The prediction of FED interest rate change using FOMC minutes

Group name: 4D-Intelli

March 18, 2019

Name	Student ID	Email
WEI Xueyang	3035542186	u3554218@connect.hku.hk
Li Diyi	3035542617	ethanli@connect.hku.hk
Sun Jiahao	3035543532	sunjh@connect.hku.hk
Peng Cheng	3035542980	u3554298@connect.hku.hk
Ho, Chon Wai	3035543025	hochon@connect.hku.hk
Cho Hyun Jin	3035424994	jcjc1992@connect.hku.hk
XIA Yuhang	3035542148	comete@connect.hku.hk

What is the FOMC minutes and why is it important for interest rate prediction?

- Over time, the Fed has substantially increased its level of transparency thereby aiming at making monetary policy more effective.
- The release of the minutes can have a sizable impact on Treasury bond yields. The impacts are largest when the tone of the minutes differs from the tone of the statement. This presumably leads markets to change their expectations of future monetary policy.
- The Fed is a highly predictable central bank and its communications have helped markets to anticipate future policy rate changes. The policy decision and communications by which the Fed or its officials explain monetary policy may have an impact on the market assessment of the future monetary policy course.

What did we do?

1. Scrape the FOMC statements/minutes from the website –Yuhang Xia, Cheng Peng
2. Data cleaning and processing –Chon Wai Ho, XueYang Wei, Cho Hyun Jin
3. Evaluate the predictive performance of FOMC minutes and analysis-Cheng Peng, Yuhang Xia
4. Summary and blog post in github - Diyi Li

Web scraping and data collection

We first start with the web scarping skills to get the FOMC minutes in the FED website. There are several different formats in the official websites from year to year, so we use different coding to deal with them. **Requests** and **Beautiful Soup** packages are the most important ones in this stage.

After we get the FOMC minutes from 1968 to 2019, we have to convert the pdf format of 484 minutes to txt format for further analysis, for which we use a Python package called “pdfminer” to work. However, this package is not perfect such that some texts in the minutes have been concatenated to a single meaningless word and some Garbled characters shows up.

Then we download the FED daily interest rate data from website macrotrends, finishing our data collection part so far.

Data cleaning

After collecting the data, we first deal with the imperfect texts with several packages of Python: re and Viterbi_segment.

By using “re” package we substitute all the non-word character to a white-space, to make sure that those garbled character would not have huge impact on our result, then we find out that some words are combined into one, for example “interestchange”, so we use the algorism of Viterbi_segment to separate them. Although these two packages are not perfect and may remain some flaws in our data, but after checking the texts files manually, we believe our results based on the cleaned data would be reliable.

Data processing

Given the cleaned data, we have tried several natural language processing and text analytics method to process it. Package Spacy, nltk and sklearn are used to tokenize the documents, remove stop words and calculate the bag of words (bow) and the tf-idf matrix.

Then we remove the words that do not show up in at least 0.5% documents among all 484 docs and scale all the documents such that we can compare them in the same level.

After that we use LSA in sklearn package to extract the 400 principle topics of the docs, and merge our document-word matrix with the interest rate data by the minutes public release date, then calculate the difference of interest rate between next release date and the release date of this time as a indicator of the movement of interest rate change, to see the minutes performance in terms of prediction.

Predictive performance analysis

We first calculate the document similarity by numpy package, seeing that a clear pattern is that basically the longer the time interval of two documents is, the less similar they are, which is understandable as we presume, since the change might be huge in longer time interval.

Then we try to use the 400 main components to run an OLS regression of interest rate change indicator as we mentioned before, but the result is not significant and trivial in absolute value. We then try to use different method to solve this problem such as using bow instead of tf-idf or using original terms instead of components, but none of them work well.

So, we decide to change our way of prediction, to see whether we can use our matrix to predict the interest rate change direction (up or down). By using logistic regression, we get some components that have P-value less than 0.05, statistically significant. Then we divide our dataset into train set and test set, and use different algorithm of machine learning to see which algorithm have the best predictive power. And it turns out the SVC is the best for all of our analysis below.

Model	Train_score	Test_score
SVR	0.01435	0.003141
SVC	0.5039	0.5670
SGD	0.01237	-0.005770
BAYES	0.3226	0.1744
LL	0	-0.01623
ARD	0.3202	0.08352
PA	-3.2414	-3.3725
TS	0.2929	0.07804
L	0.3335	0.1861

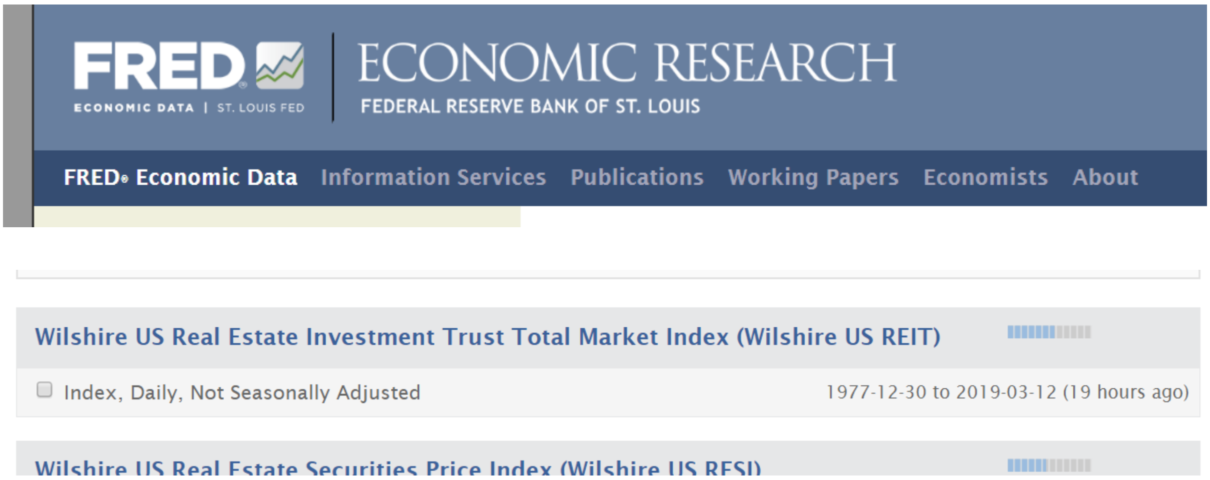
After deciding to use SVC, we then divide our dataset to three different datasets, and use train dataset and validation dataset to test different parameters for SVC model. The package model selection in **sklearn** can do that automatically for us and we test the predictive performance of the best parameter model using our test dataset and get a score of 0.69, meaning 69% accuracy of the direction prediction.

We are not satisfied with it and have tried different methods to increase the accuracy such as using bow instead of tf-idf, giving up the LSA and turning to the significant terms of our logistic regression and aggregate the 8 docs within the same year to predict the interest rate change direction from year to year, and the result is below:

methods	Best model	Validation set mean score	Test dataset mean score
Single minute interval, LSA, tfidf	SVC (c=5, linear)	0.7039	0.6907
Single minute interval, LSA, bow	SVC (c=10, linear)	0.6935	0.6289
Single minute interval, significant terms, bow	SVC (c=2, rbf)	0.7273	0.7526
Single minute interval, significant terms, tfidf	SVC (c=5, rbf)	0.7299	0.7423
Year minutes interval, significant terms, bow	SVC (c=0.1, linear)	0.8250	0.9091
Year minutes interval, significant terms, tfidf	SVC (c=10, rbf)	0.9250	1.0

As the results show, the best way for prediction is using the significant terms filtered out by logistic regression and calculated by tf-idf algorithm in year average minutes to predict the direction of interest rate movement direction by year. But the dataset contains only years from 1968 to 2018, 50 documents, so the result is not so reliable as single minute interval. Apart from the year aggregation, the most predictable one is using significant terms calculated by tf-idf algorithm, the predictive accuracy is up to 73% on average, although the test dataset score is a little lower than bow counterparty.

Then we turn our attention to industry level prediction. As we all know, the real estate industry is highly correlated with the interest rate movement, and we believe people in this industry would react actively to any minutes released by FOMC. So, we download the REIT data from 1977 from the website FRED and do the same processing and analysis as the interest rate change.



Wilshire US Real Estate Investment Trust Total Market Index (Wilshire US REIT)

□ Index, Daily, Not Seasonally Adjusted 1977-12-30 to 2019-03-12 (19 hours ago)

Wilshire US Real Estate Securities Price Index (Wilshire US RESI)

The result is even better than interest rate change direction as below:

methods	Best model	Validation set mean score	Test dataset score
Single interval, significant terms, bow for real estate	SVC (c=0.8, rbf)	0.7538	0.8209
Single interval, significant terms, tfidf for real estate	SVC (c=2, rbf)	0.8030	0.8806
Single interval, LSA, bow for real estate	SVC (c=10, linear)	0.6856	0.7463
Single interval, LSA, tfidf for real estate	SVC (c=5, rbf)	0.9091	1.0

We have also found out some key words correlated with interest rate change with real meaning and strong tendency:

For positive correlation, words like “rapidly”, “constraint”, “objection”, “tolerance”, “anxiety” are very significant, and “credit”, “reliance”, “weak”, “weakness”, “ease”, “soften”, “liquidate” are significant for negative correlation. This finding is a proof for our predictive power to some extent.

Conclusion

- It is hard to simply use FOMC minutes to predict the actual change of the Fed Rate, and every statistic model performs poorly for that.
- There is some prediction power in terms of the sign of the Fed Rate change (up or down), even conditioning on the internal information.
- The prediction power is even more significant when average the interest rate change by year, although the data would be much small and therefore might be less reliable.
- We can get a better prediction by focusing on some industry performance such as the real estate industry on which the interest rate plays an import role.

Improvements

- Our preprocessing is not perfect, meaning that some words and terms are still not what they should be in the minutes, we could clean our data manually to get a cleaner dataset.
- We use the minutes release date interval to calculate our interest rate change interval which is not so rigorous academically, we could change our variable definition to see whether we can predict one month change after minutes release or 6-month change direction instead.
- Now we only apply our findings to REIT interest rate to focus on real estate industry, and we can try other industries in the future.