Prediction of High School Graduation Rates in the U.S.

By Team Sunshine

Team Sunshine ☀

Amy Wiggington

Amy is a data analyst for a college access grant and is persistently seeking skills and tools that will help her be more efficient and impactful.

Blake Baird

Blake is a science geek working in manufacturing recruiting and is transitioning to a career as a data scientist.

Hayat Pacquette

Hayat is a former secondary math and special education teacher and is transitioning to a career as a data scientist.

Agenda

Introduction

Project Architecture

Data

Feature Engineering

Machine Learning

Deployment

Conclusion

Acknowledgements

Introduction

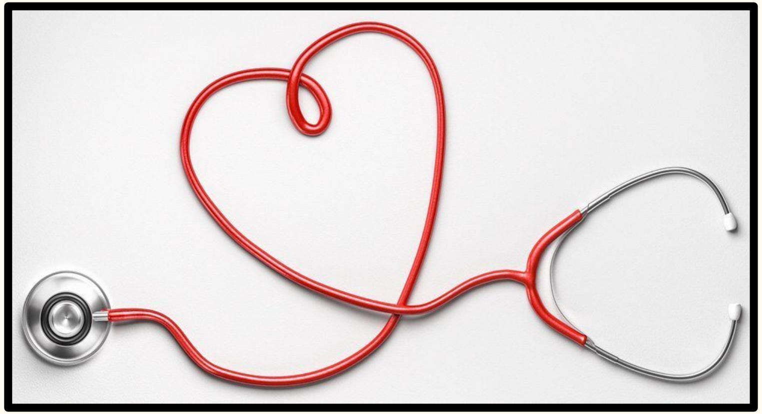
Motivation

Motivation

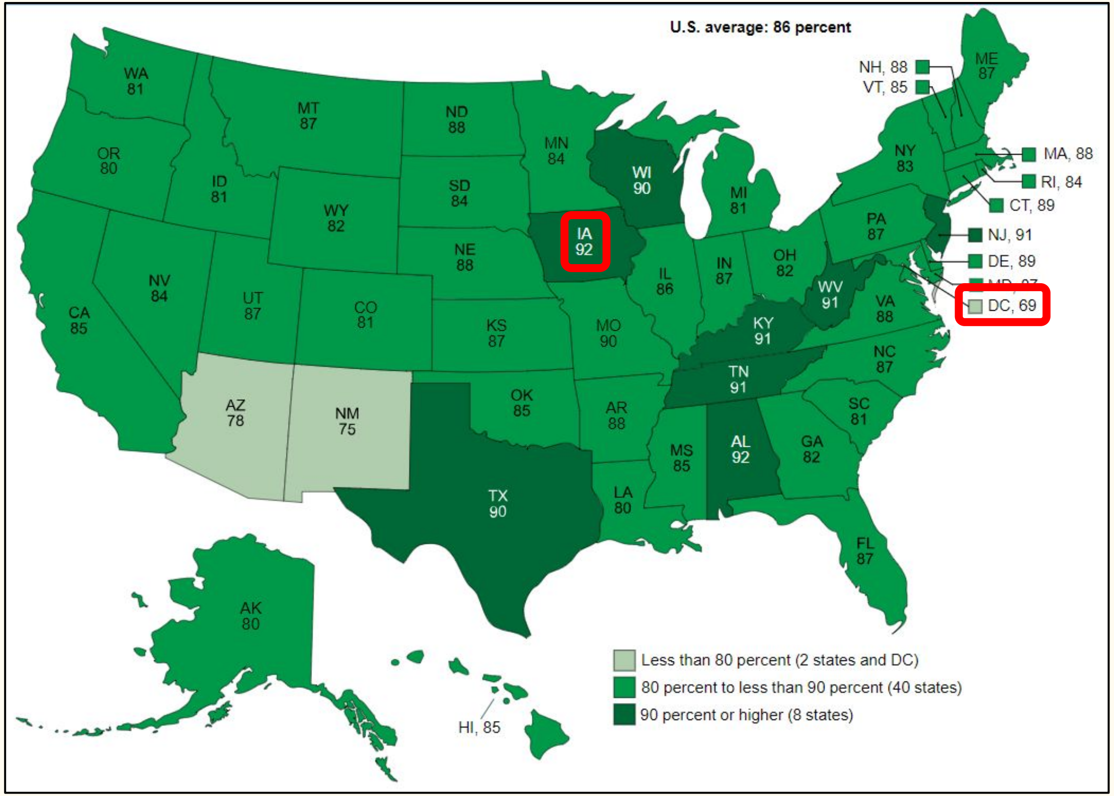
Jobs

Wages

Health

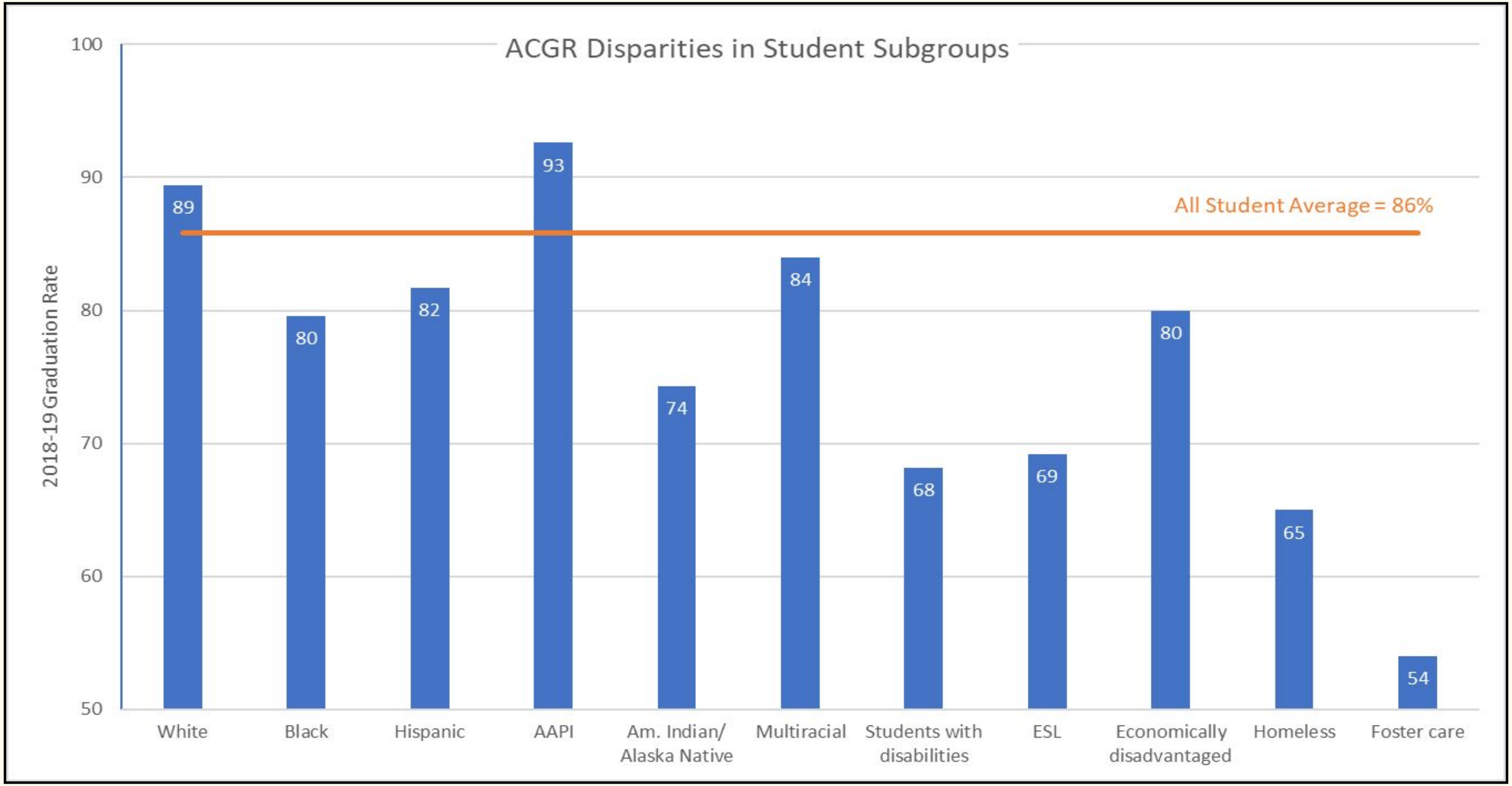


Motivation



https://educationdata.org/public-education-spending-statistics
https://nces.ed.gov/programs/coe/indicator/coi/high-school-graduation-rates

Motivation



ACGR Disparities in Student Subgroups
2018-19 Graduation Rate
100
90
80
70
60
50
All Student Average = 86%
89
80
82
93
74
84
68
69
80
65
54
White
Black
Hispanic
AAPI
Am. Indian/ Alaska Native
Multiracial
Students with disabilities
ESL
Economically disadvantaged
Homeless
Foster care

Hypothesis

We hypothesized that **economic factors** would be the most significant predictors of **graduation rate.**

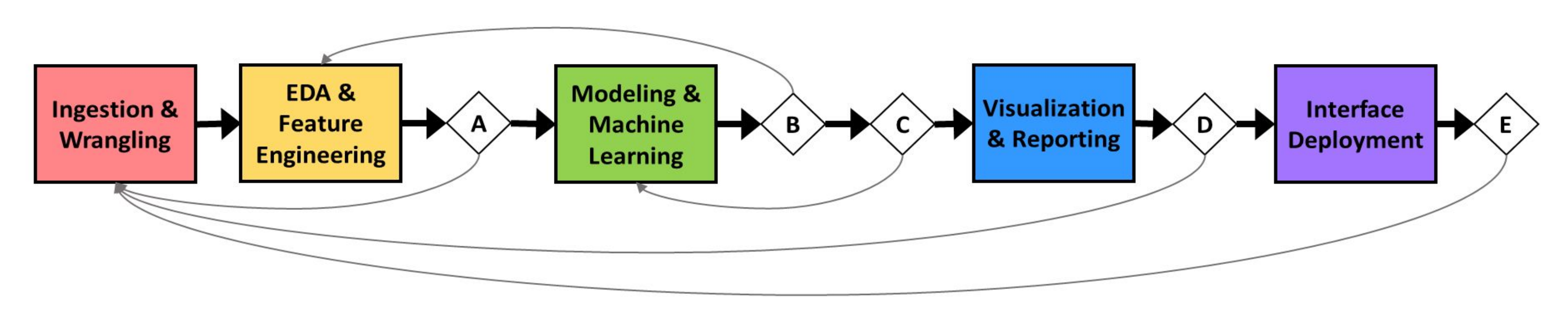
Adjusted Cohort Graduation Rate (ACGR)

Applications

- Targeting interventions for low graduation rates
- Planning parameters of new schools before building them
- Identifying key factors to investigate further

Project Architecture

Data Science Pipeline



A: Do we have enough/right type of data? Any outliers need further cleaning?

B: Could better or different data preparation methods could improve our model?

C: Do we need to make adjustments? Have we tested multiple types of models?

D: Are our models performing well and would additional data improvement it?

E: Does stakeholder feedback provide ideas for improvement?

Toolkit

Swampy (better type printing)
Vscode, vim

Amazon S3
python
scikit learn
jupyter
Joblib
pandas
matplotlib
NumPy
seaborn
bokeh
Boto 3
CONDA
Yellowbrick

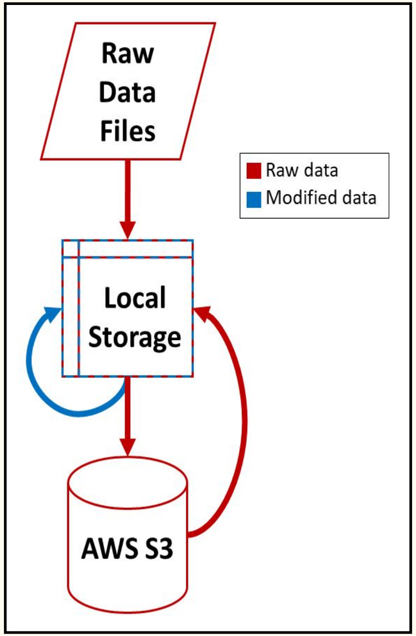
Data

Data Storage

Our choice: Amazon’s AWS S3 virtual server

- Over 2gb data of raw data and documentation
- Scalability
- Reliability
- Low cost

WORM storage



Data Sources

	Adjusted Cohort Graduation Rate [8] ---------- U.S. Department of Education	Schools Common Core of Data Directory [9] ---------- Urban Institute[1]	Assessment Participation[2] [10] ---------- U.S. Department of Education	County-Level Unemployment [11] ---------- U.S. Department of Agriculture
Size	9 CSV files 31.9 MB (Total)	1 CSV file 843.8 MB	16 CSV files (8 /subject) 1.65 GB (Total)	1 XLSX file 2.08 MB
Instances	180,232 (Total) School + Year	3,381,565 School + Year	366,400 (Total) School + Year	3277 County
Features	~29 / file e.g. School, Cohort Size, Grad Rate, Subgroups	52 e.g. School ID, Location, Type, Teachers, Enrollment	~33 / file e.g. School, Participants, Proficiency Scores	96 e.g. County, Labor Force, Unemployment
Scope	SY 2010-2011 - 2018-2019 50 states, DC, Puerto Rico, USVI, BIE School Division, School	SY 1986-1987 - 2021-2022 50 states, DC, U.S. territories, BIE, DoDEA State, Division, School	SY 2012-2013 - 2018-19, 2020-2021[3] 50 states, DC, Puerto Rico, USVI, BIE School Division, School	2000 - 2021 50 states, DC, Puerto Rico Nation, State, County

Ingestion and Wrangling

- Boto3 - ingest from cloud
- Converted range strings to numbers
- "GT50" —> 75.0
- Sanity checks
- $<$ 1 student, $<$ 1 enrollment
- Merged on 'NCESSCH', 'Year' features and 'County_Code'

Pandas api friends

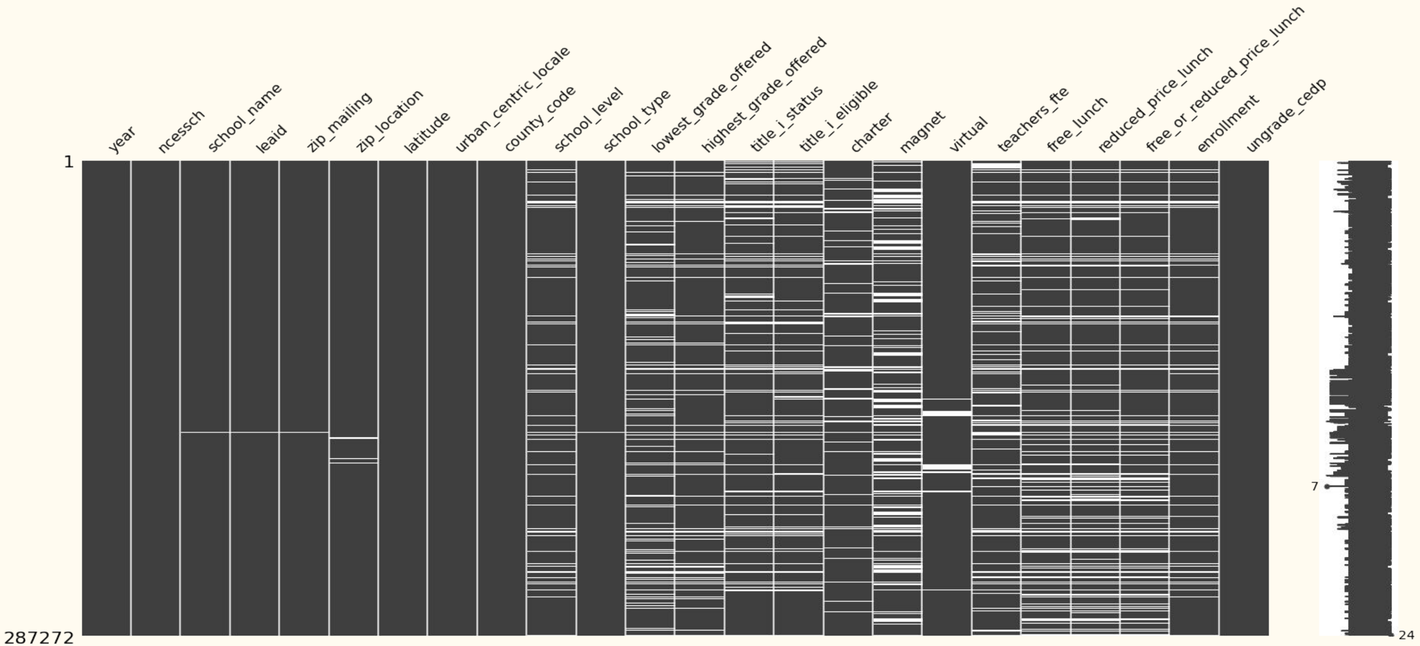
.loc (select)
.query (sql-ish select)
.concat (combine)

.map (apply function to a series)
.merge (sql-ish merge)

Data Sources

	Adjusted Cohort Graduation Rate [8] ---------- U.S. Department of Education	Schools Common Core of Data Directory [9] ---------- Urban Institute[1]	Assessment Participation[2] [10] ---------- U.S. Department of Education	County-Level Unemployment [11] ---------- U.S. Department of Agriculture
Size	9 CSV files 31.9 MB (Total)	1 CSV file 843.8 MB	16 CSV files (8 /subject) 1.65 GB (Total)	1 XLSX file 2.08 MB
Instances	180,232 (Total) School + Year	3,381,565 School + Year	366,400 (Total) School + Year	3277 County
Features	~29 / file e.g. School, Cohort Size, Grad Rate, Subgroups	52 e.g. School ID, Location, Type, Teachers, Enrollment	~33 / file e.g. School, Participants, Proficiency Scores	96 e.g. County, Labor Force, Unemployment
Scope	SY 2010-2011 - 2018-2019 50 states, DC, Puerto Rico, USVI, BIE School Division, School	SY 1986-1987 - 2021-2022 50 states, DC, U.S. territories, BIE, DoDEA State, Division, School	SY 2012-2013 - 2018-19, 2020-2021[3] 50 states, DC, Puerto Rico, USVI, BIE School Division, School	2000 - 2021 50 states, DC, Puerto Rico Nation, State, County

Dealing with Missingness

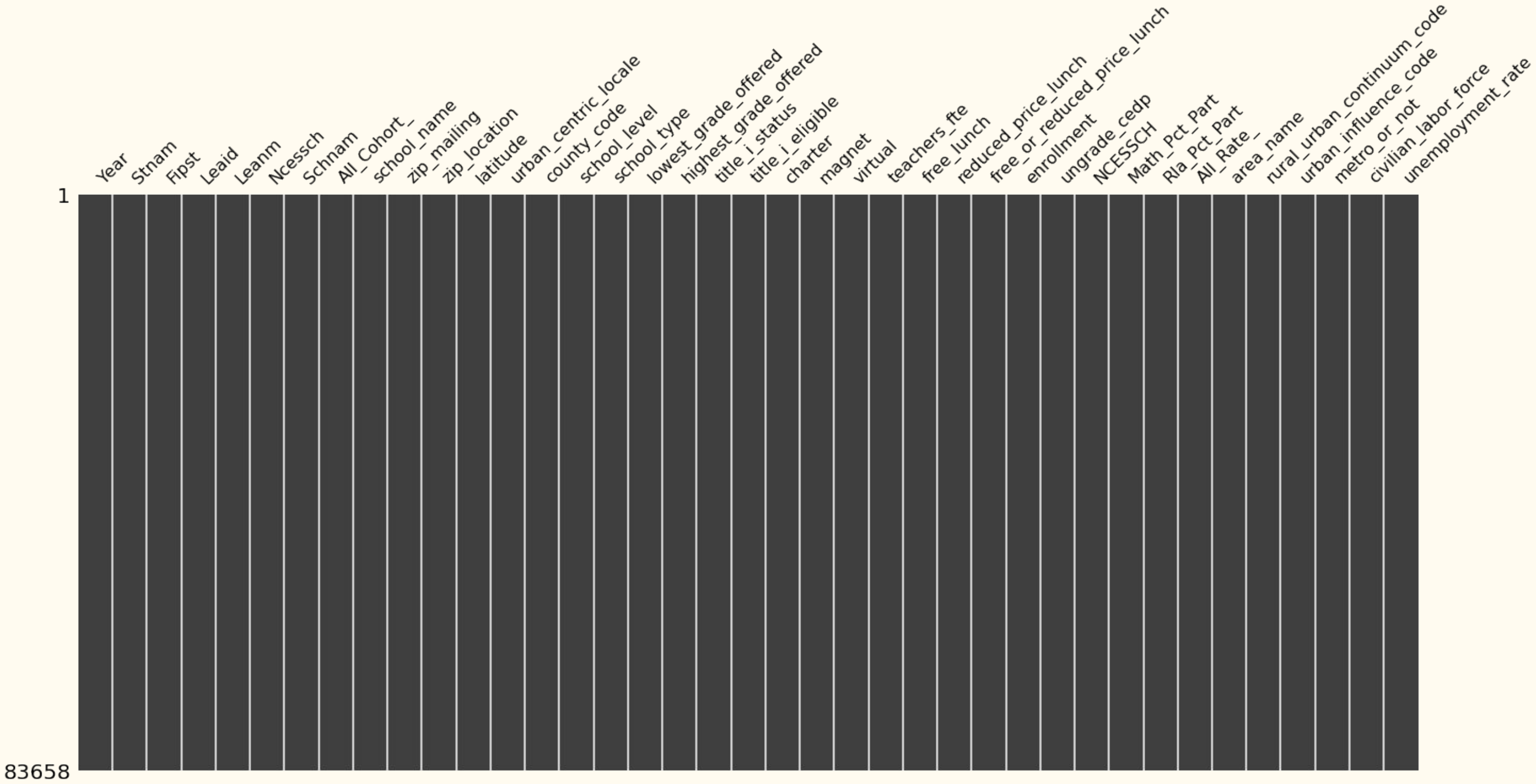


Less than 5 students (~1400 per yr)

NaN assumed non-virtual

75% of data lost

Trade-off: More features <=> Less Instances



Final shape: 40 Features and 83,658 Instances

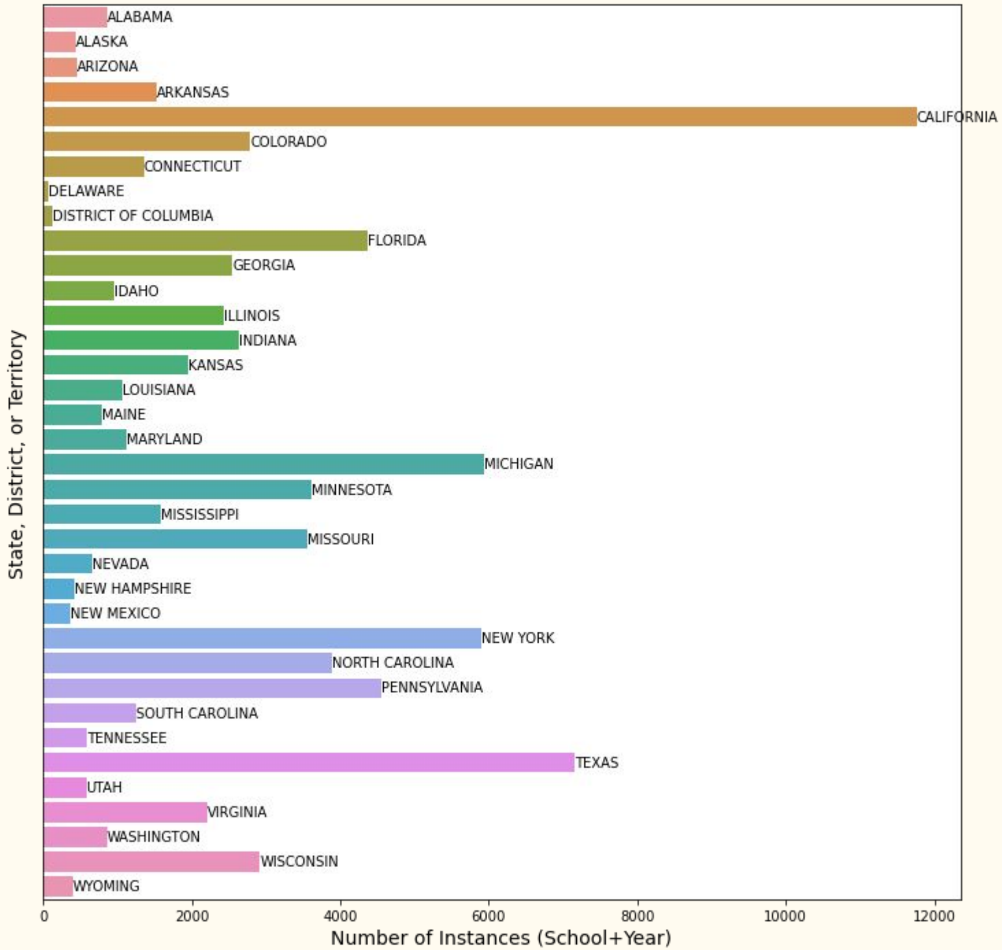
Merged on: 12-digit school id, year, county code

Exploratory Data Analysis

To gain initial insight into the data we had to do some exploratory data analysis to explore the data through numerical, tabular and graphical representations.

Sample population assessment

Plots and summary stats on interesting features

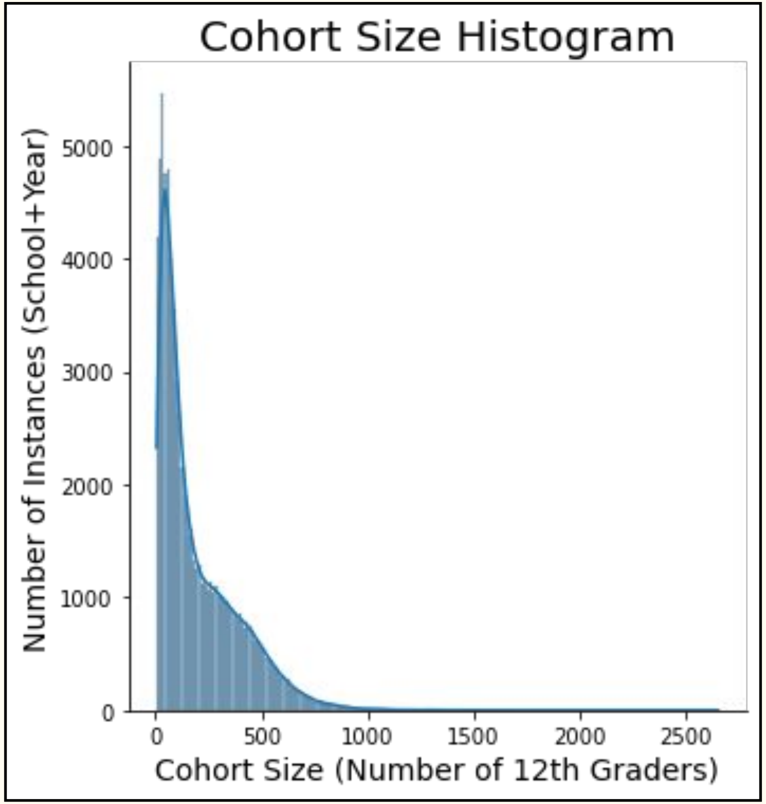


Highlights

-California and Texas highly represented

-Puerto Rico and West Virginia not present

Year	2012	2013	2014	2015	2016	2017	2018
%	14.1	14.4	11.7	12.0	15.1	16.3	16.3

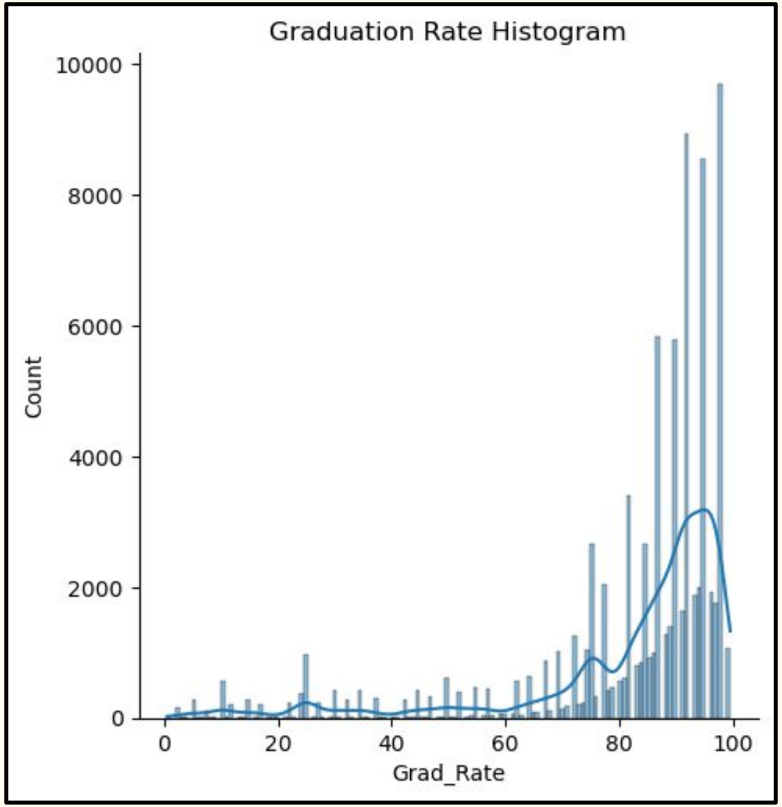


Stats

Right skewed

Mean cohort size $= 200$

Median cohort size $= 126$



Stats

Left skewed

Mean grad rate $= 82\%$

Median grad rate $= 90\%$

Description

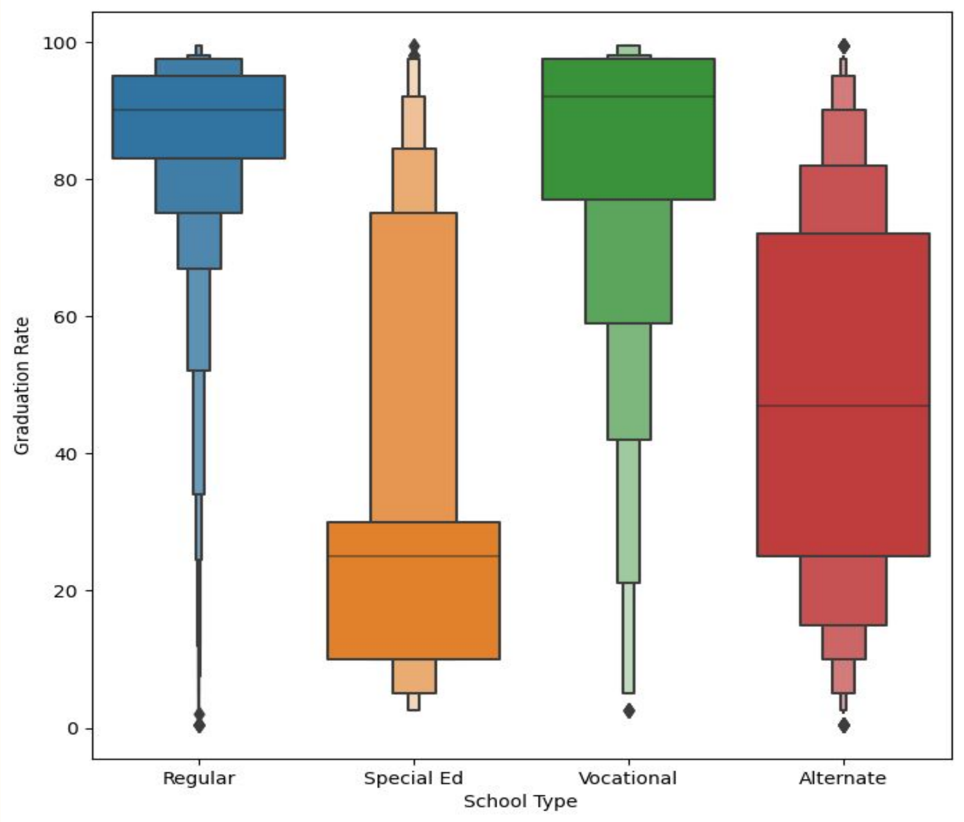
Regular- what someone thinks about as a “regular high school”; inclusive of all students

Special Education - school where students may not be able to access the general education requirements

Vocational - school where student pursues a trade in preparation for a career

Alternative - school that provides a nontraditional environments as it relates to schedules and curriculum

```
plt.title("Grad Rate by School Type")
sns.boxenplot(x=df.School_Type,y=df.Grad_Rate)
```



Feature Engineering

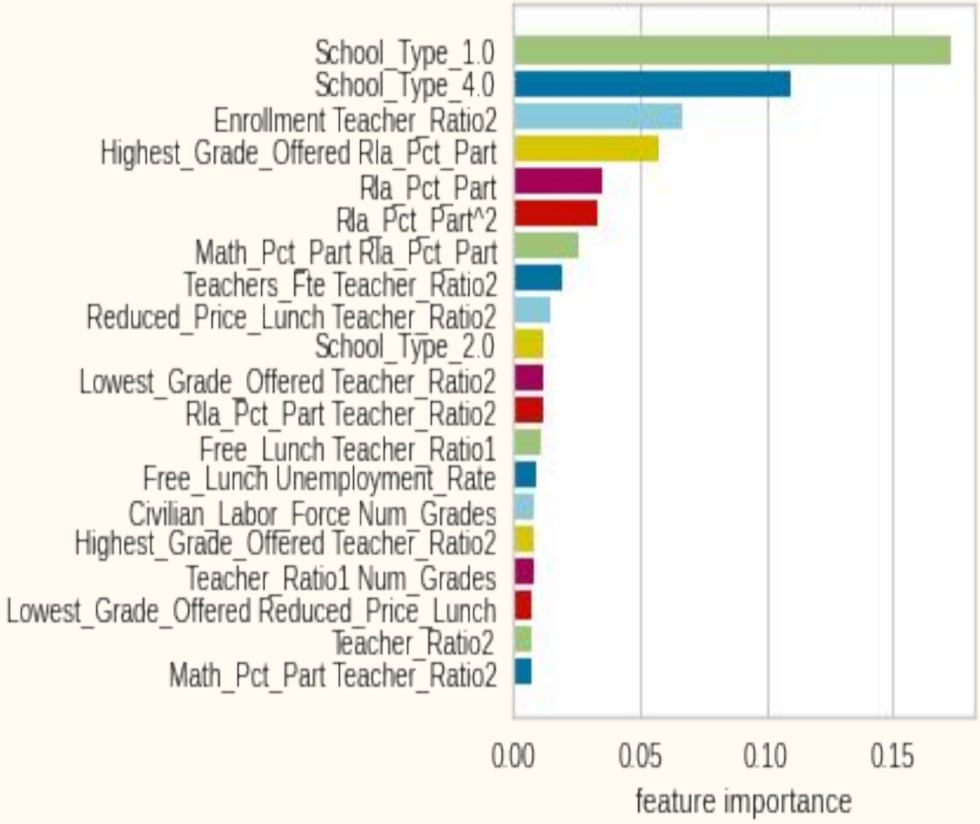
Feature Addition

Name	Description
Num_Grades	(Highest grade offered) - (Lowest grade offered)
Teacher_Ratio1	(Teacher full time equivalents) / (Total school enrollment)
Teacher_Ratio2	(Teacher full time equivalents) / (Cohort size)

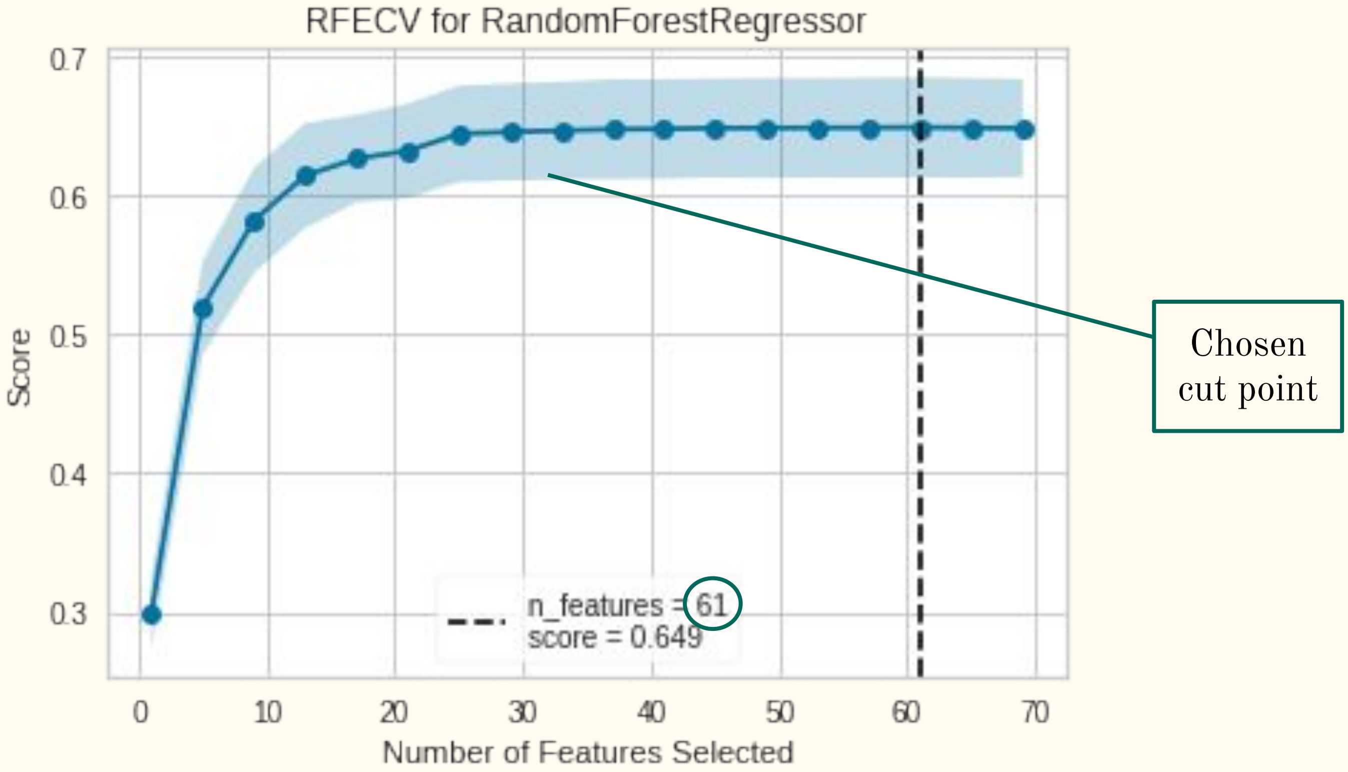
Feature Addition

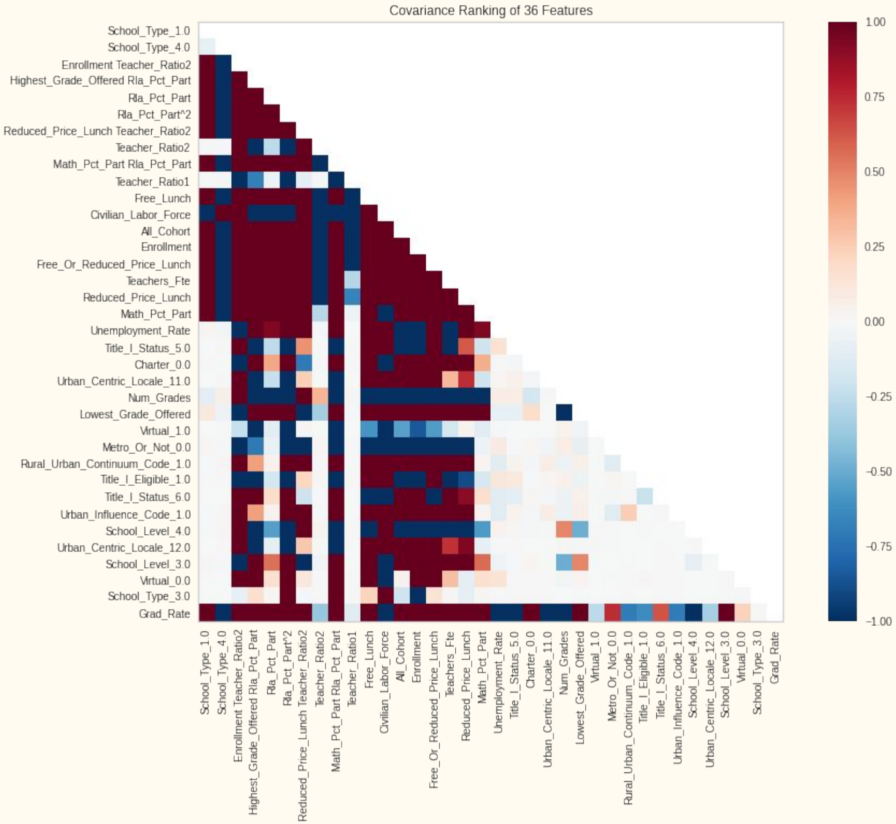
2nd-Degree Interaction Terms

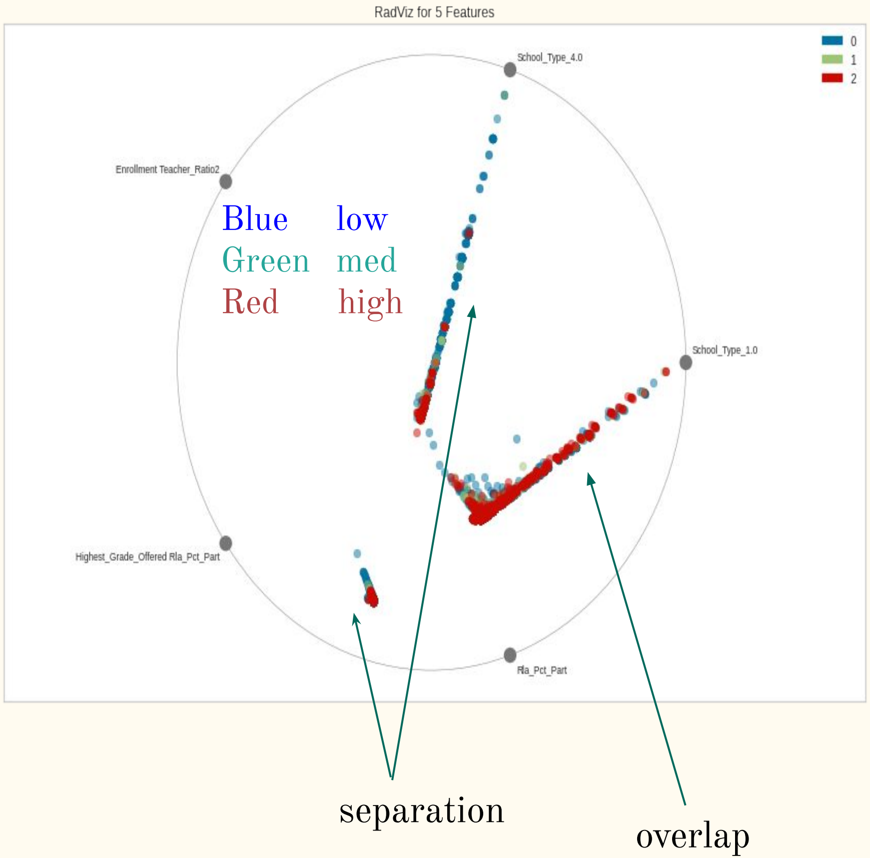
+ 5



Feature Reduction







Covariance and Radial Visualization of important features

Machine Learning

Regression Model Types

- linear model (ElasticNet)
- support vector machine (SVR)
- neural network (Multilayer Perceptron)
- boosted decision tree (AdaBoost)
- decision tree ensemble (RandomForest)

* experiments with classification too

Train-Test-Split

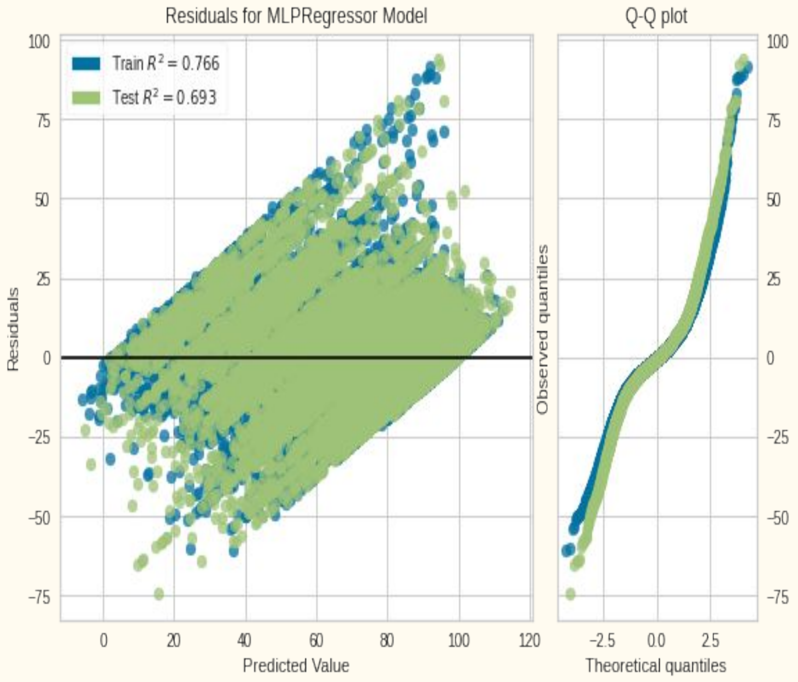
70/30 stratified on target

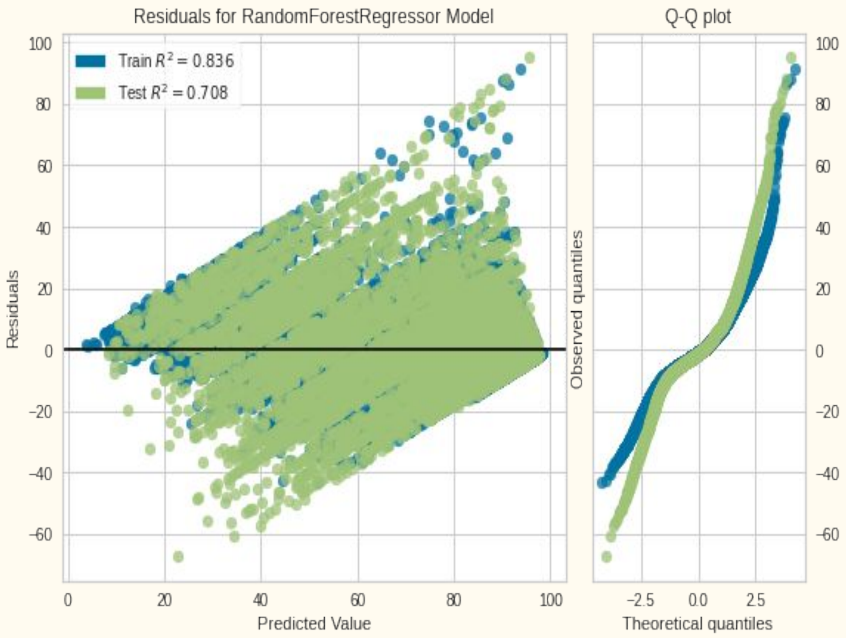
Note on Pipelines

The *sci-kit* learn Pipeline class was used for easy communication and prevention of subtle data leakage between CV folds.

feature-engine SklearnTransformWrapper class.
Keeps data as pandas dataframe not ndarray

```
pipe_mlpr = Pipeline(steps=[("scale", std_scaler),
                            ("mlpr", MLPRegressor(activation='logistic',
                            alpha=0.0002, hidden_layer_sizes=(100,),
                            solver='sgd', max_iter=1000, random_state=42))])
```


Residuals for MLPRegressor Model
Q-Q plot
Train $R^2 = 0.766$
Test $R^2 = 0.693$
Residuals
Predicted Value
Observed quantiles
Theoretical quantiles


Residuals for RandomForestRegressor Model
Q-Q plot
Train $R^2 = 0.836$
Test $R^2 = 0.708$
Residuals
Predicted Value
Observed quantiles
Theoretical quantiles

Hyperparameter Tuning

Evaluation Metrics: Accuracy (R2) and root mean squared error (RMSE). RandomForest selected as final model.

2 grid searches per model (more for random forest)

Model	Train R^2	Test R^2	RMSE
ElasticNet (Lasso penalty)	0.554	0.553	12.99
SVR	0.687	0.660	11.33
AdaBoost	0.830	0.679	11.00
Multilayer Perceptron	0.766	0.693	10.77
RandomForest	**0.836**	**0.708**	**10.48**

Model	ElasticNet	SVR	MLPR	AdaBoost	Random Forest
Params	L1 ratio 1	C 7 Epsilon 0.6 Kernel ”rbf”	Activation logistic Alpha 2E-3 Hidden layer 100x1 Solver sgd Max iterations 1000	Estimators 800 Learning rate 0.1 Loss exponential Max features 0.9 Max depth 11	Estimators 1000 Max features 0.8 Max depth 13

Validation

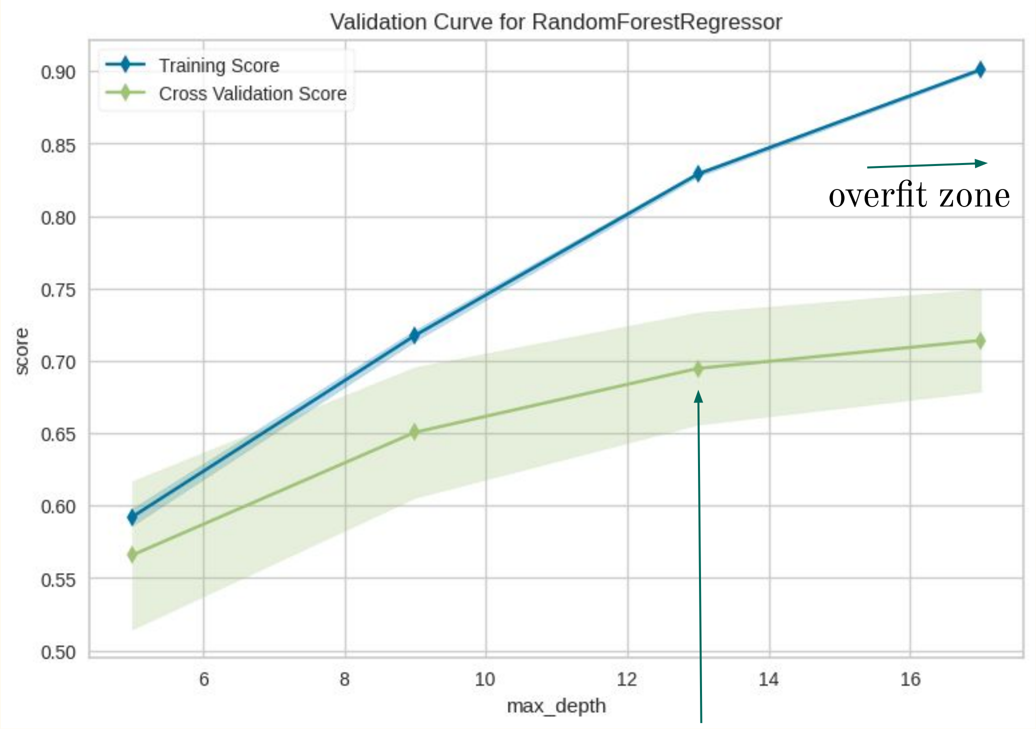
Early stopping @ depth 13

Test score still improving @ 13

Rationale for 13
- Code review said go lower
- Flawed math =)

~~1 CV fold = 8,365 instances~~
~~2^13 = 8,192~~

Validation curve supported it!

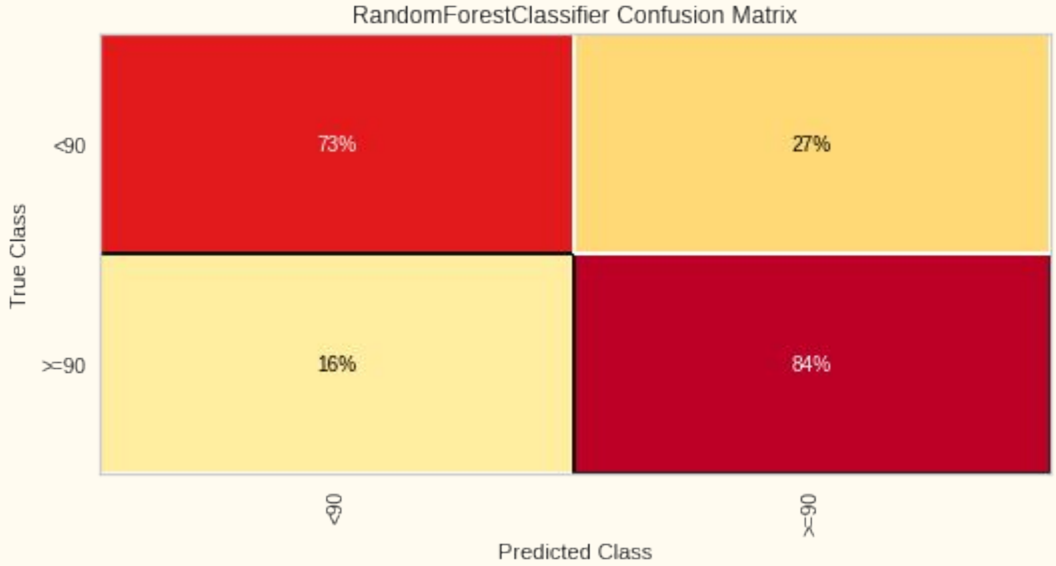


Early
stop

Classification Model

Binning @ median value => balanced classes

This split was found at the median 90% value. The F1 score was 0.789.



Trade-off

Detail of our regression model

Versus

How often we are "right"

Deployment

Deployment

Input features

Enter Rla_Pct_Part^2 [mean=8792.40 min=6.25 max=9900.25]: (Press 'Enter' to confirm or 'Escape' to cancel)

Input features to generate prediction

```python
data = {}

for feat in model.feature_names_in_:
    min_of_feat = df[feat].min()
    max_of_feat = df[feat].max()
    mean_of_feat = df[feat].mean()
    value = input("Enter {name} [mean={mean:.2f} min={min:.2f} max={max:.2f}]:".format(
        name=feat, mean=mean_of_feat, min=min_of_feat, max=max_of_feat))
    if float(value) < min_of_feat or float(value) > max_of_feat:
        print("Value outside trained range entered")
    data[feat] = value
```

[5] ✓ 2m 41.3s

```
Value outside trained range entered
```

Get prediction -> Profit

```python
model.predict(pd.DataFrame([data]))
```

[6] ✓ 0.9s

```
array([90.9541484])
```

Deployment

Feature name	Description	Source
School_Type_1.0	Regular school	CCD
School_Type_4.0	Other/Alternative school	CCD
Enrollment Teacher_Ratio2	Enrollment x Teacher ratio 2	Imputed
Highest_Grade_Offered Rla_Pct_Part	Highest grade offered x Reading participation	Imputed
Rla_Pct_Part	State reading test participation percent	DOE
Rla_Pct_Part^2	Reading participation squared	Imputed
Reduced_Price_Lunch Teacher_Ratio2	Students with reduced lunch x Teacher ratio 2	Imputed
Teacher_Ratio2	Teacher count / Cohort size	Imputed
Math_Pct_Part Rla_Pct_Part	State test participation, math x reading	Imputed
Teacher_Ratio1	Teacher count / Total enrollment in school	Imputed
Civilian_Labor_Force	County labor force	USDA
All_Cohort	Count of students who could have graduated	DOE
Enrollment	Total enrollment in the school	CCD
Free_Or_Reduced_Price_Lunch	Count of students with free or reduced lunch	CCD
Teachers_Fte	Teacher count, full time equivalents	CCD
Reduced_Price_Lunch	Count of students with reduced lunch cost	CCD
Math_Pct_Part	State math test participation percent	DOE
Unemployment_Rate	County unemployment rate	USDA
Title_I_Status_5.0	School eligible for Title I and accepts it	CCD
Charter_0.0	School is not a charter school	CCD
Urban_Centric_Locale_11.0	School is in a large city	CCD

Conclusion

Conclusion

We received approximately 70% accuracy on regression scores

A classification of the country into over and under groups around the ACGR of 90% is possible with this limited dataset at an F1 of roughly 80%.

Interesting trends we found are the following:

- Vocational and regular schools graduate much higher than special and alternative schools
- State assessment participation correlates with graduation rates fairly strongly with reading having a larger effect than math.
- Unemployment negatively correlated to graduation rate
- School size and proximity of schools to city centers is predictive

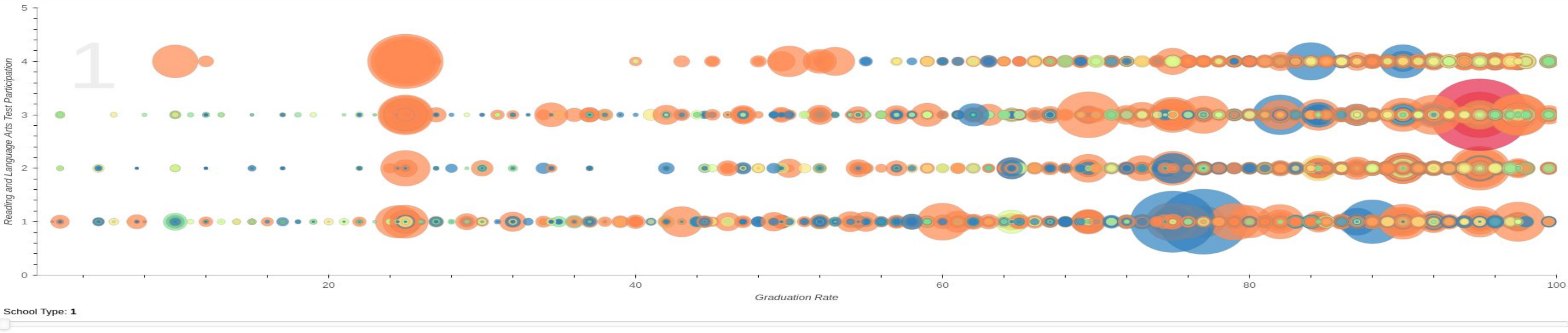
Conclusion

The most predictive factors of adjusted cohort graduation rate:

- school type (regular, vocational, special education, alternative)
- state proficiency test participation
- total teacher to cohort size ratio and its interactions
- urban versus rural environment of the school
- school lunch subsidies
- unemployment rate
- Charter school and virtual school designations

Further Work

- Racial and subpopulation data dropped because incomplete/inconsistent
- How representative is the sample?
- Check model errors for non-virtual schools.
- Further optimization, e.g. selectively applying different scaling schemes to different subsets of features instead of all at once
- Further model validation & interpretation (cv plots, partial dependence plots!!)
- Bokeh (buggy) prototype below; also had an EC2 website (not enough time to deploy)



Lessons We Learned

- ★ Group work can be challenging (and rewarding!)
- ★ Ingestion and wrangling of real data == easier said than done
- ★ EDA and Model Validation are first class citizens with the rest of the pipeline
- ★ Save feature metadata in json early (and anytime it changes!)

Acknowledgements

We would like to thank our instructors for their time both during and outside of lectures through office hours. We would also like to thank our capstone advisor for advice, suggestions and multiple check-in meetings throughout the course of this certificate program. Thank you for your time:

- Molly Morrison
- Kristen McIntyre
- Sam Goodgame
- Allen Leis
- Blake Bledar Zenuni
- Garin Kessler
- Prema Roman
- Kyle Rossetti

Image citations

- Python 3 logo
- SciKit Learn
- Yellowbrick
- Seaborn
- Matplotlib
- Jupyter Notebook
- AWS S3
- Bokeh
- Feature engine
- Pandas
- NumPy
- Boto 3