데이터 사이언스 기초(03분반)

기말 프로젝트

학과 : 빅데이터학과

학번 : 20195237

이름 : 장윤성

요약

이 프로젝트는 코스피에 등재되어 있는 세가지 기업, 삼성전자, KB 국민은행, NC소프트에 대해 뉴스 심리 지수를 활용하여 주가를 예측하는 프로젝트이다.

주가 예측은 금융 시장에서 매우 중요한 요소이며, 뉴스 심리 지수는 최근 금융 예측에 많은 관심을 받고 있다.

하여 이 프로젝트의 목표는 뉴스 심리 지수와 주가 간의 상관관계를 탐색하고, 뉴스 심리 지수를 활용한 주가 예측 모델을 개발하는 것이다.

본 프로젝트에서 먼저 python에서 제공하는 FinanceDataReader 패키지를 import하여 삼성전자와 KB 국민은행, NC소프트의 주가 데이터와 한국은행 홈페이지에 경제통계 – 심리 지수의 뉴스 심리 지수 데이터를 수집하였다. 데이터 수집 이후 뉴스 심리 지수와 주가 간의 관계를 분석하고, 주가 예측 모델을 구축하였다. 주가 예측 모델은 다중 선형 회기 모델을 사용하였다.

예측 결과 모델의 정확성은 매우 낮았다. 하지만, 그래프를 그려보니 주가와 뉴스 심리 지수의 상관관계를 확인할 수 있었다.

이 프로젝트는 나중에 금융 시장에서의 의사 결정에 도움을 줄 수 있을 것으로 기대된다.

뉴스 심리 지수를 이용해 주가 예측 모델을 만들게 된 동기

금융 시장에서는 종종 "사람들의 반대로 행동하라" 라는 말이 있습니다. 이는 주식 시장에서의 투자 결정에 대한 관점을 나타내는 말입니다. 또한, 많은 사람들이 탐욕에 눈이 멀었을 때, 주가는 떨어지는 현상을 관찰할 수 있습니다. 이러한 경험을 통해 주식 투자에는 회사의 능력 외의 다양한 요소들이 영향을 미치는 것을 알 수 있습니다.

최근 뉴스에서는 경제적, 정치적, 사회적인 측면에서 안좋은 소식들이 많았습니다. (인플레이션, 우크라이나 전쟁 등) 이에 따라 투자자들의 심리는 불안과 두려움으로 어지러워져 있는 상황입니다. 그렇기 때문에, 이러한 투자 환경에서 뉴스 심리 지수라는 새로운 개념이 주목받고 있습니다.

뉴스 심리 지수는 뉴스 내용을 분석하여 투자자들의 심리 상태를 수치화하는 지표로, 주가 예측에 유용한 정보를 제공할 수 있다는 기대를 받고 있습니다.

이와 같은 이유로, 본 프로젝트는 삼성전자, KB국민은행, NC소프트 세 가지 기업에 대해 뉴스 심리 지수를 활용하여 주가를 예측하는 프로젝트를 하기로 결정하였습니다.

사용한 데이터셋

삼성전자 DataSet

Date	Open	High	Low	Close	Volume	Change
2020-01-02	55500	56000	55000	55200	12993228	-0.01075
2020-01-03	56000	56600	54900	55500	15422255	0.005435
2020-01-06	54900	55600	54600	55500	10278951	0
2020-01-07	55700	56400	55600	55800	10009778	0.005405
2020-01-08	56200	57400	55900	56800	23501171	0.017921

KB 국민은행 DataSet

Date	Open	High	Low	Close	Volume	Change
2020-01-02	47250	47450	46450	46550	1033855	-0.02308
2020-01-03	47050	47800	46900	47150	1123490	0.012889
2020-01-06	47100	47550	46300	46600	698224	-0.01166
2020-01-07	46600	47350	46600	47000	817477	0.008584
2020-01-08	46100	46550	45800	46150	1197780	-0.01809

NC소프트 DataSet

Date	Open	High	Low	Close	Volume	Change
2020-01-02	542000	545000	539000	541000	40246	0
2020-01-03	547000	568000	542000	565000	112404	0.044362
2020-01-06	562000	587000	562000	579000	107006	0.024779
2020-01-07	583000	596000	574000	594000	84378	0.025907
2020-01-08	587000	604000	584000	604000	109267	0.016835

뉴스 심리 지수 DataSet

	A	B	C	D	E	F	G	H	I	J
1	통계표	계정항목	단위	변환	2023-05-26	2023-05-25	2023-05-24	2023-05-23	2023-05-22	2023-05-19
2	6.4. 뉴스심	뉴스심리지수		원자료	98.05	100.01	100.57	102.15	103.73	99.54

사용한 라이브러리, 함수들

Tidyverse library

read_csv

is.na

table

ggplot

gather

as.Date

merge

select

str

sapply

apply

mean

sd

lm

summary

predict

mutate

tail

plot

lines

프로젝트 진행 과정

데이터를 tibble 형식으로 불러옴

결측값 처리

뉴스 심리 지수의 데이터 가공

불러온 회사 데이터에 뉴스 심리 지수를 병합

주식 데이터 종가와 뉴스 심리 지수를 그래프로 확인

(상관관계 확인)

데이터 정규화

모델 학습

결과 확인

예측 값 복원(*sd +mean)

새로운 데이터 예측

모든 데이터 종합 데이터 시각화

결과

가설은

H0 : 뉴스 심리 지수는 유의하다.

H1 : 뉴스 심리 지수는 유의하지 않다.

로 세웠습니다.

주가 데이터를 예측한다는 것은 불가능에 가깝다고 보아 유의 수준을 0.1로 잡았습니다.

삼성전자의 summary를 보면 Open은 Close와 닮아있습니다.

삼성전자는 우리나라를 대표하는 대기업이기 때문에, 주가의 등락폭이 크지 않기 때문에 Open과 Close가 매우 닮아있는 것은 당연합니다.

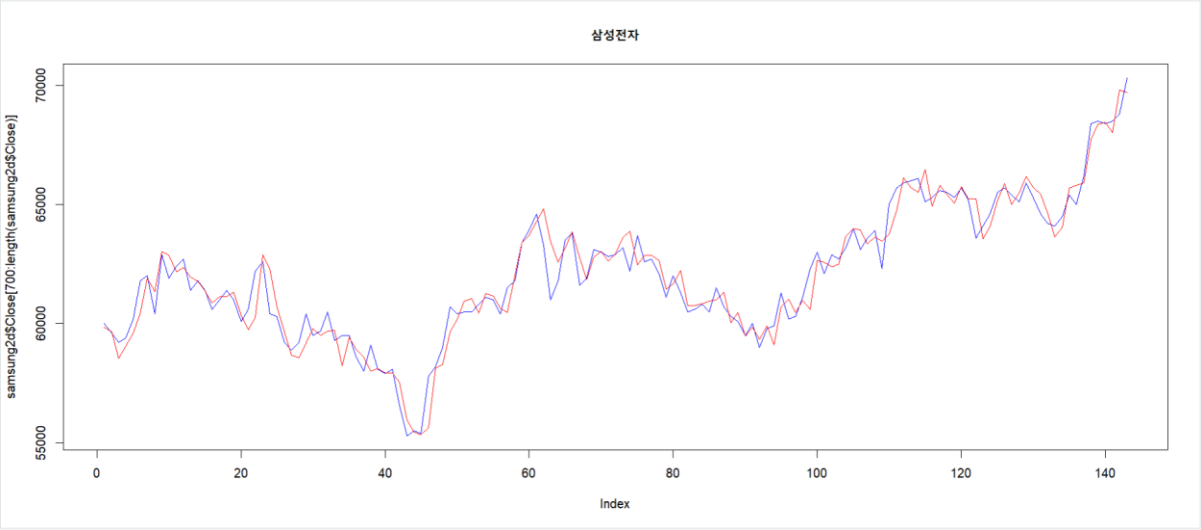
또한 기대하지 않았던 뉴스 심리 지수와의 관계도 0.05보다 작은 값을 보여주며 유의한 수준을 보여주었습니다.

KB 국민은행과 NC소프트의 summary를 보면 삼성전자와 같이 Open과 Close가 닮아있습니다.

하지만 두 데이터 모두 뉴스 심리 지수의 P-value 가 1보다 작은 값을 나타내며 큰 오차를 보였습니다.

이러한 결과를 보아 삼성전자를 제외한 두 회사의 뉴스심리지수 회기 계수는 유의하지 않다는 사실을 입증하였습니다.

모델의 정확도는 높게 나왔지만, 이는 Open과 Close가 너무 닮아 있기 때문이다. 뉴스 심리 지수와의 상관관계는 크지 않습니다.

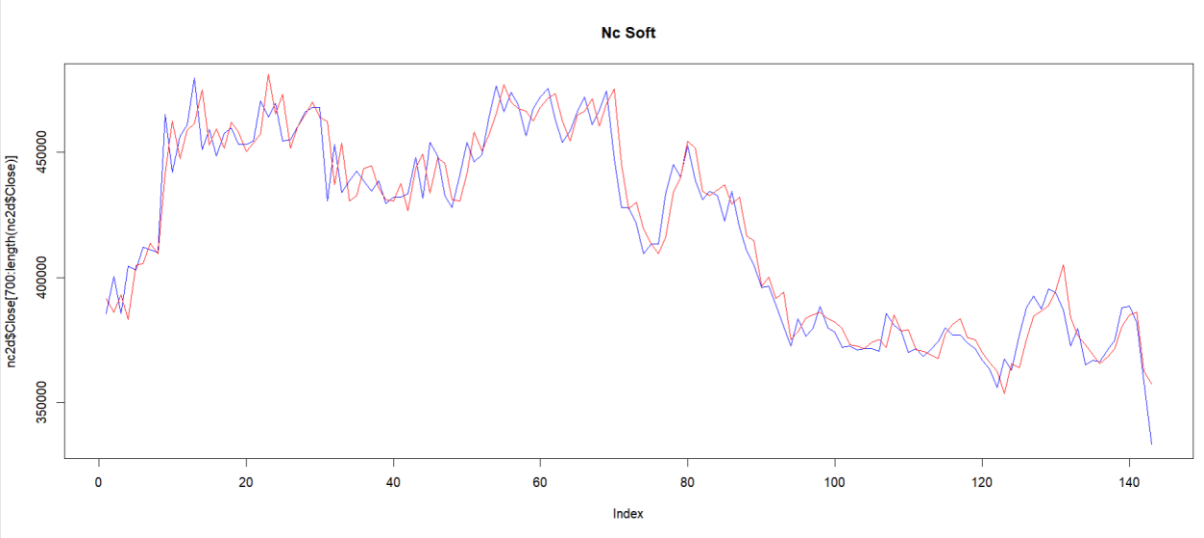


위의 그림은 삼성전자 Close에 대한 그래프와 Close를 예측한 model의 예측 값을 나타낸 그래프입니다.

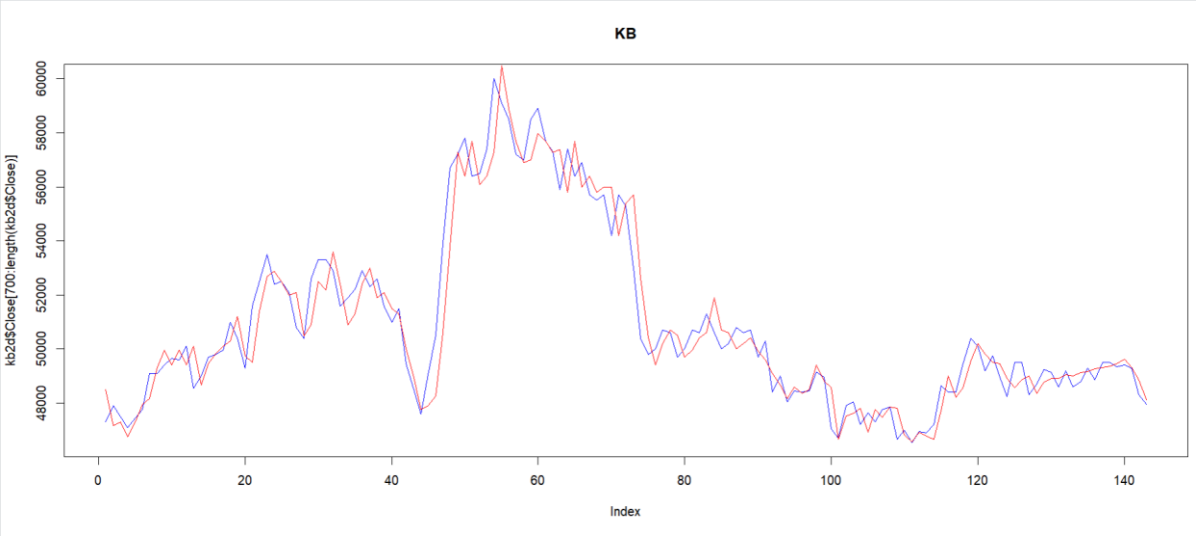
파랑색은 삼성전자의 Close, 빨강색은 model의 예측 값입니다.

자세히 보면 model의 예측 값이 실제 Close보다 오른쪽으로 밀려난 듯한 모습을 보여줍니다.

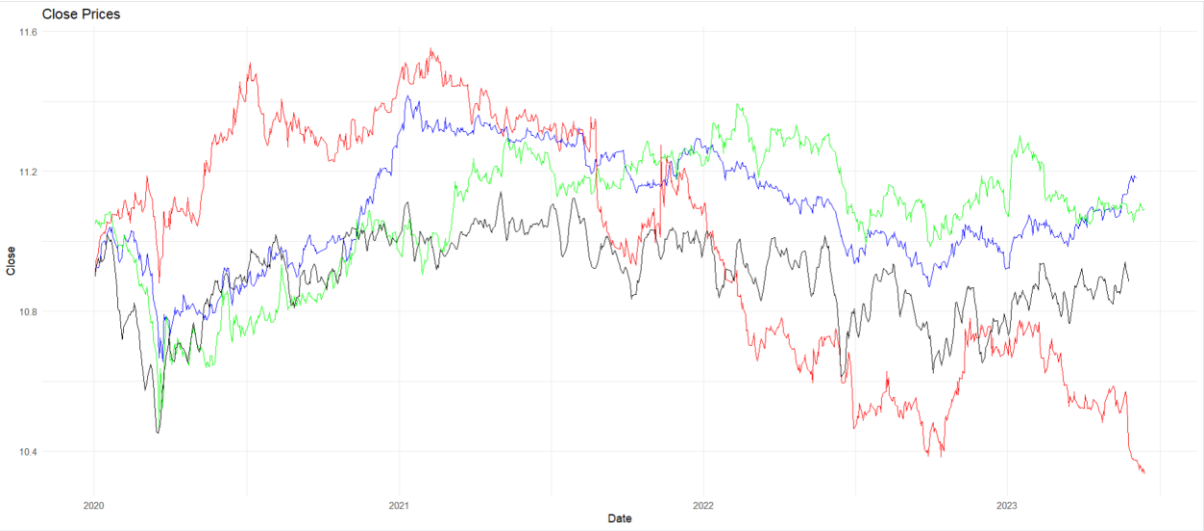
이러한 모양은 뒤의 두 회사들의 Close와 model의 예측 값의 그래프와 유사한데 이를 통해 뉴스 심리 지수는 경기 선행지표가 아닌 경기 후행 지표 라는 사실을 알 수 있습니다.



NC소프트와 model의 예측 값을 나타낸 그래프입니다.



KB 국민은행과 model의 예측 값을 나타낸 그래프입니다.



세개의 회사와 뉴스심리지수를 비슷한 값으로 맞추어 시각화해 보았습니다.

파랑색 : 삼성전자

빨강색 : NC소프트

초록색 : KB 국민은행

검정색 : 뉴스심리지수