Predicting Public School Teachers Retention Status with SASS and TFS data

http://DataLab12.github.io/

June Yu — Department of Computer Science

Dr. Jelena Tešić — Department of Computer Science

Dr. Li Feng — Department of Finance and Economics

TEXAS STATE UNIVERSITY

Motivation

How to address the teacher shortage in public school districts using the state-of-art machine learning techniques? In this work, we take a data science look at National Center for Education Statistics (NCES) survey data to identify strongest predictors of retention rate in school districts to inform the policy makers.

Research Questions

- Do teachers' demographic, education/training, or teaching experience influence their retention status?
- Do principals' age, gender, ethnicity, teaching experience affect teachers' retention?
- Are public schools' level, type, region, urban/rural location, poverty level, share of minority students correlated with the teachers' retention?
- Does incentives to recruit teachers or merit pay improve the teachers' retention?

Data Acquisition and Integrations

National Center for Education Statistics (NCES) public-use data

- Schools and Staffing Survey (SASS) studies the K-12 educator labor market in 1999-2000. The survey consists of Public/Private Teacher, School, Principal, and Public District components.
- The Teacher Follow-Up Survey (TFS 2000-2001) was conducted the year after the SASS to determine how many teachers remained or left teaching. The public survey provides 2 components - Former and Current Teacher
- The surveys include the information on the teacher/principal characteristics, working conditions, teacher compensation, hiring and retention practices, and basic characteristics of student population.

Data Integration

- All data are filtered by Public Sector before integrating to focus on public teachers
- SASS and TFS are integrated based on matching School Control Number and Teacher/Principal Control Number
- Public District Component is not integrated due to missing of matching District Control Number

Initial Exploratory Data Analysis

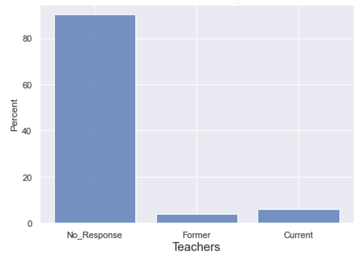

Figure 1. Ratio of Teacher Participation in TFS

From 42,086 public teachers that participated in SASS under 10% (4,156) participated in TFS, and that includes 2477 current and 1679 former teachers

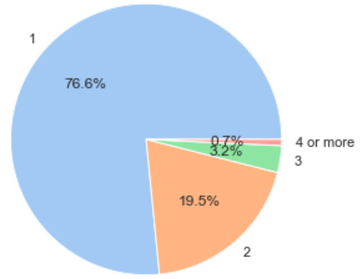

Figure 2. Number of Teachers per School

Some teachers do not have associated principal, so we end up analyzing 3,640 teachers working for 2,838 schools

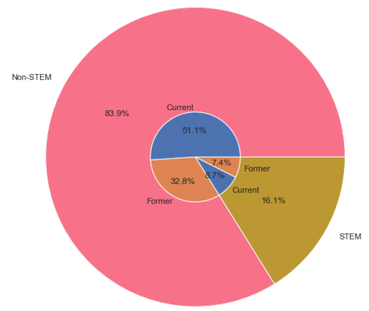

Figure 3. Teaching Assignment Field STEM vs. Non-STEM

STEM teachers are 16% of the population but tend to have higher turnover rate than Non-STEM teachers

Relevant Predictors

The hold-out set for SASS and TFS integrated data is used for **selecting** relevant predictors:

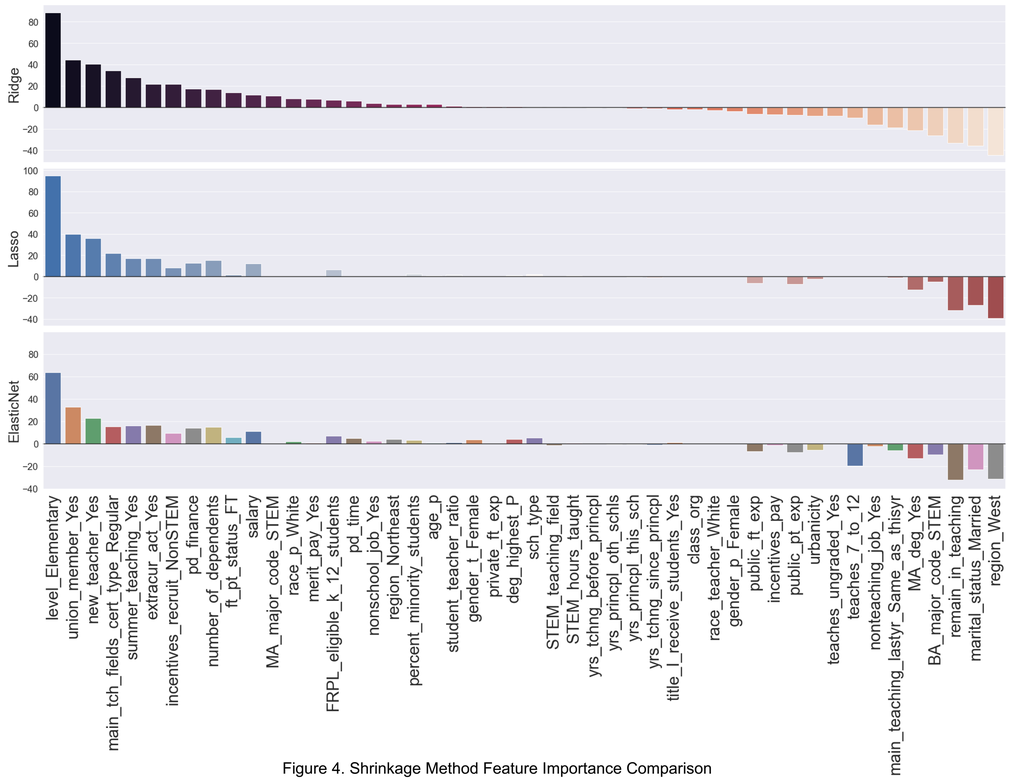

Figure 4. Shrinkage Method Feature Importance Comparison

- Step 1. Correlation Filtering: if any pair of features have high correlation, one of them is removed, or the pair is aggregated into either categorical or continuous type
- Step 2. Shrinkage with ML models: Lasso, Ridge and ElastiNet.
- Step 3. Binary to categorical values

Findings

- The initial 131 predicates are reduced to 51.
- Elementary Level is the strongest predicator.
- Type of STEM field teachers taught was not indicative of teacher's decision to quit or stay.

Machine Learning Classifier

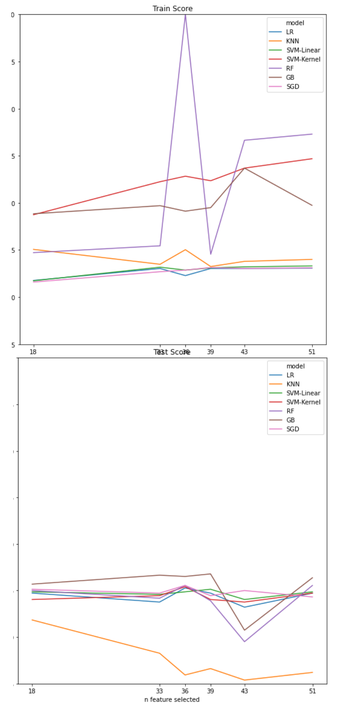

Figure 5. Seven ML Model Score Comparison on Train vs. Test Set

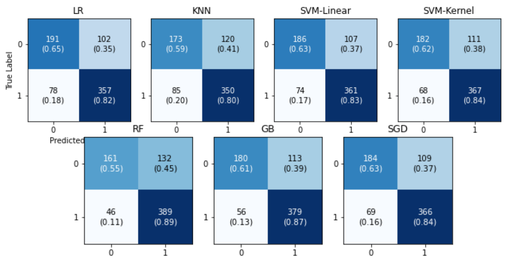

Figure 6. Seven ML Model Confusion Matrix Comparison on Test Set

- We have built multiple classifiers to predict whether the teacher will leave or stay.
- Train set and test set ratio is 8:2 w shuffling and stratification and classification used.

Next Step

- Revisit predictor selection: identify and interpret the best set of predictors for teachers' retention status
- Predictor Normalization
- Selecting the best classification model