Jules BERTRAND, Adam MOUTONNET, Aakash GROVER, Rémi LUYSSAERT, Pierre LUCAS

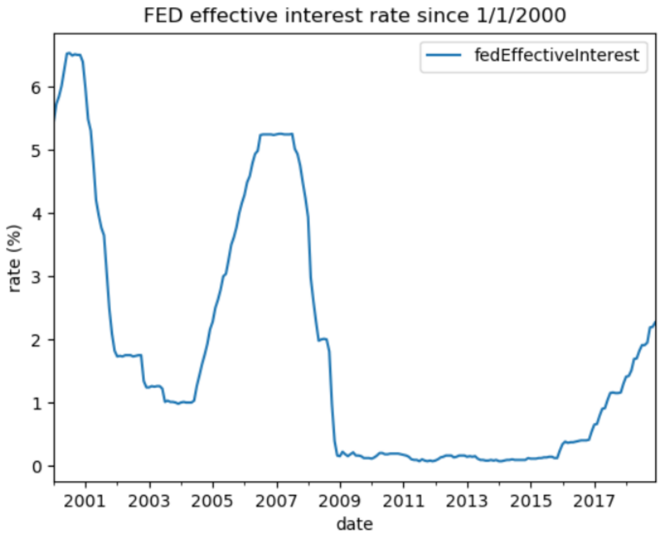
Predicting FED Monthly Effective Rate

I. Introduction

In a world where transactions are made at the speed of light, banks constantly need to balance their reserves of money. Banks lend reserve balances to other depository institutions overnight on an uncollateralized basis. The federal funds effective rate is the weighted average of all interest rates negotiated between borrowing banks and lending banks. The target is set by the federal reserve which intervenes on the markets to ensure that it is respected. In our study, we will only consider only the US federal reserve (FED). The interest of this rate is multiple:

- Banks use it to adjust their reserves. It sets the basis for short-term lending in the financial sector.
- It guides longer-term interest rates affecting household's loans interest rates and businesses borrowings.
- It makes it possible to play on the liquidity of the financial system and its propensity to lend money

There is a difference between the objective rate, fixed by the FED and toward institutions try to tends to, and the effective rate which is the real-time rate on the market. Our objective is to predict two things: whether the rate will go up, down, or stay at the same value and its real value in the future. Here is a sample of this rate since 1/1/2000:



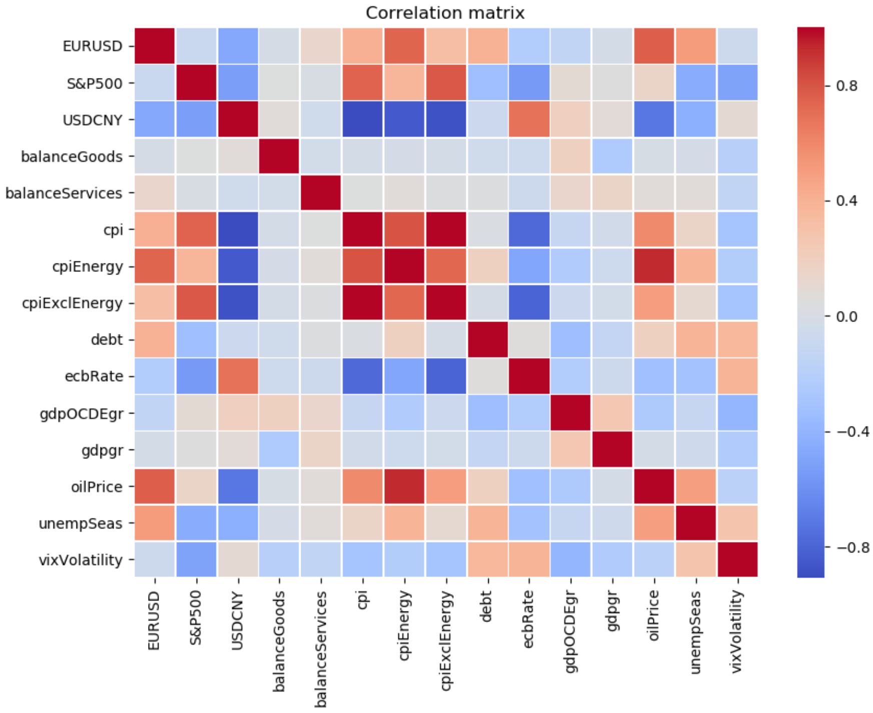
II. Data gathering and processing

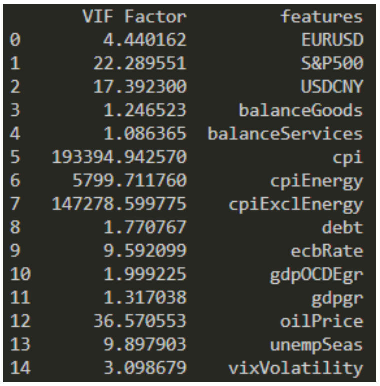
Type	Name	Source	Unit	Frequency	Availability
USA eco	US GDP growth rate	ycharts.com	% of growth	Monthly	Since 2000
USA eco	Total public debt	fred.stlouisfed.org	Millions of Dollars	Quarterly	Since 1966
USA eco	Unemployment	data.bls.gov	% of working pop	Monthly	Since 1948
USA eco	Employment/population ratio		% of pop		Since 2009
USA eco	CPI		Index		Since 1948
USA eco	CPI Energy				Since 1957
USA eco	CPI excluding energy				
USA eco	Trades balance for good	census.gov	Millions of Dollars		Since 1992
USA eco	Trades balance for services				
USA financials	S&P 500 value	macrotrends.net	Index		Since 1927
USA financials	Vix index (volatility)				Since 1990
Global eco	OECD GDP growth rate	data.oecd.org	% of growth	Quarterly	Since 1948
Global eco	ECB rate	ecb.euroa.eu	% (interest rate)	Monthly	Since 1999
Fx rate	EUR/USD Fx rate	finance.yahoo.com	EUR/USD		Since 2000
Fx rate	CNY/USD Fx rate		CNY/USD		Since 1992
Global eco	Oil Price	fred.stlouisfed.org	$ per barrel		Since 1987
Politics	POTUS	Wikipedia	R for Republican D for Democrat	Monthly	Since 1897
Politics	House of Representative				
Politics	Federal Reserve Chairman				Since 1914

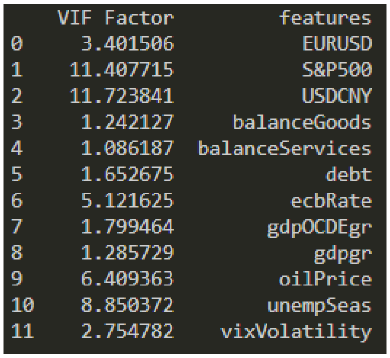
On the previous table, you can see what kind of data we scraped, on which website, its unit, its frequency and since when it is available. Few things to notice:

- Since "Employment/population ratio" is the only data not available since 1/1/2000, we will get rid of this one, and keep only the data after 1/1/2000 for our dataset.
- We have two data whose frequency is quarterly; we need to find a way to "turn it monthly":
  - For the OECD growth rate, the growth rate over 1 month equal to $\sqrt[3]{1+r}-1$ with r the quarterly growth rate.
  - For the total public debt, we'll first convert it to a growth rate by diving the difference of two consecutive values by the first one, and subtracting 1. Then we'll apply the same method as for the OECD growth rate to have it monthly.
- Since trade balances in Millions of dollars have high values, we'll convert them to a growth rate to.
- Some of the data is not fully available for year 2019. Therefore, we will consider only the data until 12/31/2018 (included).

Now that we have a complete database, we will try to find some multi collinearity inside it. The following heatmap correlation matrix and the upper-right VIF analysis directly show us the collinearities.




```
        VIF Factor         features
0         4.440162           EURUSD
1        22.289551           S&P500
2        17.392300           USDCNY
3         1.246523     balanceGoods
4         1.086365  balanceServices
5    193394.942570              cpi
6      5799.711760        cpiEnergy
7    147278.599775    cpiExclEnergy
8         1.770767             debt
9         9.592099          ecbRate
10        1.999225        gdpOCDEgr
11        1.317038            gdpgr
12       36.570553         oilPrice
13        9.897903        unempSeas
14        3.098679    vixVolatility
```


```
    VIF Factor         features
0     3.401506           EURUSD
1    11.407715           S&P500
2    11.723841           USDCNY
3     1.242127     balanceGoods
4     1.086187  balanceServices
5     1.652675             debt
6     5.121625          ecbRate
7     1.799464        gdpOCDEgr
8     1.285729            gdpgr
9     6.409363         oilPrice
10    8.850372        unempSeas
11    2.754782    vixVolatility
```

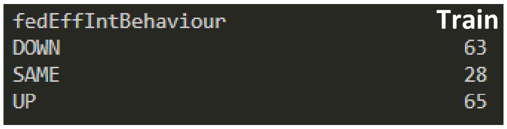
Of course, cpi, cpiEnergy and cpiExclEnergy are highly correlated among them, as well as oilPrice and USDCNY with cpi. We will then remove the three cpi variables, to have a proper VIF analysis (lower-right one). USDCNY and S&P500 are still highly correlated, but we deem them to be important and keep them anyway.

Now that we have a non-correlated data, let's discuss the time-series analysis. In order to predict the behavior of the rate as well as its value at time t, we will make the assumption that any other data at time t (such as unemployment, debt…) is unavailable. Hence, only the data of time t-1, t-2, …, t-n can be taken into account (with **n** the number of time steps used to do our prediction) as features to predict the observed interest rate at time t. We then processed the data to have a database (that depends on n) with the observation of the interest rate at t+1 **Y** linked to their corresponding features **X** that are the values of every variables at time t-1, …, t-n. Note that a variable *fedEffIntBehaviour* has been created and take value *DOWN, UP or SAME* according to the behavior of the fed interest rate.

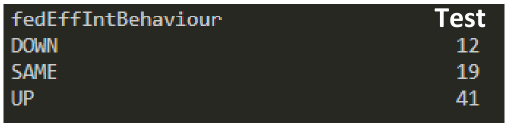
We choose to take data from 2012 included and before as our training set, and data after 1/1/2013 (included) as our test set.

III. First problem: a classification to predict the behavior

Here our vector of observations **Y** will be our variable *fedEffIntBehaviour*, with possible values *DOWN, UP or SAME*. We need to solve a multi-class classification problem. We will use three multi-class classification models: LDA, Random Forest and Decision Tree. The baseline model, which will always predict *UP*, has an accuracy of 56.9%.

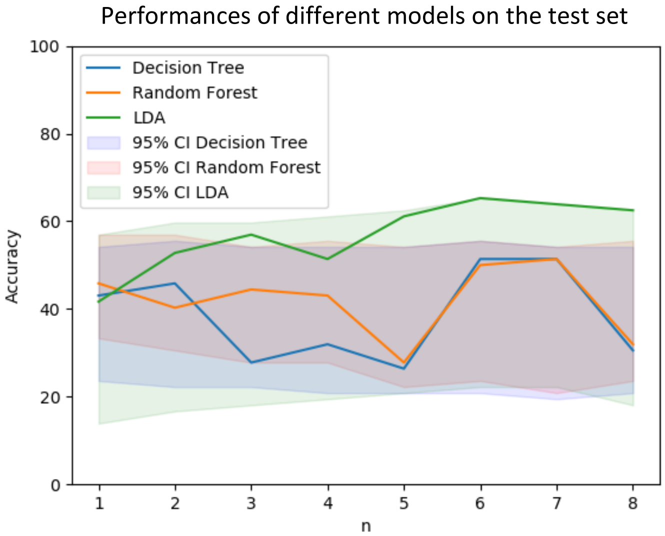


```
fedEffIntBehaviour          Train
DOWN                           63
SAME                           28
UP                             65
```

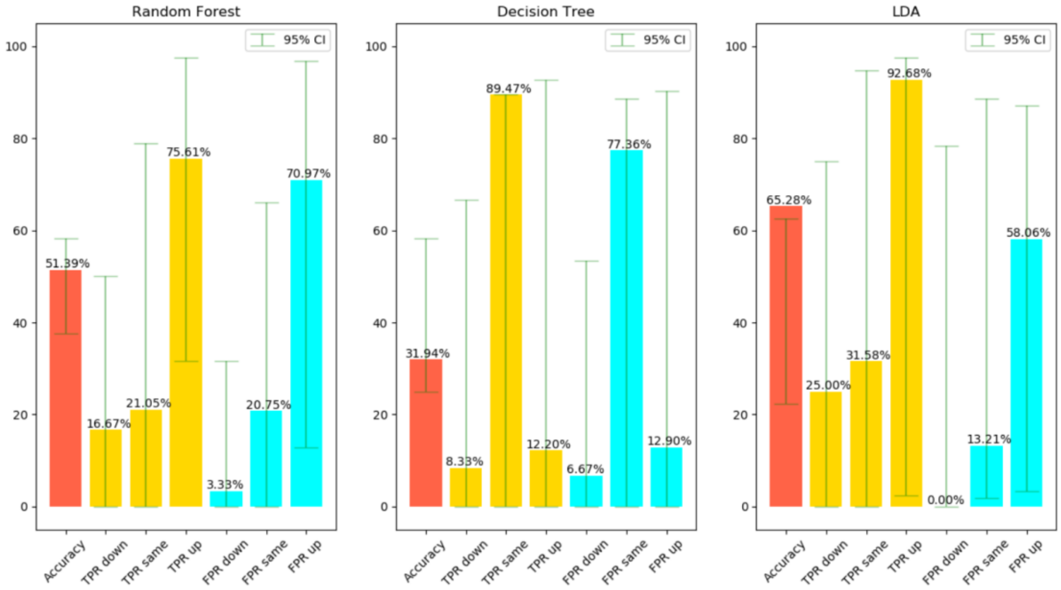


```
fedEffIntBehaviour           Test
DOWN                           12
SAME                           19
UP                             41
```

Our first objective will be to find which value of **n** to take in order to have the best results. This value will be validated thanks to the performances (accuracy) on the test set of our models with python default settings. Then we will consider that this **n** is the best value for our problem, and try to refine our models by cross-validating their parameters. Maybe it would have been a great idea to add **n** to the cross-validation grid, but it would have been far too long for computers to compute it, that's why we adopted this technique.



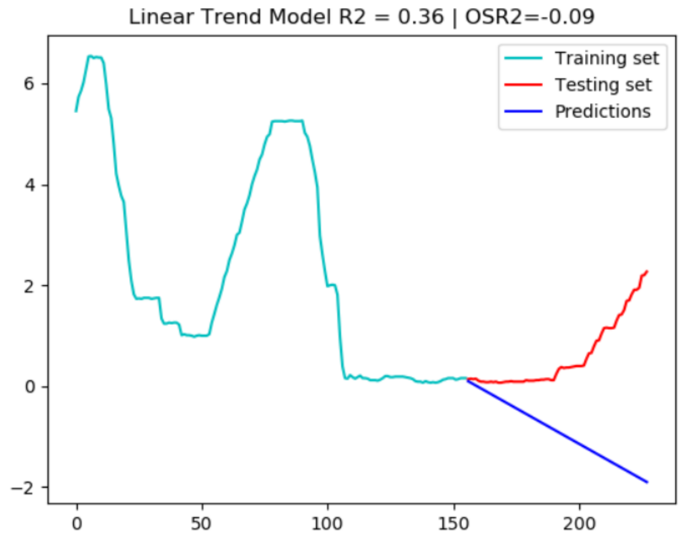
We see here that the best accuracy of 65.28% is reached by the LDA for n=6, and the two other models reach good accuracy of ≈50% too. We are aware that the confidence interval is very large for such values of n, and we will try to reduce it by 5-fold cross-validating parameters of the Decision-Tree and of the Random-Forest. Here are the results we obtain:



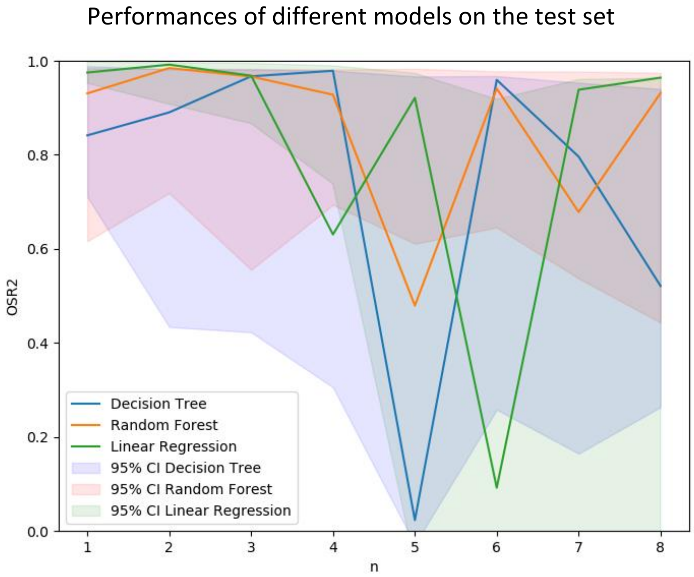
The only improvement here is that the accuracy confidence interval size decreased a lot thanks to the cross-validation for Decision Tree and Random Forest. However, concerning the true positive rate and the false positive rate for each class, the confidence interval is really large. Unfortunately, this can be explained by the fact that we clearly don't have the same class ratio in the training set and in the test set, and even if the LDA does a pretty good job, we still have very large confidence interval, even after cross-validation.

IV. Second problem: a regression to predict the future value

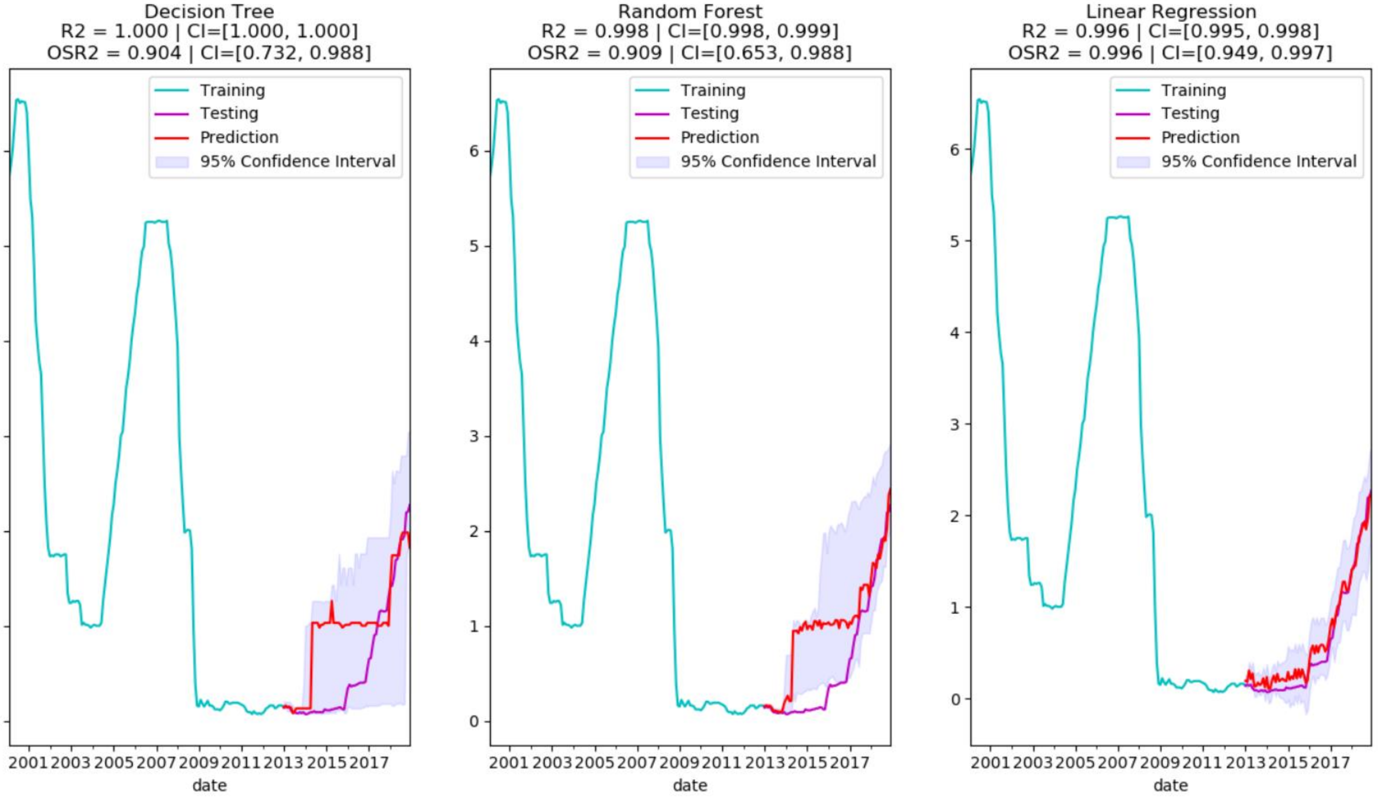
Here our vector of observations **Y** will be our variable *fedEffectiveInterest*. We face a regression problem and therefore we will use three regression models: Linear Regression (which is in fact an auto-regressive model), Random Forest and Decision Tree. The baseline model, will be a linear trend one:



Even if the R2 is not so bad, this baseline model is not spreading very well in the test set. As for the classification problem, our first goal is to find the best **n**:



Here the best OSR2 of 0.996 (and the best confidence interval) is reached by the linear regression model for n=1. Globally, for each model n=1 allows better results with good confidence interval. Now let's analyze the performance for n=1 after after cross-validating:



Even after cross-validating, none of the models manage to beat the performances of the linear regression. We clearly see that the predicted value is near the real curve, and that the real curve is within the (pretty narrow) confidence interval. However, one need to keep in mind that this problem is easier than the classification problem, because as the variations are not very fast and abrupt a model predicting approximately the value of the previous point will get good results anyway.

V. Interpretation and Conclusion

Our best model is the linear regression to predict the Fed's effective rate. Despite its strong performance (R2 and OSR2 > 0.99), the 95% confidence interval for predictions is very wide, too wide to allow banks, individuals or companies to use our model. These confidence intervals are mainly due to the amount of data available:

- Some features are too recent to be used (not enough data points), while they are very interesting, such as the EUR/USD exchange rate or the Employment/population ratio that gives us the population's participation rate.
- When we look at the Fed's effective interest rate curve over a longer period (1960-2018), we see that it follows long term cycles (10-25 years). These are macroeconomic cycles, characterized by features such as economic growth, growth in international trade, the presence of speculative bubbles in financial markets, unemployment rates, private debt (companies and households), etc. All these features are interdependent but not fully correlated, and we need more cycles, i.e. data over a longer period of time, to analyze in detail which indicators influence the Fed interest rate and at what time (eg. Unemployment at t-3 and speculative bubble at t-1).
- It is possible to build a more complex model, possibly a neural network. But again, the number of data points was too limited in our case to be able to try.

Thus, the limiting factor for our project was really the number of data points we had compared to the number of features we tried to include in the model. However including all these features was clearly necessary to obtain a good model, and it makes sense in the real world as macroeconomics rely on a lot of different parameters.