3

Метою роботи є підвищення ефективності фінансової активності на фондовому ринку, зокрема при прогнозуванні цін акцій компанії Goldman Sachs Group, одного з найбільших в світі інвестиційних банків, який займається торгівлею цінними паперами та інвестиційним менеджментом.

4

Тема актуальна, оскільки:

1. Все більша кількість людей цікавляться ринком цінних паперів, дана робота дає змогу початківцям дослідити зв’язок технічного аналізу і прогнозування
2. Інвесторам важливо знати, як буде розвиватися ситуація на ринку
3. Результатами прогнозу можуть користуватись досвідчені аналітики, які на практиці комбінують інструменти аналізу ринку з використанням машинного навчання

5

Дві основні точки зору щодо можливості ефективно прогнозувати вартість цінних паперів виражені в гіпотезі випадкового блукання та гіпотезі ефективного ринку.

Перша стверджує, що зміна цін акцій відбувається хаотично і майбутнє фондового ринку не може бути спрогнозовано за допомогою наукових методів.

Друга гіпотеза більш оптимістична. Відповідно до неї всі ринкові тренди і рух в поточний момент моментально відображаються в котируваннях цінних паперів. Вартість ринкового актива відображає публічну інформацію, зокрема інформацію, представлену в пресі, звітах компаній, політичних заявах, економічних та фінансових показниках.

В ході виконання робрти було припущено, що гіпотеза ефективного ринку – вірна.

6

Для досягнення мети роботи необхідно:

1. Проаналізувати можливості отримати прогноз поведінки ціни акцій на основі економічних показників
2. Дослідити математичні методи прогнозування даних та обрати найкращий для даної задачі
3. Виконати збір необхідних даних та побудувати ознаки
4. Зібрати щоденні економічні новини, що стосуються обраної компанії, та класифікувати їх за настроєм
5. Проаналізувати та протестувати отримані дані, понизити розмірність, виконати трансформації з метою підготовки даних до застосування у моделі
6. Реалізувати та обучити обрану модель, отримати прогноз, зробити висновки щодо отриманих результатів та
7. Розробити веб-застосунок

7

Потрібно зрозуміти, що впливає на те, чи буде ціна акцій GS рухатися вгору або вниз. Необхідно включити якомога більше інформації.

Кореляція активів – це міра того, як інвестування рухаються одне по відношенню до іншого на окремому часовому проміжку. Можна виділити такі активи: показники схожих компаній, ціни на сировинні товари, курси іноземних валют, економічні показники, фондові індекси. Компанія Goldman Sachs залежить від багатьох зовнішніх факторів, включаючи своїх конкурентів, клієнтів, стан глобальної та національної економіки, геополітичну ситуацію, фіскальну політику і взаємодіє з ними.

Також були зібрані щоденні фінансові новини, що стосуються інвестиційного банку, за період 10-ти років з веб-сайту ділової газети Financial Times для подальшої класифікації.

8

Було обрано і порівняно п’ять методів, які можна застосувати для прогнозування фінансової інформації. Зокрема: векторна авторегресія, багатовимірна лінійна регресія, модель авторегресійного інтегрованого ковзаючого середнього (ARIMA – можно ли сократить), прихована марківська модель, штучні нейронні мережі.

Необхідно було розглянути моделі, які гарно зарекомендували себе при роботі з багатовимірними фінансовими часовими рядами, головним пріорітетом у виборі є можливість методу знаходити нелінійні залежності, приховані патерни та глобальні і локальні тренди.

За сукупністю характеристик та властивостей було прийнято рішення застосувати рекурентні нейронні мережі, зокрема їх модифікацію LSTM.

9

Серед програмних рішень було оглянуто онлайн-сервіс StocksNeural та програмне забезпечення з відкритим вихідним кодом Stocksight. Також існує варіант звернення до фахівця або самостійний аналіз, але при таких підходах існує ризик обману або помилки. Суттєвим недоліком рішень є відсутність комбінування різноманітних вхідних даних, корисних для прогнозування або недостатня їх кількість, а також обмежений доступ до програм.

Тому підхід до прогнозування, який полягає в використанні якомога більшої кількості ознак для навчання моделі представлений в даній дипломній роботі і є найбільш вдалим та актуальним.

10

При створенні ознак був виконаний технічний аналіз, в результаті якого створені технічні індикатори. Технічні індикатори – це алгоритми, що дозволяють отримати інформацію про майбутні ціни за допомогою котирувань за певний період. Існує два основних типи індикаторів: індикатори тенденції, які підтверджують напрямок тренду та осциляторні індикатори, які дають змогу передбачити зміну напрямку тренду. Використано декілька найбільш ефективних індикаторів кожного типу, список приведений в роботі.

11

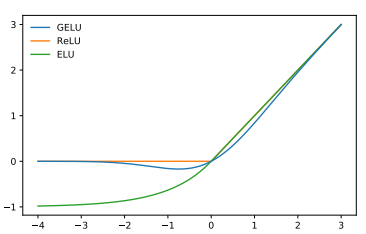
Перетворення Фур'є застосовується при аналізі часових рядів з метою виділення локальних та глобальних трендів та зменшення шуму в даних. Складні сигнали будь-якої форми можливо представити сумою синусоїдальних коливань – гармонік. Якщо занулити амплітуду високочастотних складових, а потім на основі зміненого спектра відновити вихідну функцію, виконавши зворотне перетворення Фур'є, то вона стане більш гладкою за рахунок видалення високочастотної компоненти. В роботі було проведено спектральну обробку часового ряду прогнозованих даних для створення корисних для моделі ознак.

12

На зміни в ціні акцій компанії можуть впливати фінансові новини. Однією з причин прийняття певного рішення про інвестиції можуть бути новини, пов’язані з об’єктом інвестицій.

Текст новин очищено та оброблено для подання в модель. Класифікація за настроєм на негативні, нейтральні та позитивні фінансові новини відбувається за допомогою переднавченої на фінансовому тексті NLP моделі finBERT.

(функція активації GeLu - Gaussian Error Linear Units)



13

На слайді бачимо отриманий датасет з результатами класифікації та стовбчасту діаграму з кількістю значень кожного класу.

14

Для зменшення розмірності даних застосовано **метод головних компонент Principal Component Analysis**.

Завдання алгоритму – знайти підпростори меншої розмірності, в ортогональній проекції на які дисперсія (середньоквадратичне відхилення) є максимальним.

Переваги PCA:

* Видаляє корельовані ознаки
* Підвищує продуктивність алгоритму
* Зменшує перенавчання

Недоліками PCA є те, що:

* Незалежні змінні стають менш інтерпретованими
* Перед застосуванням РСА необхідно виконати стандартизацію або центрування ознак
* Втрата деякої частини інформації

15

Для прогнозування фінансових числових рядів було обрано нейронну мережу довгої короткострокової пам’яті LSTM.

Один з методів реалізації рекурентної нейронної мережі – багато-до-багато, тобто many-to-many, що дозволяє подавати послідовності на вхід моделі та отримувати послідовності в результаті прогнозування.

LSTM розроблений для **уникнення проблеми довгострокової залежності**. Ключовий елемент – стан ячейки – елемент, що проходить крізь весь ланцюг. Інформація може переміщуватись по ньому, що позитивно впливає на знаходження віддалених зв’язків, прихованих поведінкових особливостей.

16 (схема)

На слайді можна побачити схему, яка відображає послідовність дій, виконаних при створенні, обробці ознак для отримання фінального результату.

17

Представлено корр матриці активів, вони мають високу кореляцію, якої можна буде позбутися за допомогою методу головних компонент.

18

Виконано кодування категоріальних змінних методом LabelEncoding.

Дані:

* розділено на тренувальну, валідаційну та тестову вибірки
* стандартизовано (тренувальні та тестові окремо, використовуючи тільки тренувальні)
* Розмірність датасету понижено методом головних компонент з 81 до 31
* Дані стаціонаризовано
* Досліджено на стаціонарність розширеним тестом Діки-Фуллера (рівень значимості 0.05). Всі ознаки, включаючи цільову – стаціонарні.

19

Оскільки LSTM може приймати та повертати послідовності, на вхід моделі будуть подвавтися дані за попередні 60 днів, на їх основі робиться прогноз на 30 днів.

Для цього необхідно переупорядкувати дані. Незалежні змінні розміщені разом з зсунутими назад наступними залежними змінними (рис 1)

З кожною ітерацією вікно здвигається на 1 значення, створюючи окремий елемент послідовності виборок. (рис. 2)

В результаті для подання в модель отримано дані розмірності (рис. 3)

20

Побудована нейронна мережа типу LSTM. Конфігурація нейронної мережі детальніше описана в роботі. Слід зазначити що НМ має два приховані шари по 120 вузлів, на вхід подаються елементи вибірки розмірністю 60х32, дропаут 0.2:

Отримано прогнозовані дані, які зворотньо перетворені до часового ряду. Результати прогнозування візуалізовані

Підбор гіперпараметрів нейронної мережі проводився у вкладених циклах з послідовню заміною значень та навчанням різних моделей. Для оцінки моделей в процесі підбору гіперпараметрів та оптимальної побудови використані метрики:

* середній квадрат похибки MSE;
* середня абсолютна похибка MAE;
* коефіцієнт детермінації R2;
* коваріація;
* коефіцієнт кореляції Пірсона.

Оцінки прогнозу фінальної моделі приведені в таблиці.

21

Для спрощення візуальної оцінки користувачем та демонстрації отриманих даних був розроблений веб-інтерфейс. В ньому реалізована можливість використовувати технічні індикатори та їх комбінації, а також отримання прогнозу акцій компанії Goldman Sachs. Для розробки був використаний фреймворк мови Python для розробки веб-додатків Dash та бібліотека Plotly

22

Висновки

В ході виконання дипломної роботи був реалізований підхід до прогнозу часового ряду цін акцій певної компанії, який базується на зборі якомога більшої кількості даних, які впливають або відображають рух на ринку акцій.

* З можливих математичних методів обрано найбільш підходящий для даної задачі - модифікація RNN модель довгої короткострокової пам’яті LSTM
* Досліджено проблемну область, зібрані необхідні дані. Для створення додаткових ознак застосовано перетворення Фур’є та класифіковано фінансові новини за настроєм.
* Вхідні дані оброблені та протестовані, розмірність понижена з застосуванням методу головних компонент. Тренувальна та валідаціна вибірки включають в себе дані з 1 січня 2010 до 31 серпня 2019. Прогнозування відбувається на період з 1 вересня 2019 до 2 квітня 2021.
* Нейронна мережа обучена, результат прогнозування оцінений метриками. Розроблено веб-застосунок на Dash, який є інтерактивним та дозволяє користувачу оцінити результат прогнозування і взаємодіяти з інформацією.

Програмна реалізація відбувалась в середовищах Jupyter та PyCharm Community, було використано такі бібліотеки мови програмування Python, як pytrends, numpy, ta, BeautifulSoup, transformers, TyTorch, sklearn, keras (tensorflow).

23

дякую за увагу.