Sameer Kurien

Predicting state population growth by region and energy consumption

Predicting migrantion increases by energy consumption and division

- Energy usage can reflect changes in the population.
- **United States Energy, Census, and GDP** 2010-2014, through Kaggle.
- Census (region, division, population, migration) information by state.
- Energy (consumption, production, expenditures, price) by type of energy.
- Hypothesis: The changes in energy use over the years can be used to predict shifts in the population of a state.

Data

RangeIndex: 52 entries, 0 to 51 Columns: 191 entries, StateCodes to RNETMIG2014 dtypes: float64(99), int64(91), object(1) memory usage: 77.7+ KB

	StateCodes	Region	Division	Coast	Great Lakes	TotalC2010	TotalC2011	TotalC2012	TotalC2013	TotalC2014	...	RINTERNATIONALMIG2013
0	AK	4.0	9.0	1.0	0.0	653221	653637	649341	621107	603119	...	3.203618
1	AL	3.0	6.0	1.0	0.0	1931522	1905207	1879716	1919365	1958221	...	1.165832
2	AR	4.0	8.0	0.0	0.0	1120632	1122544	1067642	1096438	1114409	...	2.141877
3	AZ	3.0	7.0	0.0	0.0	1383531	1424944	1395839	1414383	1422590	...	1.090035
4	CA	4.0	9.0	1.0	0.0	7760629	7777115	7564063	7665241	7620082	...	4.207353
5	CO	4.0	8.0	0.0	0.0	1513547	1470445	1440781	1470844	1477177	...	2.074200
6	CT	1.0	1.0	1.0	0.0	764970	739130	725019	754901	750019	...	4.753602
7	DC	3.0	5.0	0.0	0.0	190529	183806	172963	175560	178929	...	5.871584
8	DE	3.0	5.0	1.0	0.0	250212	272568	273728	273716	274013	...	2.608949
9	FL	3.0	5.0	1.0	0.0	4282673	4141711	4029903	4076406	4121680	...	5.783717

10 rows × 191 columns

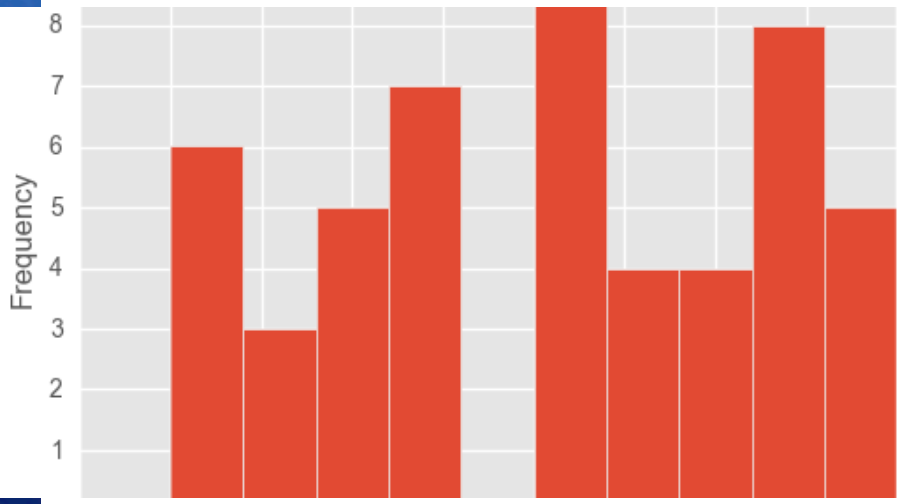
Variables

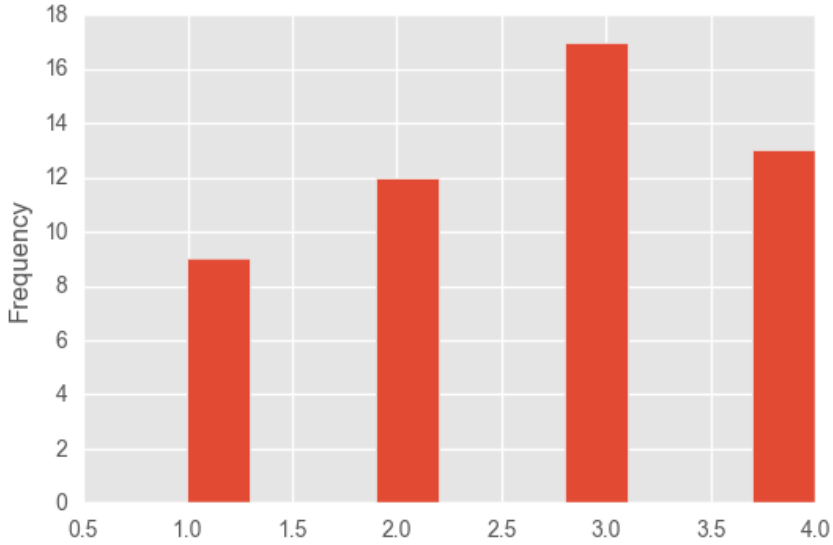
- 'StateCodes', 'Region', 'Division', 'Coast', 'Great Lakes', 'TotalC2010', 'TotalC2011', 'TotalC2012', 'TotalC2013', 'TotalC2014', 'TotalP2010', 'TotalP2011', 'TotalP2012', 'TotalP2013', 'TotalP2014', 'TotalE2010', 'TotalE2011', 'TotalE2012', 'TotalE2013', 'TotalE2014', 'TotalPrice2010', 'TotalPrice2011', 'TotalPrice2012', 'TotalPrice2013', 'TotalPrice2014', 'TotalC10-11', 'TotalC11-12', 'TotalC12-13', 'TotalC13-14', 'TotalP10-11', 'TotalP11-12', 'TotalP12-13', 'TotalP13-14', 'TotalE10-11', 'TotalE11-12', 'TotalE12-13', 'TotalE13-14', 'TotalPrice10-11', 'TotalPrice11-12', 'TotalPrice12-13', 'TotalPrice13-14', 'BiomassC2010', 'BiomassC2011', 'BiomassC2012', 'BiomassC2013', 'BiomassC2014', 'CoalC2010', 'CoalC2011', 'CoalC2012', 'CoalC2013', 'CoalC2014', 'CoalP2010', 'CoalP2011', 'CoalP2012', 'CoalP2013', 'CoalP2014', 'CoalE2010', 'CoalE2011', 'CoalE2012', 'CoalE2013', 'CoalE2014', 'CoalPrice2010', 'CoalPrice2011', 'CoalPrice2012', 'CoalPrice2013', 'CoalPrice2014', 'ElecC2010', 'ElecC2011', 'ElecC2012', 'ElecC2013', 'ElecC2014', 'ElecE2010', 'ElecE2011', 'ElecE2012', 'ElecE2013', 'ElecE2014', 'ElecPrice2010', 'ElecPrice2011', 'ElecPrice2012', 'ElecPrice2013', 'ElecPrice2014', 'FossFuelC2010', 'FossFuelC2011', 'FossFuelC2012', 'FossFuelC2013', 'FossFuelC2014', 'GeoC2010', 'GeoC2011', 'GeoC2012', 'GeoC2013', 'GeoC2014', 'GeoP2010', 'GeoP2011', 'GeoP2012', 'GeoP2013', 'GeoP2014', 'HydroC2010', 'HydroC2011', 'HydroC2012', 'HydroC2013', 'HydroC2014', 'HydroP2010', 'HydroP2011', 'HydroP2012', 'HydroP2013', 'HydroP2014', 'NatGasC2010', 'NatGasC2011', 'NatGasC2012', 'NatGasC2013', 'NatGasC2014', 'NatGasE2010', 'NatGasE2011', 'NatGasE2012', 'NatGasE2013', 'NatGasE2014', 'NatGasPrice2010', 'NatGasPrice2011', 'NatGasPrice2012', 'NatGasPrice2013', 'NatGasPrice2014', 'LPGC2010', 'LPGC2011', 'LPGC2012', 'LPGC2013', 'LPGC2014', 'LPGE2010', 'LPGE2011', 'LPGE2012', 'LPGE2013', 'LPGE2014', 'LPGPrice2010', 'LPGPrice2011', 'LPGPrice2012', 'LPGPrice2013', 'LPGPrice2014', 'GDP2010Q1', 'GDP2010Q2', 'GDP2010Q3', 'GDP2010Q4', 'GDP2010', 'GDP2011Q1', 'GDP2011Q2', 'GDP2011Q3', 'GDP2011Q4', 'GDP2011', 'GDP2012Q1', 'GDP2012Q2', 'GDP2012Q3', 'GDP2012Q4', 'GDP2012', 'GDP2013Q1', 'GDP2013Q2', 'GDP2013Q3', 'GDP2013Q4', 'GDP2013', 'GDP2014Q1', 'GDP2014Q2', 'GDP2014Q3', 'GDP2014Q4', 'GDP2014', 'CENSUS2010POP', 'POPESTIMATE2010', 'POPESTIMATE2011', 'POPESTIMATE2012', 'POPESTIMATE2013', 'POPESTIMATE2014', 'RBIRTH2011', 'RBIRTH2012', 'RBIRTH2013', 'RBIRTH2014', 'RDEATH2011', 'RDEATH2012', 'RDEATH2013', 'RDEATH2014', 'RNATURALINC2011', 'RNATURALINC2012', 'RNATURALINC2013', 'RNATURALINC2014', 'RINTERNATIONALMIG2011', 'RINTERNATIONALMIG2012', 'RINTERNATIONALMIG2013', 'RINTERNATIONALMIG2014', 'RDOMESTICMIG2011', 'RDOMESTICMIG2012', 'RDOMESTICMIG2013', 'RDOMESTICMIG2014', 'RNETMIG2011', 'RNETMIG2012', 'RNETMIG2013', 'RNETMIG2014'

egion	Division	Coast	Great Lakes	TotalC2010	TotalC2011	TotalC2012	TotalC2013	TotalC2014	...	RINTERNA
.0	9.0	1.0	0.0	653221	653637	649341	621107	603119	...	3.203618
.0	6.0	1.0	0.0	1931522	1905207	1879716	1919365	1958221	...	1.165832
.0	8.0	0.0	0.0	1120632	1122544	1067642	1096438	1114409	...	2.141877
.0	7.0	0.0	0.0	1383531	1424944	1395839	1414383	1422590	...	1.090035
.0	9.0	1.0	0.0	7760629	7777115	7564063	7665241	7620082	...	4.207353
.0	8.0	0.0	0.0	1513547	1470445	1440781	1470844	1477177	...	2.074200
.0	1.0	1.0	0.0	764970	739130	725019	754901	750019	...	4.753602
.0	5.0	0.0	0.0	190529	183806	172963	175560	178929	...	5.871584
.0	5.0	1.0	0.0	250212	272568	273728	273716	274013	...	2.608949
.0	5.0	1.0	0.0	4282673	4141711	4029903	4076406	4121680	...	5.783717

mns

Division distribution

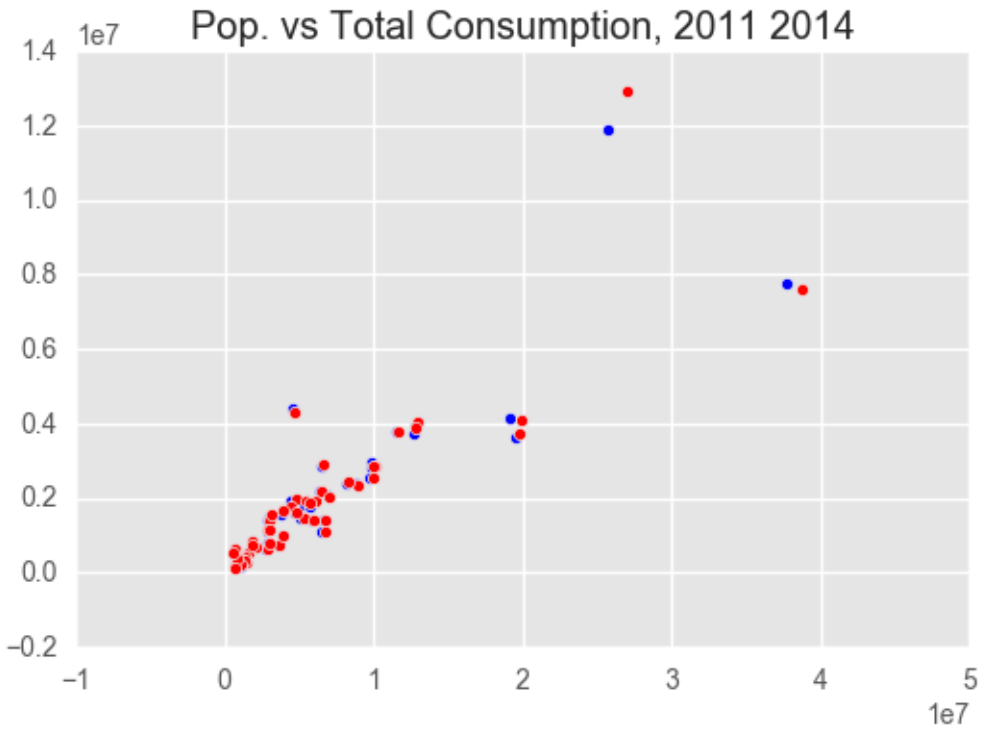




Region Distribution

- Region: The number corresponding to the region the state lies within, according to the 2010 census. (1 = Northeast, 2 = Midwest, 3 = South, 4 = West)
- Division: The number corresponding to the division the state lies within, according to the 2010 census. (1 = New England, 2 = Middle Atlantic, 3 = East North Central, 4 = West North Central, 5 = South Atlantic, 6 = East South Central, 7 = West South Central, 8 = Mountain, 9 = Pacific)

Exploratory Plotting



Plotting the various energy and population features seemed to indicate no significant change over the time period covered by the dataset.

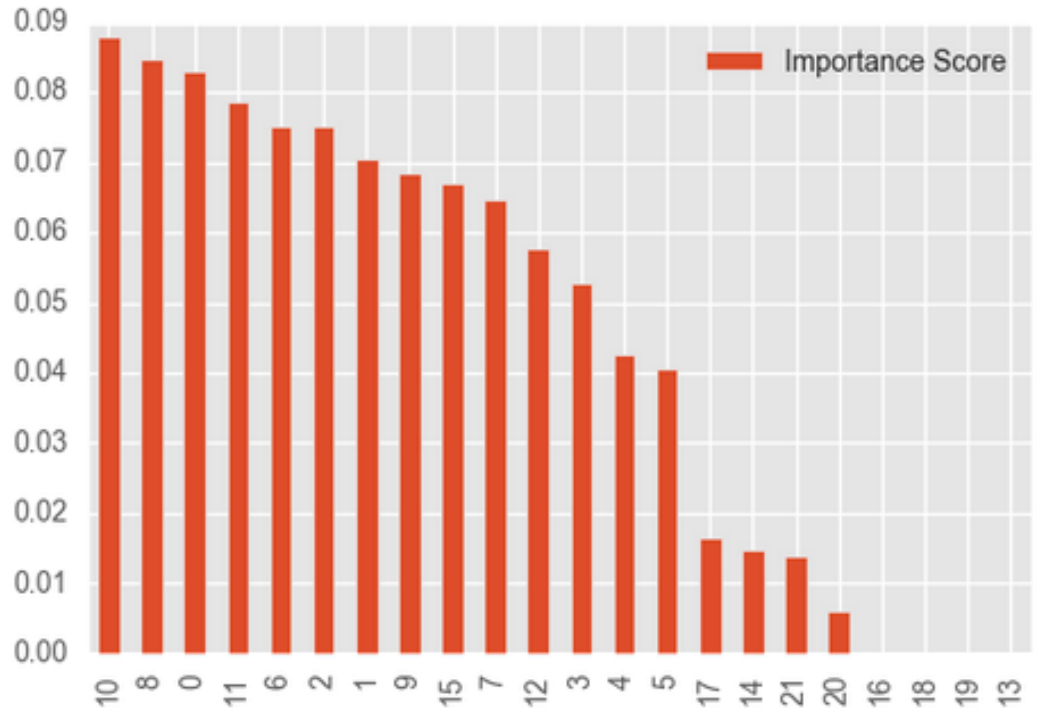
Random Forest

-Decided to determine importance of the remaining features through a random forest
-Divisions were one-hot encoded
-Rate of Net Migration 2014 was transformed with boolean PopulationIncreased2014
-Forest was then cross validated up to 100 trees

```
feature_cols = ['TotalC2014', 'BiomassC2014', 'CoalC2014',
       'ElecC2014', 'FossFuelC2014', 'GeoC2014', 'HydroC2014',
       'NatGasC2014', 'LPGC2014', 'POPESTIMATE2014', 'RBIRTH2014',
       'RDEATH2014', 'RNATURALINC2014',1.0, 2.0, 3.0, 4.0, 5.0, 6.0,
       7.0, 8.0, 9.0,]
X = df[feature_cols]
y = df['PopulationIncreased2014']
```

Random Forest Results

	Features	Importance Score
10	RBIRTH2014	0.087906
8	LPGC2014	0.084668
0	TotalC2014	0.082936
11	RDEATH2014	0.078689
6	HydroC2014	0.075130
2	CoalC2014	0.075093
1	BiomassC2014	0.070543
9	POPESTIMATE2014	0.068483
15	3	0.066869
7	NatGasC2014	0.064709
12	RNATURALINC2014	0.057672
3	ElecC2014	0.052814
4	FossFuelC2014	0.042683
5	GeoC2014	0.040576
17	5	0.016351
14	2	0.014876
21	9	0.013921
20	8	0.006079
16	4	0.000000
18	6	0.000000
19	7	0.000000
13	1	0.000000



Conclusions

- The random forest, which hit 70.7% accuracy with 51 trees,
- the Rate of Birth, Oil Consumption and Total Energy Consumption scores were the most powerful predictors of increased migration
- East North Central (3) division (15 on the table) a stronger influence compared to the rest of the divisions, which all fell below the population and energy features.

Next Steps

- First steps would be increasing the workable data, possibly by breaking down each state into multiple rows for each year
- Check how GDP factors influence the migration rate
- Initial plan was regression based, more complex, fell apart after several mistakes were corrected.