Early Childhood Education Impacts

Does participation in SABER early childhood goals help reduce a countrys primary school dropout rate?

Dataset and Research Question

Dataset

World Bank Education Statistics (1970-2017)

- Education Enrollment and Attainment
- Education Assessment and Learning Outcomes
- Economic and Labor Indicators
- Population and Health Statistics

This dataset is sparsely populated

- Min Year: 4%
- Max Year: 27%

Almost 900,000 Rows

Min 35,000 Values

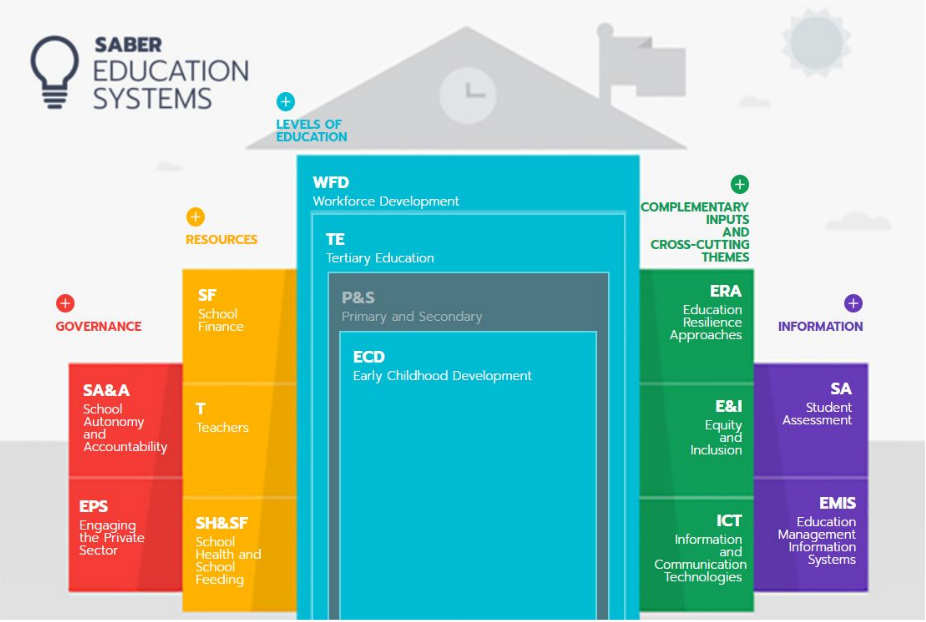
Max 240,000 Values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 886930 entries, 0 to 886929
Data columns (total 70 columns):
Country Name         886930 non-null object
Country Code         886930 non-null object
Indicator Name       886930 non-null object
Indicator Code       886930 non-null object
1970                 72288 non-null float64
1971                 35537 non-null float64
1972                 35619 non-null float64
1973                 35545 non-null float64
1974                 35730 non-null float64
1975                 87306 non-null float64
1976                 37483 non-null float64
1977                 37574 non-null float64
1978                 37576 non-null float64
1979                 36809 non-null float64
1980                 89122 non-null float64
1981                 38777 non-null float64
1982                 37511 non-null float64
1983                 38460 non-null float64
1984                 38606 non-null float64
1985                 90296 non-null float64
1986                 39372 non-null float64
1987                 38641 non-null float64
1988                 38552 non-null float64
1989                 37540 non-null float64
1990                 124405 non-null float64
1991                 74437 non-null float64
1992                 75543 non-null float64
1993                 75793 non-null float64
1994                 77462 non-null float64
1995                 131361 non-null float64
1996                 76807 non-null float64
1997                 73453 non-null float64
1998                 84914 non-null float64
1999                 118839 non-null float64
2000                 176676 non-null float64
2001                 123509 non-null float64
2002                 124205 non-null float64
2003                 130363 non-null float64
2004                 128814 non-null float64
2005                 184108 non-null float64
2006                 140312 non-null float64
2007                 137272 non-null float64
2008                 134387 non-null float64
2009                 142108 non-null float64
2010                 242442 non-null float64
2011                 146012 non-null float64
2012                 147264 non-null float64
2013                 137509 non-null float64
2014                 113789 non-null float64
2015                 131058 non-null float64
2016                 16460 non-null float64
```

Research Questions

Early Performance of SABER Programs
(Systems Approach for Better Education Results)

- Does participation in the SABER early childhood goals lead to improved outcomes for children?
    - Improved outcome?
    - Measuring SABER participation?
    - Other meaningful indicators?
    - What's missing?

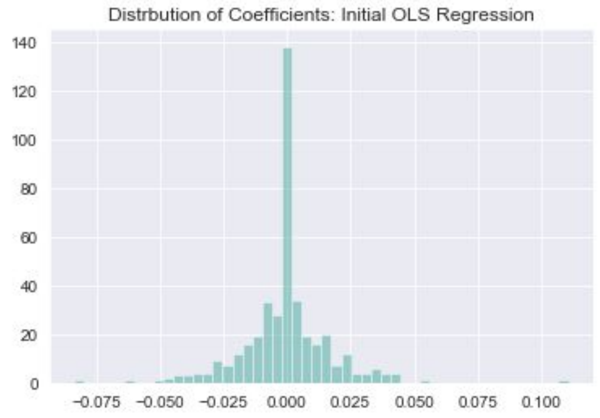


Feature Selection and Engineering

Feature Selection

Stage 1: Feature Density, Outcome Variable Selection and Initial Model

- 413 Variables > 2 Values in 150+ Countries
- Outcome: Avg. Change in Rate of Out-of-School Primary-aged Children
- Initial Models: overfit, no standout indicators



Feature Engineering

Null Handling, Change Variables, and Current Variables

- Backfill, forward fill

```
# Make a copy of the data frame that has only the features for the model, backfill then frontfill any NaNs
features_df = features_df.fillna(method='bfill', axis=1)
features_df.iloc[:, 2:] = features_df.iloc[:, 2:].fillna(method='ffill', axis=1)
```

- Numpy Mean of (Diff)

```
# For each row, take the mean of the year to year differences, ignoring NaNs
for i in range(len(features_arr)):
    features_arr[i] = np.append(features_arr[i], np.nanmean(np.diff(features_arr[i][2:])))
```

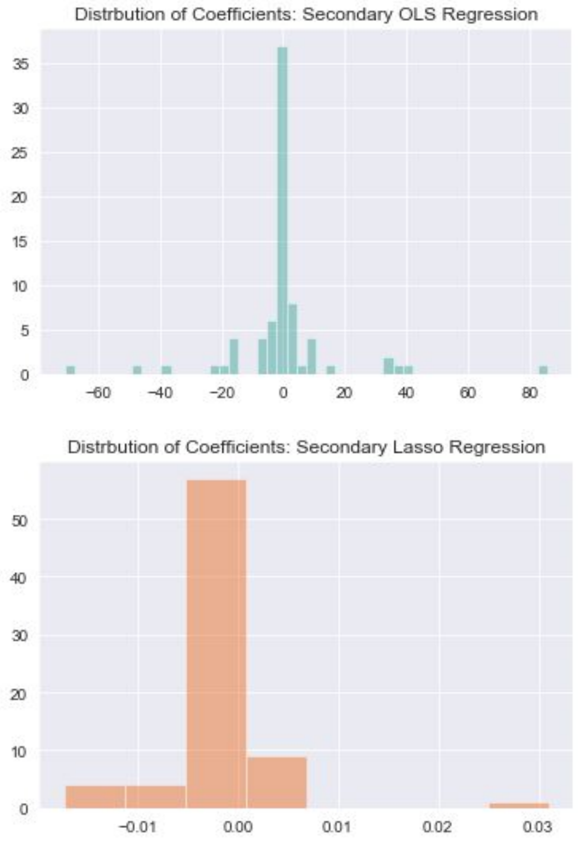
- 2015 (if NaN work backwards)

```
# For each row, check find the most current year where the value is not NaN
for i in range(len(features_arr)):
    for x in features_arr[i][-2:2:-1]:
        if not isnan(x):
            features_arr[i] = np.append(features_arr[i], x)
            break
```

Feature Selection

Stage 2: SelectKBest and Lasso Regression

- Select 75 best
- Models no longer overfit
- Lasso reduces 50+ variable coefficients to 0
    - All but 5 < 0.01



Feature Selection

Stage 3: Indicators of Theoretical Interest

SP.POP.TOTL: Population, total
SE.PRM.AGES: Official entrance age to primary education (years)
SE.COM.DURS: Duration of compulsory education (years)
SH.DYN.MORT: Mortality rate, under-5 (per 1,000)
SL.UEM.TOTL.ZS: Unemployment, total (% of total labor force)
SL.TLF.TOTL.FE.ZS: Labor force, female (% of total labor force)
UIS.FEP.2.GPV: Percentage of students in lower secondary general education who are female (%)
UIS.GOER.56: Gross outbound enrolment ratio, all regions, both sexes (%)
NY.GNP.PCAP.PP.CD: GNI per capita, PPP (current international $)
UIS.ROFST.1: *Rate of out-of-school children of primary school age, both sexes (%)*
* GNI = total domestic and foreign output within country
*PPP = purchasing power parity

Dropped Countries

Afghanistan, American Samoa, Andorra, Aruba, Austria, Bermuda, Bosnia and Herzegovina, British Virgin Islands, Brunei Darussalam, Cayman Islands, Channel Islands, China, Congo, Dem. Rep., Curacao, Czech Republic, Dominica, Faroe Islands, French Polynesia, Gabon, Gibraltar, Greenland, Guam, Haiti, Hong Kong SAR, China, Iraq, Isle of Man, Jamaica, Kosovo, Libya, Liechtenstein, Madagascar, Macao SAR, China, Malawi, Maldives, Micronesia, Fed. Sts., Monaco, Nauru, New Caledonia, Northern Mariana Islands, Puerto Rico, Singapore, Sint Maarten (Dutch part), Slovak Republic, Somalia, South Africa, St. Martin (French part), St. Lucia, Turkmenistan, Turks and Caicos Islands, Virgin Islands (U.S.)

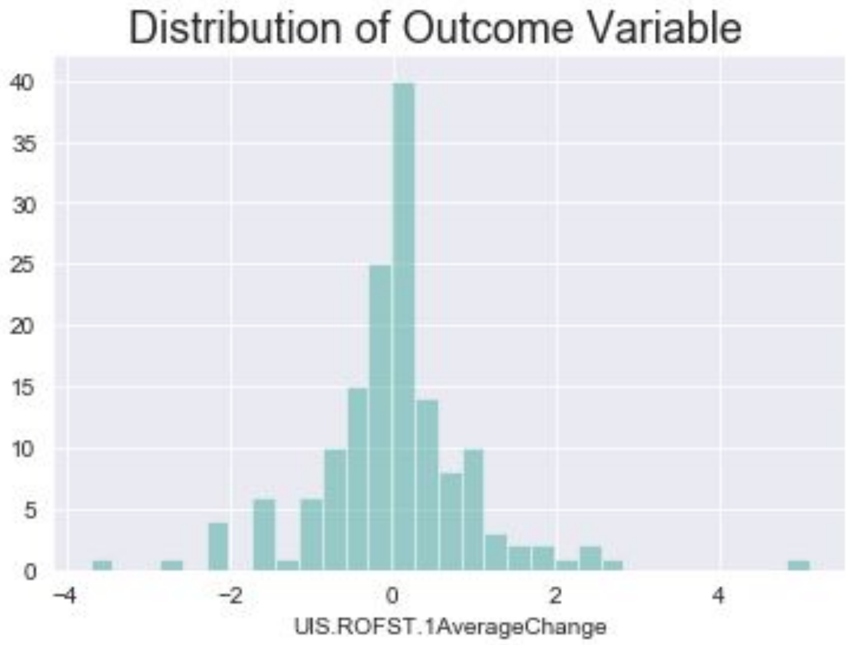
Model Building and Validation

Model Performance

Model	Features	Parameters	Model Score
OLS	All - Scaled	None	0.17
OLS	Change	None	0.08
OLS	Current	None	0.14
Random Forest Regressor	All - Scaled	{criterion: mae, min_impurity_decrease: 0.001, n_estimators: 200}	-0.12
Random Forest Regressor	Change	{criterion: mse, min_impurity_decrease: 0.01, n_estimators: 200}	-0.21
Random Forest Regressor	Current	{criterion: mae, min_impurity_decrease: 0.01, n_estimators: 100}	-0.14
Gradient Boosting Regression	All - Scaled	{learning_rate: 0.0001, n_estimators: 500}	-0.03
Gradient Boosting Regression	Change	{learning_rate: 0.001, n_estimators: 100}	-0.02
Gradient Boosting Regression	Current	{learning_rate: 0.0001, n_estimators: 100}	-0.04

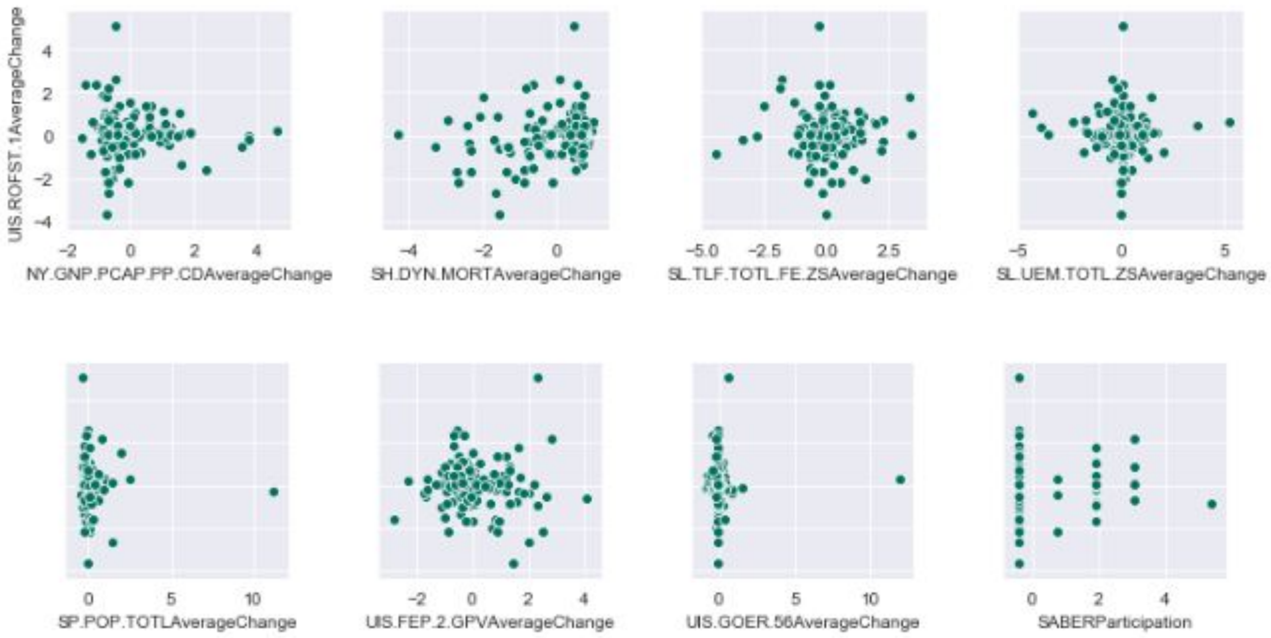
Feature Distributions and Model Assumptions

Distribution of Outcome Variable



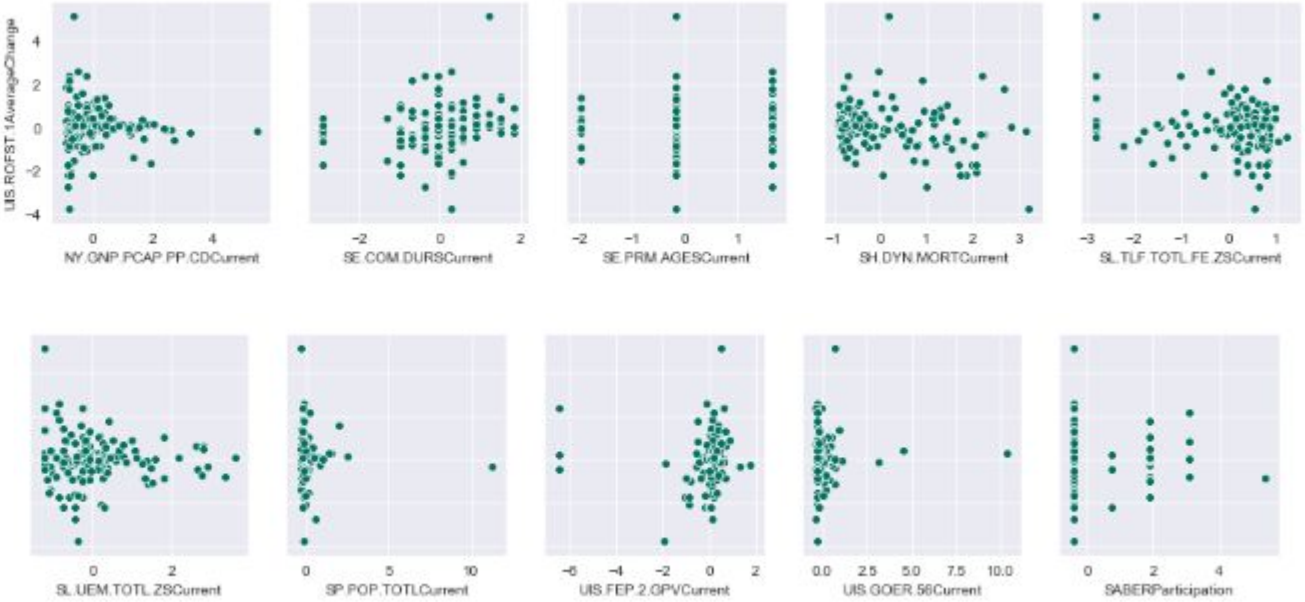
Feature Distributions and Model Assumptions

Linear Regression Model Assumptions: Linear Relationship Change Features



Feature Distributions and Model Assumptions

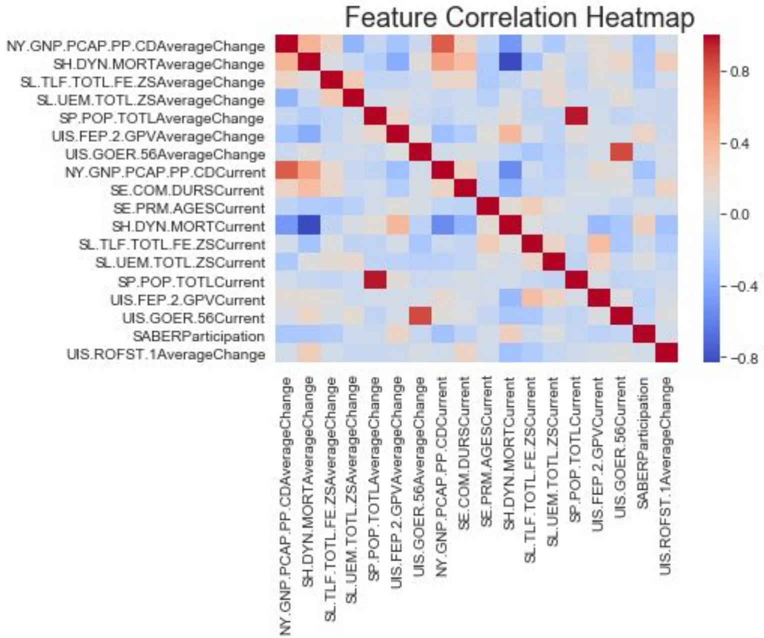
Linear Regression Model Assumptions: Linear Relationship Current Features



Feature Distributions and Model Assumptions

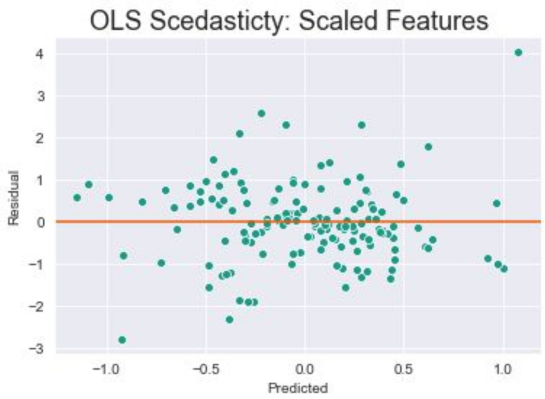
Linear Regression Model Assumptions: Low Multicollinearity

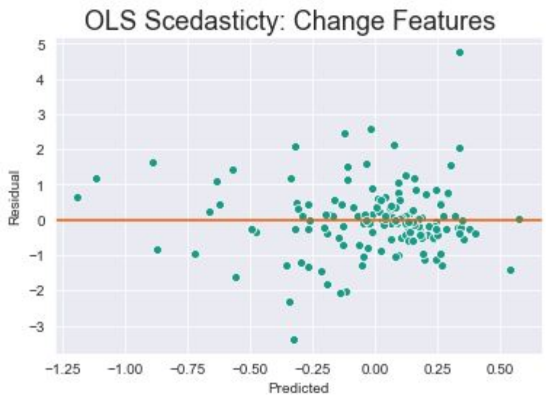
- Relatively low collinearity
- Keep all
  - Model all
  - Model avg change features
  - Model current features

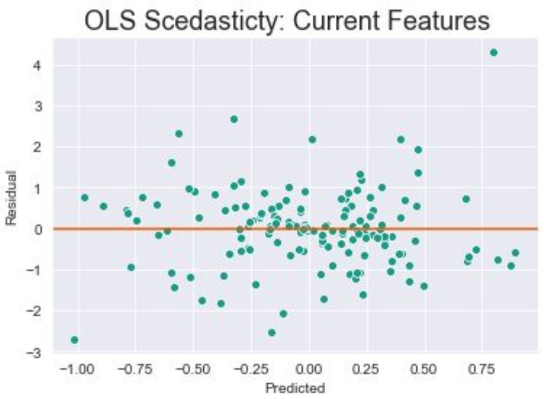


Feature Distributions and Model Assumptions

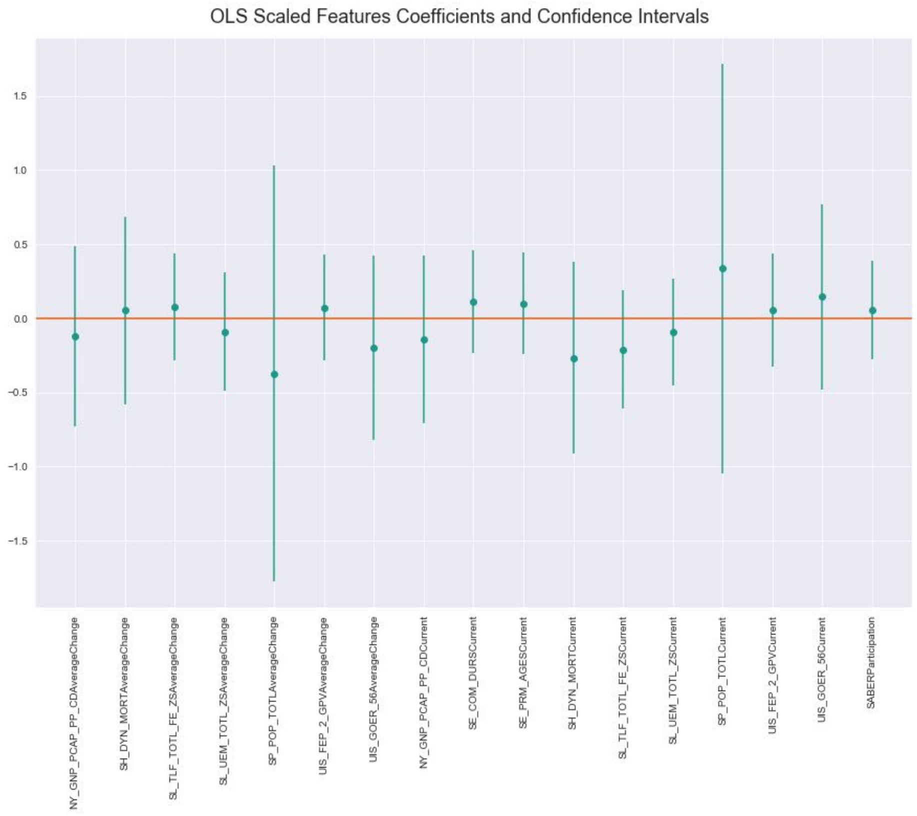
Linear Regression Model Assumptions: Homoscedasticity



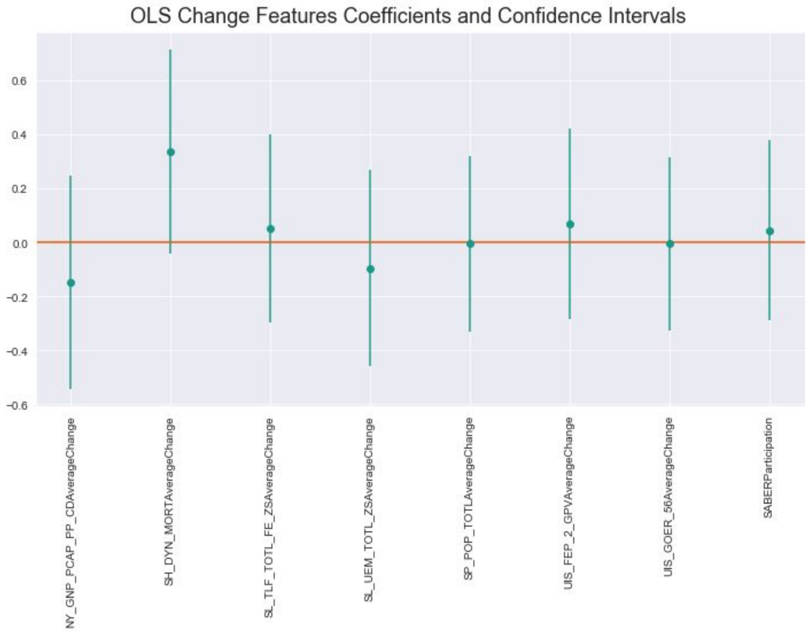


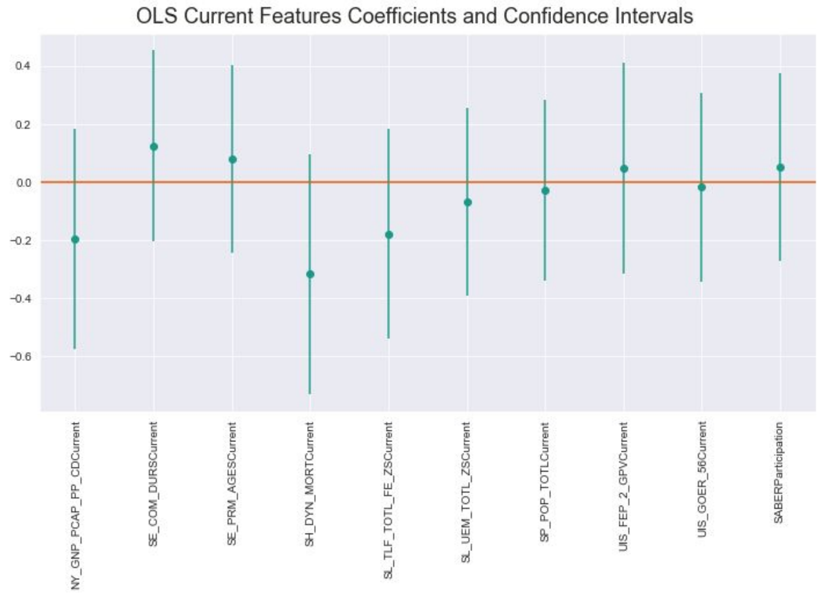


Linear Regression Features Coefficients

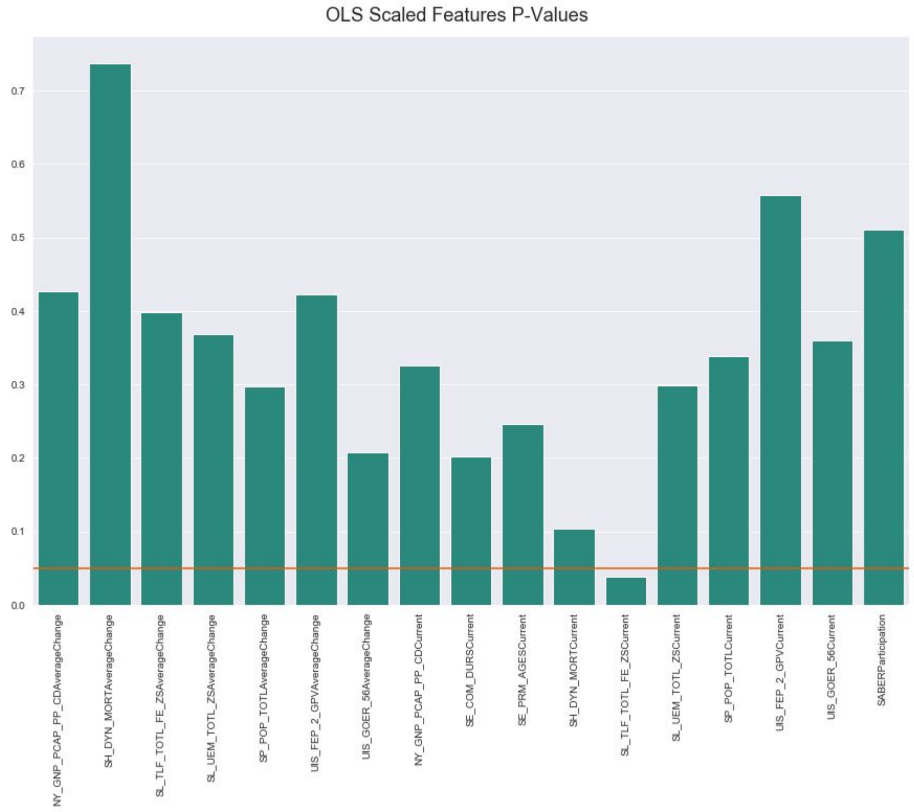


Linear Regression Features Coefficients

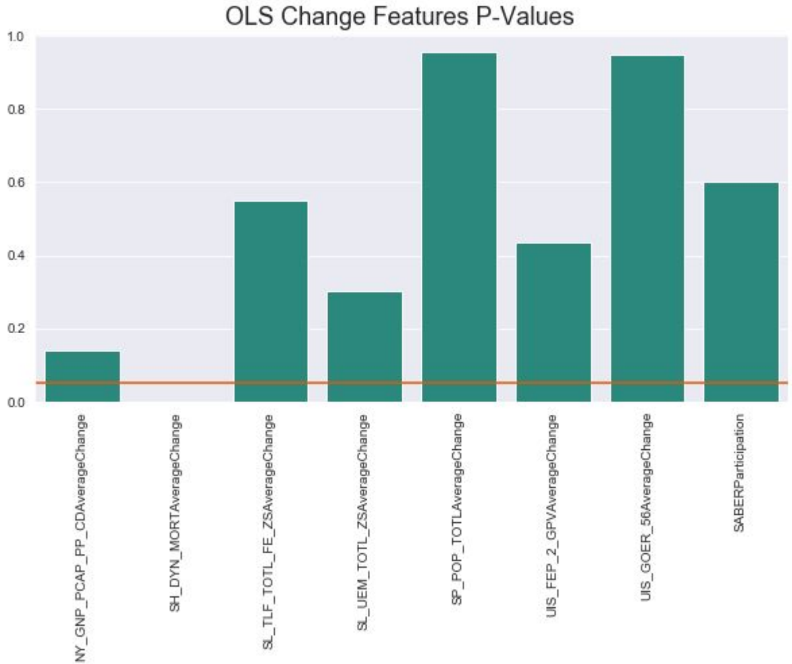


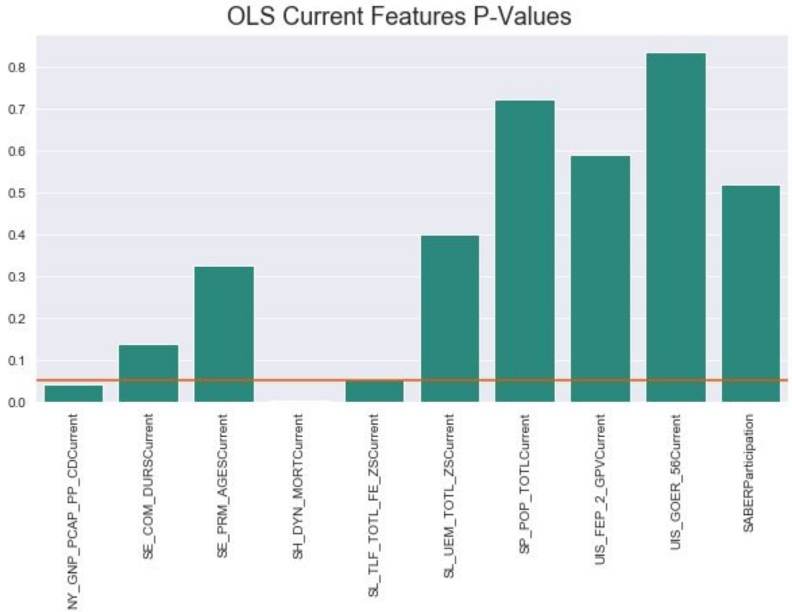


Linear Regression Feature Significance

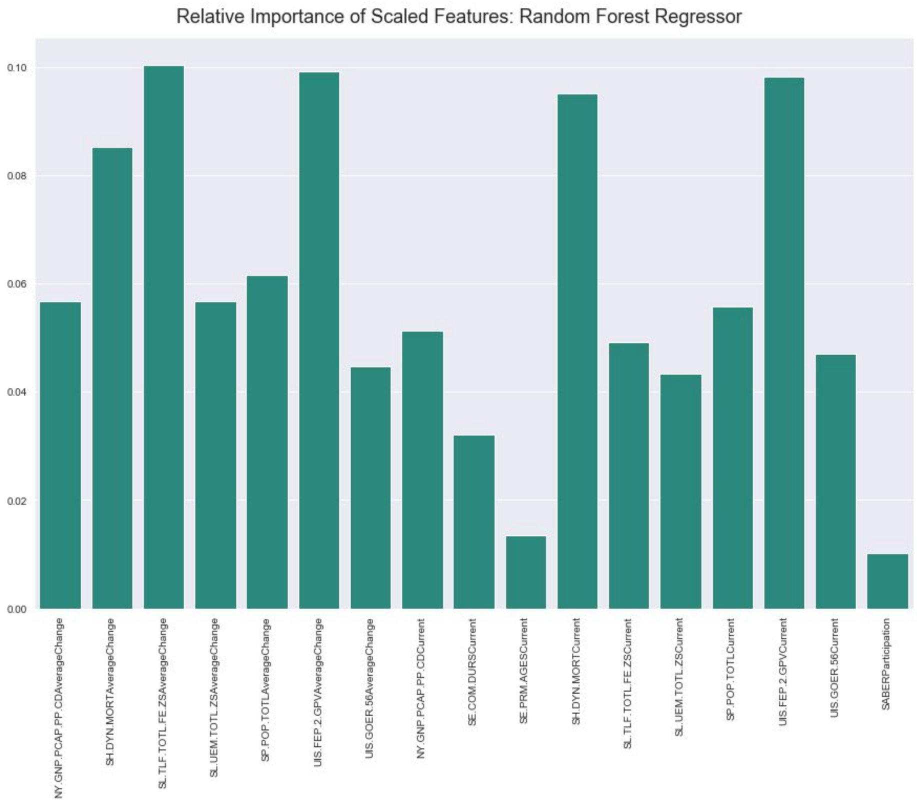


Linear Regression Feature Significance

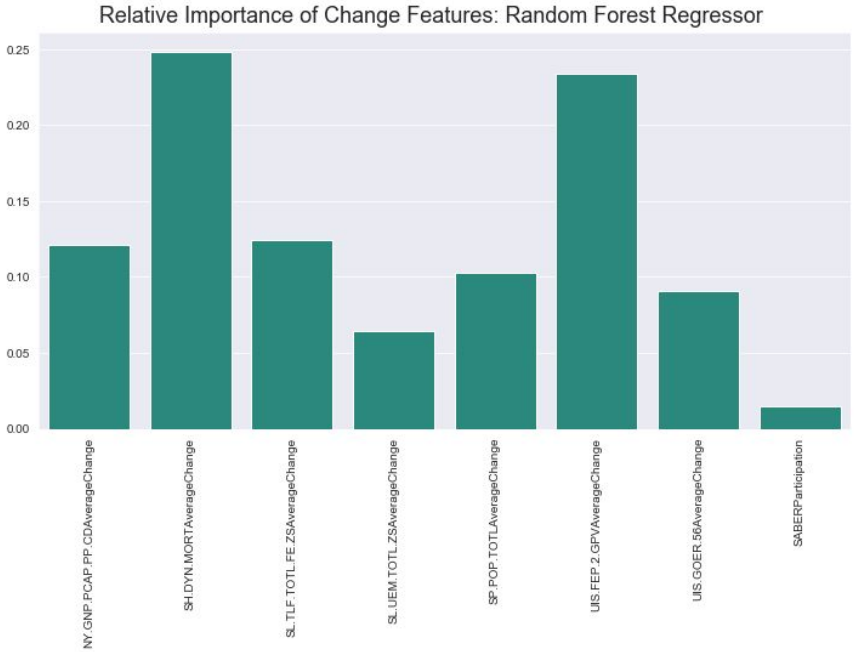


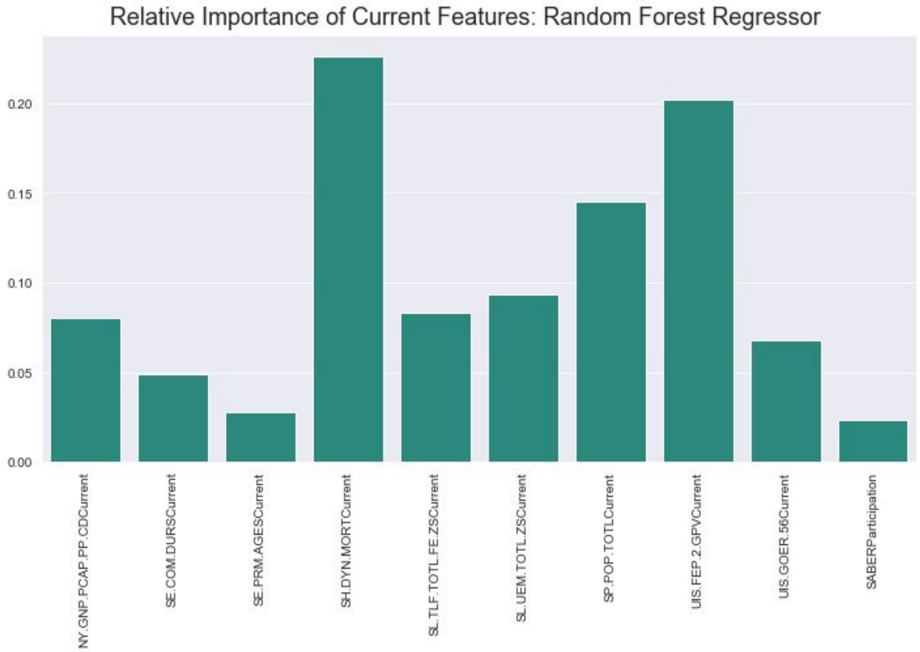


Random Forest Regressor Results

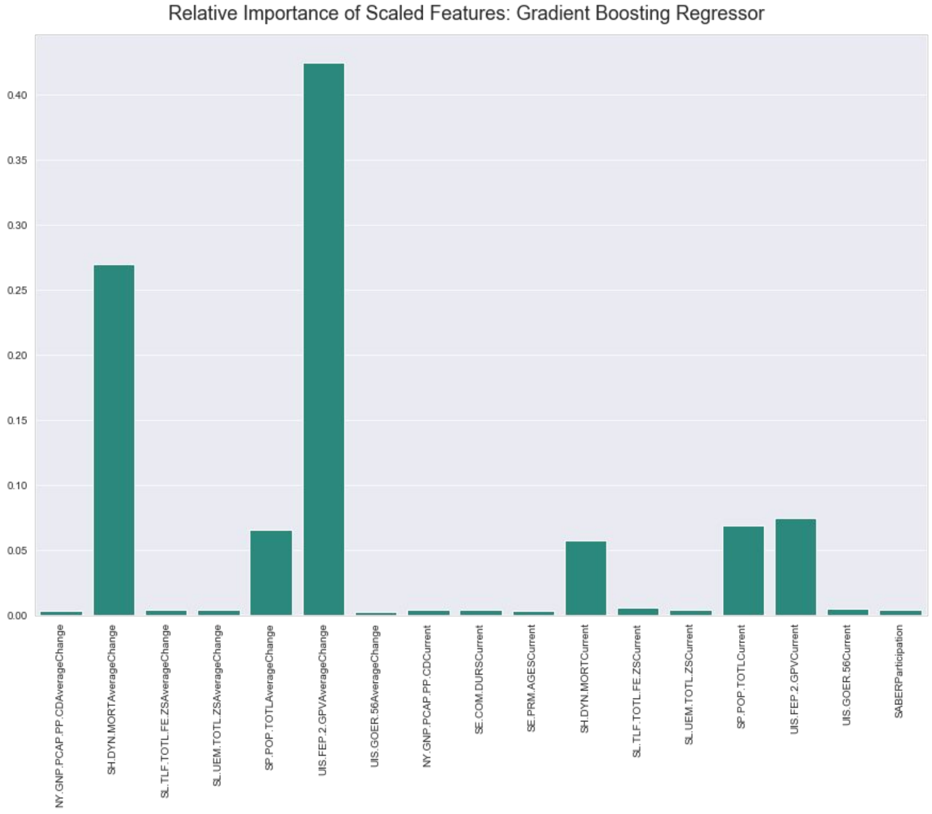


Random Forest Regressor Results

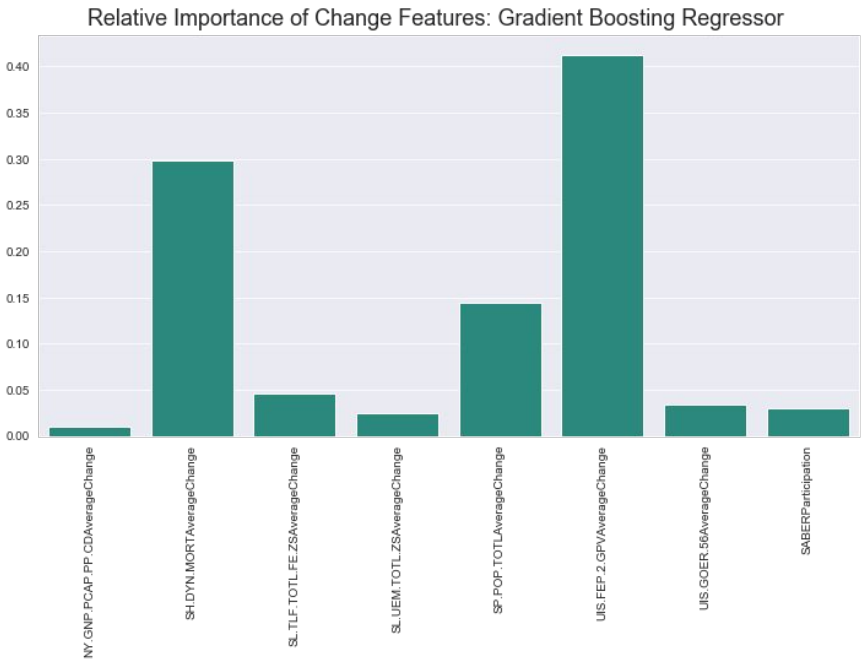


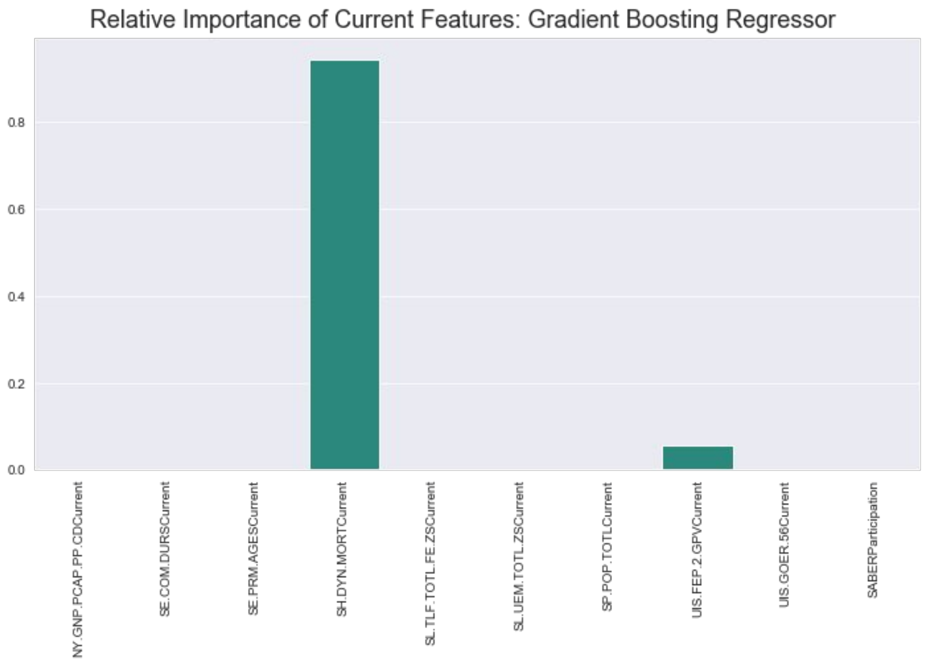


Gradient Boosting Regressor Results



Gradient Boosting Regressor Results





Conclusion

Assumptions and Shortcomings

- Data is sparse
- Problem is complex
- Features don't meet all linear regression assumptions

Conclusions and Next Steps

- Little statistical significance for SABER participation 2010-2015
- Correlation with:
  - Labor force, female (% of total labor force)
  - Mortality rate, under-5 (per 1,000)
  - Percentage of students in lower secondary general education who are female (%)
- Get change in score data with next measurement year

Questions? Comments? Concerns?