Classifiers: NYC Poverty Status

A Data Science Project (in the making)

FLATIRON SCHOOL

Project Overview

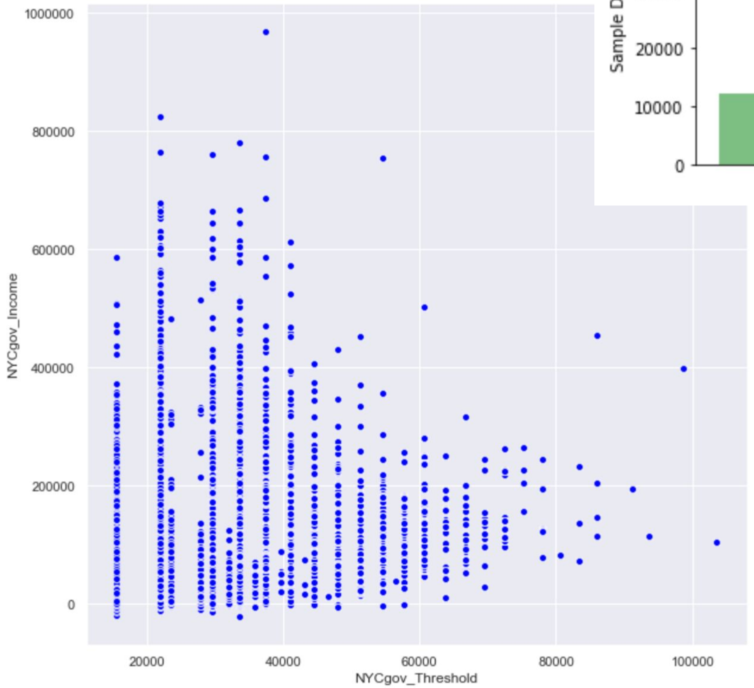
Aim: To identify classifier models to predict whether an individual is or is not officially in poverty according to NYC government terms.

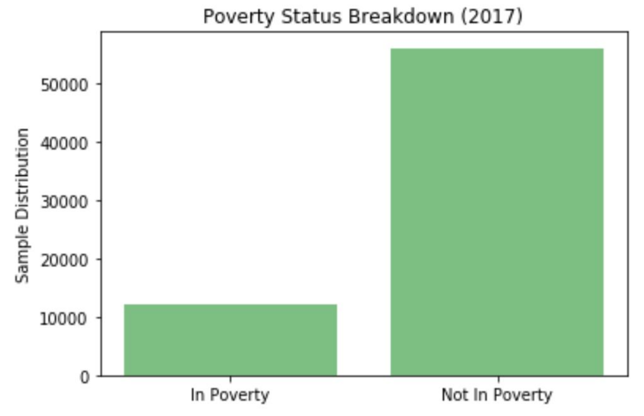
- Citywide poverty rate fell to 19% in 2017 from 20.6% in 2014...but
- 839,705 city students (74% of total student population) qualify for free or reduced-priced lunches, a common poverty marker and the highest percentage in over five years.
- One reason: new classification practices that have improved the ability to identify low-income kids.
- One of a growing class of city-wide datasets (via NYC OpenData)

Starting off: 68,094 total samples *(55,985 vs. 12,109)*

- Calculation: Whether "Income" <> "Poverty Threshold"
  (National index adjusted NYC cost of living/housing)

- Sample scales to entire NYC population
  Multiplied by specific weights (Individual/Household)



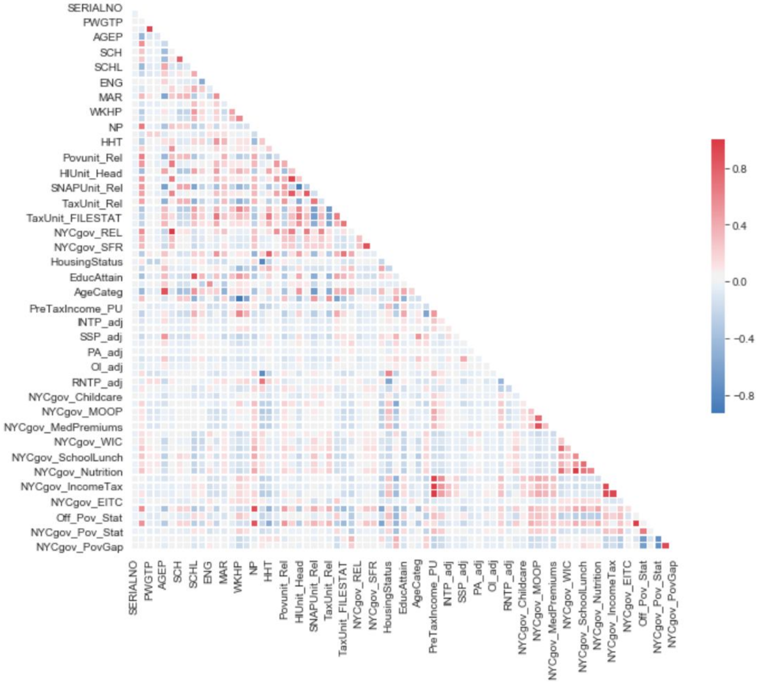


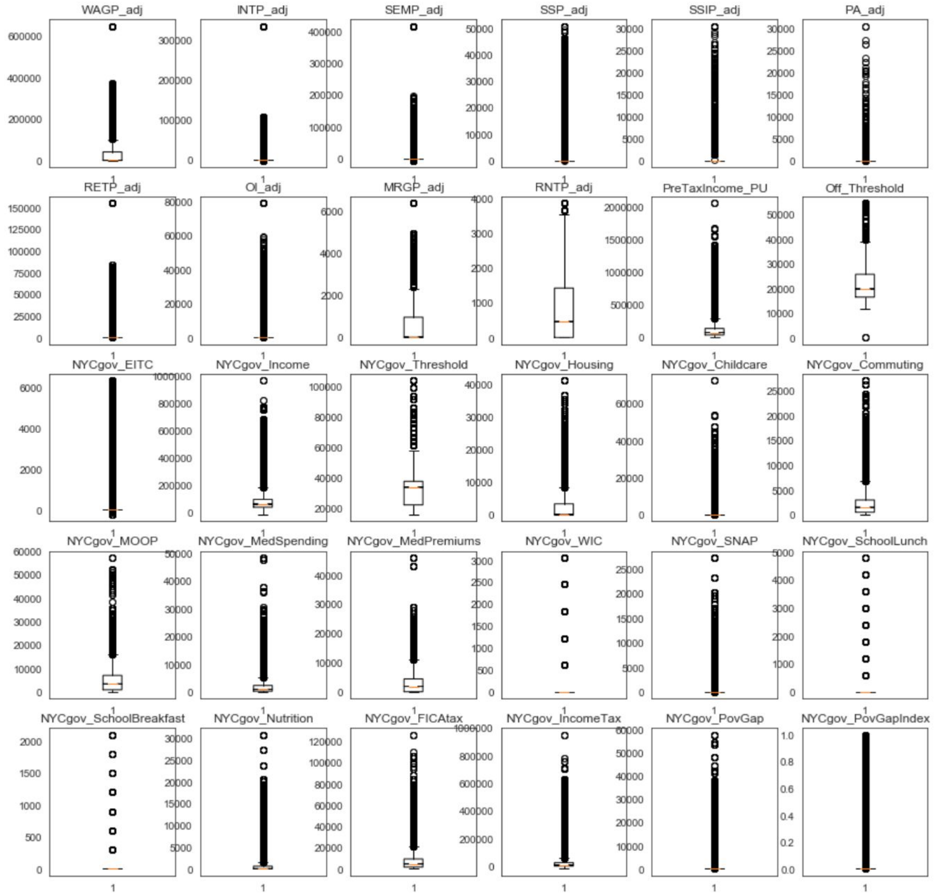
Sources & Tools

- From:
  - NYC Open Data ("NYCgov_PovertyStatus_2017")
  - (to join with) City-wide Demographic, Income data (over time)
- Used:
  - Models:
    - Logistic Classifier
    - KNN
    - Decision Tree
    - Random Forest
    - Adaboost & XGboost
  - Tools: SMOTE, PCA, GridSearchCV

Dirty Data!

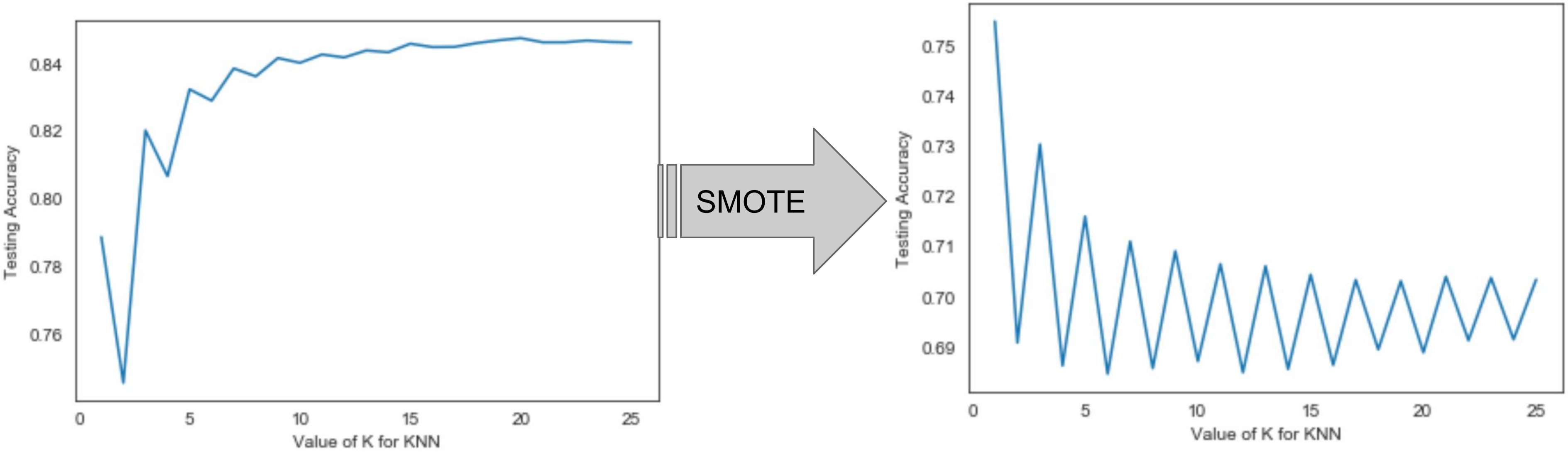
- Holes/Inconsistencies, Outliers, and Multi-indexed variables



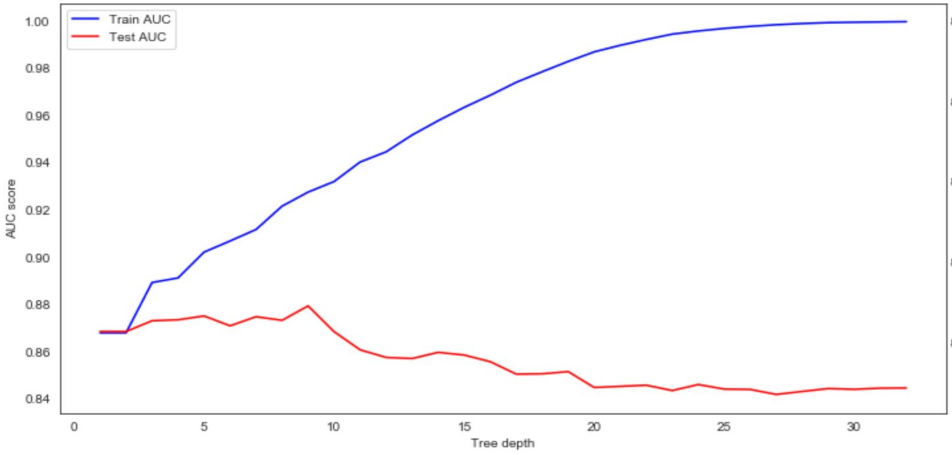


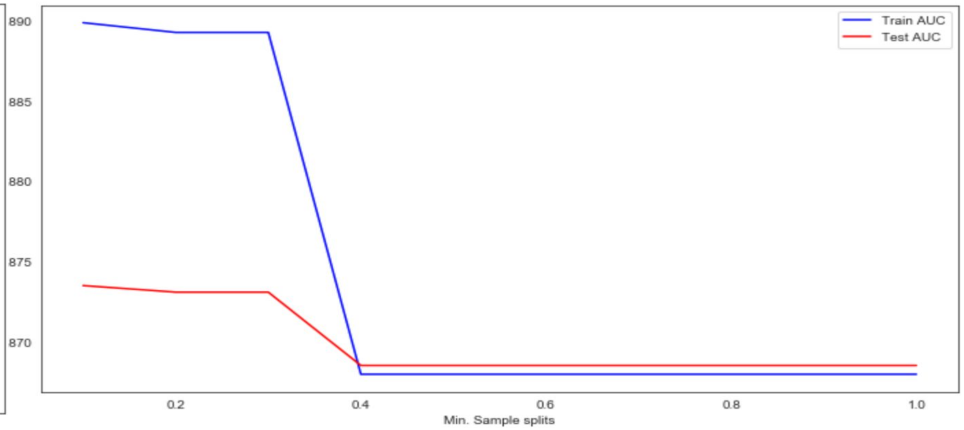
Issues with Logistic Regression, KNN Models

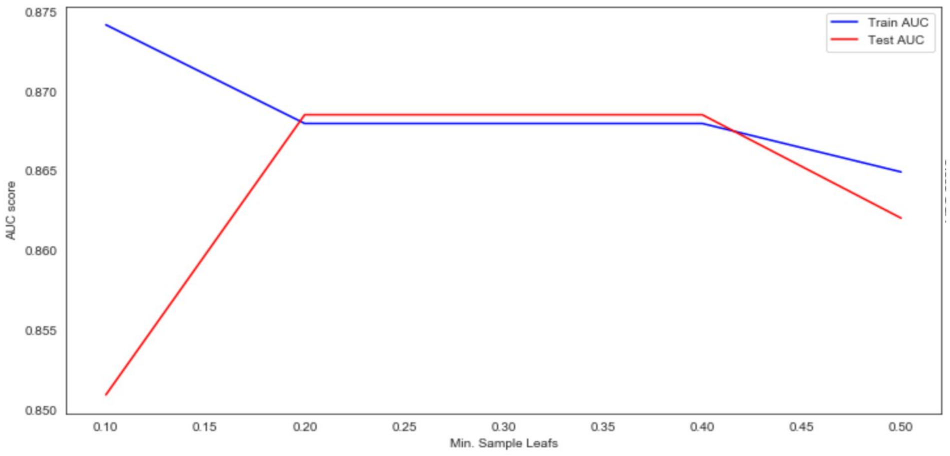
- Feature Selection (Extreme Under- / Over- Fits)
- Confounding & Multicollinearity Concerns
- Class Imbalance (below)
- Need/Opportunities for much more investigating/tuning

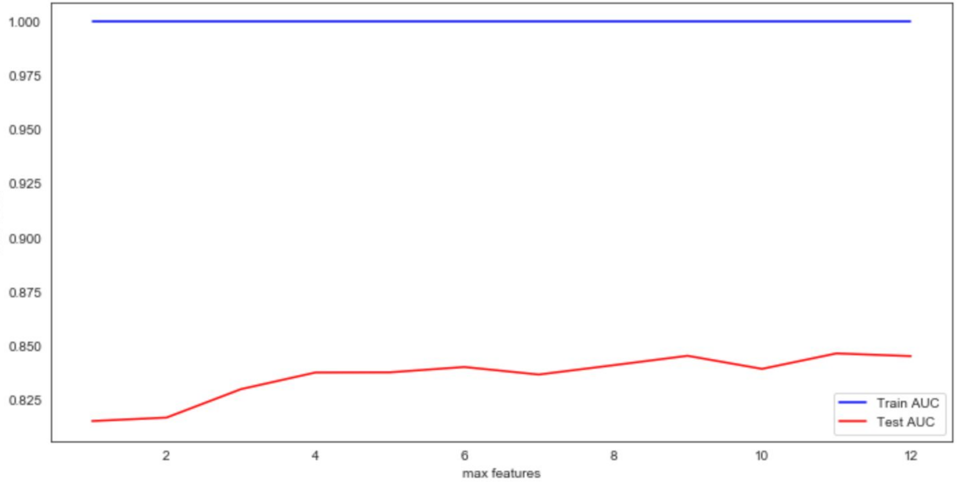


Decision Tree: Hyper-Parameters

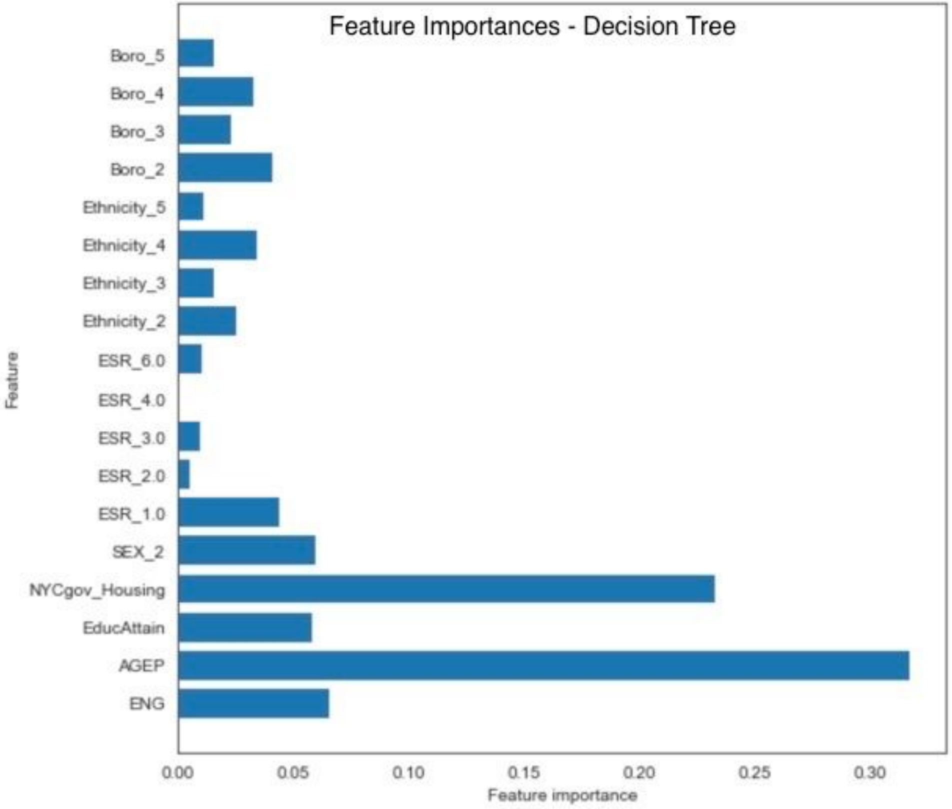


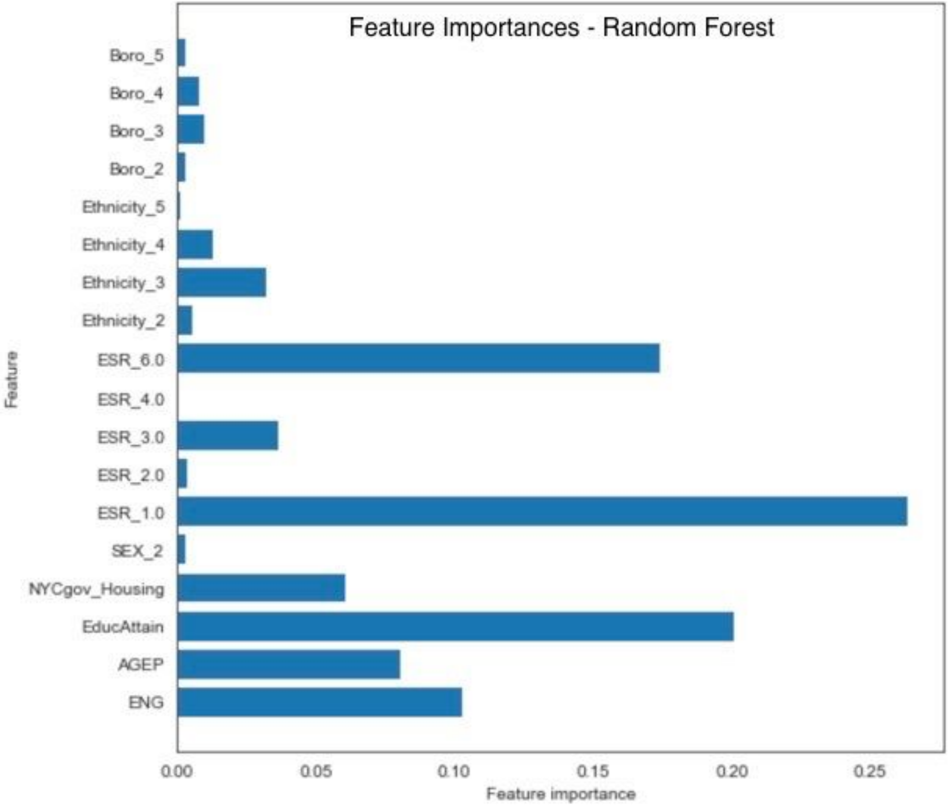






Choosing the Best Features for each Model





Final Metrics

Logistic Regression	Predicted: Negative	Predicted: Positive
Actual Negative	105	2,912
Actual Positive	91	13,916

Decision Tree	Predicted: Negative	Predicted: Positive
Actual Negative	1,873	4,078
Actual Positive	4,429	23,667

ADABOOST	Predicted: Negative	Predicted: Positive
Actual Negative	670	2,288
Actual Positive	494	13,572

KNN	Predicted: Negative	Predicted: Positive
Actual Negative	2,981	36
Actual Positive	19	13,988

Random Forest	Predicted: Negative	Predicted: Positive
Actual Negative	1,377	5,109
Actual Positive	3,398	24,163

XGBOOST	Predicted: Negative	Predicted: Positive
Actual Negative	606	2,352
Actual Positive	365	13,701

- Currently, Inconclusive Results

  Cross Validation metrics were okay, but not much better than Dummy Classifier

  - To be continued:
    - (even) deeper dive into data
    - Many more iterations of tuning hyper-parameters

		Prec	Recall	Acc	F1
Dummy Classifier		0.82	1	0.82	0.90
Logistic Regression	Training	0.82	0.99	0.82	0.90
	Testing	0.82	0.99	0.82	0.90
KNN	Training	0.81	0.95	0.80	0.89
	Testing	0.78	0.97	0.78	0.89
Decision Tree	Training	0.97	0.98	0.95	0.97
	Testing	0.85	0.84	0.75	0.85
Random Forest	Training	0.82	1.00	0.92	0.90
	Testing	0.83	1.00	0.82	0.90
Adaboost	Training	0.85	0.96	0.83	0.90
	Testing	0.86	0.96	0.84	0.91
Xgboost	Training	0.86	0.97	0.84	0.91
	Testing	0.85	0.97	0.84	0.91

Conclusions

- Much, Much Room for Improvement
  - For me: more to come!
    - Maximize Precision or Recall?
      (which is more tolerable- FP or FN?)
  - For NYC’s system:
    - Slim down + Optimize bureaucracy-laden data structure
    - Potential as Keystone / Connector Data
      - Predictive Power, Systemic “Nudges”

Any Questions?