# Introduction

By better identifying which school districts are at risk, government agencies and non-profit organizations could better utilize resources to enhance student achievement. The purpose of this project is to predict which school districts will close within five years. The final deliverable is a machine learning model that can be used by government (local, state, and federal) and non-profit organizations to better identify at-risk school districts and implement interventions tailored towards those school districts’ needs.

# Background

Primary and secondary education in the United States is a large system. As displayed in Figure 1 below, the total spending on primary and secondary education is comparable to the federal defense budget.

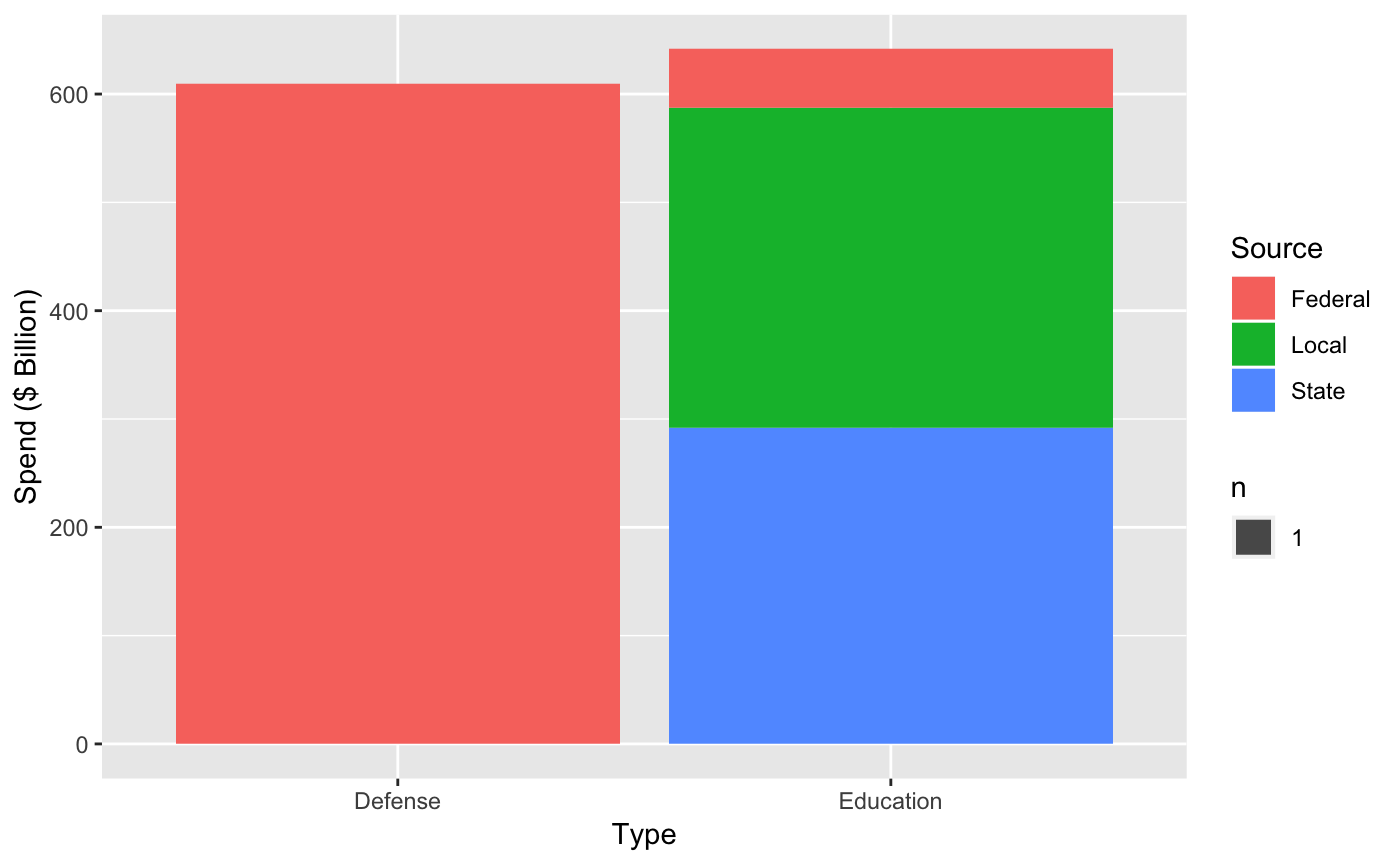


Figure 1: Total Education Spend and Federal Defense Spend were comparable in FY2014

Given the scale of the operation, incremental improvements have a large potential impact. Of the 18,465 school districts in 2009-2010, only 1,110 were closed by 2014-2015 (six percent), but the total cost of the closing school districts was over $4 billion. By better identifying which school districts are at risk, government agencies and non-profit organizations could better utilize resources to enhance student achievement.

The National Center of Education Statistics (NCES) manages the Common Core of Data, which is a publicly available database of information about primary and secondary schools used by the US Department of Education. The dataset contains detailed information about school districts’ financials and demographics from 1990-2015. With the Common Core of Data, machine learning techniques could be used to better identify at-risk school districts.

# Results

A gradient boosting based model achieved recall of 0.848 and 0.232. The key model parameters were:

* Only 15 features selected based on features correlated with label
* Features scaled to values between 0 and 1
* Balanced weighting to prioritize school districts that will close within five years
* Missing values for each feature imputed with median of entire dataset

Although other models generated higher recall or higher precision, there was tradeoff between recall and precision. Further model performance enhancement would require a definitive understanding of error costs.

# Approach

The three main tasks of this project were:

1. Acquire and preprocess the Common Core of Data.
2. Perform exploratory data analysis to identify factors correlated with whether school districts close.
3. Create and tune machine learning models that predict whether school districts close.

## Acquire and preprocess the Common Core of Data.

Only the 2009-2010 and 2014-2015 Common Core of Data were used for this project. Data for years after 2015 are still being completed by the NCES. Given that labels represent information five years into the future, the 2009-2010 data are the most recent collection of labelled data available.

The dataset is stored on NCES’s website as delimited files. Files were organized by year and type of data they contained (i.e. fiscal data 2009-2010). Along with each file, there was documentation that provided the character delimiter, column specifications, and number of records. Each file was downloaded onto a local folder (data) and renamed with a naming scheme (filetype\_fiscalYYYY.txt.zip).

After each file was downloaded, they were processed by a Python script. The main processing steps were:

1. Encode missing and irrelevant values.
2. Aggregate school demographic data into school district demographic data
3. Create label representing whether school district is still operational in five years.

## Perform exploratory data analysis to identify factors correlated with whether school districts close.

Exploratory data analysis was completed using R programming. The main steps involved were:

1. Remove columns with many missing or non-relevant values.
2. Modeled distribution of remaining columns.
3. Tested columns for correlation with whether school districts close within five years.

The numerical features that are correlated with whether school districts will close within five years are:

* total\_students: total number of students at the school district
* total\_schools: total number of schools at the school district
* teachers\_total: total number of teachers at the school district
* total\_revenue: total revenue (USD) of the school district
* total\_federal\_revenue: percentage of total revenue coming from federal government
* total\_state\_revenue: percentage of total revenue coming from state government
* total\_local\_revenue: percentage of total revenue coming from local government and other local sources
* total\_expenditure; total spending of school district
* total\_salaries: total salaries of all staff within school district
* minority\_students: percentage of school district students who are minorities

These correlations were verified by two-sample t-tests on bootstrap samples. The total number of male students at the school district was also tested, but it was not statistically significant.

The categorical features that are correlated with whether school districts will close within five years are:

* lowest\_grade: lowest grade level offered at the school district
* highest\_grade: highest grade level offered at the school district
* metro\_micro: whether the school district is located in a metropolitan (population over 50,000) or micropolitan (population between 10,000 and 50,000) area
* charter\_status: whether the school district is comprised of all charter schools, some charter schools, or no charter schools

These correlations were verified by chi-squared tests Whether the school district was managed by the Bureau of Indian Education was also tested, but it was not statistically significant.

## Create and tune machine learning models that predict whether school districts close.

Because the data was imbalanced (only six percent of 2009-2010 school districts closed in 2014-2015), accuracy was not an adequate evaluation metric. Recall and precision were used for model selection, and recall was prioritized over precision based on the assumption that false positive errors were more costly than false negative errors.

Before any supervised learning methods were tested, k-means clustering was first applied on the dataset. School districts were separated into clusters by the model, and then clusters with higher frequency of school districts closing within five years were identified. The results are shown in Figure 2. Even without using labels, clustering identified school districts more likely to close within five years. In addition, the clusters at higher risk were also significantly smaller than other clusters. Recall was 0.398 and precision was 0.411. These observations suggested that supervised learning methods will yield better results.



Figure 2: Even without using labels, clustering identified school districts more likely to close.

Next, the following iterative process was used to test multiple models from supervised learning methods.

1. Split data into train-development-test sets.
2. Build model using sklearn pipelines.
3. Evaluate model (calculate recall and precision on development set).

Logistic regression was first used to establish a baseline, and subsequent models were tested with the intention of incrementally improving recall and precision. Table 1 shows the results from the models tested. Highlighted in bold are model parameters that contributed to changes in model results.

Table 1: Gradient Boosting yielded the best results.

|  |  |  |
| --- | --- | --- |
| **Model** | **Recall** | **Precision** |
| Logistic Regression   * 155 features, 387 with dummy variables | 0.043 | 0.429 |
| Logistic Regression   * 155 features, 387 with dummy variables * **Min-Max Scaling** | 0.116 | 1.000 |
| Logistic Regression   * 155 features, 387 with dummy variables * Min-Max Scaling * **Balanced Weighting** | 0.841 | 0.221 |
| **XG Boost**   * 155 features, 387 with dummy variables * Min-Max Scaling * Balanced Weighting * PCA * **Imputation (median)** | 0.810 | 0.239 |
| XG Boost   * **15 features**, 95 with dummy variables * Min-Max Scaling * Balanced Weighting | 0.848 | 0.232 |

Gradient boosting yielded the best performance. After gradient boosting was implemented, parameter tuning did not yield definitively better results and instead yielded tradeoffs between recall and precision.

Next, in an attempt to improve model performance beyond gradient boosting, automated machine learning was tested. This was implemented using the TPOT python library which is an automated machine learning approach based on genetic algorithms. Table 2 shows the results of TPOT creating machine learning pipelines based on multiple objectives.

Table 2: Recall-Precision tradeoff was also observed in models created by TPOT.

|  |  |  |
| --- | --- | --- |
| **Model** | **Recall** | **Precision** |
| AutoML with TPOT   * Optimize recall | 1.0 | 0.057 |
| AutoML with TPOT   * Optimize f1-weighted | 0.4 | 0.84 |
| AutoML with TPOT   * Optimize roc-auc | 0.371 | 0.886 |

TPOT did not generate definitely better models than gradient boosting. Although TPOT did generate a model with perfect recall score on the test set when optimized for recall, the precision was extremely low. The other model generated by TPOT also displayed recall-precision tradeoff.

# Recommendations

As mentioned earlier, further model performance enhancements would require a more definitive understanding of false positive and false negative error costs which would result in more objective evaluation of models. One possible method of measuring error costs would be to calculate financial costs of intervention policies that would be implemented.