바이오 기사 감성분석을 통한 주가 예측 Prophet 모델 개선

팀명 : 스파이더 구성원: 두히가체구(김주희), jang_123(장석규)

주제: (과제2) 해외 뉴스 데이터를 이용한 투자 정보 분석

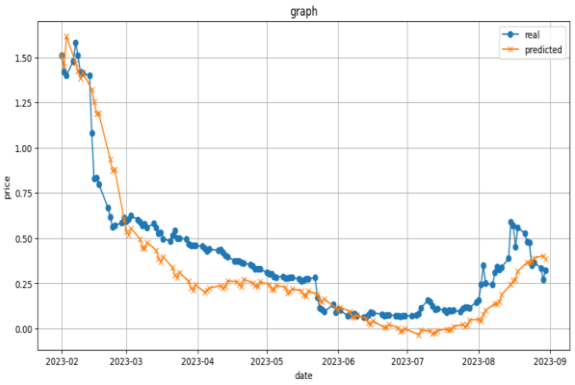
1. 분석 배경

Prophet 모델은 매주, 매월의 비선형적 트렌드와 휴일 등의 효과를 가산하여 예측을 진행하는 모델이다. Prophet 모델은 날짜 변수와 측정값만을 이용하여 만들 수 있는 모델이기 때문에 구현이 쉽고 연산속도가 빠르다. 그러나 주식 예측에 해당 모델을 사용할 경우 주가에 변동을 줄 수 있는 이벤트에 대처하기 어렵고 오차가 발생한다. 이러한 한계점을 극복하고자 관련 티커 종목의 해외 뉴스 데이터를 이용하여 감성분석을 진행하고, 해당 감성 score를 이용해 prophet 모델을 보완하여 예측 오차를 줄이기 위해 해당 프로젝트를 진행하게 되었다.

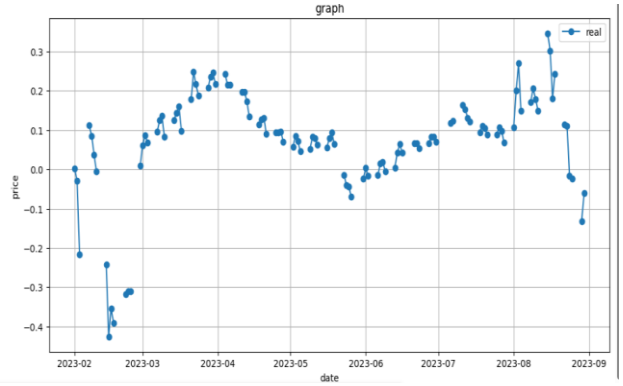
2. 분석 과정

2-1. prophet 모델 제작

Prophet 모델의 이점을 살려 날짜와 종가만을 이용하여 모델을 제작한다. 1월은 온전히 학습의 용도로 사용하고 2월부터는 학습 세트를 해당 날짜의 전날까지 계속해서 1일씩 늘려가며 예측을 진행한다. 해당 과정을 거치는 이유는 1~8월까지의 데이터만 주어졌기에 train set이 부족하여 생길 수 있는 모델의 부정확함을 극복하기 위함이다. 이후 실제값과 예측 값과의 차이를 새로운 칼럼으로 정의한다.



<그림 1. 티커코드가 TTOO인 종목에 대한 실제값(파란색)과 예측값(주황색)>



<그림 2. 예측값과 실제값 간의 차이>

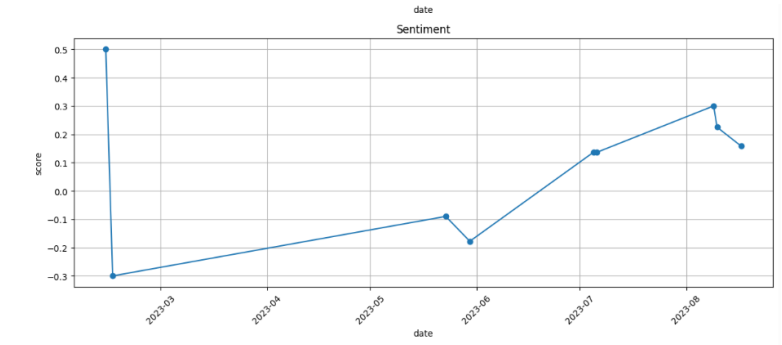
2-2. 뉴스 감성 분석

-뉴스 불용어 제거 및 토큰화: 뉴스 제목와 내용을 하나의 열로 합쳐서 분석을 진행하였다. 뉴스

데이터에서 각 문장을 소문자로 변환하고, 감정 분석에 사용되지 않는 불용어를 최대한 제거하였다. Word_tokenize 함수를 사용하여 단어를 토큰화하였다.

-Word2Vec 워드 임베딩 이용한 모델 학습: Word2Vec 모델을 초기화하고, 토큰화한 단어 데이터를 사용하여 Word2Vec 모델을 학습시켰다. Word2Vec은 단어 간의 의미적 관계를 고려한 워드 임베딩을 생성하여 유사한 단어들이 비슷한 임베딩 벡터를 갖게 된다. 원본 텍스트 데이터에 비해 상대적으로 낮은 차원의 임베딩 벡터를 생성하여 고차원 데이터의 차원을 감소시켰다. 생성된 임베딩 벡터를 사용하여 단어 유사성을 계산하는 자연어 처리 작업을 진행하였다.

-감정 분석: 각 문장에서 불용어를 제외하고 토큰화한 단어 데이터를 가지고 단어별 감정 점수를 계산하여 감정 점수를 합산하였다. 대량의 텍스트 데이터에서 텍스트에 포함된 감정을 자동으로 분석할 수 있다. 빠르게 대량의 데이터를 처리하고 감정 정보를 추출할 수 있는 감정 분석 코드를 작성하였다. 각 단어의 감정 점수를 계산하고, 이를 조합하여 문장 또는 문서의 감정 특성을 추출하였다. 텍스트 데이터의 감정 수치를 쉽게 파악할 수 있는 감정 분석 모델을 개발하였다.



<그림 3. 티커코드가 TTOO인 종목에 대한 감정 분석 시각화>

2-3. 오차 보완

해당 감정 점수에 파라미터 조정을 통해 얻은 상수 값을 곱한뒤 예측값에 더해주었을 때, 성능 향상이 가능하다는 것을 rmse를 통해 증명했다.

```
Root Mean Squared Error (RMSE): 0.15278785608215006
```

```
Root Mean Squared Error (RMSE): 0.1480762152458108
```

3. 기대 효과

해당 예에 대한 RMSE의 차이는 0.004로 굉장히 작지만 주어진 기사데이터의 수가 종가 데이터의 수에 비해 굉장히 적은 것을 감안하면 정확도 향상의 측면에서 의의가 있다. 이로써 기사정보를 반영하는 prophet모델이 사용될 수 있다.