This project aims to analyze/predict the US Treasury yields given macroeconomic factors.

Background information:
US treasuries are essentially "I owe you" notes from the US government. When the government needs funding, it auctions off debt of various maturities to the public. There are 14 regularly issued types of treasuries and their maturities can range anywhere from 4 weeks to 30 years (see details below). This project only explores bills and nominal coupons.

Type	Tenor	Type	Tenor
Bills	4-Week	Nominal Coupons	7-Year
Bills	13-Week	Nominal Coupons	10-Year
Bills	26-Week	Nominal Coupons	30-Year
Bills	52-Week	Treasury Inflation Protected Securities	5-Year
Nominal Coupons	2-Year	Treasury Inflation Protected Securities	10-Year
Nominal Coupons	3-Year	Treasury Inflation Protected Securities	30-Year
Nominal Coupons	5-Year	Floating Rate Notes	2-Year

The treasury yield can be interpreted as the return on investment, expressed as a percentage, on the US government's debt. It is the interest rate that the US pays to borrow money. Since treasuries are awarded to investors at competitive auctions, the treasury yields not only indicate the borrowing cost of the government, but also tell us how investors feel about the economy.

In this project, I hope to gage how the borrowing cost of the government has fluctuated given various economic factors. Explanatory variables include GDP, CPI, unemployment rates, amount outstanding, proximity to the debt ceiling and deficit levels. Response variables are the treasury yields (I focused on the 4-week yield). I chose these explanatory variables because they are common features used to evaluate the overall health of the market and often mentioned in the Federal Reserve Minutes.

Data:
Most of the data was readily available on the Federal Reserve's website. All csv files have been uploaded on the repository as well.

- Treasury yields (target, monthly ts) – constant maturity data, monthly averages http://research.stlouisfed.org/fred2/series/DGS10
- Unemployment rate (monthly ts) – represents the number of unemployment as a percentage of the labor force http://research.stlouisfed.org/fred2/series/UNRATE
- CPI (monthly ts) – measure of the average monthly change in the price for goods and service paid by urban consumers between any two time periods http://research.stlouisfed.org/fred2/series/CPIAUCSL
- Debt limit (irregular ts) – history of debt limit containing date/amounts http://www.whitehouse.gov/sites/default/files/omb/budget/fy2013/assets/hist07z3.xls
- Deficit (annual ts) – deficit at the end of the fiscal year http://research.stlouisfed.org/fred2/series/FYFSD/

- RGDP (quarterly ts) – inflation adjusted value of the goods and services produced by labor and property located in the United States http://research.stlouisfed.org/fred2/series/GDPC1
- Amount outstanding (monthly ts) – amount outstanding for each US treasury security. Data obtained using Federal Reserve (spreadsheet format)

Data Pre-processing:
I had to first convert all the raw Excel files from the websites into csv files and make sure that each file contains only a header and data. For the debt limit file, I manually processed the "Description" column to obtain a date and an amount. Then, I created an "import" function in python to read in the csv files.

Special pre-processing was required for GDP and CPI: these terms are often measured on a year-over-year basis, so I converted the raw numbers to a growth rate (%).

I converted the data into a dataframe with a date index (monthly). After some initial exploration, I realized that many of the values are missing. For example, GDP is published on a quarterly basis, debt limit has irregular time intervals, and CPI is on a monthly basis. In order to standardize the dataframe, I used the functions "interpolate" and "bbfill" to fill in the NaN values.

One problem that I was interested in is to see whether the proximity to the debt ceiling generated any fear to investors. To measure that, I added a column called "prox2ceiling" in the dataframe, which can be interpreted as the number of months until the debt ceiling. The detailed algorithm of this calculation is explained in code.

Lastly, I truncated the dataframe to analyze the time between 1990-09-01 and 2014-09-01, since data is more accurate and complete during this time.

Initial Exploration:
I first constructed a scatter matrix to see how the variables are related. (See Fig 1 and Fig 2)

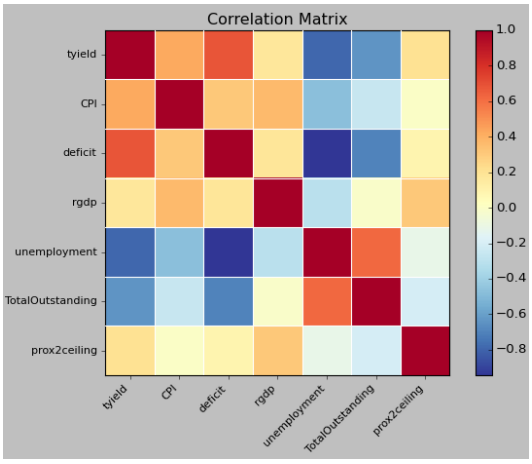

Figure 1

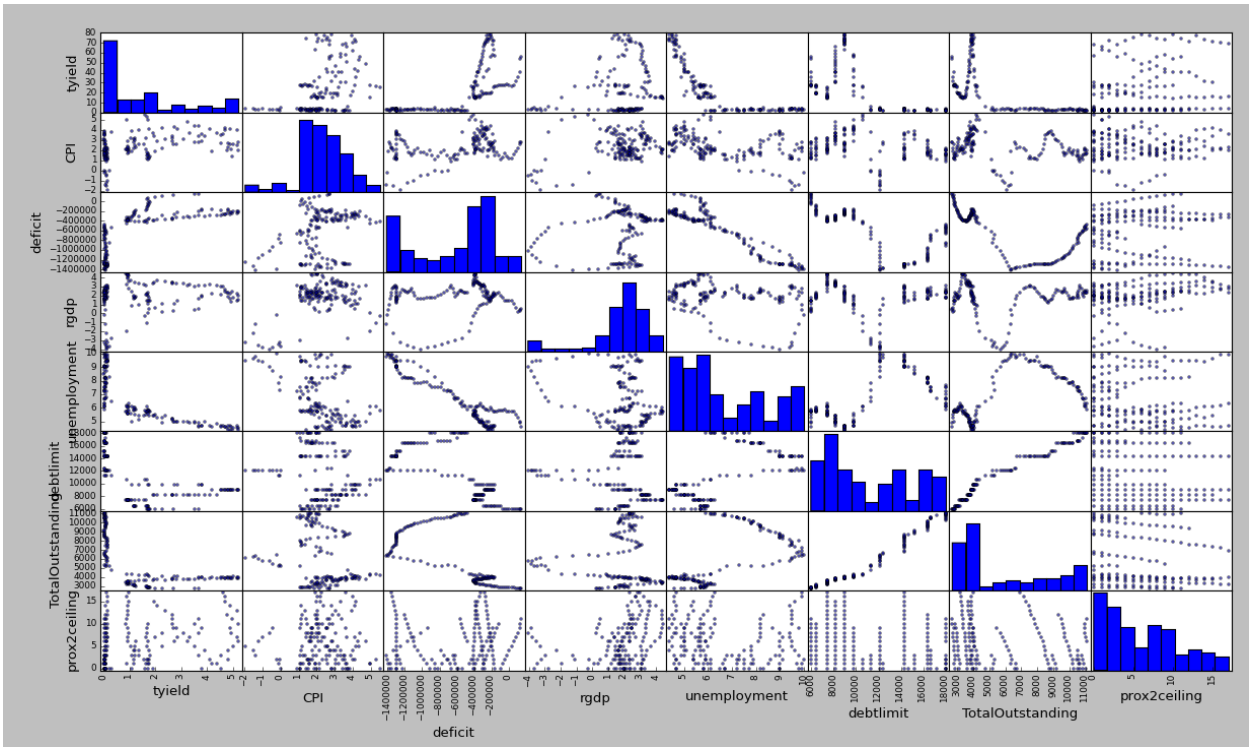

Figure 2

Here's what I noticed:

- Many of the explanatory variables exhibit multicolinearity
- Total debt outstanding has a correlation with deficit
- Debt limit is positively correlated to debt outstanding
- CPI year-over-year growth rate seems to have no relation to yield, unless it is very high or very low (GDP growth rate exhibits the similar behavior when it is very low)
- When unemployment rate is high, the yield is substantially lower

I also wanted to play around with the prox2ceiling column to see whether the debt ceiling had any effects on the yields. My hypothesis is:

*closer to debt ceiling = more likely US will default = lower price = higher yield

To test out this hypothesis, I first graphed the 4-Week Yields (see Fig 3). Each line segment represents a different debt limit. Warmer colors represent more recent dates. If the hypothesis holds, then I would expect to see line segments with negative slopes. However, the 4-Week Yields did not appear to have a definite pattern.

I then decided to apply the same logic to the other treasury yields (maybe investors don't care too much about short-term treasuries). I graphed # months to next debt ceiling vs. treasury yields for each type of the treasury. Surprisingly, I still did not observe any patterns.

This initial graph analysis may indicate that the debt ceiling is simply a rubber stamp – investors are not spooked by debt ceiling threats.

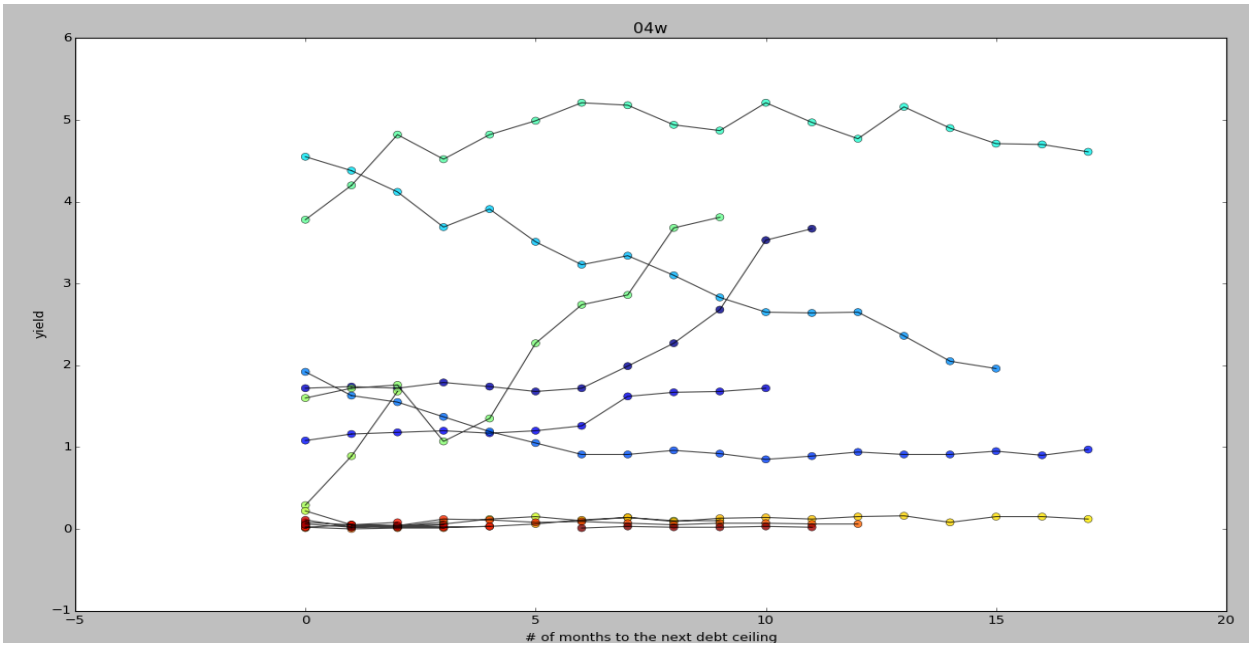

Fig 3

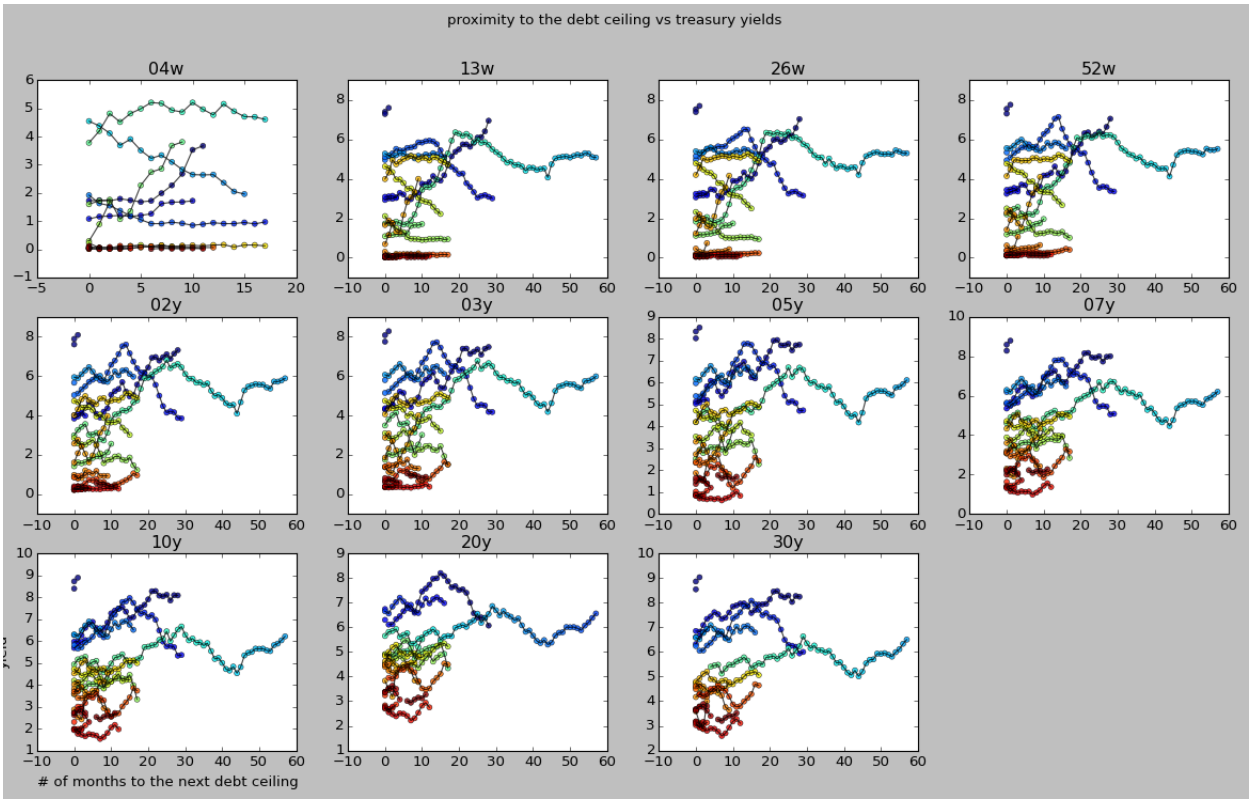

Fig 4

Modeling and Conclusions:

Part 1:

Since this is a continuous/supervised machine learning problem, I first tried linear regression. I used *statsmodels* to run a regression model on all the explanatory variables and the 4-Week Yield as my target variable. I also plotted my predictions (red) vs. actual observations (blue).

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.

Fig 5

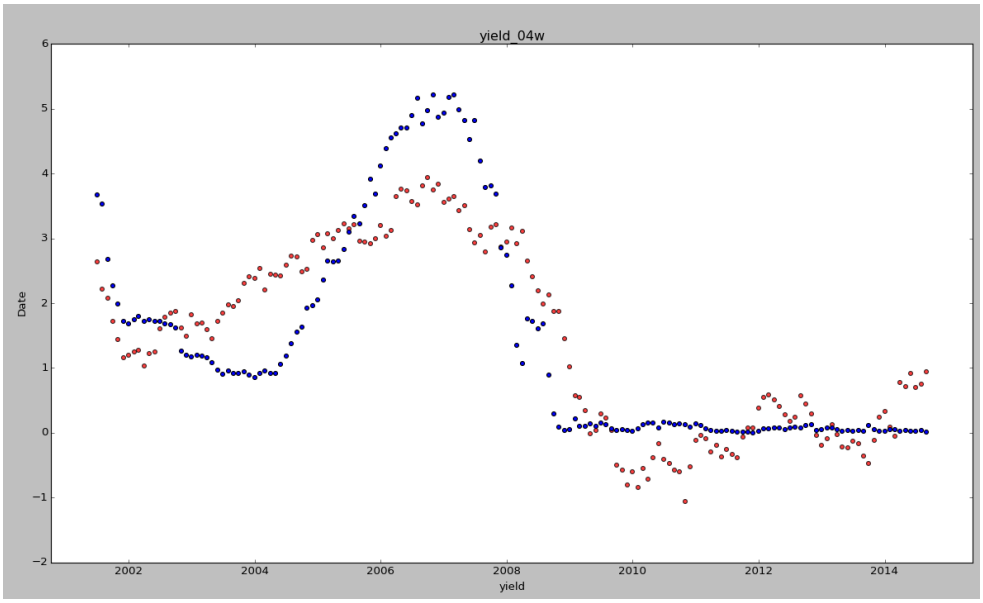

Fig 6

From Fig 6, the predicted yields are hovering true observations. However, this regression failed to capture micro trends, especially between 2003 and 2008. The 4-Week Yield also stabilized in late 2008 (because of the housing market bubble), which the model also failed to capture.

From Fig 5 OLS Summary, the R-squared value of 0.742 is not too bad. However, a message warns that there is strong multicolinearity. I decided to add more interaction terms. Because the p-value for CPI is greater than the significance level of 0.05, I've decided to throw it out. The prox2ceiling variable is also pretty big, but I decided to keep it, since it's less than 0.05.

I ran a new model: *tyield ~ deficit:unemployment + deficit:TotalOutstanding + deficit + rgdp + unemployment + TotalOutstanding + prox2ceiling*. Resulting model has a better R-squared value of 0.893; interaction term deficit:TotalOutstanding, prox2ceiling and TotalOutstanding are now above the significance level. This strengthens the graphic analysis conclusion that the debt ceiling date does not have a significant effect on the yields.

I tweaked the model more and tried running a regression with fewer terms: *tyield ~ deficit:unemployment + deficit + rgdp + unemployment*. The resulting model still has multicolinearity problems. The R-squared value improved to 0.889, and all of the explanatory variables are now above 0.05. Fig 7 shows the new predicted (red) vs actual (blue) observations.

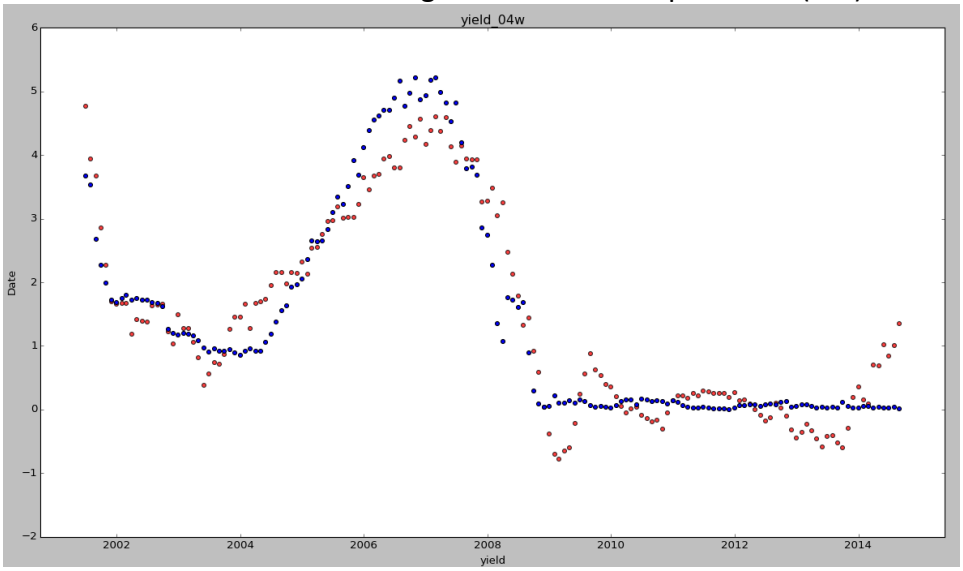

Fig 7

At this point, I wanted to see if this regression applies to treasuries that have different maturities, say the 10-Year notes (Fig 8 & Fig 9). The R-squared value of 0.630 for the 10-Year Yields (using the same regression model as the 4-Week) was significantly smaller. The plot of predicted (red) vs actual (blue) observations also demonstrates that this model may not be the best fit. The micro-trends here are missed almost completely.

In conclusion: linear regression might work for each type of security. However, strong multicolinearity still exists. Coupled with the fact that there could be other explanatory factors

not in the existing models, linear regression may not be the best model in predicting and analyzing yields.

OLS Regression Results

Dep. Variable:	tyield	R-squared:	0.630
Model:	OLS	Adj. R-squared:	0.625
Method:	Least Squares	F-statistic:	120.9
Date:	Thu, 18 Dec 2014	Prob (F-statistic):	4.32e-60
Time:	13:55:39	Log-Likelihood:	-419.62
No. Observations:	289	AIC:	849.2
Df Residuals:	284	BIC:	867.6
Df Model:	4		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	0.3962	0.518	0.765	0.445	-0.623	1.415
deficit:unemployment	-4.629e-08	8.67e-08	-0.534	0.594	-2.17e-07	1.24e-07
deficit	6.287e-06	7.48e-07	8.400	0.000	4.81e-06	7.76e-06
rgdp	0.0939	0.038	2.489	0.013	0.020	0.168
unemployment	1.0533	0.094	11.167	0.000	0.868	1.239

Fig 8

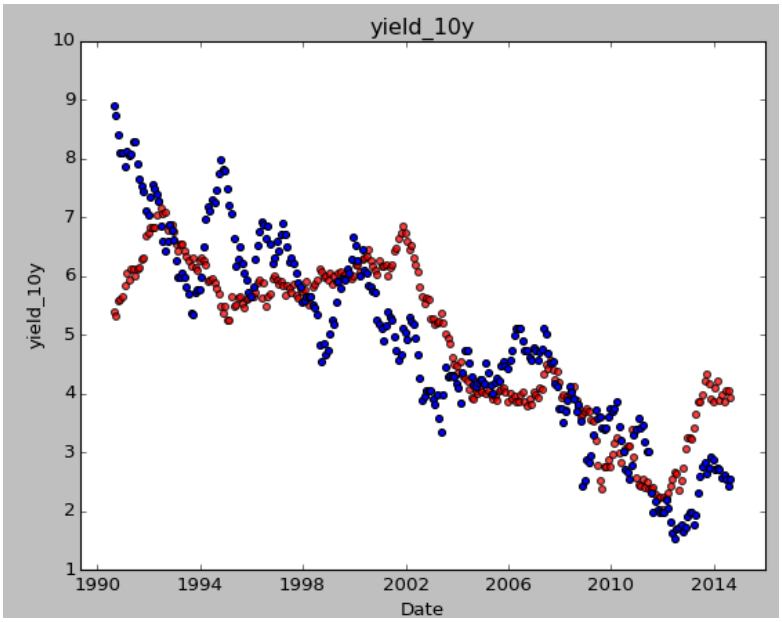

Fig 9

Part 2:

The other modeling approach I decided to try was a Decision Tree. That way, I don't have to manually decide which explanatory variables are important, what are the appropriate cutoff levels, etc. I used the *DecisionTreeRegressor*, *tree,* and *grid_search* in this part of the project.

First, I wanted to see what happens if I have a small tree for the 4-Week Yields. I set max_depth = 2 and created a DecisionTreeRegressor (Fig 10). Next, I used the GridSearchCV function (using CV = 5) to search through max_depth = [1, …, 10] to find the optimal depth based on the mean-squared-error (Fig 11). Because I used mean-squared-error as my scoring metric, GridSearchCV

flipped the signs. Depth = 5 (max MSE of negative 2.5) turns out to be the best depth for the decision tress.

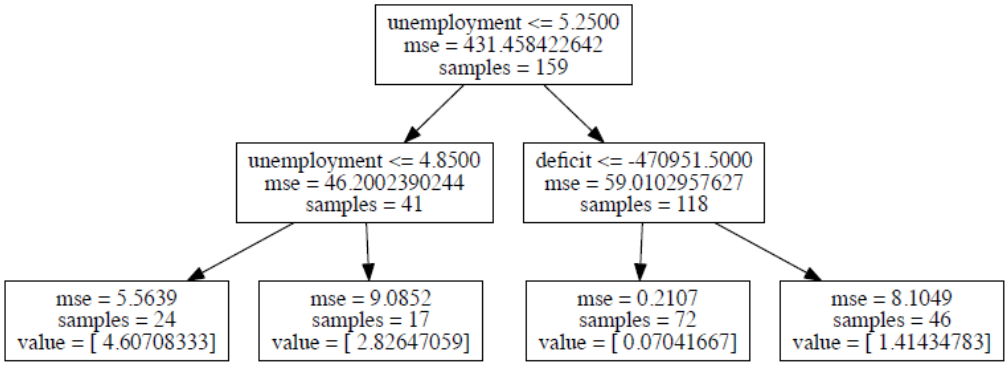

Fig 10

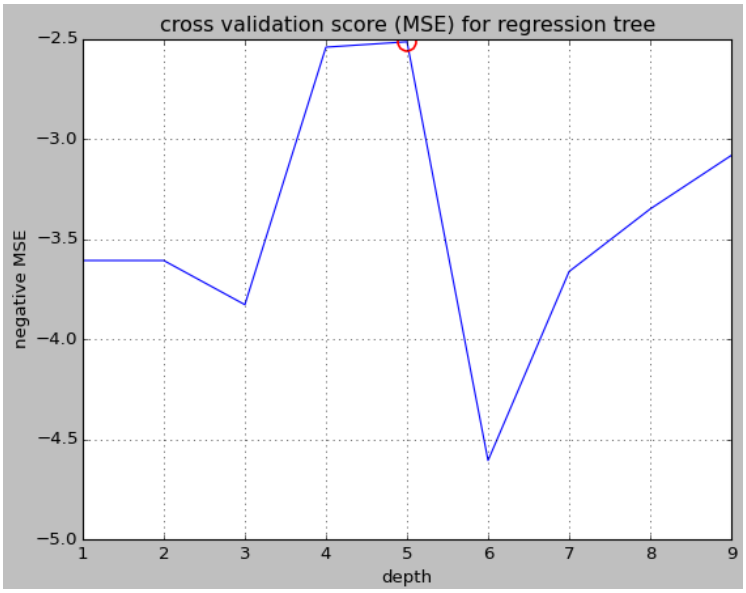

Fig 11

After fitting the entire data to a new decision tree of max depth = 5, I checked to see what are the most important features. The tree indicates that unemployment rate, deficit and debt outstanding are the most important factors in determining the 4-Week Yields.

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

I plotted the decision tree vs. the actual observations (Fig 12). The best decision tree (max depth of 5) follows the observations pretty closely. Max depth of 2 is clearly under-fitting the data, and max depth 9 is clearly over-fitting the data. One drawback to this approach is that yields are not easily predictable after 2008. Fig 13 shows part of the decision tree with max_depth = 5. This tree is consistent with the last and indicates that unemployment, debt outstanding, and deficit are important factors in determining treasury yields.

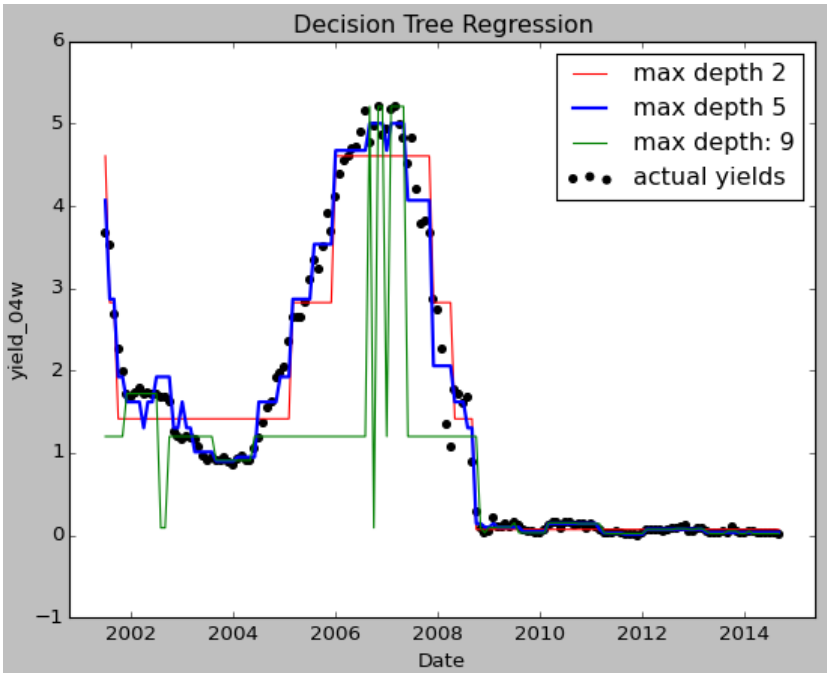

Figure 12

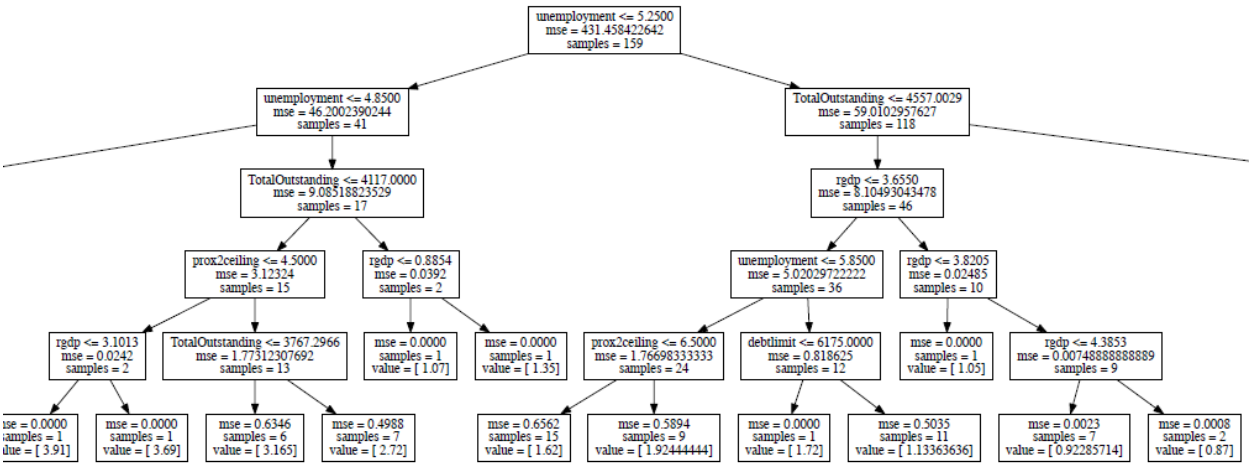

Figure 13

I repeated the same decision tree analysis for the 10-Year Yields to see if the same results hold. Sure enough, the decision tree yielded much better results than the linear regression model (Fig 14). The feature_importances_ scores are listed below. Similar to the 4-Week Yield tree model, the most important factors are debt limit, unemployment rate, and amount outstanding. The unemployment rate, surprisingly, is no longer the most important factor.

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

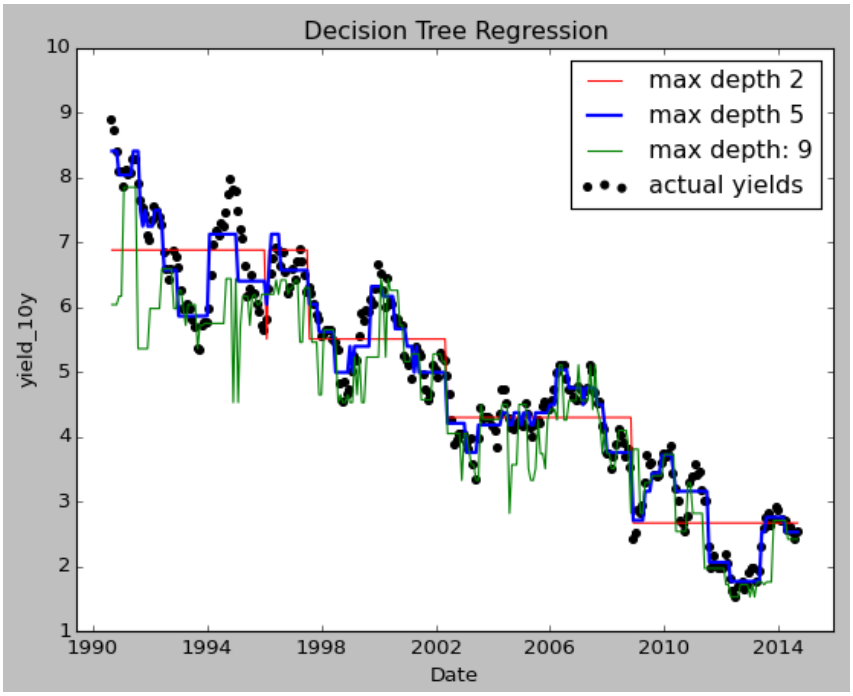

Figure 14

I repeated the same procedure for the 13-Week Yields and 10-Year Yields. Results are very similar. In conclusion, I think the decision tree model is a good fit for this project. The model provides good insights to what economic factors (at what levels) may affect the treasury yields. Furthermore, it confirms prior results that proximity to the debt ceiling does not really have an effect on the treasury market.

This result can allow debt managers to monitor yields more effectively and reduce borrowing costs by understanding investor behaviors.

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: unable to fix multicolinearity problems, unable to factor in ‘time’ in the model or use time series tools in Python, need to factor in more explanatory variables

Key learnings:

- Correlation does not necessarily mean causation!
- Let machine do the job: sometimes it may not be feasible to manually construct a model. Using machine learning algorithms are great approaches to solve a data problem, especially when we have no idea where to even begin.
- Cross validation: it is important to perform cross validation to make sure the model is appropriate
- Data cleaning takes a lot of time: the hardest part of this project was getting the data points to line up. In real world scenarios, it is rare to come across a good/clean dataset.
- Use graphics: graphics are one of the most convincing tools that a data scientist can use to get a point across
- Learn from open source materials: there are a ton of resources available online, and a great online community that can answer all sorts of data problems