Predicting Graduation Outcomes for NYC Public Schools

NATHANIEL LANIER

BROWN UNIVERSITY

DATA1030

PRESENTATION DATE: 12/03/2020

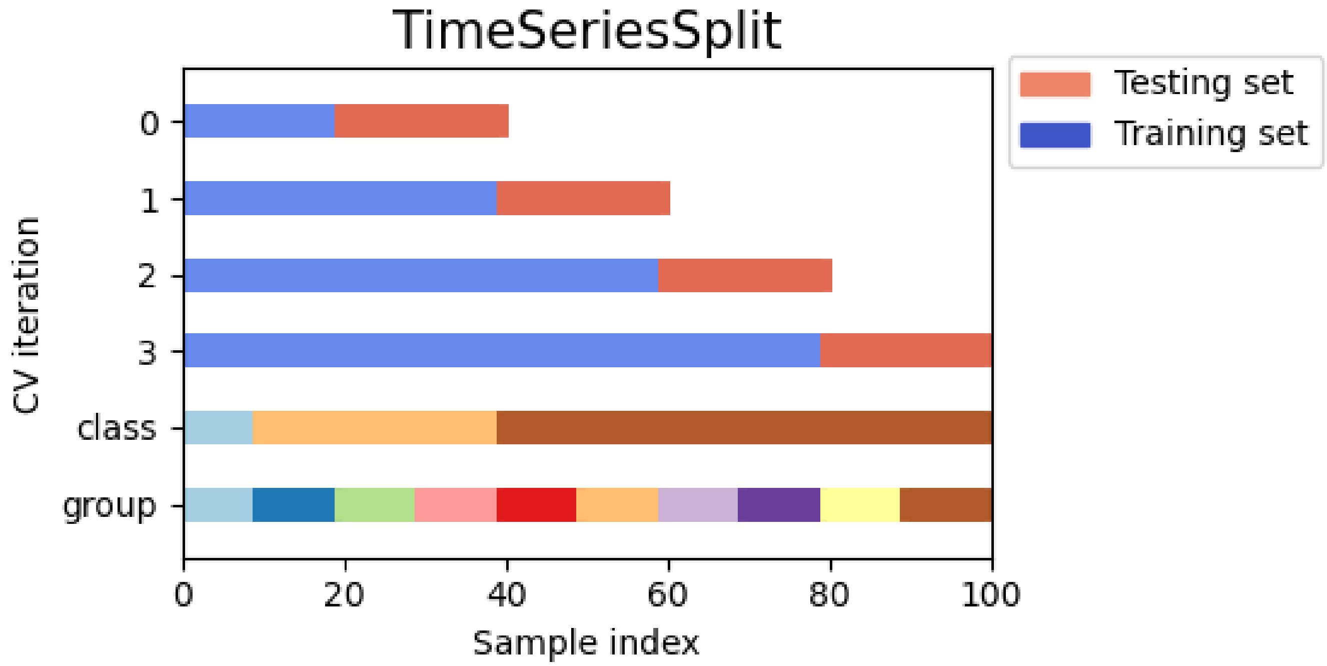
HTTPS://GITHUB.COM/NCL11/DATA1030_PROJECT

Problem: Predict Cohort Graduation Percentage Outcomes

- High School graduation has a large impact on income and overall life outcomes
- Accurately predicting graduation rates can help the city allocate resources accordingly
- Two types of diplomas in NY state: Regents and Local
- The target variable is graduation percentages broken down on the school level, so it is a regression problem
- The data is sourced from NYC OpenData: https://opendata.cityofnewyork.us/
- Data cleaning, dealing with missing data and adding Income column was necessary before we began splitting
- Consistent increase in graduation percent every year

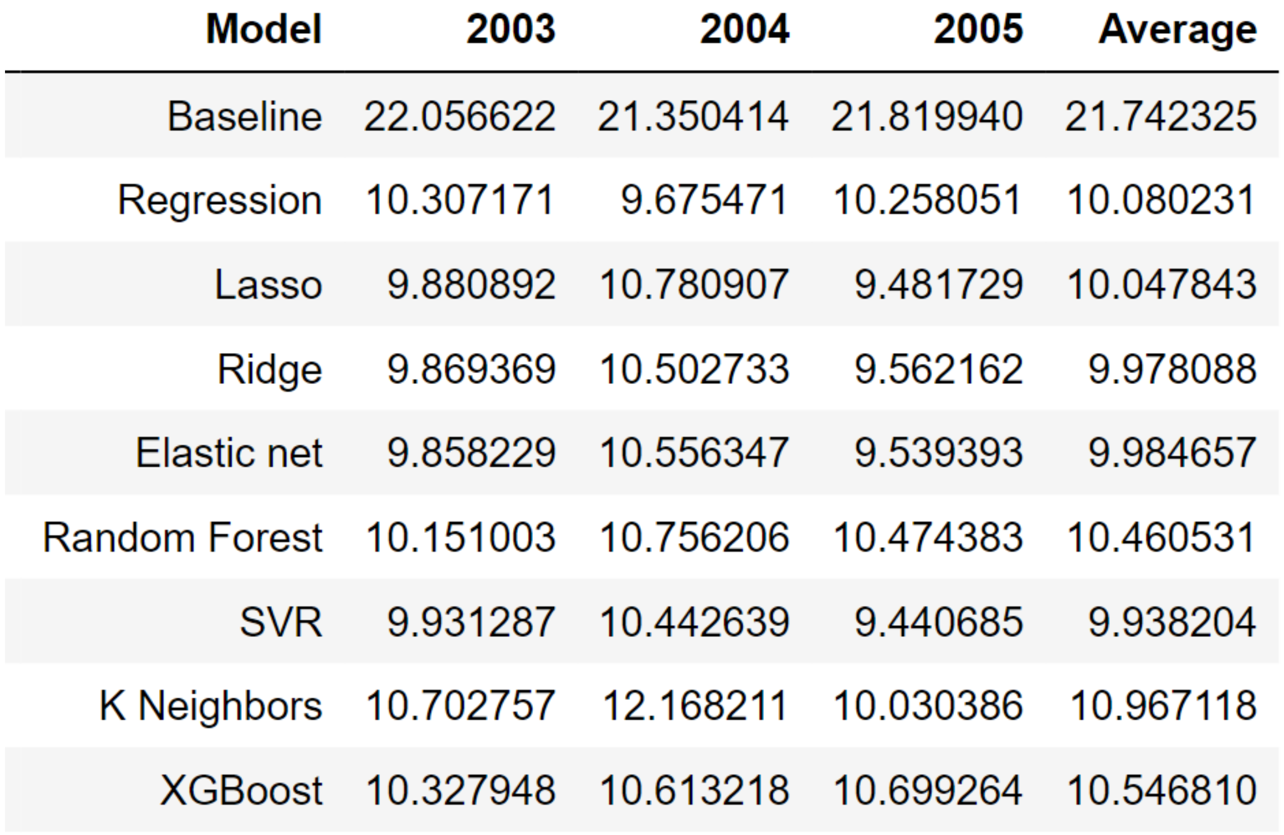
Splitting and Preprocessing

- Times series split
- Most other variables were relatively evenly distributed so time was the only variable taken into consideration when splitting
- Number of schools is fixed so no need to perform well on unseen schools
- Source for Graphic: https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_indices.html#sphx-glr-auto-examples-model-selection-plot-cv-indices-py



Models Tested

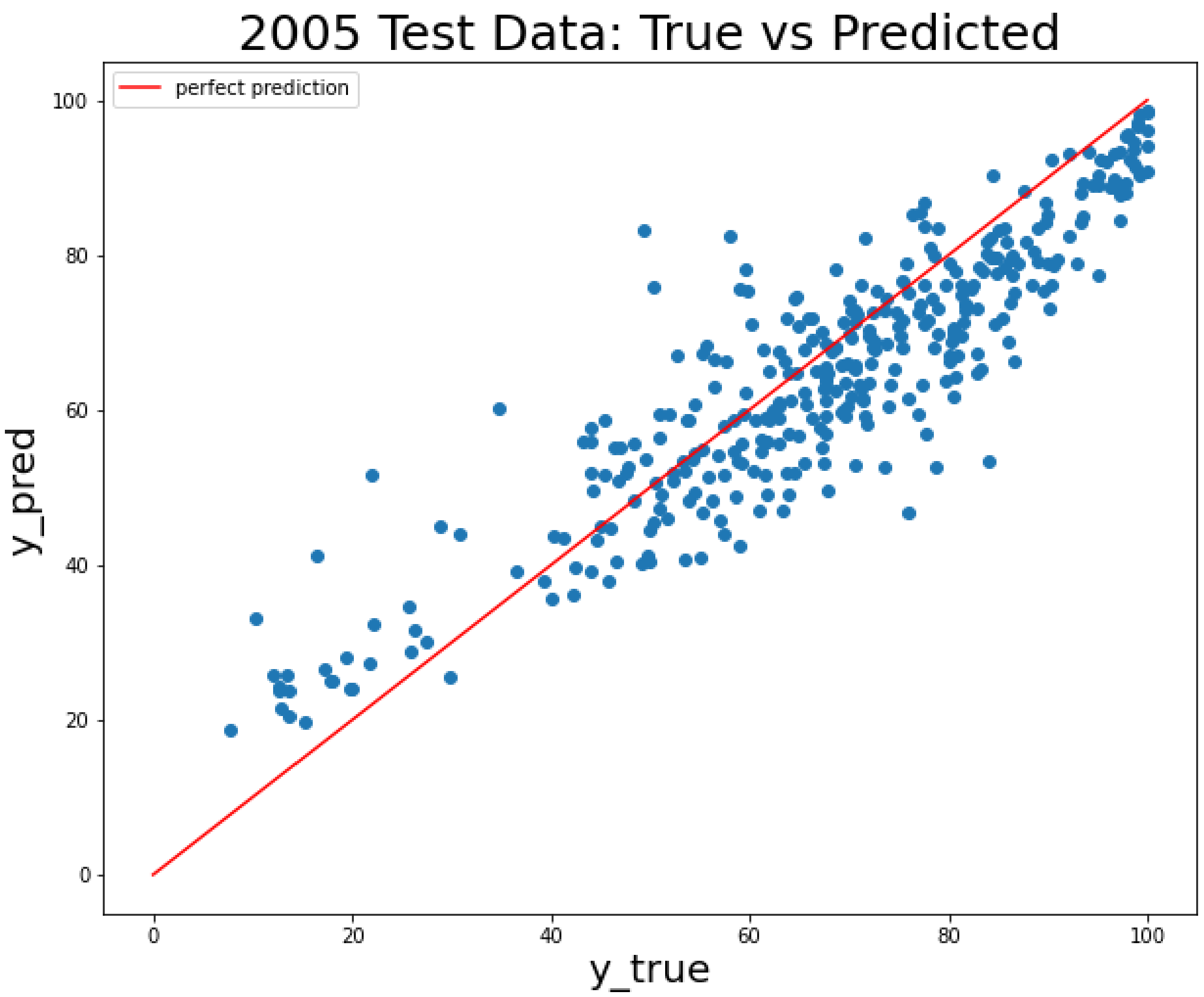
- Scores measured in RMSE
- SVR had the best average and best single year test score- kernel: rbf, gamma: 2^-13, C:128
- Lasso had the second-best single year test score- alpha: 0.1
- Elastic Net was third
- Regular regression performed surprisingly well



Model	2003	2004	2005	Average
Baseline	22.056622	21.350414	21.819940	21.742325
Regression	10.307171	9.675471	10.258051	10.080231
Lasso	9.880892	10.780907	9.481729	10.047843
Ridge	9.869369	10.502733	9.562162	9.978088
Elastic net	9.858229	10.556347	9.539393	9.984657
Random Forest	10.151003	10.756206	10.474383	10.460531
SVR	9.931287	10.442639	9.440685	9.938204
K Neighbors	10.702757	12.168211	10.030386	10.967118
XGBoost	10.327948	10.613218	10.699264	10.546810

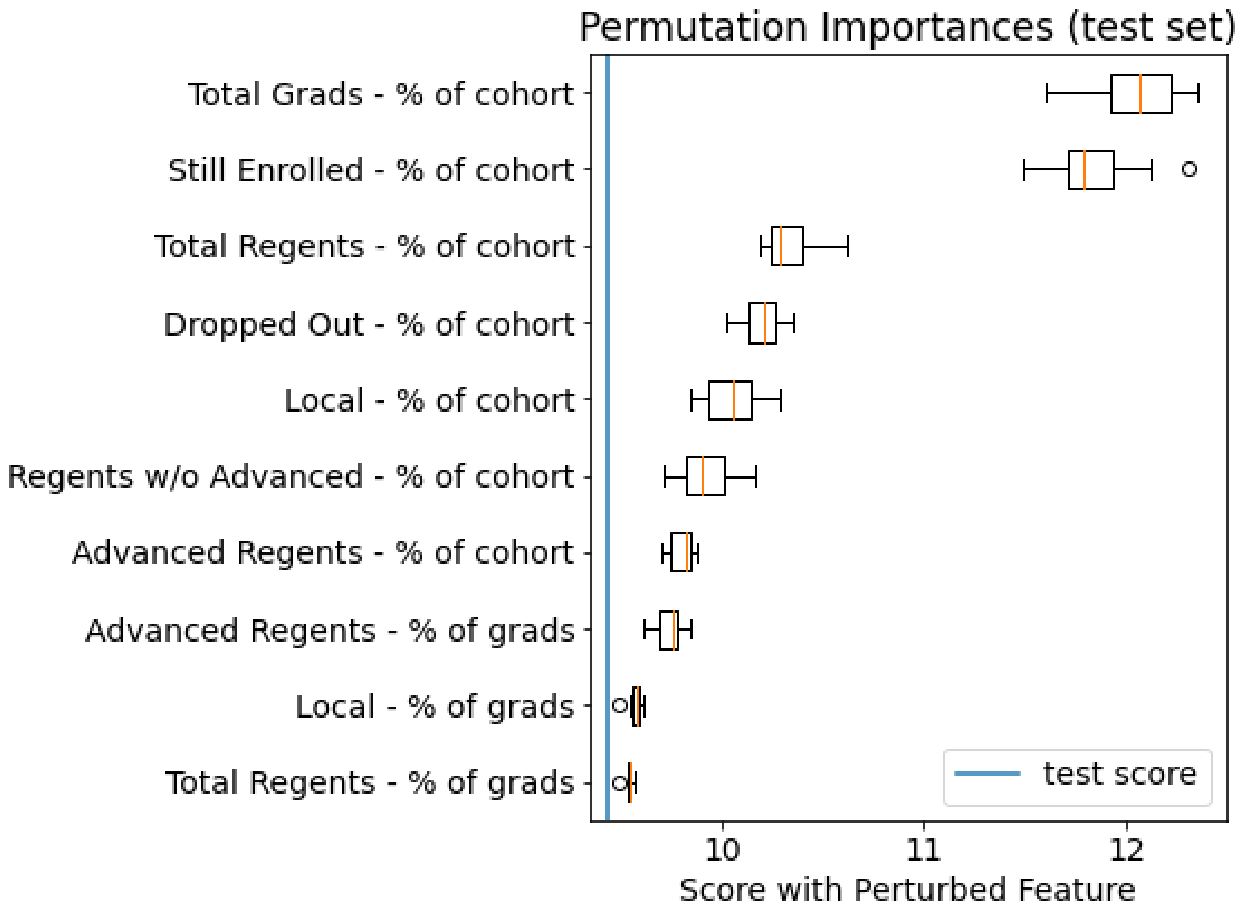
SVR Performance

- Overall score is decent but some of the outlier predictions are concerning
- Seems to underpredict on average
- Likely caused by fact that response is non-stationary



Feature Importance

- Total Grads - % of Cohort and Still Enrolled - % of cohort were two most important by a significant margin
- Interestingly there were several features that slightly improved the score when they were permuted



Outlook

- Next step is to get more recent data but we are confident the pipeline built in this project could be reused
- In order to improve the model we would like additional features such as some metric of parental involvement, more accurate income data, income data for school employees etc. which will likely cost additional money
- Developing local feature importance measurements will be important to explain level of funding to school administrators
- Given the relatively low RMSE we think this model could play an important role in resource allocation decision making

Thanks For Listening!

QUESTIONS?