Conclusion

December 11, 2019

1 Conclusion

(https://github.com/QuantCS109/TrumpTweets)

1.1 What we accomplished

1.1.1 Data

We had to work with two different kinds of data: text from Trump's tweets and market data.

We obtained, parsed, cleaned, and made Trump's tweet data easily accessible through the TweetData class.

We obtained high quality market datasets, and learned juggle with futures and options data.

We scraped hundreds of options settlement files, and set different classes to be able to obtain information as necessary, such as in the FuturesData, RatesCure, FuturesCurve, and VolCurve classes.

We had to find futures and options expirations for each of 13 different asset classes and all the caveats for each to solve our problem.

1.1.2 Volatility

We learned how to obtain information such as implied volatility and gamma, out of options prices, and looked for ways to summarize this information in a way relevant to our problem.

1.1.3 Text

We worked built WordEmbeddings using a neural network Skipgram model and used the embeddings to generate topics.

We used Vader for generating sentiments of tweets.

1.1.4 Features

We designed dozens of features to try to predict market direction, both from futures and options data, as well as from Trump's tweets.

1.1.5 Modeling

We looked at issues that are particular to finance, such as low signal-to-noise ratio, high variance in models, time series analysis, and data leakage, We attempted to use information that was available at the time of making trading decisions. We looked at appropriate measures for performance.

1.1.6 Feature Importance

Keeping model interpretability was highly important for us. Looking at the top 10 predictors and the amount they explained in the total model is extremely powerful for us. This leaves the door open for further analysis and ask new questions. To quote Marco's First Law of Backtesting- "Backtesting is not a research tool. Feature importance is."

1.1.7 Teamwork

We had a team working on opposite sides of the globe one in New York and two in UAE. We come from different backgrounds, nationalities and work experiences. It was a very fun and interesting aspect of the project! Powered by Github and Canvas of course.

1.2 What we missed

The gamma features weren't as significant as other features, which was disappointing given the amount of work involved in obtaining them. Theres room to learn what could have been done to make them more relevant as in theory they seem like very powerful features to have.

We predicted 1 day direction of markets. I think that this is sub-optimal, and a larger trading time frame would have probably been more suitable. The issue with predicting more than 1 day return is your returns start overlapping. Two contiguous 5 day returns will have 4 days of overlap, and so assumptions like independence start breaking. We read about some approaches to fix this, so as using bagging with each bag only using non-interlapping data, but we didn't have time to explore this facet.

We built in the capability of the model to not trade if it wasn't sure enough of market direction, but we didn't explore how it would compare to our results since it would have taken an extensive cross-validation.

We would've loved to craft 'hand-curated' topics by feeding in 'seed words' and extracting topics from them.

We realize how important is to approach statistical modeling with patience, caution and an inquisitive mindset. It's extremely easy to have a wrong input in the model and have completely erroneous results.

2 Literature Review / Related Work

2.1 References

2.1.1 Options

Stochastic Calculus for Finance II, Continuous-Time Models. Steven E. Shreve, Springer Options theory, stochastic models, Black-Scholes proof.

Inside Volatility Arbitrage, Alireza Javaheri, Wiley Options theory, volatility surface, greeks

The Mathematics of Arbitrage, Delbaen, Schachermayer, Springer Fundamental theorem of asset pricing

Wikipedia - Black-Scholes

Black Model formulas

Gamma-calculation: https://quant.stackexchange.com/questions/32974/proof-of-gamma-profit-formula

2.1.2 Trading

Advances in Financial Machine Learning, Marcos Lopez de Prado, Wiley Log-loss as the appropriate cross-validation measure for hyper-parameter tuning in finance.

F1-score to complement accuracy when measuring model performance

Time-Series cross-validation with embargo

Feature importance as the most important measure of performance at the model building stage

Focus on reducing variance and not overfitting in financial models https://jpm.pm-research.com/content/15/3/30

2.1.3 Tweeter resources

https://developer.twitter.com/

http://www.trumptwitterarchive.com/

https://github.com/bpb27/twitter_scraping

2.1.4 Word2Vec

Wikipedia https://en.wikipedia.org/wiki/Word2vec

Udacity AI for Trading Course https://github.com/udacity/deep-learning-v2-pytorch/tree/master/word2vec-embeddings

Visualising MNIST http://colah.github.io/posts/2014-10-Visualizing-MNIST/

Intro to Recurrent Neural Networks from Udacity: https://www.youtube.com/watch?v=UNmqTiOnRfg

Word2Vec tutorial from Chris Mccormick http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Original Google paper on Word2Vec from Mikolov et al. https://arxiv.org/pdf/1301.3781.pdf

Second paper on Word2Vec from Mikolov et al http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

TSNE visualiztion of embeddings: https://towardsdatascience.com/google-news-and-leo-tolstoy-visualizing-word2vec-word-embeddings-with-t-sne-11558d8bd4d

2.1.5 Vader from NLTK:

for our sentiment features it was interesting to see the shift in Trump's Twitter activity as discussed at The Outline website.

https://www.nltk.org/_modules/nltk/sentiment/vader.html

https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

https://theoutline.com/post/2445/trump-s-first-real-tweet-was-on-july-6-2011?zd=1&zi=blsl6n2o

SciKit Learn: https://scikit-learn.org/

Matplotlib: https://matplotlib.org/

Seaborn: https://seaborn.pydata.org/

Similar projects A project from this class, two years ago, found that 5 minute returns was the most predictive out of short term returns. Instead of focusing on the most predictive return, we looked at the difference between 15 minute returns and 5 minute returns as well as an average of 1, 5, and 15 minute returns when we created some of our features.

https://pdfs.semanticscholar.org/af67/ae4c3ac357679c10ddc394df52d392432f63.pdf.