An analysis of school district fiscal data and its implication on graduation rate

Chun Xie, Xuehan Chen, Yimin Xiao, Yue Wang

Agenda

- Introduction
- Preliminary analysis
- Models
- Conclusions and future work
- Problems so far

Problem description and data source

Education finance ------- student achievement

Education finance ---?--- student graduation

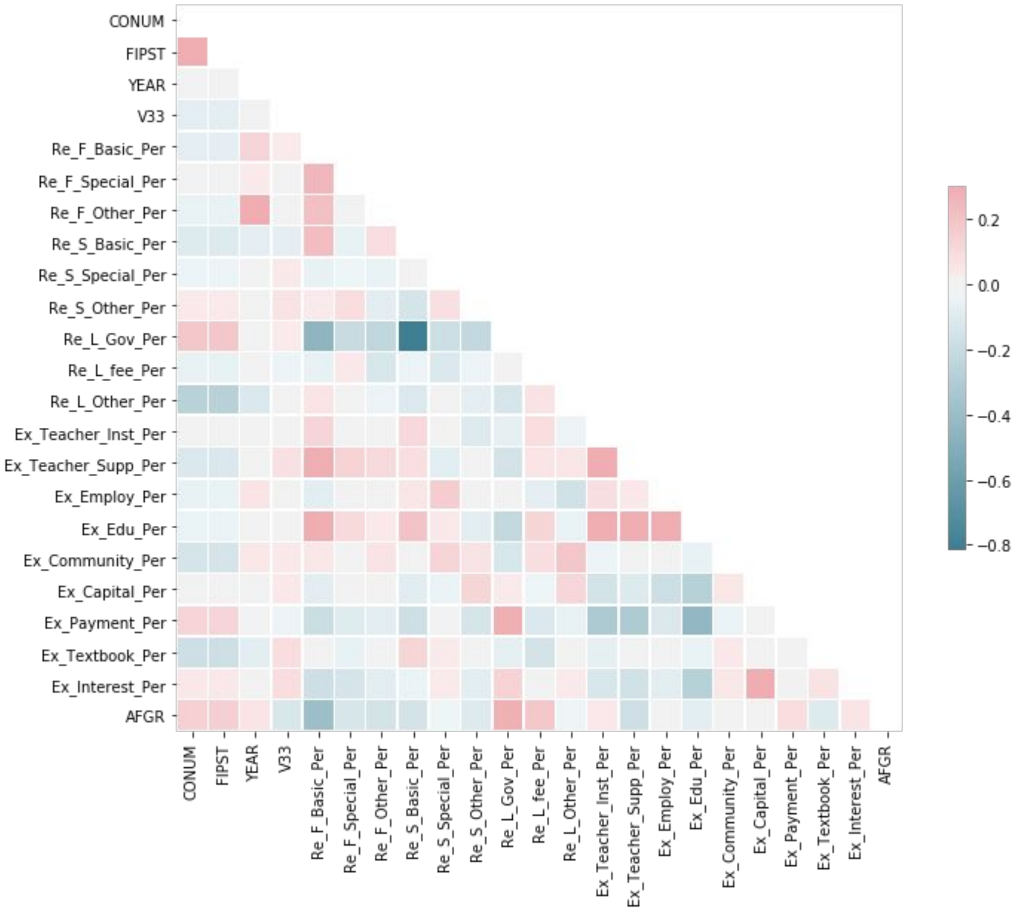
National Center for Education Statistics

Feature description

80,000+ rows
200+ columns

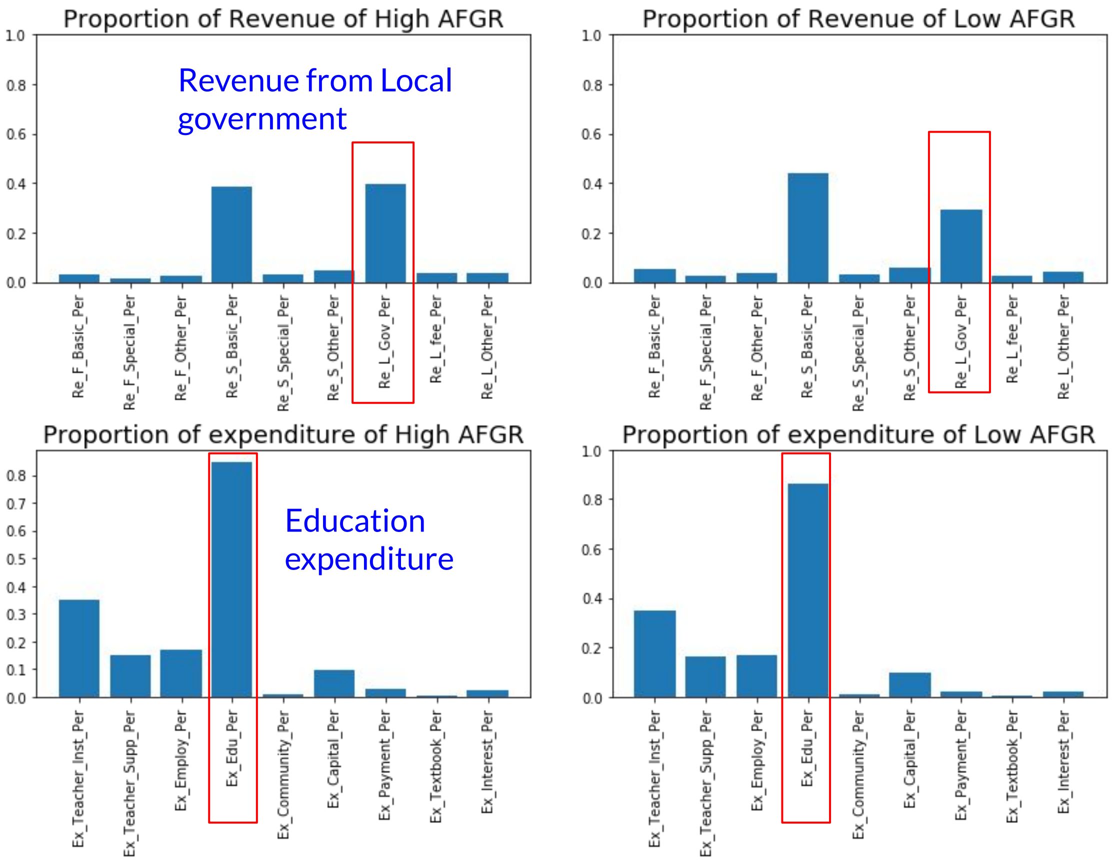
School Identification				
1	LEAID	LEAID		
School Characterization				
2-7	(Use the original)	SCHLEV	AGCHRT	CONUM FIPST YEAR V33
Revenue				
8	Re_F_Basic	Federal Level	Basic/Staff	C14 C16 C17 C25
9	Re_F_Special		Special program	C15 C19 B11 B10 B12
10	Re_F_Other		Others/Not specified	C20 C36 B13
11	Re_S_Basic	State Level	Basic/staff	C01 C04 C10 C12 C38
12	Re_S_Special		Special program	C05 C06 C07 C08 C09
13	Re_S_Other		Others/Not specified	C11 C13 C35 C39
14	Re_L_Gov	Local Revenue	Government/tax/school system	T02 T06 T09 T15 T40 T99 D11 D23
15	Re_L_fee		Sales and services (student fees)	A07 A08 A09 A11 A13 A15 A20 A40
16	Re_L_Other		Other income	U11 U22 U30 U50 U97 C24
Expenditure				
17	Ex_Teacher_Inst	Teacher's salary and Employee benifit	Instruction (basic)	Z33
18	Ex_Teacher_Supp		Support Services	V11 V13 V15 V17 V21 V23 V37 V29
19	Ex_Employ		Employee benefit	Z34
20	Ex_Edu	For elementary /secondary education	Instruction expenditure	E13
			Support Services	TCURSSVC E11
			Other	V60 V65
21	Ex_Community	For community		TNOELSE
22	Ex_Capital	Capital outlay expenditures		TCAPOUT
23	Ex_Payment	Payments	Payments to state government	L12 M12 Q11
			Payments to private schools	V91
			Payments to charter schools	V92
24	Ex_Textbook	Textbook		V93
25	Ex_Interest	Interest on debt		I86
Graduation rate				
26	AFGR	Average freshman graduation rate		AFGR

Correlation



Preliminary analysis

- District graduation rate(AFGR)
- Proportion of Revenue
- Proportion of Expenditure



Model description

- Linear regression: use the original continuous number as target
  Logistic regression: classify the target variable into two categories
- Feature engineering: percentage of each subcategory
- Three different feature sets

Evaluation metric

- Validation data and test data
    - Validation: determine the complexity of the model
    - Test: measure generalization performance
- RMSE and AUC
    - RMSE: for regression models
    - AUC: for classification models

Validation and testing performance

- Linear regression
  - Validation: the best model has a RMSE of 11.4394
  - Testing: the best model has a RMSE of 11.6663
- Logistic regression
  - Validation: the best model has an AUC level of 0.7676
  - Testing: the best model has an AUC level of 0.7409

Inference and conclusion

- Positive relationship:
  - Revenue from students' fee (Re_L_fee_Per)
  - Expenditure on teacher salary (Ex_Teacher_Inst_Per)
- Negative relationship:
  - Revenue from federal funding (Re_F_Basic_Per)
  - extbo _Per)

In [39]:
```
negative_fis.head()
```
Out[39]:

	features	weight
17	Ex_Textbook_Per	-22.926003
1	Re_F_Basic_Per	-21.142299
11	Ex_Teacher_Supp_Per	-4.597501
3	Re_F_Other_Per	-2.976202
6	Re_S_Other_Per	-1.451370

In [38]:
```
positive_fis.head()
```
Out[38]:

	features	weight
8	Re_L_fee_Per	13.406366
10	Ex_Teacher_Inst_Per	2.045284
7	Re_L_Gov_Per	1.350755
24	YEAR10	0.517996
19	SCHLEV_2	0.493301

The best model is model 1

```
In [38]: pipe_model_best = Pipeline(stages=[
             feature.VectorAssembler(inputCols=['V33','Re_F_Basic_Per', 'Re_F_Special_Per', 'Re_F_Other_Per', 'Re_S_Basic_Per',
                     'Re_S_Other_Per', 'Re_L_Gov_Per', 'Re_L_fee_Per', 'Re_L_Other_Per', 'Ex_Teacher_Inst_Per',
                     'Ex_Teacher_Supp_Per', 'Ex_Employ_Per', 'Ex_Edu_Per', 'Ex_Community_Per', 'Ex_Capital_Per',
                     'Ex_Payment_Per', 'Ex_Textbook_Per', 'Ex_Interest_Per','SCHLEV_2', 'SCHLEV_3', "SCHLEV_5", 'YEAR8',
                     'YEAR9', 'YEAR10', 'AGCHRT2', 'AGCHRT3'],outputCol='features'),
             feature.StandardScaler(withMean = True, inputCol = 'features', outputCol = 'Std_features' ),
             regression.LinearRegression(featuresCol='Std_features',labelCol='AFGR')
         ]).fit(dummy_df)
```

```
In [36]: pipe_model_best.stages[2].coefficients
```

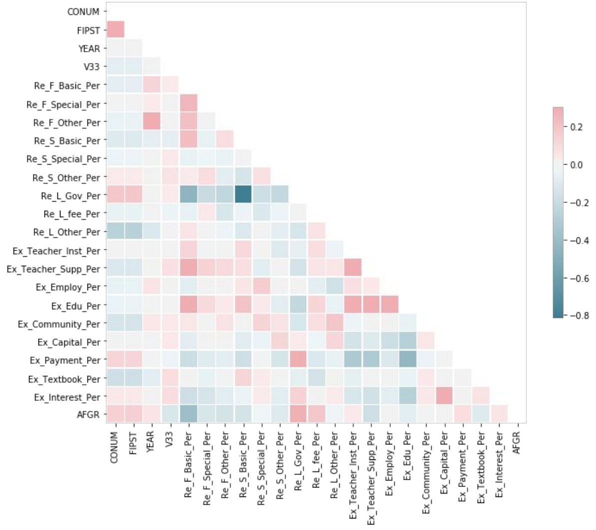
```
Out[36]: DenseVector([-1.2129, -4.3864, -0.4822, -1.3433, 0.0666, -0.3026, -0.8357, 1.1513, 1.671, 0.0518, 0.851, -1.6094, -0.
         5916, 0.4681, 0.1491, -0.2201, -0.2314, -0.898, -0.2159, 2.4186, 2.4347, 0.3271, 0.1429, 0.6276, 2.0327, 2.5305, 3.03
         17])
```

Future work

- Additional models to try: SVM and random forest
- Explore the relationship between graduation rate and poverty level
- Revenue amount received per capita: fiscal data / student amount

Problems found so far and plans to solve them

- Additional feature engineering method: PCA
- Interaction variable
    - School level and fiscal data
    - School type and fiscal data
    - Year and fiscal data



Thank you!

Q&A