

DOI:

К ВОПРОСУ О ВОЗМОЖНОСТИ ПРЕДСКАЗАНИЯ ЗНАЧЕНИЙ ФОНДОВЫХ ИНДЕКСОВ С ПОМОЩЬЮ ДАННЫХ О ТОНАЛЬНОСТИ СООБЩЕНИЙ В TWITTER

Гилязова А.А.

Институт проблем управления им. В.А. Трапезникова РАН,

Россия, г. Москва, ул. Профсоюзная д.65

giliazova@mail.ru

Аннотация: В данной работе исследуется возможность предсказания значений фондовых индексов, используя машинное обучение и данные о тональности сообщений в социальной сети Twitter.

Ключевые слова: тональность, Twitter, фондовый индекс, индекс Доу-Джонса, GSPC.

Введение

Предсказание движений фондового рынка с целью максимизации прибыли является задачей, интересующей инвесторов и учёных. Есть две основные теории об образовании фондовых индексов. Во-первых, фундаментальный анализ или анализ внутренней стоимости (Fundamental or Intrinsic Value Analysis (IVA)). Его главное допущение в том, что паттерн поведения цен на индивидуальный актив имеет тенденцию снова проявляться в будущем. Это предполагает возможность предсказывать будущие движения цен, глядя на их поведение в прошлом. Несмотря на это Eugene, (1965) [1] обнаружил, что последовательные изменения фондовых цен между периодами независимы. [2]

Во-вторых, гипотеза эффективного рынка (Efficient Market Hypothesis (EMH)). Эта теория утверждает, что рынок является эффективным в отношении какой-либо информации, если некоторая цена «полностью отражает» эту информацию. [3] Вся доступная информация уже включена в цену актива. Гипотеза эффективного рынка предполагает независимость в движении цен. Участники рынка не смогут делать выводы о будущей цене актива, глядя только на прошлую цену, поскольку изменения цен происходят случайным образом. [2] Тем не менее, мало внимания уделялось объёму торгов, поддерживающему классическую теорию рыночного обмена. Исследуя объём торгов по дням, можно собрать немного больше информации о цене актива, поскольку можно обнаружить позитивную взаимосвязь между объёмом и ценой. [2, 4]

Согласно [5], эта теория утверждает, что цены фондового рынка, в основном, двигаются новой информацией, т.е. новостями, а не текущими или прошлыми ценами. Коль скоро новости непредсказуемы, то цены фондового рынка должны следовать паттерну случайного блуждания и не могут быть предсказаны с точностью выше 50 процентов. [6] В [7] утверждается, что некорректно считать эквивалентными гипотезу эффективного рынка и гипотезу случайного блуждания ввиду необходимости поправки на риск.

В статье 2011-го года “Twitter Mood Predicts the Stock Market” (Настроения в Твиттере предсказывают фондовый рынок) [6] в Journal of Computational Science (Журнал вычислительной науки) авторы Johan Bollen, Huina Mao и Xiaojun Zeng заявили об использовании сделанных с помощью проприетарного алгоритма оценок коллективных настроений в социальной сети Twitter для достижения точности 86,7% в предсказании колебаний значений промышленного индекса Доу-Джонса – взвешенного по цене фондового индекса для акций 30 крупнейших компаний США. [8] Авторы объединились с хедж-фондом Derwent Capital Markets и приобрели десятки миллионов долларов. Менее чем через год после основания этот хедж-фонд был закрыт и продал с аукциона свои активы приблизительно за 120 000 фунтов стерлингов, намного ниже заявленной точки безубыточности в 350 000 фунтов стерлингов и ориентировочной цены в 5 миллионов фунтов стерлингов. [8, 9] Их статья привлекла огромное внимание со стороны учёных, собрав 2648 цитат, в том числе 248 в высококвалифицированных статьях, по данным semanticscholar.org на 15.06.2020. [10]

Та статья первоначально появилась в 2010 году в arXiv, онлайн-архиве для статей перед публикацией. И получила внимание таких медиа как Huffington Post [11], Time [12, 13], CNN [14] и других. Возможно, этот факт способствовал ослаблению обычных практик контроля качества научных публикаций. [8]

Статья была опубликована в Journal of Computational Science, от Elsevier, основанном в 2010 г. Она является не только самой высокоцитируемой статьёй этого журнала, но и превосходит следующую за ней по числу цитат в Web of Science более чем в 12 раз, а по числу цитат в Google более чем в 30 раз. [8]

Существует несколько безуспешных попыток воспроизвести эти результаты на данных другого временного периода, но они оставляют открытой возможность, что найденный в [6] эффект настроений в Twitter был нивелирован после загрузки той статьи в arXiv. [8]

Но в статье 2017-го года [8] была предпринята безуспешная попытка воспроизвести эти результаты на данных о твитах, включавших 90 процентов данных авторов оригинальной статьи [6]. Предполагается, что выдающиеся результаты, полученные в [6], были основаны на особенностях проприетарного алгоритма для подавления шума. Обнаружено, что анализ общих настроений в Twitter не полезен для предсказания ежедневных рыночных показателей и что видимость предсказательной силы получена за счёт набора статистических отклонений: (1) отклонение из-за множественных сравнений в текстовых данных высокой размерности, (2) отклонение из-за сложного алгоритма обработки естественного языка и длинного окна отбора проб, (3) предвзятость публикации.

В данной работе результаты [8] были перепроверены на тех же данных, но с помощью других методов машинного обучения, а также с добавлением на рассмотрение индекса GSPC.

Результаты

Данные для исследования взяты на сайте Kaggle [15].

В них на период с 19.07.2007 по 31.12.2008 даны значения индексов Доу-Джонса и GSPC (S&P 500), а также оценки тональности сообщений в Twitter.

Список названий столбцов в данных:

- Date: дата.
- tone: данные о тональности сообщений Twitter.
- calm: данные о тональности спокойствия в сообщениях Twitter.
- tone_Z10: данные о тональности, нормализованные с $k=10$. По сути, это центрированный Z-score тональности по 21 наблюдению.
- tone_Z1: данные о тональности, нормализованные с $k=1$. По сути, это центрированный Z-score тональности по 3 наблюдениям.

- DJI: значение цены закрытия для индекса Доу-Джонса на заданный день. Раннее закрытие обрабатывается также, как и в обычные рыночные дни.
- DJIvolumek: объём индекса Доу-Джонса на заданный день, в тысячах акций.
- DJIforwardreturn: относительное изменение индекса Доу-Джонса за один рыночный период. Например, значение 0,01 соответствует увеличению индекса Доу-Джонса на 1%, отсчитывая от цены закрытия текущего дня к цене закрытия следующего дня. Есть наложение информации между данными о тональности твитов и данным показателем: данные о тональности наблюдаются перед полночью EST, в то время как инвестировать по данному показателю нужно было бы до 4PM EST.

- GSPC: цена закрытия для индекса GSPC на заданный день. Раннее закрытие обрабатывается также, как и в обычные рыночные дни. Статья [8] не рассматривает данный индекс.
- GSPCvolumek: объём индекса GSPC на заданный день, в тысячах акций.
- GSPCforwardreturn: относительное изменение индекса GSPC за один рыночный период, отсчитывая от цены закрытия текущего дня к цене закрытия следующего дня.

При проведении расчётов рассматривались различные варианты деления этих столбцов.

Данные X для обучения модели сглаживались скользящим окном высотой в 10 строк:

```
X1 = data[['tone','calm']]
X2 = data[['tone']]
X3 = data[['tone_Z10','calm']]
X4 = data[['tone_Z10']]
X5 = data[['tone_Z1','calm']]
X6 = data[['tone_Z1']]
```

Целевые значения:

```
y1=data['DJI']
y2=data['DJI_volume_k']
y3=data['DJI_forward_return']
y4=data['GSPC']
```

```
y5=data['GSPC_volume_k']
y6=data['GSPC_forward_return']
```

Таким образом, рассматривались 36 вариантов сочетаний (X, y).

Строки, соответствующие выходным дням, для которых нет значений индексов, были отброшены. Расчёты проводились с помощью языка Python с библиотеками Pandas, Sklearn.

Данные для обучения составляли 80% всех данных, а тестовые 20%. Выбранное значение показателя random_state равнялось 23. Подбор значений гиперпараметров осуществлялся с помощью RandomizedSearchCV со значениями параметров cv = 5, scoring='neg_mean_absolute_error'.

В Таблице 1 представлены наилучшие результаты среди 36 вариантов, полученные для разных моделей. Сравнение производилось по значениям коэффициента детерминации r2_score.

Таблица 1. Результаты расчётов

Название модели	Лучший r2_score	explained_variance_score	X, y
AdaBoostRegressor	0,089	0,102	X3, y4
BaggingRegressor	0,059	0,062	X1, y5
ExtraTreesRegressor	0,196	0,197	X1, y5
GradientBoostingRegressor	0,152	0,156	X1, y5
HistGradientBoostingRegressor	0,029	0,044	X5, y2
RandomForestRegressor	0,247	0,250	X1, y5

Как видно из Таблицы 1, большинство моделей показало наилучший результат в сочетании X1, y5, т.е. при предсказании 'GSPC_volume_k' с помощью данных о тональности и о спокойствии в наборе твитов.

Тем не менее, даже эти наилучшие результаты обладают слишком малой прогнозной силой, чтобы их можно было использовать практически. На Рисунке 1 показана диаграмма рассеяния для наилучшей модели, т.е. RandomForestRegressor для сочетания X1, y5.

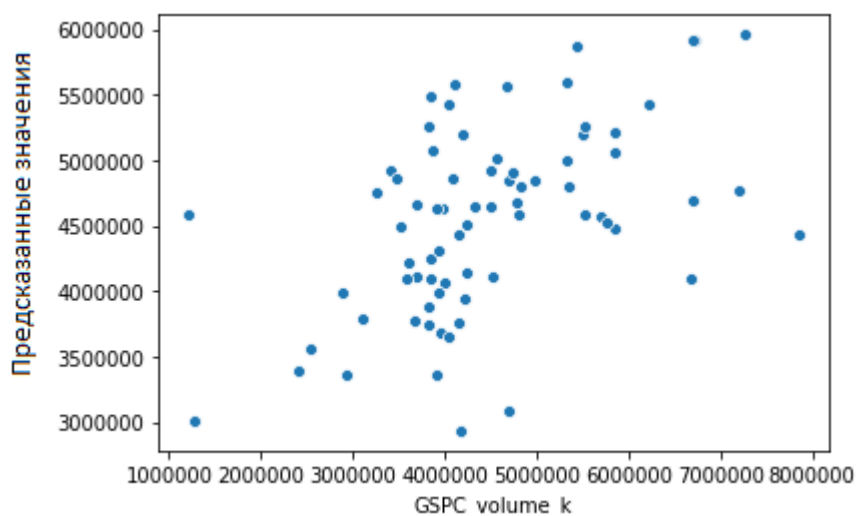


Рисунок 1. Диаграмма рассеяния для наилучшей модели.

Заключение

В данной работе результаты [8] были перепроверены на тех же данных, но с помощью других методов машинного обучения, а также с добавлением на рассмотрение индекса GSPC. Полученные результаты свидетельствуют о том, что данные о тональности сообщений в Twitter не могут надёжно предсказывать значения фондового индекса Доу-Джонса и GSPC.

Литература

1. Eugene and 1965, F. (1965) 'The behavior of stock-Market Prices', Chicago Journals, 38(1), pp. 34–105.
2. Gustavo Porras. Social Mood Impact on Financial Decision Making: A Study of Twitter Sentiment on Stock Index Volume / Master of Science in Finance, National College of Ireland, 2019.

3. Гипотеза эффективного рынка // Википедия. [2020]. Дата обновления: 12.05.2020. URL: <https://ru.wikipedia.org/?oldid=106988880> (дата обращения: 17.06.2020).
4. *Le, Quang Tiep and Mehmed, Mustafa*. The relationship between trading volume, stock index returns and volatility: Empirical evidence in Nordic countries. // Master Thesis in Finance, Lund University – Spring 2009
5. *Qian, Bo, Rasheed, & Khaled*. (2007) *Applied Intelligence* 26, 25–33.
6. *Bollen, Johan, Huina Mao, and Xiaojun Zeng (BMZ)*. 2011 (TMP). Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2(1): 1–8. <https://arxiv.org/pdf/1010.3003.pdf>
7. *Lachanski, Michael*. 2016. Did Twitter Mood Really Predict the DJIA?: Misadventures in Big Data for Finance. *Penn Journal of Economics (PennJOE.com)* 1(2): 8–48.
8. *Michael Lachanski and Steven Pav*. Shy of the Character Limit: “Twitter Mood Predicts the Stock Market” Revisited. *ECON JOURNAL WATCH* 14(3). September 2017: 302–345.
9. *Mackintosh, James*. 2012. Last Tweet for Derwent’s Absolute Return. *Financial Times*, May 24.
10. <https://www.semanticscholar.org/paper/Twitter-mood-predicts-the-stock-market-Bollen-Mao/e498784edf2c02fe0b228479f88120f08b381cb6>
11. http://www.huffingtonpost.com/2010/10/18/post_556_n_766591.html
12. <http://newsfeed.time.com/2010/10/21/how-can-twitter-predict-the-stock-market/>
13. <http://techland.time.com/2011/03/24/let-twitter-tell-you-where-to-invest-your-money/>
14. <http://archives.cnn.com/TRANSCRIPTS/1010/19/qmb.01.html>
15. <https://www.kaggle.com/rtatman/stock-market-twitter-summary-stats>