Проект по предметот Вовед во науката за податоци на тема:

Prediction financial performance of the companies

Автори:
Никола Талевски 211009
Евгенија Попчановска 211011
Ана Костадиновска 211006

Ментор:
д-р Димитар Трајанов

Септември 2024 година

Вовед

Целта на проектот е предвидување на ESG индексите на компании кога се дадени цената на акцијата и сентимент на новости. Обратно, правиме предвидување и на цената на акцијата за даден сентимент на новости и ESG индекс. Линк до репозиториумот на проектот: github.com/popchanovska/PredictionFinancialPerformance.

Собирање на податоци

Статии

Во нашиот проект го користиме следниот пребарувач (scraper) за статии: github.com/lewisdonovan/google-news-scraper. Претставува Node.js пакет за пребарување на статии на Google News платформата врз основа на search query или специфични теми на Google News. Овој пакет користи Puppeteer за преземање на статиите.

- Објаснување за дадениот scraper и како се користи има тука: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/1.%20Scraping/News/scraper-google-news.pdf

- Скрипта која се извршува за да се соберат податоци од низа компании има тука: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/1.%20Scraping/News/scraper-google-news.js

- Резултатите се во JSON формат и се наоѓаат на следниов линк: https://github.com/popchanovska/PredictionFinancialPerformance/tree/main/1.%20Scraping/News/data

ESG индекси

- Скрипта која се извршува за да се соберат ESG индексите од yahoo finance: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/1.%20Scraping/ESG%20ratings/scraping_esg.ipynb

- Скрипта која се извршува за да се собери вкупниот ESG индекс од yahoo finance: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/1.%20Scraping/ESG%20ratings/Full_ESG_Rating_scrape.ipynb

- Резултатите од двете горенаведени скрипти се наоѓаат тука: https://github.com/popchanovska/PredictionFinancialPerformance/tree/main/1.%20Scraping/ESG%20ratings/data

Цени на акциите

- Цените на акциите на секоја компанија се симнати во целост од yahoo finance. Резултатите се наоѓаат овде: https://github.com/popchanovska/PredictionFinancialPerformance/tree/main/1.%20Scraping/Stocks/data

News sentiment analysis

- News Reader: Дадената скрипта го користи JSON фајлот со статии* за да креира подобрено податочно множество во DataFrame тип: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/2.%20Title%20analysis/NewsReader.ipynb

- Title Classification: Дадената скрипта ги лаберира насловите од вестите како: government, social, environment, neutral: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/2.%20Title%20analysis/TitleClassification.ipynb

- Sentiment Analysis: Дадената скрипта пробува повеќе модели за класификација на насловите од статиите како позитивни или негативни: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/2.%20Title%20analysis/news-scraper-and-sentiment-analysis.ipynb

Резултати од претходно наведените скрипти има на овој branch во соовдетните фолдери: https://github.com/popchanovska/PredictionFinancialPerformance/tree/main/2.%20Title%20analysis/data

Креирање на финално податочно множество

- Со помош на оваа скрипта се создава податочното множество за насловите од вестите за компаниите. Тука, различните множества за вестите како Title Classification Sentiment Analysis се спојуваат во едно множество, заедно со мета податоците за вестите: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/news_data.ipynb

  Резултатите од оваа скрипта се наоѓаат тука: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/data/news_data.csv

- Со помош на оваа скрипта, во едно податочно множество се спојуваат сите цени на акцците за секоја компанија: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/stocks-concat.ipynb

  Резултатите може да се видат тука: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/data/stocks.csv

- Со помош на оваа скрипта, претходно генерираните скрипти за вестите, цените на акциите и ESG ratings, се генерира ново податочно множество со податоци за секоја вест за секоја компанија, вредноста на акциите на таа компанија во тој ден, и посебните Environment, Social и Government ratings за таа копманија: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/data-merge.ipynb

  Резултатите се наоѓаат овде: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/data/data.csv

- За крај, на претходното множество се додаваат и вкупните ESG ratings со помош на оваа скрипта: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/ESG_merge.ipynb

  Резултатите од оваа, а воедно и финалното множество кое се користи во моделите се наоѓа овде: https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/3.%20Dataset%20creation/data/data_with_esg.csv

XGBoost модели

1. На овој branch има скрипта која ја прави финалната верзија на податочното множество. Тоа се состои од цената на акцијата за секој ден, и процент на: позитивни government вести, негативни government вести, позитивни environment

вести, негативни environment вести, позитивни social вести и негативни social вести за соодветниот ден. Дополнително има скрипта со неколку XGBoost модели со различни параметри за предвидување на E, S и G индексите. https://github.com/popchanovska/PredictionFinancialPerformance/tree/main/5.%20Additional%20code.

2. Со помош на генерираното множество од последниот сегмент, креиран е модел за предвидување на вкупниот ESG rating за секоја компанија базиран на анализата на насловите на вестите и цените на акциите во определени денови. https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/4.%20Models/ESG/XGBoost_model_for_predicting_ESG.ipynb
3. Креирани се два модели кои ја предвидуваат цената на акциите на компанијата во определен ден базирани на претходното множество:
   А) Првиот модел е креиран да ја предвидува цената на акциите на компанијата базиран на акциите на компанијата од претходниот ден, ESG индексите и рејтингот, како и анализата на насловите на вестите за тој определен ден.
   Б) Вториот модел, поради сомневање за overfitting, ги изоставува вредностите на акциите од претходниот ден, додека го користи истото множество.

   Двата модели може да се најдат овде:
   https://github.com/popchanovska/PredictionFinancialPerformance/blob/main/4.%20Models/Stocks/XGBoost_model_for_predicting_stock_price.ipynb