Анализ влияния новостей на стоимость акций

Малюгин Руслан

December 8, 2022

Аннотация

В данной работе приведен обзор инструментов для анализа и сбора новостей и информации о стоимости акций. Рассматривается несколько решений для обработки и дальнейшего анализа текста. Производится анализ и проверка гипотез о влиянии новостей на изменение стоимости акций. Так же реализован код для мониторинга новостей и получения прогноза в реальном времени.

1 Обзор инструментов

1.1 Сбор новостей

В качестве источника новостей был выбран сайт CNBC. Этот сайт один из самых популярных, соответственно здесь большой охват аудитории и, возможно,новости отсюда влияют на людей при покупке и продаже ценных бумаг.

У сайта нет официального API, но есть библиотека для работы - cnbc. Она имеет лимиты на запросы в месяц, но позволяет получать новости о конкретных компаниях.

В дальнейшем работа ведется с этой библиотекой.

1.2 Обработка новостей

Для обработки новостей были попробованы 2 метода: построение эмбеддингов новости и после обучения модели на них, и использование предобученной модели Roberta. Это модель, разработанная компанией Twitter для анализа тональности. Для работы первого метода необходим датасет для обучения, в связи с чем был написан бот для парсинга новостей и составления датасета. Для хорошей работы потребуется время для сборы данных, так что на данный момент анализ ведется с помощью второго метода.

2 Гипотезы о влиянии новостей

Для проверки были выдвинуты следующие гипотезы: при негативном окрасе новости цена акции падает, и наоборот, при позитивном возрастает, и то, что изменение цены в случае негативного или позитивного окраса новости сильнее, чем в случае нейтрального изменения.

Для эксперимента были выбраны компании Google, Meta, Apple и Tesla, так как они являются наиболее популярными и про них больше новостей.

2.1 Гипотеза Negative vs Positive

В качестве H_0 будем предполагать, что в случае оценки новости как Negative, через сутки цена упадет, а в случае Positive - вырастет. Обозначим μ_t - цена в момент выхода новости, μ_{t+1} - цена через день. Таким образом имеем следующие бернуллиевские величины:

$$\xi_i = \begin{cases} 0,\ Prediction\ =\ Negative,\ \mu_t < \mu_{t+1} \\ 1,\ Prediction\ =\ Negative,\ \mu_t > \mu_{t+1} \\ 1,\ Prediction\ =\ Positive,\ \mu_t < \mu_{t+1} \\ 0,\ Prediction\ =\ Positive,\ \mu_t > \mu_{t+1} \end{cases}$$

В качестве гипотезы рассматриваем $H_0 : \theta\ >\ 0.5\ vs\ H_1 : \theta < 0.5$

Для проверки гипотезы воспользуемся биномиальным тестом. Он состоит в том, что если $\xi_i \sim Bern(\theta)$, то $\sum_{i=1}^{n} \xi_i \sim Bin(n, \theta)$. После этого мы можем рассмотреть 0.95 квантиль распределения и сделать вывод.

В экспериментах были получены следующие величины:

$\hat{\theta} = 0.63,\ n = 27$

После проверки гипотез было получено p-value на уровне 0.94, что говорит нам о том, что мы более уверены в первой гипотезе, нежели в ее отвержении(важно помнить, что мы не можем утверждать о принятии первой гипотезы, так как выборка была рассмотрена небольшая).

2.2 Гипотеза Neutral vs NonNeutral

В данном случае будем замерять изменении цены при нейтральном окрасе новости и изменении цены при положительном или негативном окрасе. Обозначим за $\xi_i = |\mu_t - \mu_{t+1}|$ - изменение цены, $\overline{\xi_{NN}}$ - среднее изменение при положительном или негативном окрасе (NN = NonNeutral), и $\overline{\xi_N}$ - среднее изменение при нейтральном окрасе. Будем рассматривать гипотезы $H_0 : E(\xi_{NN} - \xi_N) = 0 \ \ vs \ H_1 : E(\xi_{NN} - \xi_N) > 0$

Для теста воспользуемся предположением о нормальности выборок(это неслишком строгое требование и в целом все хорошо работает и в случае распределений, близких к нормальному) и используем t-test.

T-test состоит в следующем. Пусть у нас есть 2 случайные величины X_1 и X_2 с матожиданиями M_1, M_2, и размерами выборок n_1 и n_2 соответственно. Будем проверять гипотезу $H_0 : M_1 - M_2 = 0$. Введем статистку $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}$, где s_1, s_2 - выборочные дисперсии, $\overline{X}_1$, $\overline{X}_2$ - выборочные матожидания. При справедливости нулевой гипотезы эта статистика имеет распределение Стьюдента. После этого, аналогично первого случая сравниваем с квантилями и выносим решение.

В ходе нашего эксперимента были получены $\overline{\xi_{NN}} = 1.37$ и $\overline{\xi_N} = 1.17$. После проверки было получено p-value на уровне 0.67, что говорит о том, что вероятно, новости все-таки как-то влияют на изменении цены(важно помнить, что мы не можем утверждать о принятии первой гипотезы, так как выборка была рассмотрена небольшая).

3 Тестирование модели с новостными признаками

Для тестирования был выбран датасет с ценами на акции Apple, собранный за несколько лет. Его большой минус - это интервалы длиной день. Но в данной работе нам важно не построить модель для точного предсказания, а проверка того, что новости положительно влияют на прогнозирование.

Для экспериментов была выбрана базовая модель CatBoost. В качестве признаков будем иметь: временные(день, месяц, час), прошлые(допустим, это будет 10 прошлых измерений), новостные признаки(2 признака - вероятность Negative и Positive, выданные моделью, в случае отсутствия новостей = ставим нули). Таргет будет бинарный - цена вверх или вниз.

После обучения модели были получены следующие показатели качества по метрике accuracy: 0.682 и 0.696. Как показываю результаты эксперимента, даже в такой простой модели, добавление новостных признаков, помогло нам улучшить качество, хоть и не сильно, но при более серьезном подходе к построению модели, эмбеддингов, сбору данных можно получить более хороший результат.

4 Заключение

В ходе проделанной работы были опробованы многие библиотеки для работы с новостями и финансовыми данными. Реализован метод для сбора датасетов с новостями. Собраны и найдены датасеты с новостями и финансовыми данными. Описаны возможные подходы для решения задачи. Проверены гипотезы о значимости новостей на курс акций. После этого были построены модели и зафиксировано улучшение качества модели при добавлении новостных признаков.

В дальнейшем планируется собрать хороший датасет и испробовать другие методы работы с текстами, а так же более мощный модели для предсказания.

References

[1] [**https://link.springer.com/article/10.1007/s42001-019-00035-x**]

[2] [**https://arxiv.org/abs/0809.2792**]

[3] [**https://econpapers.repec.org/article/gamjecomi/v_3a8_3ay_3a2020_3ai_3a4_3ap_3a107-_3ad_3a**

[4] [**https://www.mdpi.com/2227-7099/8/4/107/htm**]

[5] [**https://www.diva-portal.org/smash/get/diva2:1637576/FULLTEXT01.pdf**]

[6] [**http://statistica.ru/theory/t-kriterii/**]

[7] [**https://github.com/topics/cnbc-api**]

[8] [**https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment**]

[9] [**https://www.kaggle.com/datasets/lorilaz/apple-news-headline-sentiment-and-stock-info**]