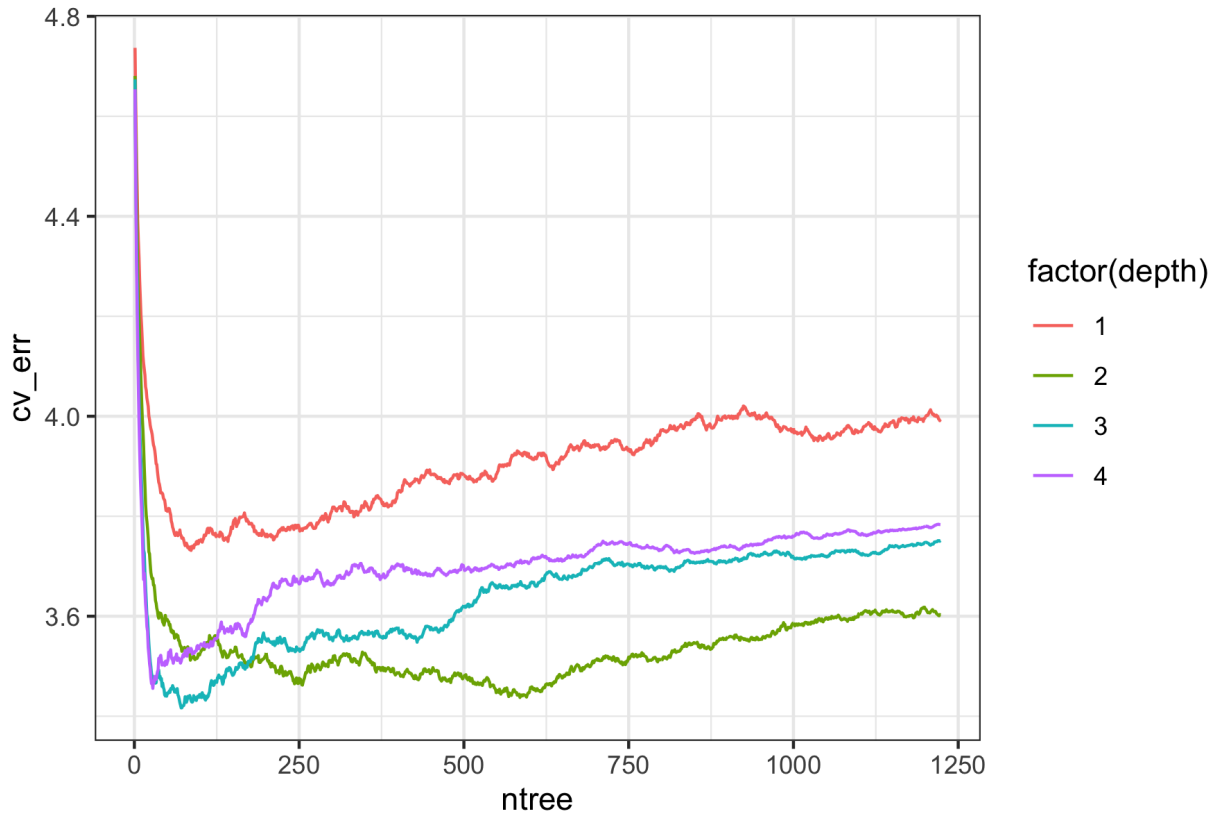

Figure 13: CV error over different iterations of boosting

Table 5: Relative influence of variables in boosting.

Variable	Relative Influence
single_mom	18.26
married	10.24
exp_21_30	7.42
mbsal_21_30	6.97
advanced_educ	6.72

Table 6: Root-mean-squared error of the five models, compared with the intercept-only method

Method	Test RMSE
Intercept-only	2.07
Ordinary Least Squares	1.75
Ridge Regression	1.88
Lasso Regression	1.86
Elastic Net Regression	1.88
Random Forest	1.72
Boosting	1.72

The top 5 relative influence of each variable selected by boosting is shown in Table 5

5 Conclusions

5.1 Method comparison

Table 6 shows how each model performed.

5.2 Takeaways

TODO

5.3 Data and Analysis Limitations

One major limitation is the need for more county-level data, such as adding in other factors that may contribute to dropout (i.e., juvenile crime rate, health-related factors, etc.), as well as a more careful analysis to highly correlated explanatory variables. Another limitation was the lack of data from more recent years on school dropouts, or even for teacher compensation for that matter. The latter proved to be the most limiting, as only two time periods between 2005-2007 were available. Although the County Health Rankings & Roadmaps has plethora of factors that could be added to this analysis, their data only goes back to 2010, and it would bring bias to include this dataset here.

Another limitation is that the per capita income is over the past 12 months in 2009 inflation-adjusted dollars, while our teacher compensation dataset corresponds to 2006-2007.

5.4 Follow-ups

Although the models that were created in this study had a lower root-mean-squared error than the intercept model, it was still noticeably high and that may be due to lack of other important factors that may correlate with school dropouts. So, as mentioned above, it is highly recommended to take into consideration other county-level factors. Additionally, the teacher compensation study was taken place over two time-periods, so using both years may prove to be more insightful.

Citations: https://nces.ed.gov/fastfacts/display.asp?id=16