2016 Post-Election Analysis by School District

GENERAL ASSEMBLY - DATA SCIENCE

RYAN COHEN

- Examining the 2016 election results at the county level based on school district diversity across the United States of America.
- Are there school districts within counties where the demographic data suggests that they would vote in a way that was inconsistent with the other school districts within the same county?

DEVELOPING STORY

TOPIC/PROBLEM

DATA SOURCES

DATA SETS USED IN THIS MODEL

1. AT&T Data for Diplomas - Graduation Statistics

- Contains high school cohort and graduation statistics merged with 2010 census data for school districts in the US. (9907, 580)
- Obtained from https://datafordiplomas.devpost.com/details/resources which contains data wrangled by Data for Diplomas and Everyone graduates center at John Hopkins.

2. United States General Election 2016 by County

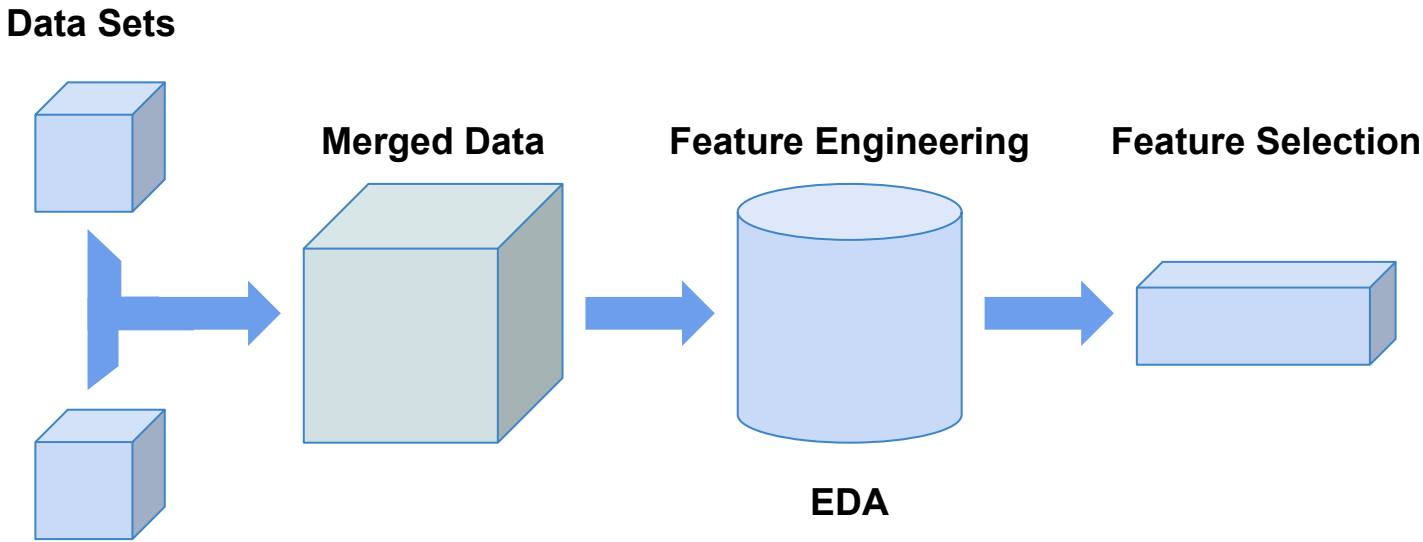
- Contains voting results (integer), by county (FIPS #), and by candidate (each candidate has its own row per county). (15565, 8)
- Obtained from https://data.world/aaronhoffman/us-general-election-2016, which contains scraped data from the NY times website.

Merging the Graduation/Census and Election Data Together:

- Feature Engineering
  1. Common CountyFips column in the graduation/census data set to merge on in the election data set.
  2. Columns that describe the demographic data.
     Ex. percentage of students of a specific race attend a school district vs how many students of that race in the cohort of the entire county.
  3. Region binaries for the following US Regions: Northeast, South, Midwest, Wild West, and Pacifica.
- Feature Selection/Tuning
  1. Dimension reduction via decision tree on all 600+ features
  2. Exploratory data analysis (heatmaps, boxplots, scatterplots)
  3. Hand-picked

EXPLORATORY ANALYSIS

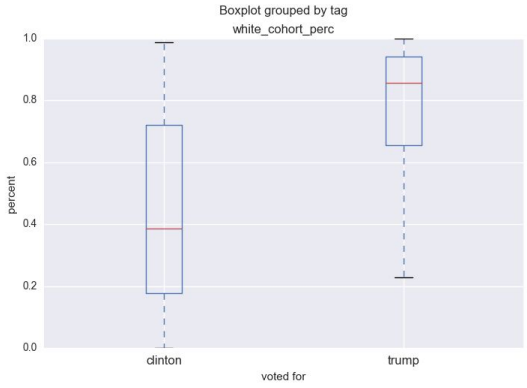
DATA PRE-PROCESSING STEPS



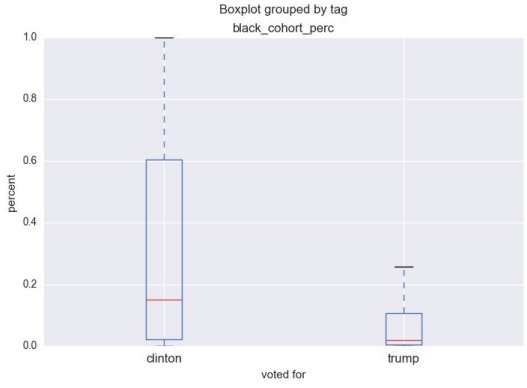
EXPLORATORY ANALYSIS

DATA PRE-PROCESSING STEPS

VISUALIZATIONS

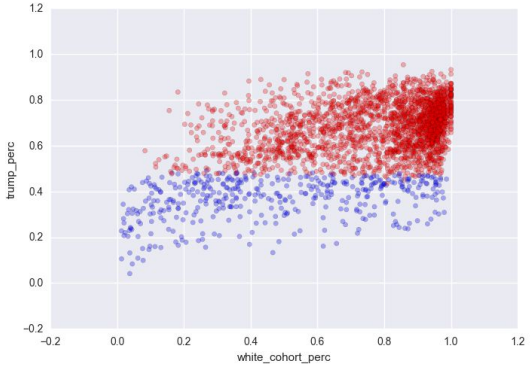


Distribution for each candidate based on how much % of the student cohort for the county is white.

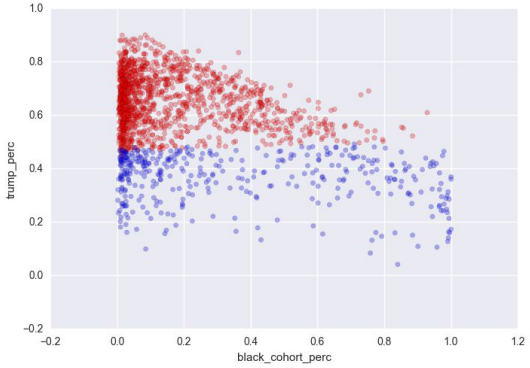


Distribution for each candidate based on how much % of the student cohort for the county is black.

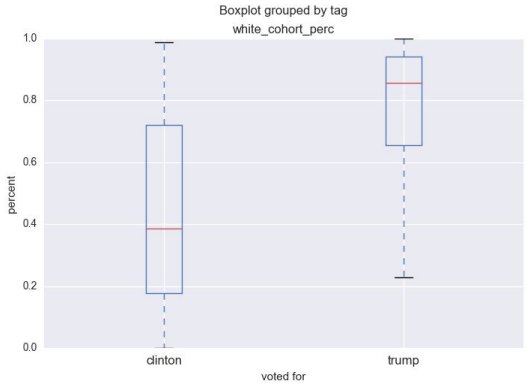
Scatter for the same thing as the boxplot. The higher % of the cohort is white, more Trump.



Scatter for the same thing as the boxplot. The higher % of the cohort is black, less Trump.

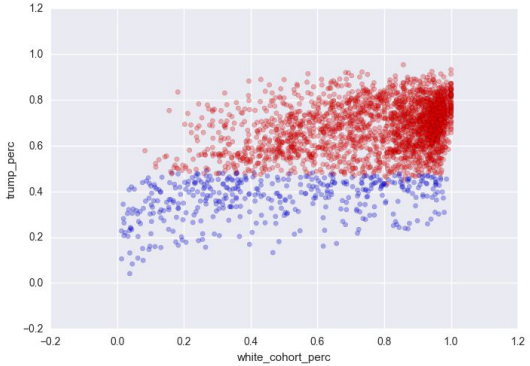


VISUALIZATIONS



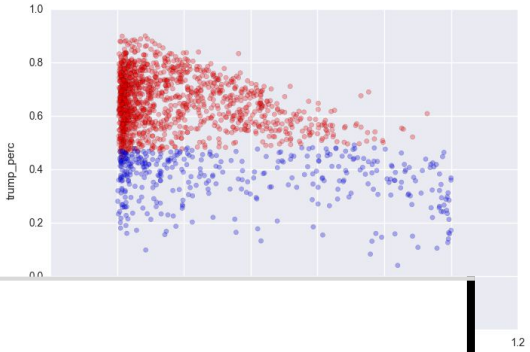
Distribution for each candidate based on how much % of the student cohort for the county is white.

Scatter for the same thing as the boxplot. The higher % of the cohort is white, more Trump.



Distribution for each candidate based on how much % of the student cohort e county ck.

Scatter for the same thing as the boxplot. The higher % of the cohort is

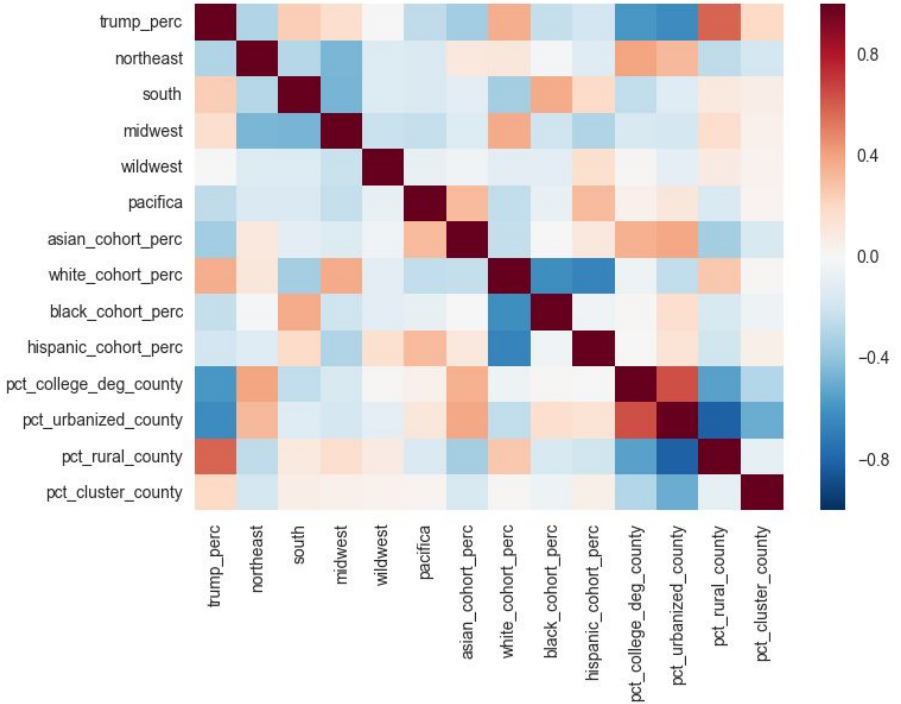


RACE FOR THE WHITE HOUSE

IS IT ALL ABOUT RACE?

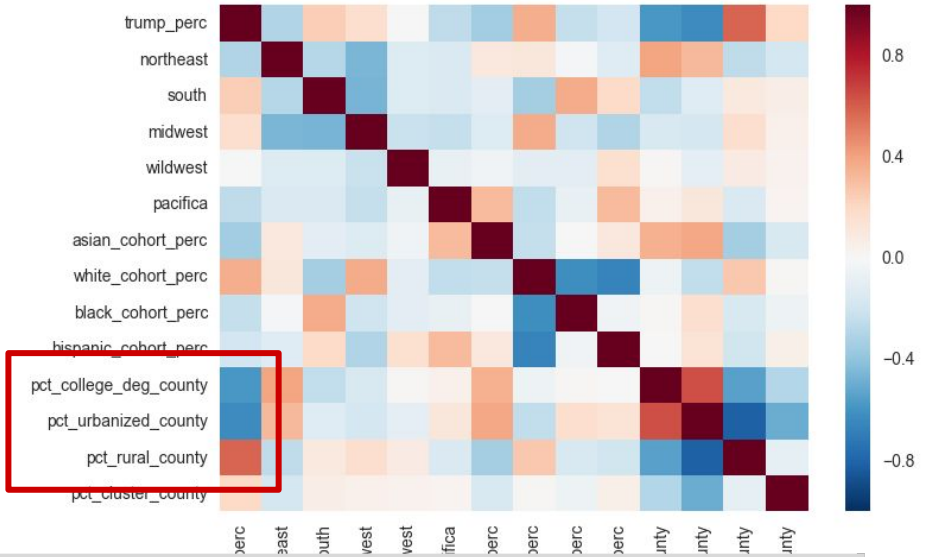
HEATMAP

- Colorful heatmap
- Correlations between features used in the model.
- Pct_college_degree, pct_urbanized_county, and pct_rural_county have strong correlations with the % of votes that the_Donald received.



HEATMAP

- Colorful heatmap
- Correlations between features used in the model.
- Pct_college_degree, pct_urbanized_county, and pct_rural_county have strong correlations with the % of votes ...ald received.



BREAKING NEWS

IS IT ALL ABOUT RACE? (IT'S NOT)

Hypothesis: Based on demographic data, a machine can predict the county results of the presidential election better than the null accuracy and there are school districts with polarizing cohort compositions than that of their peers that will fall into the errors of the prediction model.

Method: producing a machine learning model that will take in features related to the county level demographics and predict whether or not that school district would vote for the_Donald or not.

UP NEXT...

SOON: 2016 ELECTION PREDICTION MODELS

Features (inputs): 'all_cohort_perc_county', 'asian_cohort_perc', 'white_cohort_perc', 'black_cohort_perc', 'pct_college_deg_county', 'pct_rural_county', 'pct_cluster_county', 'northeast', 'south', 'pacifica', 'midwest'

Target (response): Binary classification for county voting majority for Trump (1) or not (0). [‘tag_1’]

Null Accuracy (score to beat): 0.7373953868136355

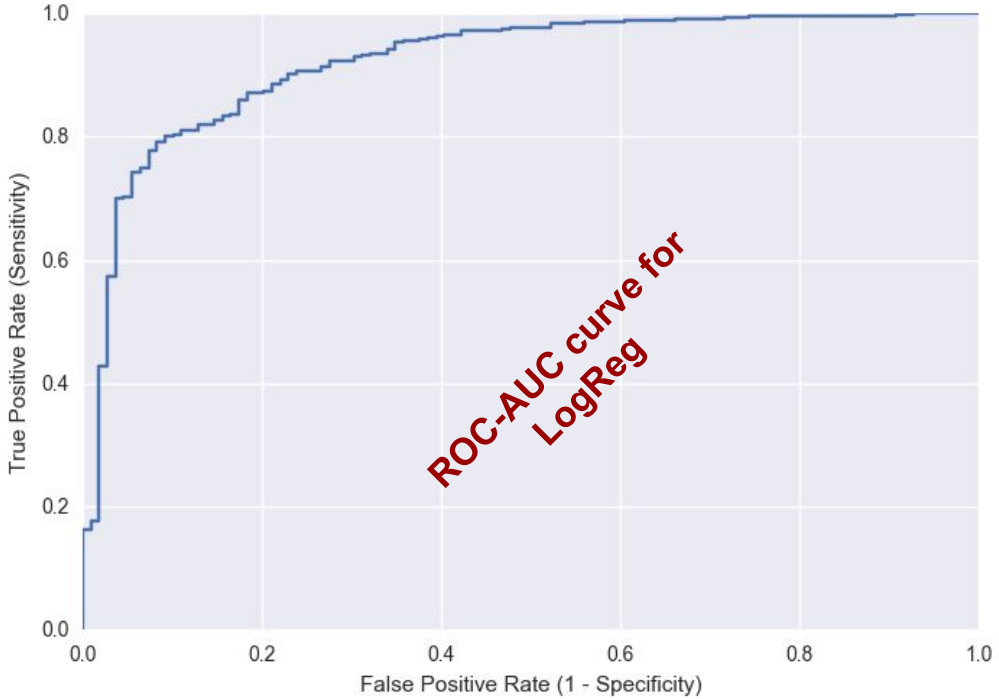
Train: on county level demographics

Predict: on school district level demographics

HOW?

MACHINE LEARNING MODELS

MODEL EVALUATION



GINI index for feature importance

Features	Importance
white_cohort_perc	0.431560
pct_college_deg_county	0.271108
asian_cohort_perc	0.136634
all_cohort_perc_county	0.050772
pct_rural_county	0.030418

MACHINE LEARNING MODELS

Voting Classifier: (cross validated 10x)

- **RandomForestClassifier** - accuracy: 0.843339323478
- **DecisionTreeClassifier** - accuracy: 0.830989728204
- **LogisticRegression** - accuracy: 0.849257013678

Voting classifier prediction accuracy on school district data: **0.878138395591**

Using these predictions, county level accuracy is calculated by:

- Columns grouped by CountyFips #
- Multiply predicton by % of county that the cohort represents
- Add the weighted predictions together per county
- Weighted pred > .5 = 1

	Confusion Matrix	Predicted 0	Predicted 1
Actual	**0**	1855	718
	1	472	6753

COUNTY LEVEL PREDICTION

Prediction accuracy: **0.941472172352**

		Predicted	
	Confusion Matrix	**0**	**1**
Actual	**0**	346	52
	1	111	2276

COUNTY LEVEL PREDICTION

Prediction accuracy: **0.9414721**

Null Accuracy: **0.7373953**

Predicted

Confusion Matrix	0	1
0	346	52

BREAKING NEWS

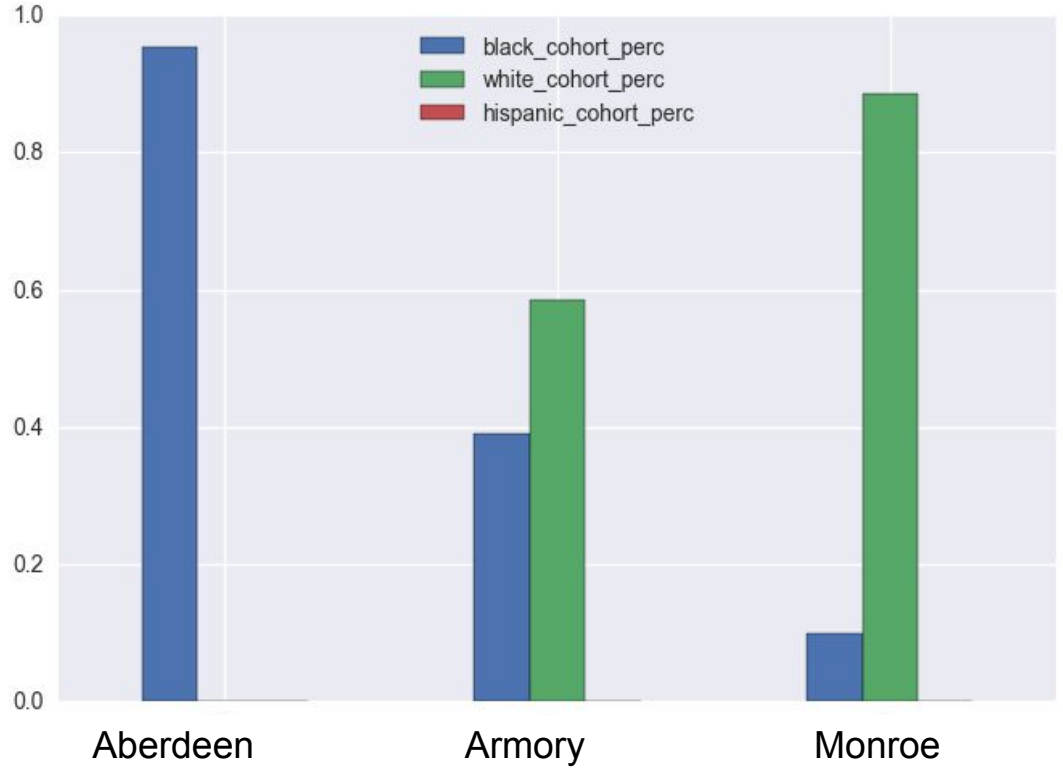
MACHINE BEATS NULL ACCURACY!

- Error Analysis: investigating why the model predicted incorrectly.

CountyFips	State	County Name	School District	Trump Perc	Target	Prediction
28095	MS	Monroe	Aberdeen	0.640454	1	**0**
28095	MS	Monroe	Armory	0.640454	1	1
28095	MS	Monroe	Monroe County	0.640454	1	1

INVESTIGATION

WHERE DID WE GO WRONG?



INVESTIGATION

SUSPICIOUS COHORT DIVERSITY

- There are school districts where their cohort diversity is not representative of the county voting results.
- School Districts/Counties with a large % of hispanics did not vote as a monolith.
- Percentage of adults 25 years or older within the county that hold a college degree was one of the most “important” feature to split on for the decision trees.
- Data was not complete, most likely from the graduation rates data set, which may skew some of the predictions by county.

CONCLUSIONS

WHAT DID I LEARN?

- Gathering the rest of the missing data (mostly school district data) will increase model accuracy.
- Adding 2012 and 2014 electorate data as well to improve model.
- Figuring out more features to tune the model.
- Digging deeper into the incorrect model predictions.
- US map visualizations.

DEVELOPING STORY

WHAT IS NEXT?