THE 3 E'S

USING EDUCATION, EMPLOYMENT, AND ECONOMIC DEVELOPMENT INDICATORS TO PREDICT MATH PISA PERFORMANCE

AGENDA

- Introduction
- Data Cleaning
- Feature Selection
- Intro to Modeling
- Examples of Modeling Graphs
- Best Performing Model
- Worst Performing Model
- Lessons learned and Use Case
- Conclusion

INTRODUCTION

- PISA
  - Worldwide study to evaluate education systems
  - Measure education performance of 15 year old students on math, science, and reading
- World Development Indicators Used:
  - Economy - Gross domestic product (GDP), inflation, etc.
  - Education - Enrollment rates, ratio of teachers to students, vocational versus primary school enrollments
  - Employment - Wage & salaried workers, unemployment rates based on different criteria

WHERE DOES MY DATA COME FROM?

- World Development Indicator Dataset from Kaggle: https://www.kaggle.com/worldbank/world-development-indicators
- The World Development Indicator Dataset includes over a thousand annual indicators for roughly 247 countries. The dataset has information for 55 years (1960 to 2015).

- PISA data: https://pisadataexplorer.oecd.org/ide/idepisa/
- The PISA data website allows us to select from 100s of countries. It measures the performance of 15 year old students in the subjects of math, science, and reading. The data used in my project brings in the PISA scores from 2003 to 2015

DATA CLEANING

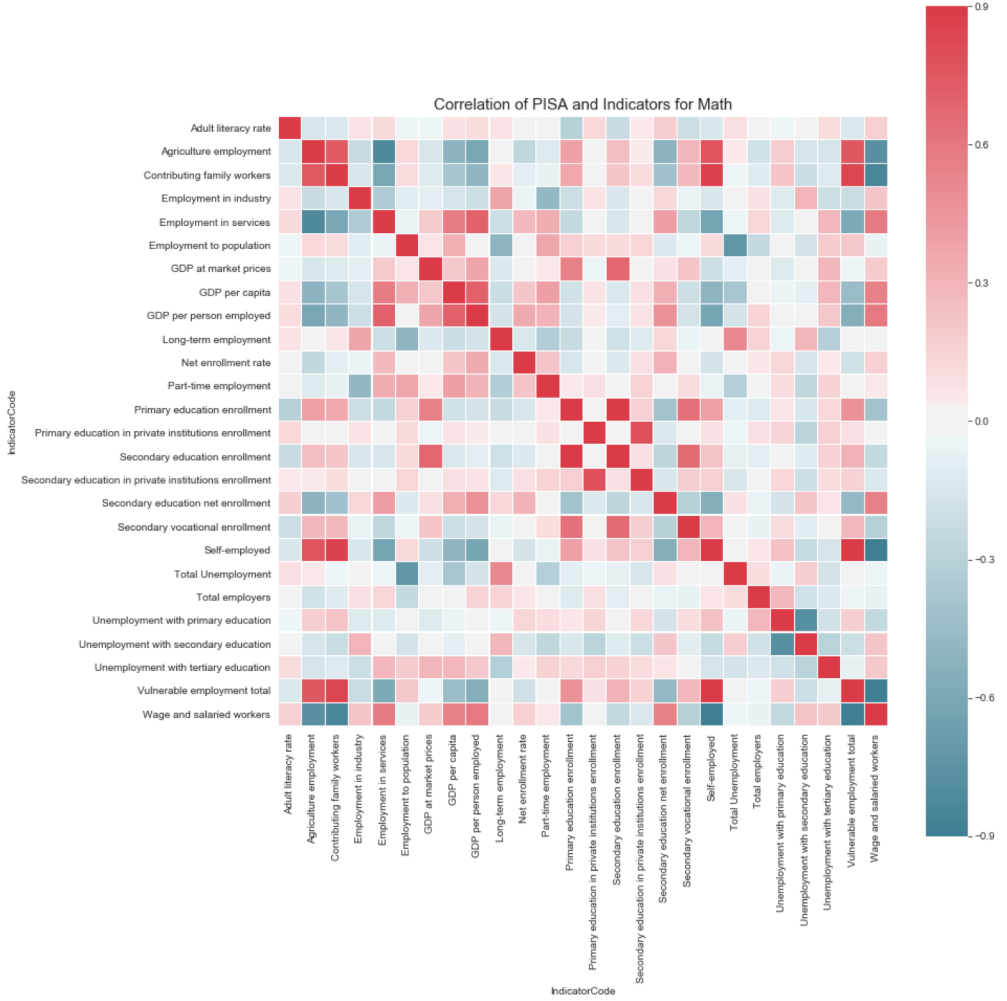
- Decision was made to focus on one PISA area (math) as opposed to all three due to collinearity
- Columns that were not relevant from merged datasets were dropped
- Replaced null values with PISA averages for countries
- Merged data sets
- Removed years that did not exist in PISA
- Aggregating and formatting data in correct shape to ensure wide format

FEATURE SELECTION

- The 2 datasets are quite large and contains a lot of collinearity
- We started with thousands of development indicators, but through data cleaning, have been able to narrow down to 28 features (indicators)
- Usage of cluster and heat maps provide visuals of highly correlated features
  - Helpful in determining if features might be repetitive of each other
  - First look at what features related to each other and starts to show trends

FEATURE SELECTION

- Heat Map
  - Allows visual of features and if collinearity exists
  - Able to revisit data and clean up further to reduce collinearity
  - Shows areas of high vs low correlation

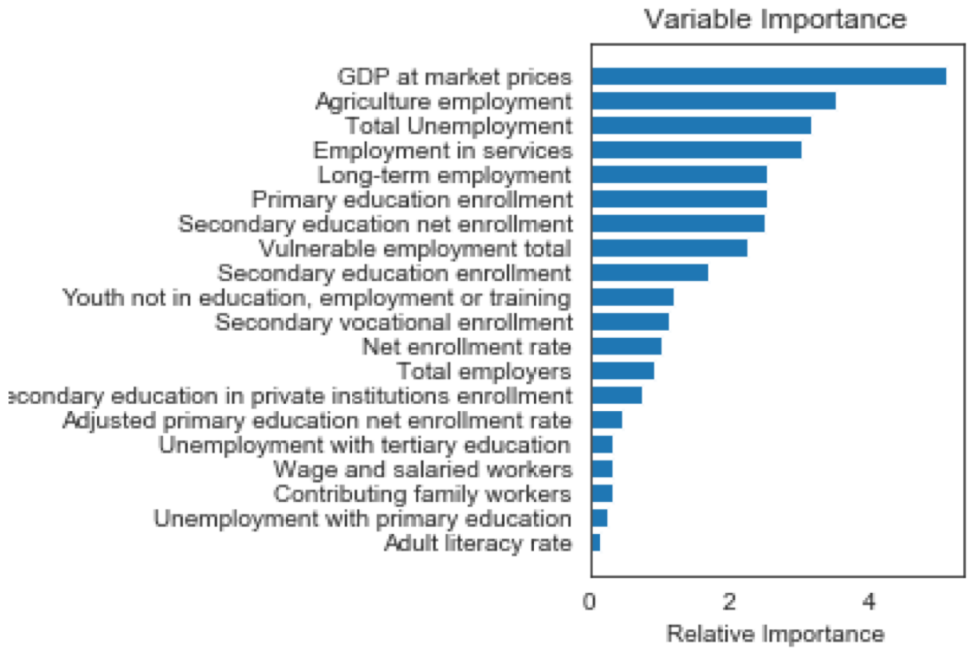


INTRO TO MODELING

- Y is PISA Score (continuous on scale of X to X – find it and put it in there – have that weird graph from old version I can show)
- X is 3 E's World Indicators
- Scores are measured by R^2
- Types of models used for the 3 E's
  - Random Forest Regression
  - Linear Regression
  - Lasso Regression
  - Ridge Regression
  - Gradient Boost Regression
  - K Nearest Neighbors Regression

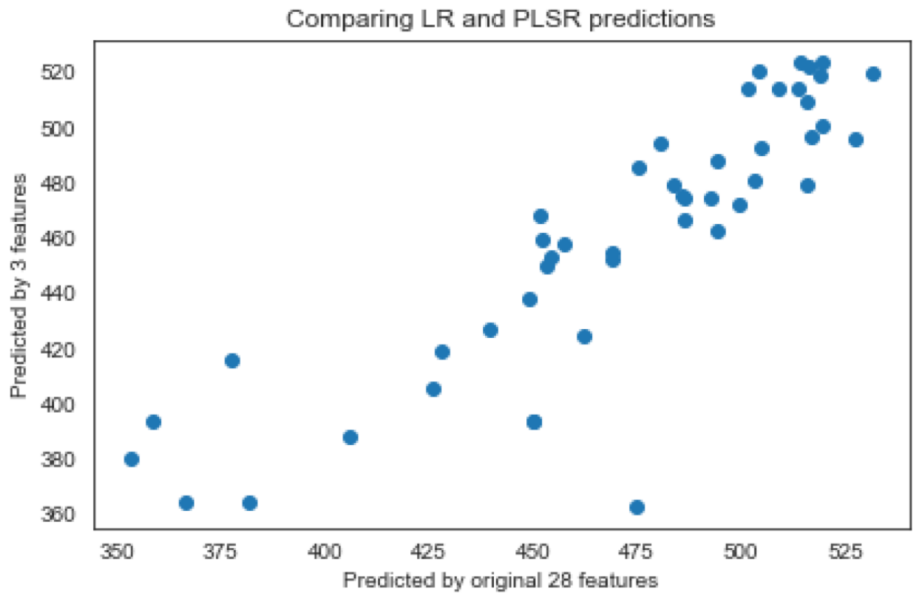
EXAMPLES OF MODELING GRAPHS

- Gradient Boost
  - Measure importance of variance features by counting how many times a features is used over the course of many decision trees
  - Model shows the employment features are at the top



EXAMPLES OF MODELING GRAPHS

- Linear Regression Comparison to Partial Least Squares Regression
  - Comparison of performance of original 28 features against reduction to 3 features
  - Looking at the model, the accuracy is not the best, but data is not suffering from overfitting



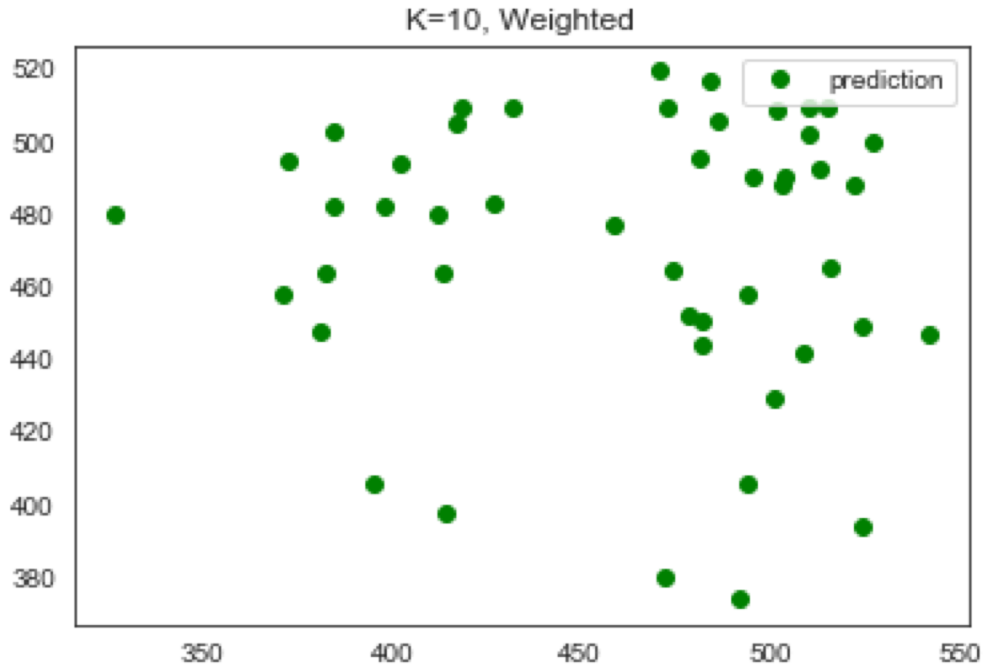
BEST PERFORMER – RANDOM FOREST

- Ranking of features consistent with what we have seen in other models
- Delta between training and testing data is not high
  - Overfitting has been corrected

```
Feature ranking:
1. feature 9 GDP per person employed (0.401219)
2. feature 8 GDP per capita (0.256522)
3. feature 12 Part-time employment (0.063141)
4. feature 6 Employment to population (0.042381)
5. feature 14 Primary education in private institutions enrollment (0.038278)
6. feature 27 Youth not in education, employment or training (0.029899)
7. feature 2 Agriculture employment (0.017246)
8. feature 25 Vulnerable employment total (0.016594)
9. feature 4 Employment in industry (0.015480)
10. feature 18 Secondary vocational enrollment (0.012791)
11. feature 5 Employment in services (0.012680)
12. feature 10 Long-term employment (0.011111)
13. feature 17 Secondary education net enrollment (0.009446)
14. feature 23 Unemployment with secondary education (0.008223)
15. feature 24 Unemployment with tertiary education (0.007525)
16. feature 19 Self-employed (0.007499)
17. feature 22 Unemployment with primary education (0.007267)
18. feature 13 Primary education enrollment (0.006949)
```

WORST PERFORMER – K NEAREST NEIGHBORS

- Working with features that are different from each other
- Attempt to predict math PISA score against development indicator
- There are no neighbors because of this
- Model does not return good data



LESSONS LEARNED AND USE CASE

Lessons Learned

- When modeling, reduce collinearity.
- Reduce features - should not have more features than rows
- Eliminate redundant features
- Use of correlation plots to provide a visual of features that are duplicates and should be removed
- Establish specific timeframe you want the data to pull from to provide more accurate averages

Use Case

- App geared towards expats
- Analysis to aid countries on economic spend versus investment in different areas for growth – focus on PISA. If country was looking to improve PISA, etc – edit this blurb

CONCLUSION

- More data does not equal accurate models
  - Leads to overfitting due to collinearity
  - Inconsistent trends
  - Significant deltas between testing and training data sets
- What surprised me about the features?
  - Economy and Employment indicators ranked higher in predicting higher math PISA scores than Education
  - Models such as ridge, showed results that certain indicators actually result in lower PISA scores that were not expected
    - Higher self employment shows lower PISA scores

THANK YOU