Predicting Treasury Yields

Using linear regression and decision trees to analyze US Treasury Yields

Background information

- US treasuries = “I owe you”
- Gov’t needs funding -> auctions -> raise money
- 14 regularly issued types of treasuries:
  - **Bills: 4/13/26/52 weeks**
  - **Nominal Coupons: 2/3/5/7/10/30 years**
  - Treasury Inflation Protected Securities: 5/10/30 years
  - Floating Rate Notes: 2 year

What is yield?

- Return on investment in US gov't debt
- Expressed as %
- Interest rate the US pays to borrow money
- Investor's outlook on the economy
- Example:
  - High yield
  - Higher borrowing cost for the gov't
  - Better return for investors
  - Better economic outlook

Questions

- What factors do investors look at when determining yields?
- Did the debt ceiling crisis spook investors?

Data

- Federal Reserve, White House, and Treasury websites
  - Treasury yields (bills and nominal coupons)
  - Unemployment rate
  - CPI
  - Debt limit
  - Deficit
  - GDP
  - Amount of debt outstanding
- Spreadsheets

Data Pre-Processing

- Convert Excel files to csv files
- Header/data fix
- Debt limit/CPI/GDP files required special fixes
- *Import* function

```
# file import function
def file_import(filename, truncateBefore = '1989-12-31'):
    output = pd.read_csv(filename, index_col = 0, parse_dates = True)
    output.columns = [filename[:-4]]
    output.index.names = ['Date']
    output.index = [date+pd.tseries.offsets.MonthBegin(n=-1) if date.day!=1 else date for date in output.index ]
    output = output.truncate(truncateBefore)
    return output
```

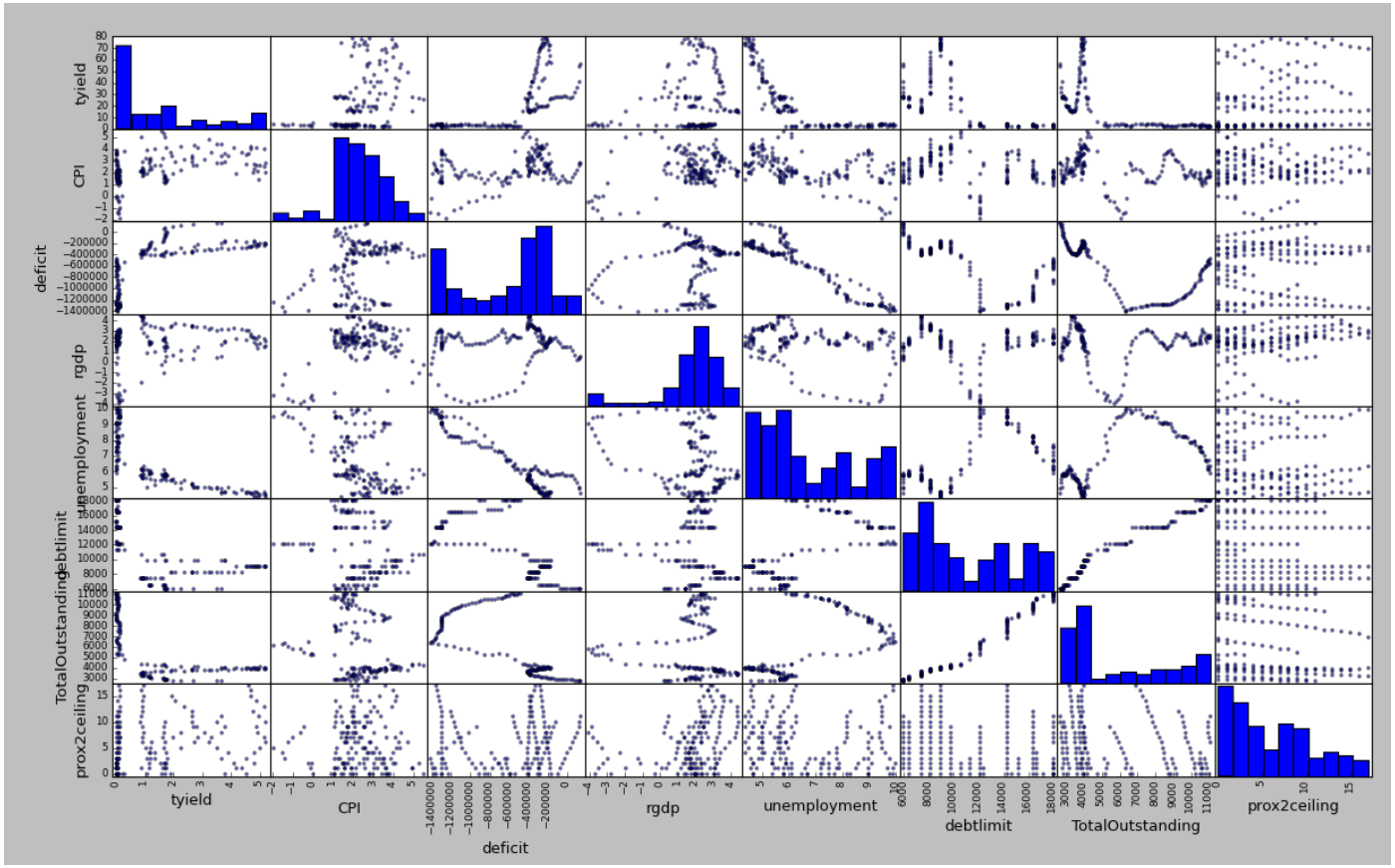
Data Pre-Processing Cont.

- Form a dataframe: date index = monthly
- Missing values: backfill, pad, interpolate

```
>>> raw_data
                CPI  deficit      rgdp  unemployment  debtlimit  TotalOutstanding  prox2ceiling
1990-01-01      NaN      NaN       NaN           5.4     3122.7               NaN        3122.7
1990-02-01      NaN      NaN       NaN           5.3        NaN               NaN           NaN
1990-03-01      NaN      NaN       NaN           5.2        NaN               NaN           NaN
1990-04-01      NaN      NaN       NaN           5.4        NaN               NaN           NaN
1990-05-01      NaN      NaN       NaN           5.4        NaN               NaN           NaN
1990-06-01      NaN      NaN       NaN           5.2        NaN               NaN           NaN
1990-07-01      NaN      NaN  1.725641           5.5        NaN               NaN           NaN
```

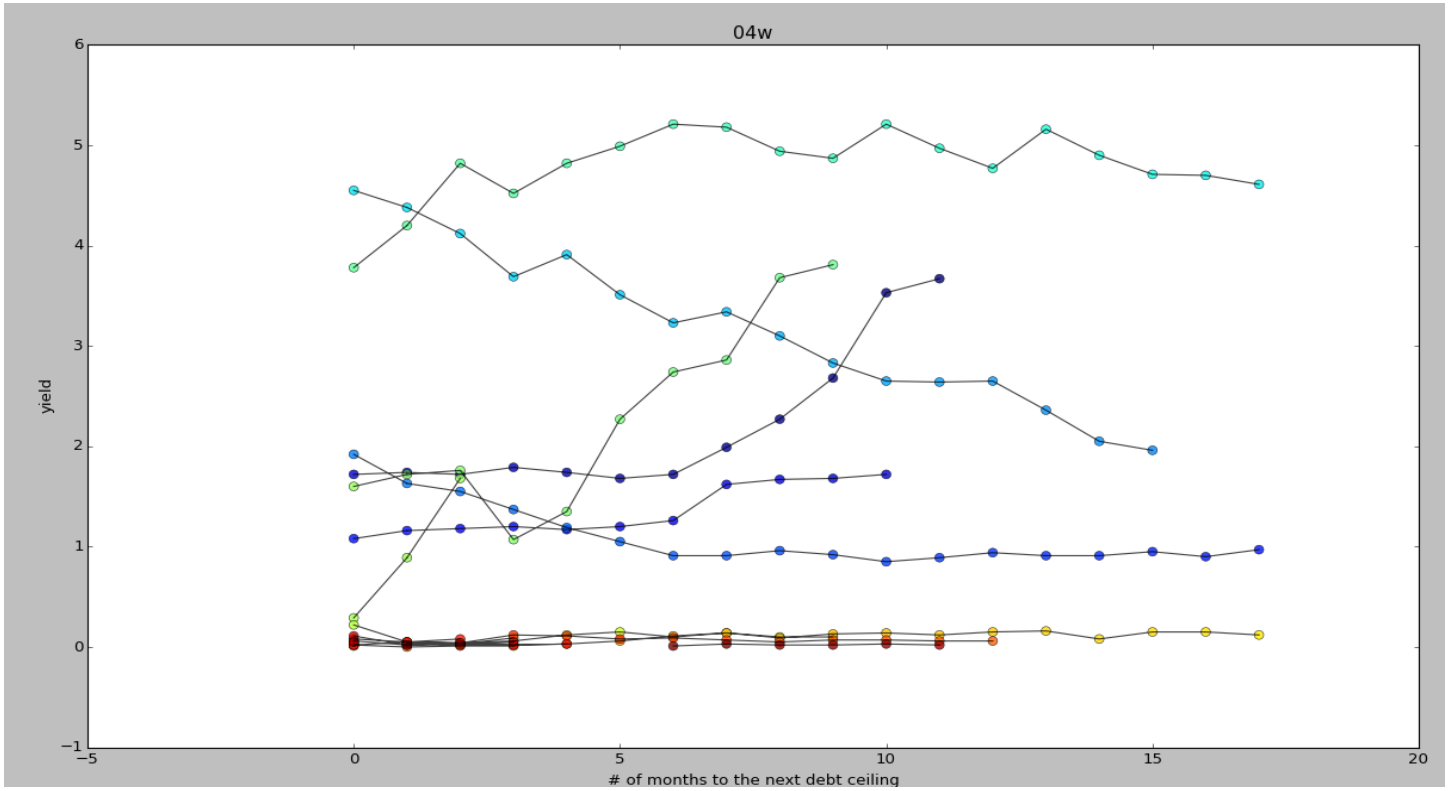
- Calculate *prox2ceiling*
- Time frame: 1990-09-01 to 2014-09-01

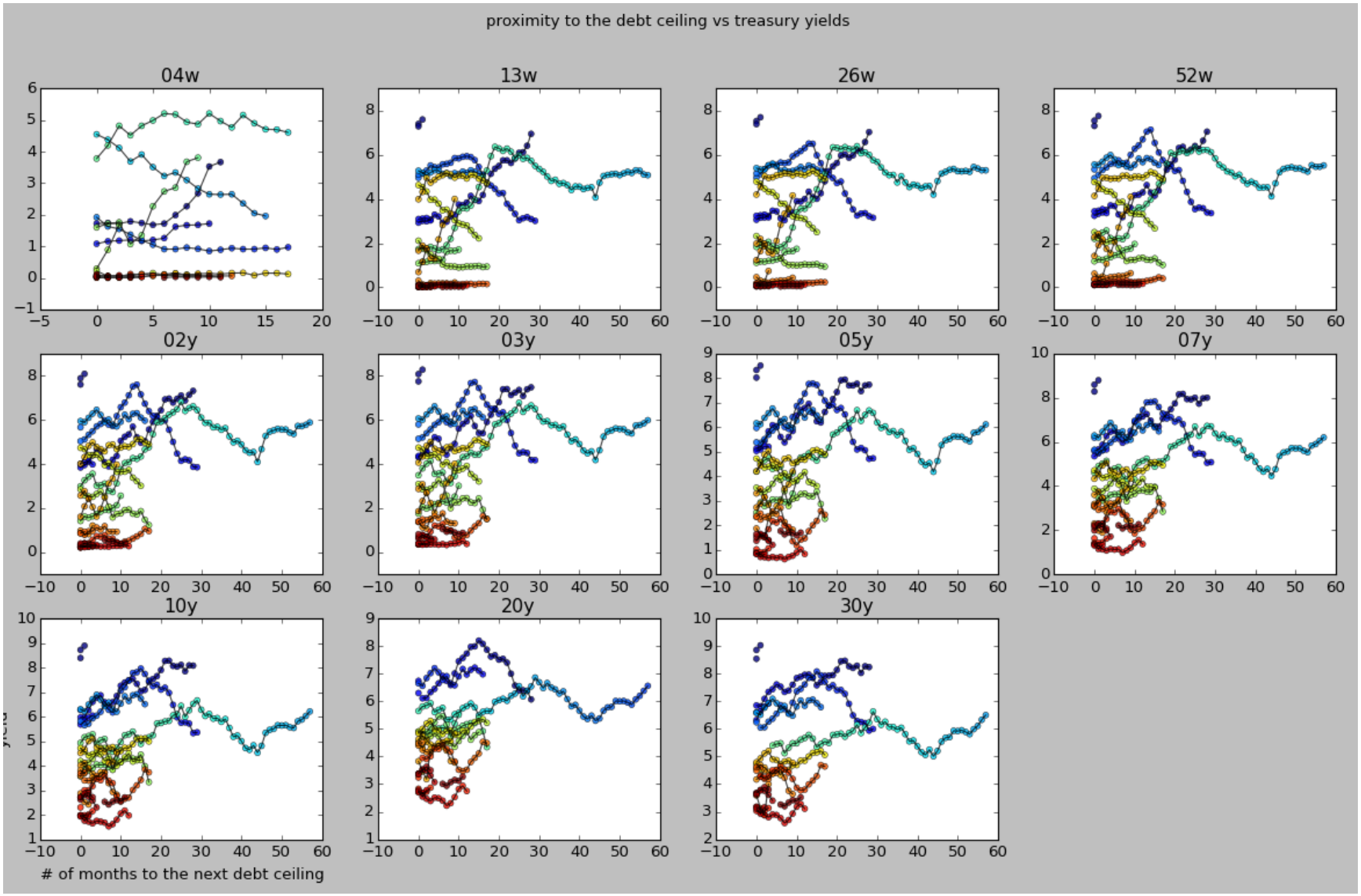
Initial Exploration



Initial Exploration Cont.

Hypothesis: closer to debt ceiling = more likely US will default = higher yield




proximity to the debt ceiling vs treasury yields
04w
13w
26w
52w
02y
03y
05y
07y
10y
20y
30y
# of months to the next debt ceiling

Modeling: Linear Regression

- Continuous supervised – use *statsmodels*
  - *est = smf.ols(formula='tyield ~ CPI + deficit + rgdp + unemployment + TotalOutstanding + prox2ceiling', data=full_data).fit()*

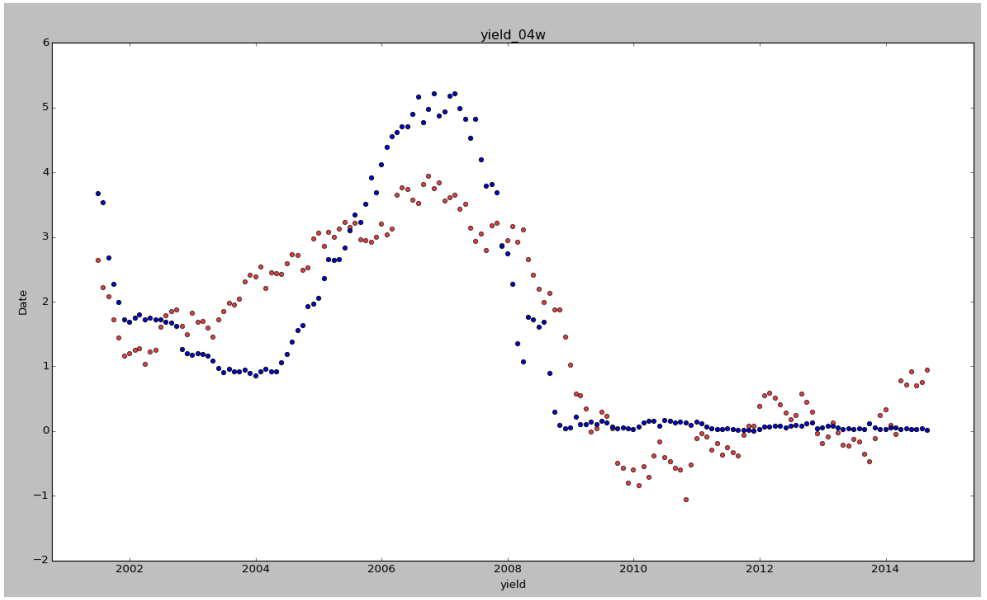
OLS Regression Results

Dep. Variable:	tyield	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.731
Method:	Least Squares	F-statistic:	72.67
Date:	Tue, 25 Nov 2014	Prob (F-statistic):	3.69e-42
Time:	16:13:54	Log-Likelihood:	-197.42
No. Observations:	159	AIC:	408.8
Df Residuals:	152	BIC:	430.3
Df Model:	6		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	10.2957	0.849	12.123	0.000	8.618	11.974
CPI	-0.0472	0.069	-0.682	0.497	-0.184	0.090
deficit	-3.62e-06	5.72e-07	-6.327	0.000	-4.75e-06	-2.49e-06
rgdp	-0.1123	0.047	-2.402	0.018	-0.205	-0.020
unemployment	-1.4631	0.149	-9.796	0.000	-1.758	-1.168
TotalOutstanding	-0.0002	3.59e-05	-5.842	0.000	-0.000	-0.000
prox2ceiling	0.0327	0.016	2.032	0.044	0.001	0.064

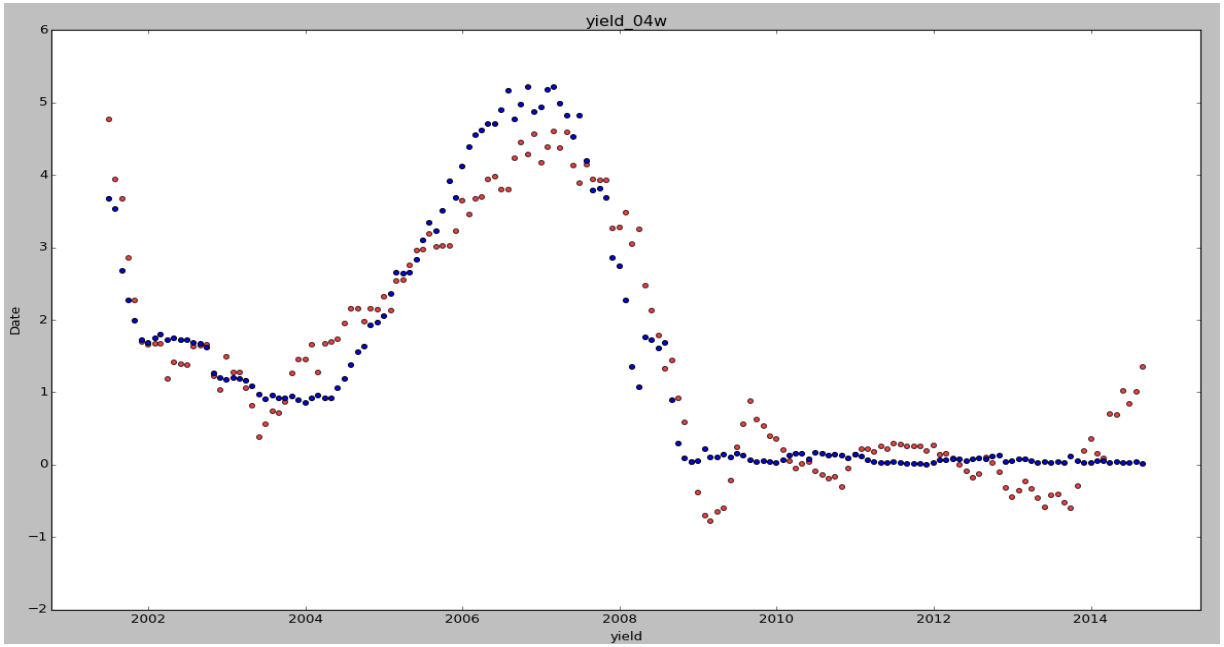
Omnibus:	3.303	Durbin-Watson:	0.125
Prob(Omnibus):	0.192	Jarque-Bera (JB):	2.329
Skew:	-0.118	Prob(JB):	0.312
Kurtosis:	2.456	Cond. No.	9.64e+06

Warnings:
[1] The condition number is large, 9.64e+06. This might indicate that there are strong multicollinearity or other numerical problems.



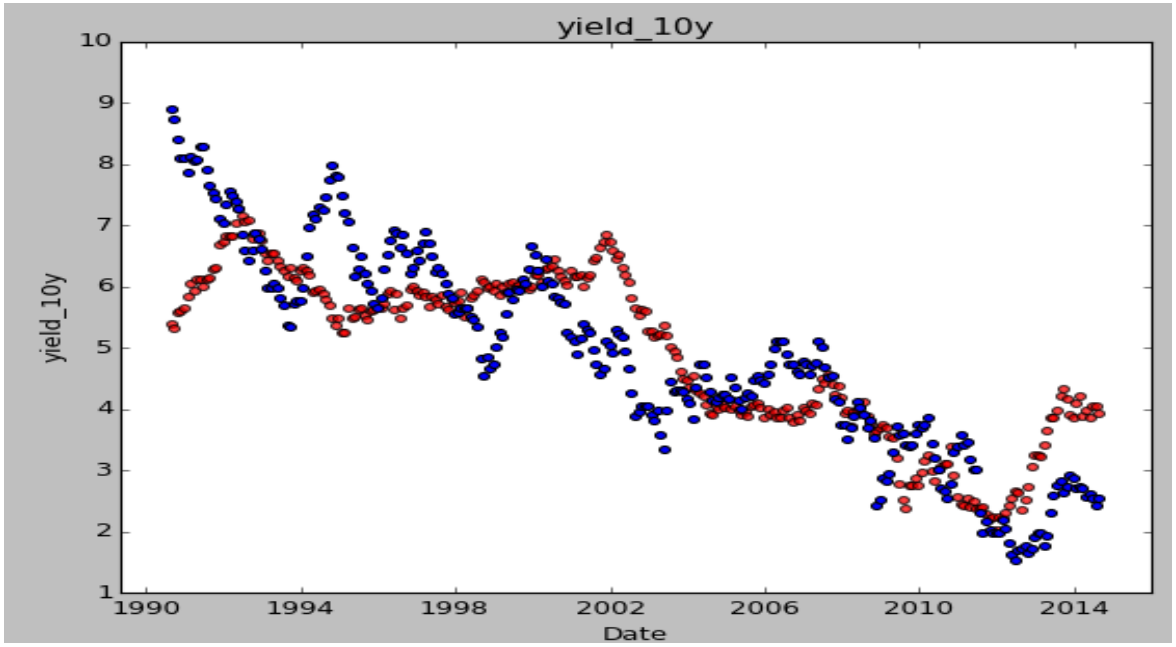
Trial and Error

- est = smf.ols(formula='tyield ~ deficit:unemployment + deficit + rgdp + unemployment', data=full_data).fit()



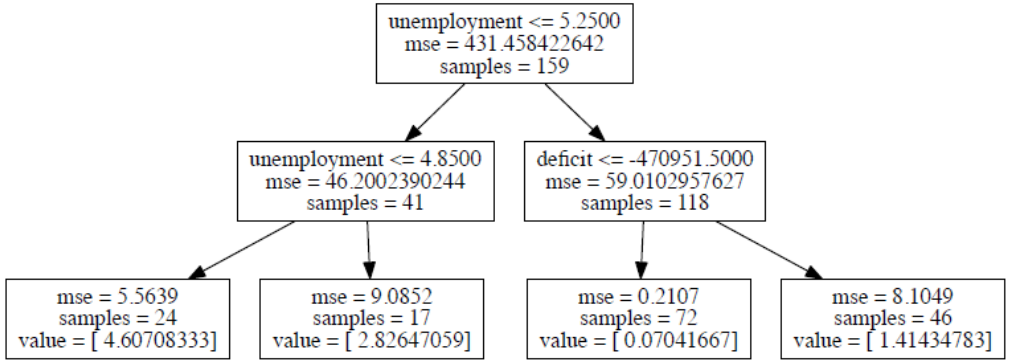
Bad results for 10-Year Yields

- est = smf.ols(formula='tyield ~ deficit:unemployment + deficit + rgdp + unemployment', data=full_data).fit()

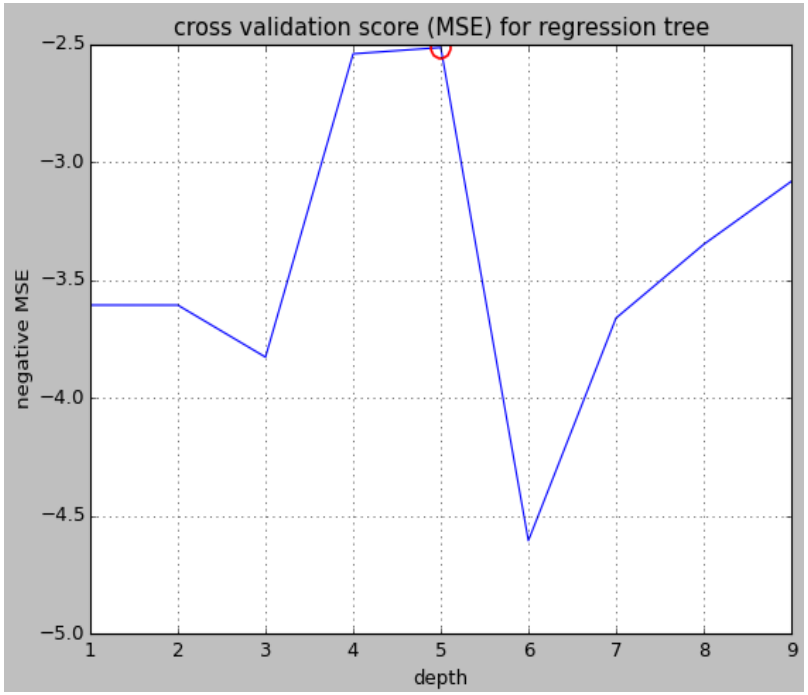


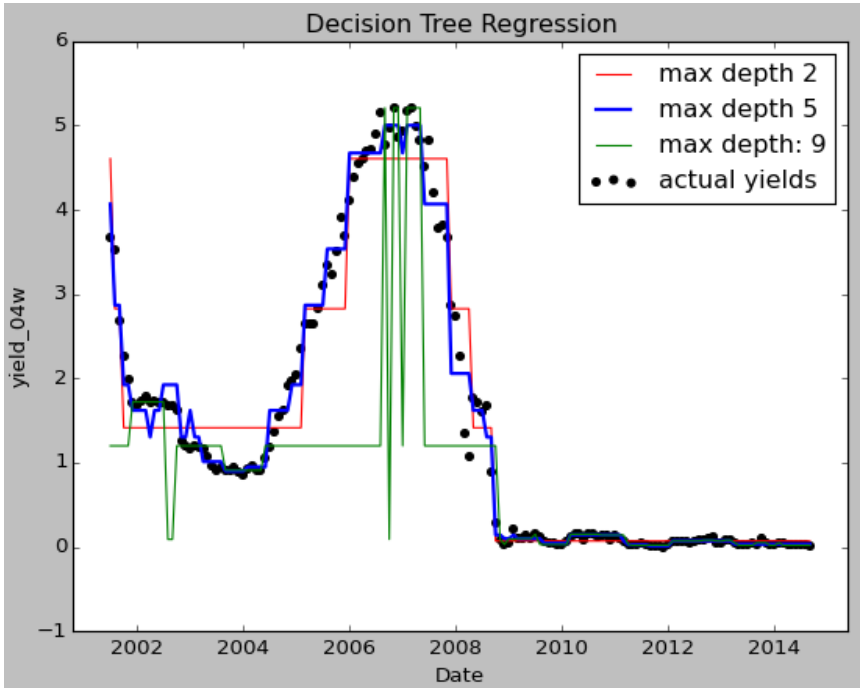
Modeling: Decision Tree

- DecisionTreeRegressor, tree, grid_search
- Max_depth = 2



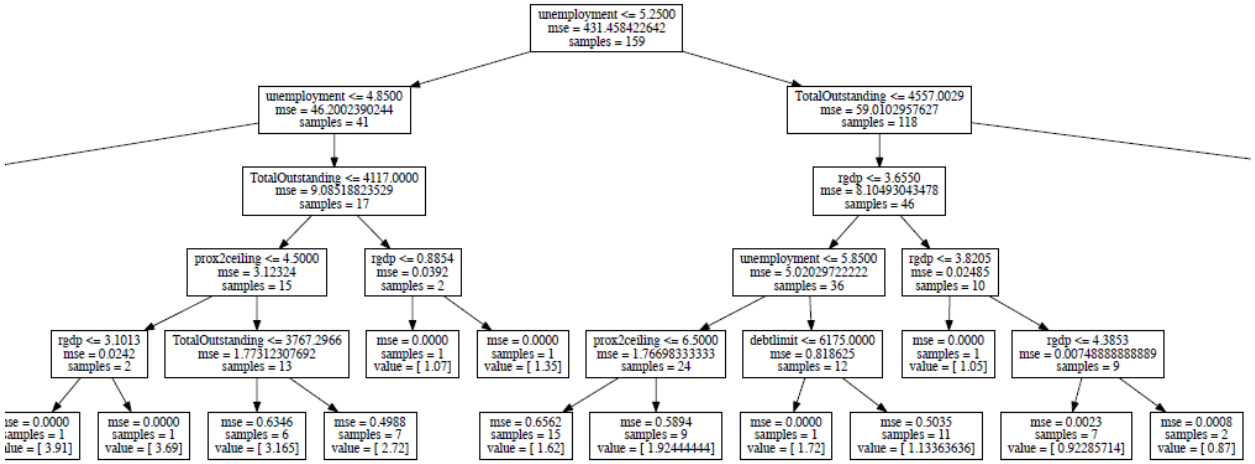
Find the best max_depth





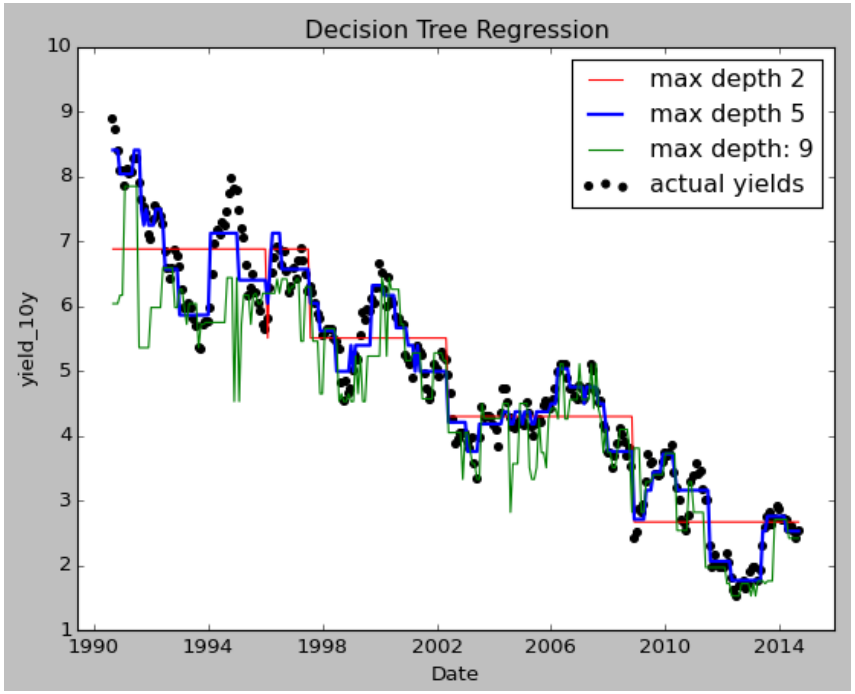
Important features: max_depth = 5

CPI	Deficit	Rgdp	Unemployment	Debtlimit	Outstanding	Prox2ceiling
4.31e-05	1.19e-01	7.16e-03	8.42e-01	6.53e-03	1.93e-02	5.67e-03
7	2	4	1	5	3	6



Repeat for 10-Year Yields

CPI	Deficit	Rgdp	Unemployment	Debtlimit	Outstanding	Prox2ceiling
0.00201	0.01666	0.01285	0.01870	0.78724	0.15852	0.00399
7	4	5	3	1	2	6



Conclusions

- Debt ceiling threat = not too important
- Factors investors consider when looking at yields: unemployment rate and debt outstanding
  - Deficit and debt limit may be important too depending on the maturity of the security

Possible Extensions

- Produce a dataframe with a daily time series
- Focus on 2008 and after (after the financial crash)
- Factor in other explanatory variables such as the Dow Jones Industrial Average or the federal funds rate
- Use more comprehensive linear regression models to tackle multicolinearity
- Use k-means clustering to further examine the explanatory variables
- Factor in “Time” as an explanatory variable – run time series models (Moving Average?)

Challenges and Successes

- Successes
  - Data concatenation
  - Visualization
  - Decision tree modeling
- Challenges
  - Multicolinearity problems
  - Factor in ‘time’ in the model
  - Explore more explanatory variables

Key Takeaways

- Correlation does not necessarily mean causation!
- Use machine learning algorithms to tackle problems
- Cross validation
- Data cleaning requires a lot of time
- Use graphics
- Open source materials available