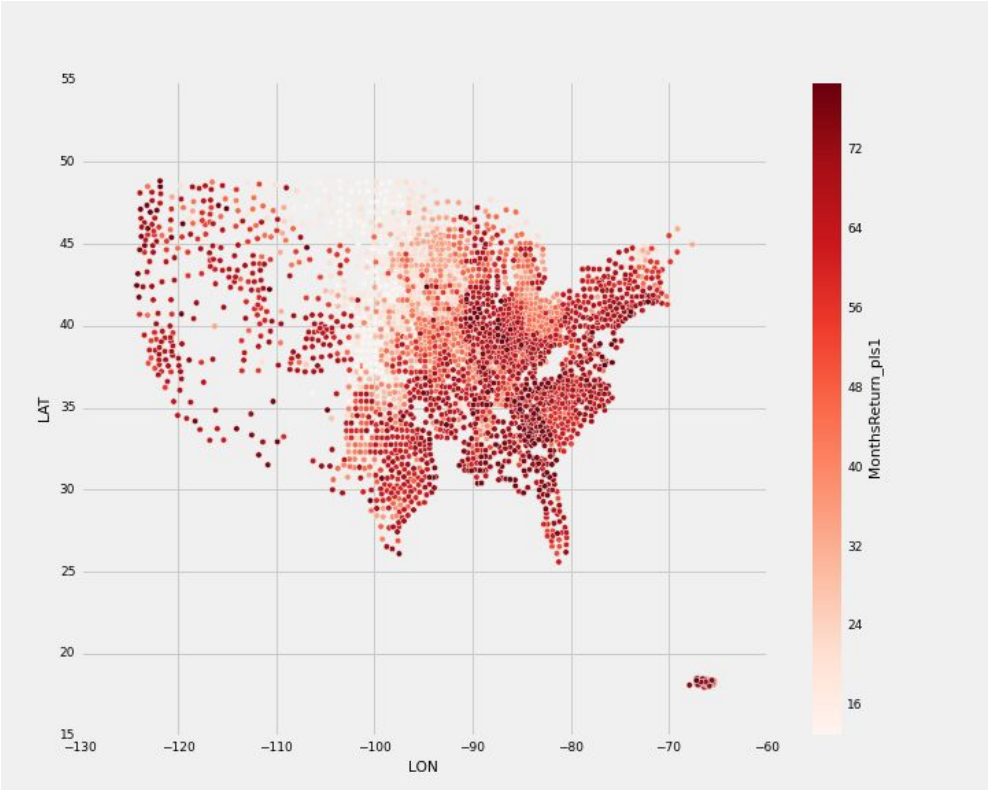
Predicting Local Experience after the Great Recession

Kyle Alden -- DAT8

The problem

Localities experienced the Great Recession in vastly different ways. Some economies quickly returned to their pre-recession health, while others have yet to fully regain their footing 6 years after the official end of the recession.



The question

How many months will it take for any US county to return to within 1 percent of its pre-recession unemployment rate?

and b) Can we predict whether a country has returned by July 2015 or not?

Response variable

Step 1

Step 2

Step 3

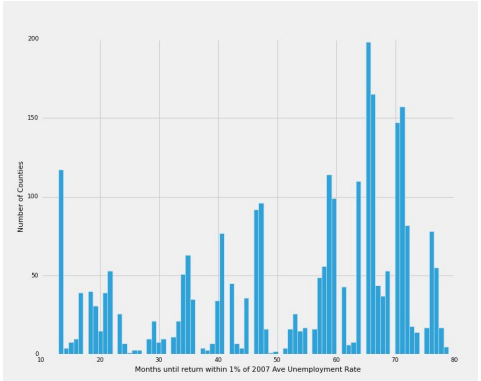
The Source

My data comes from the Bureau of Labor Statistics and includes the estimated unemployment rate for every county or county equivalent between 2006 and July 2015.

Calculate my response

I found the first month that the unemployment rate dropped below the average rate in 2007 plus one percent for each county. I used a 3 month moving average to help ignore anomalous months.

Explore



Seasonal!

Feature Variables

Step 1 → Step 2 → Step 3

Types of Sources

- Demographic (Census)
- Education (Census)
- Industrial Typology (USDA)
- Economic (HH income, stimulus received)
- More to come!

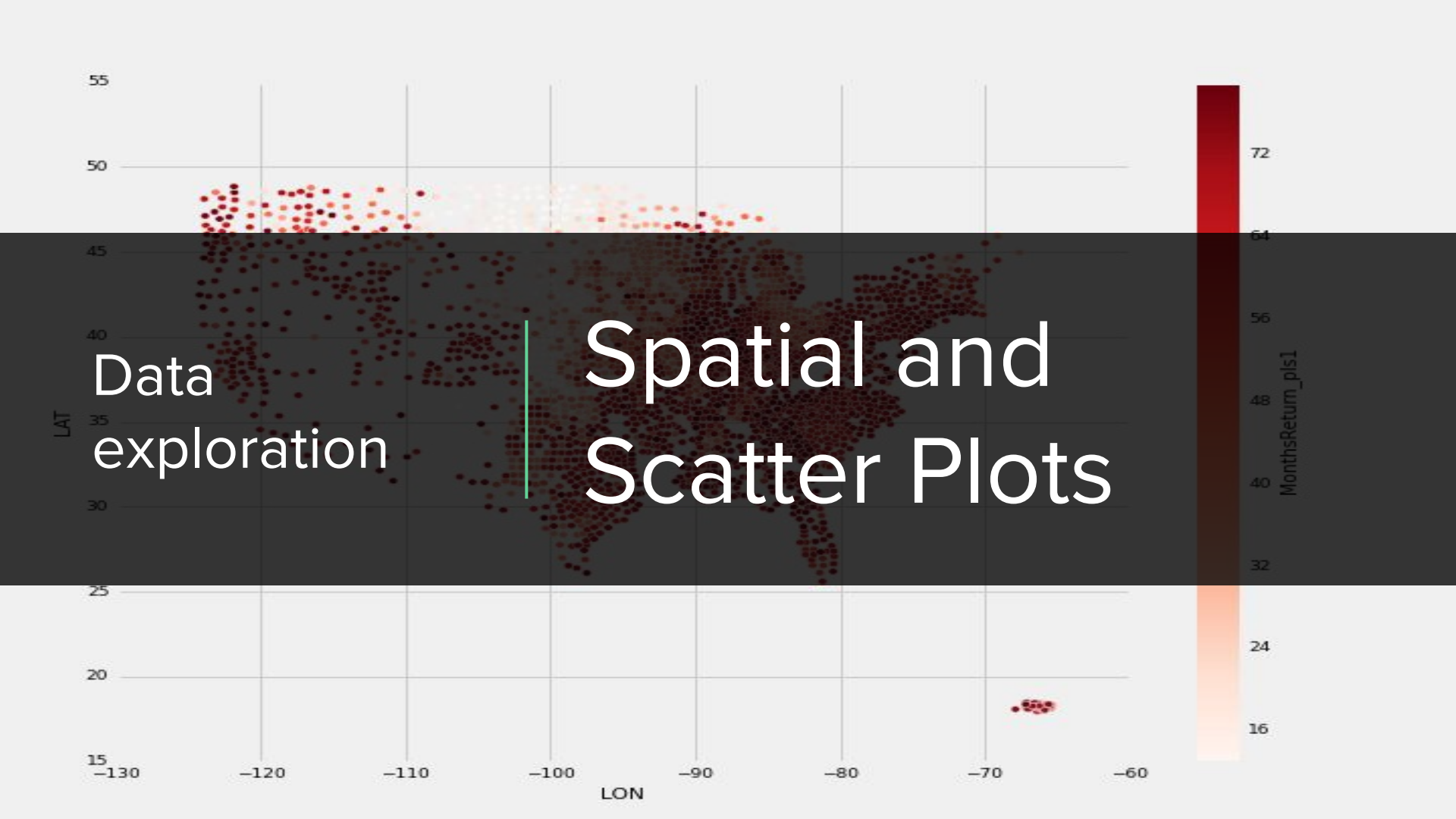
Feature Management

Merges were used (as well as some old school spreadsheet work) to join my feature data to the response dataset. Each county has a unique FIPS code or I used concatenated county, state names.

Feature Issues + Engineering

In several cases data was not available for every county.

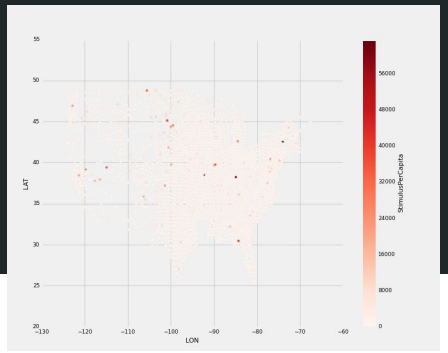
I am working to develop additional features including one that represents the biggest increase for each county.



Data exploration

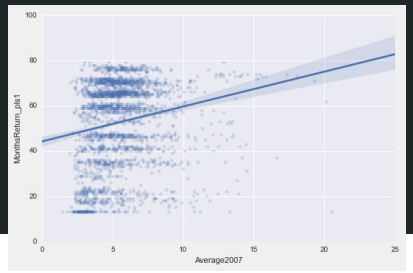
Spatial and Scatter Plots

Explore Trends



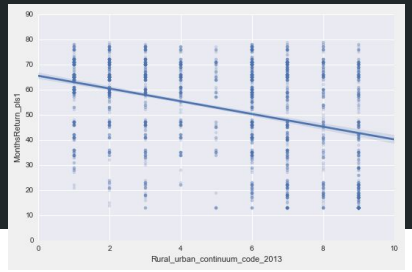
Stimulus Per Capita

There are significant outliers in our data: Albany, NY received over $62,000 per resident, while most counties got less than $1,000



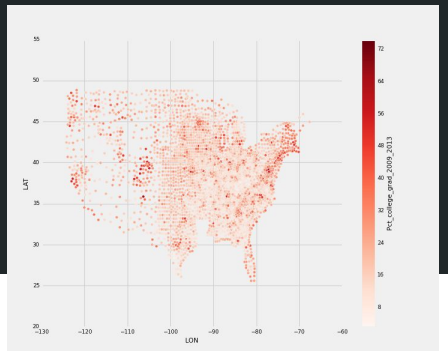
2007 Rate

Counties with higher rates in 2007 have taken longer to return than ones with lower rates.



Urban-Rural Continuum

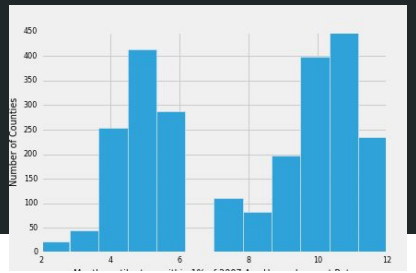
Rural counties seem to have (surprisingly) taken less time to return than urban counties, I doubt this relationship is linear though.



% College Grad

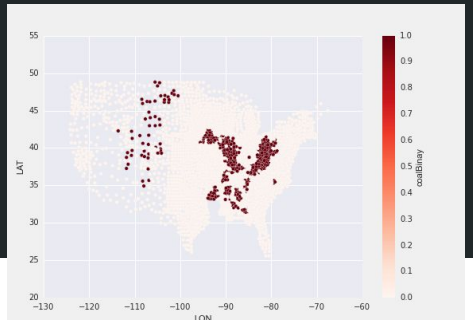
There doesn't seem to be a simple relationship between counties with many college grads and their comeback time.

Explore Trends Part 2



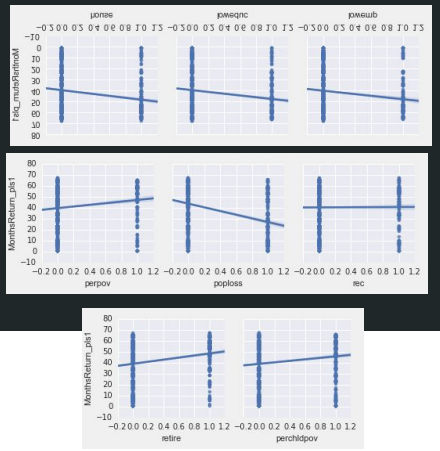
Month Returned

There is a clear monthly trend with two yearly peaks -- especially when I removed the first month in the dataset.



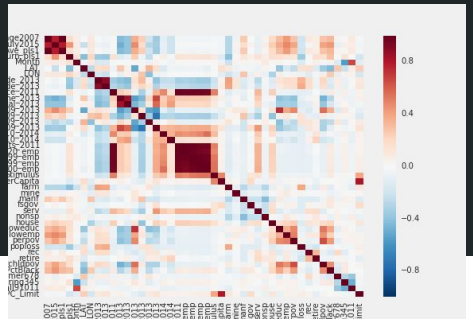
Coal Counties

The price of coal has dropped as regulation has increased over the previous several years, probably indicating a slow return to pre-recession employment



USDA Policy Codes

Some of the many USDA policy code types (persistent poverty, population loss, etc) were useful in determining when a county would return.



Correlation Matrix

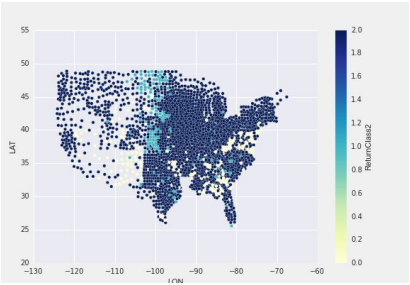
My most useful graphic was the above correlation matrix.

Classification Question:

Has a county returned to within 1% of pre-recession unemployment or not?

Overall Params

Used 3 feature sets:
1_ Domain Knowledge
2_ Limited 1
3_ All
Null accuracy = **.87878**



KNN

Accuracy = .899

True Positives: 648
True Negatives: 35
False Positives: 57
False Negatives: 1

Logistic Regression

Accuracy = .872

True Positives: 661
True Negatives: 1
False Positives: 91
False Negatives: 6

Decision Trees

Accuracy = .888

True Positives: 659
True Negatives: 15
False Positives: 77
False Negatives: 8

Naive Bayes

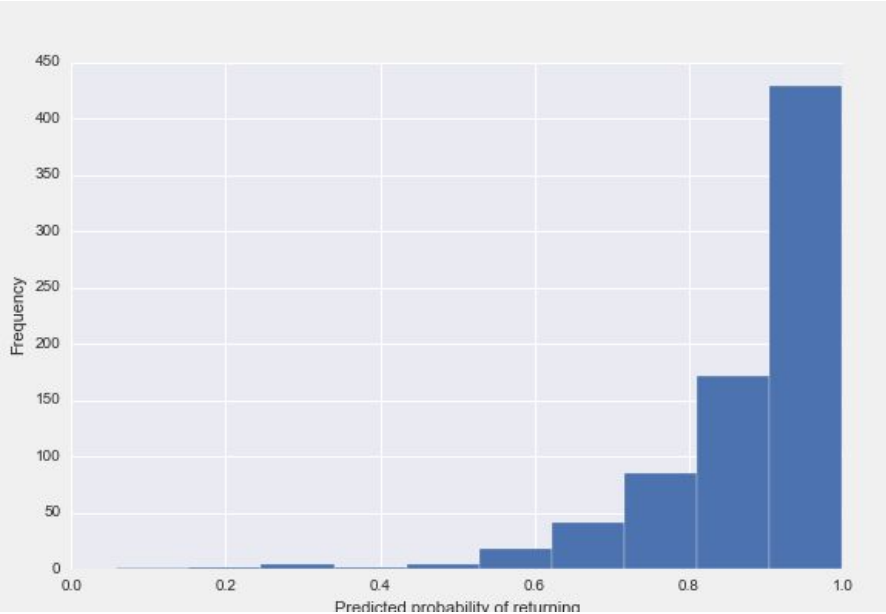
Accuracy = .86

True Positives: 653
True Negatives: 13
False Positives: 79
False Negatives: 14
--Limited list

Classification Problem Problems

Nearly all of my classification approaches had a hard time predicting negatives (counties that didn't return to 0). I was able to set thresholds, which helped improve how skewed my model was, but did not improve their accuracy much better than null.

Logistic Regression: Predicted probability of returning to within 1% of pre-recession unemp rate



Regression Question:

How long will it take for a county to return to within 1% of pre-recession unemployment?

Overall Params

2631 counties have returned to within 1%.

Null RMSE = **18.773**

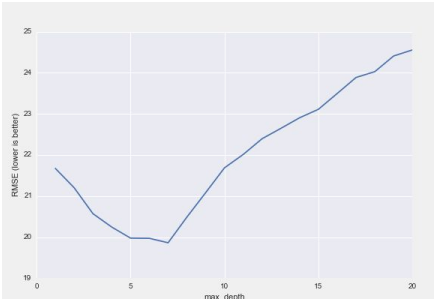
RMSE of just Average of 2007 = 18.31

Linear Regression

I tried several regression models with different sets of features. The one that performed best was the one where I included all features.
Best RMSE = 14.27

Regression Trees

Again, all features performed best.

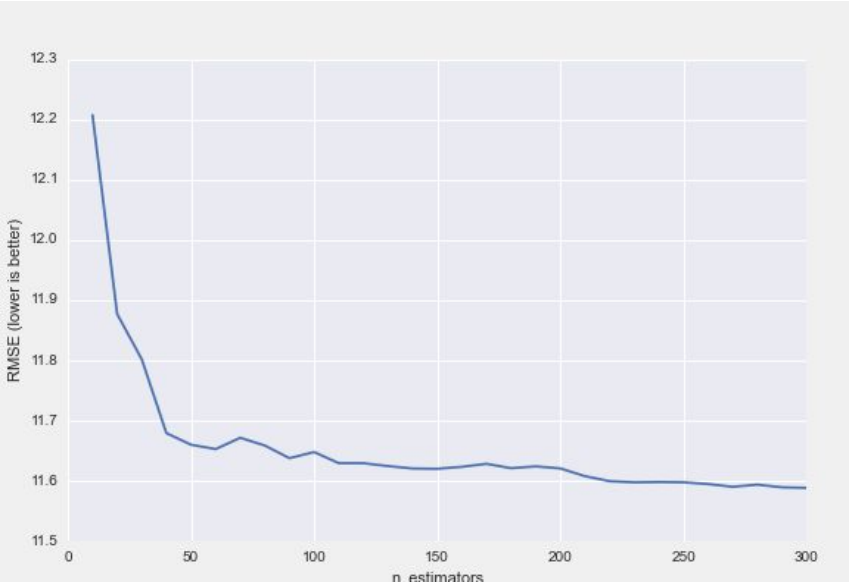


Best RMSE = 13.6

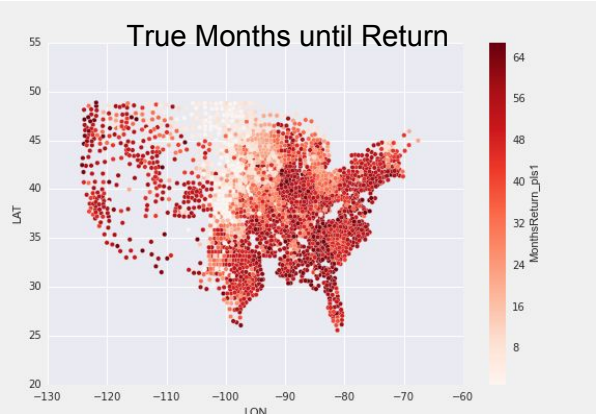
Random Forests Regressor

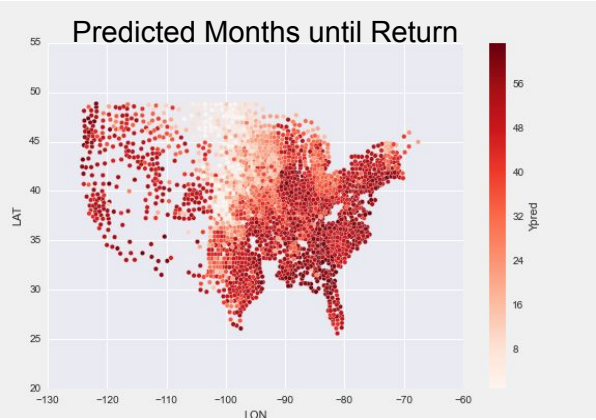
My best model of the regression methods was the random forests regressor where the best RMSE was 11.6.
I used 5-fold cross validation to calculate the best number of estimators and I then used 5-fold cv to find the best number of max features.

Calculating Best Number of Estimators

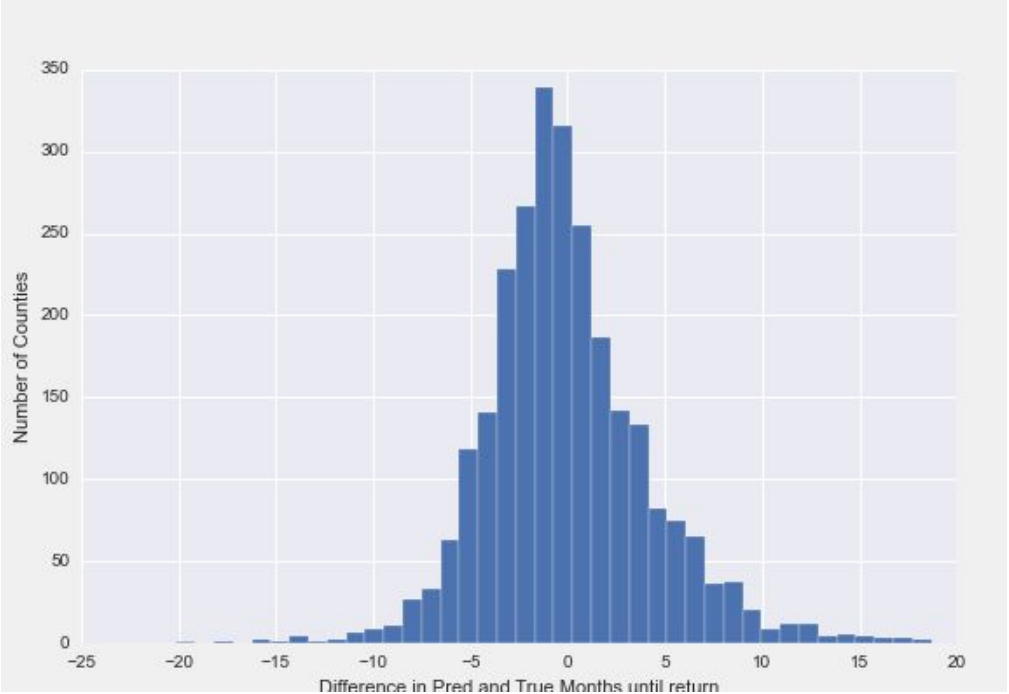


Results of Best Random Forests Model





23% correct within 1 Month
57% correct within 3 Months
88% correct within 8 Months



What next?

Residuals don’t seem to be very spatially clustered, which is a sign that there may not be many additional spatial variables that are out there that could explain the outstanding differences.

We could always find more features.

I would love to have figured out my classification model better.