УДК 004.8

ПРОГНОЗИРОВАНИЕ ЦЕН АКЦИЙ НА БИРЖЕ С ИСПОЛЬЗОВАНИЕМ РАЗРОЗНЕННЫХ ИСТОЧНИКОВ.

Газизулин А. Ф. (Университет ИТМО)
Научный руководитель – к.т.н., доцент Ключев А. О.
(Университет ИТМО)

Введение. Инвестиции и торговля на фондовом рынке могут быть сложными, но при правильном анализе и возможности прогнозирования приносят прибыль. На цены акций влияют как несколько технических индикаторов, так и множество других источников информации: новости, аналитика, настроение пользователей и др. В теории финансов есть гипотеза эффективного рынка, она гласит, что цены активов не могут полностью зависеть от устаревшей информации, а рыночные цены реагируют на новую информацию, например статьи финансовых новостей. Так, например авторы работ [1-3] предлагают использовать разрозненные источники: экономические новости, временные ряды, технические индикаторы, данные о клиентах, их действиях с акциями и используют в своих исследованиях следующие алгоритмы: LSTM, ANN, SVM и RNN.

Основная часть. В результате проведения обзора работ у зарубежных коллег, была сформулирована следующая проблематика: необходимо предсказывать вариацию цены акции компании, а также интерпретировать результаты предсказания в удобной для пользователя форме.

Для решения данного вопроса предлагается методика, в которой в качестве исходных данных будут выступать:

- ценах OHCL (свечные), по интересующей компании;
- новости компаний;
- новости политики;
- новости экономики.

Для обработки и анализа данных использовались как стандартные методы первичной обработки, например такие как отчистка или замена пропущенных значений, так и методы отбора предикторов по их значимости: SVM-RFE, Gradient Search, Random Search. Для текстовых данных будет использоваться следующие методы обработки: токенизация, удаление/добавление стоп-слов, удаление часто/редко встречающихся слов (опционально), стемминг, нормализация. Также для текстовых данных будет проведён частотный анализ с помощью TF-IDF, размерность которого будет уменьшена с помощью метода главных компонент PCA для отсеивания наименее полезных слов в словаре.

В качестве методов обучения были выбраны несколько алгоритмов с возможностью интерпретирования модели: LR – линейная регрессия, ANN, а также SVM, для сравнения результатов, т.к. он всегда показывает средние результаты и может использоваться в качестве опорного метода. Все обученные модели будут оцениваться стандартными методами, такими как матрица ошибок (accuracy), MSE, RMSE и MAE.

Выдвинута гипотеза, что новостные статьи в совокупности с данными по ценам акций дают лучшую точность предсказания для модели машинного обучения. Для этого была сформулирована цель и поставлены задачи для её достижения:

- использовать различные типы данных, связанных с предметной областью в качестве обучающего набора;

- для преобразования данных использовать совокупность приёмов и методов, которые перекрывают недостатки методов из сводной модели;
- обучить несколько моделей на выбранных алгоритмах, проанализировать модели с помощью совокупности (или отдельных) методов, чтобы дать достаточно точную оценку.

Первоначальный набор преобразованных данных с получасовым тиком содержал: цену открытия/закрытия, объём торгов, максимальную цену открытия, минимальную цену открытия, новостная статья, разбитая на частотный словарь TF-IDF и с уменьшенной размерность методом PCA, а также преобразованные в числовой формат дата, неделя, месяц, час и минуты. Последнее необходимо для объединения котировок и новостей, т.к. новости не так регулярны, как котировки, то после объединения появляются незаполненные места, поэтому они заполняются предыдущей новостью до следующей новости, т.е. получается, что как новость вышла она на протяжении какого-то времени влияет на цену акции.

После обработки данных, была построена тепловая корреляционная карта. По ней определялись данные, которые слабо влияют на целевую переменную, и которые в процессе эксперимента можно исключить из выборки.

Далее для трех выбранных алгоритмов: линейная регрессия, SVM и нейронная сеть с помощью модуля случайного леса подбирались гиперпараметры, применение которых к модели даёт наилучший прирост точности. После этого, при настроенных параметрах алгоритмов строилась карта корреляции данных с целевой переменной, и исходя из её результатов из набора исключались некоторые столбцы.

В результате обучения и тестирования моделей, лучшие результаты показали модели с алгоритмами SVM и логистической регрессии с точностью 55% и 63%. С ANN был получен результат в 52 %. Для их обучения и тестирования из набора данных были исключены столбцы минимальной цены закрытия и выбором 4 главных компонент PCA для TF-IDF словаря новостных статей.

Выводы. Данное исследование позволит улучшить точность предсказания цен акций компаний, так как в ходе исследования была выявлена корреляция между данными из новостных источников и целевой переменной. Таким образом в совокупности с техническими индикаторами и свечными данными, которые значительно коррелируют с предсказываемой величиной, и данными из разрозненных источников можно заметно увеличить точность предсказания стоимости ценной бумаги.

Список использованных источников

1. Weng B., Ahmed M. A., Megahed F. M. Stock market one-day ahead movement prediction using disparate data sources // Expert Systems with Applications. 2017. Т. 79. P. 153-163.
2. Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P. & Pandey, N. Explainable stock prices prediction from financial news articles using sentiment analysis // PeerJ Computer Science. 2021. Т. 7, P. 1-21.
3. Al-Nefaie A. H., Aldhyani T. H. H. Predicting Close Price in Emerging Saudi Stock Exchange: Time Series Models // Electronics. 2022. Т. 11. №. 21. P. 3443.