#1

안녕하십니까? 주식 4조의 발표를 맡은 박준배라고 합니다.

#2

저희 조는 김윤지님, 남슬아님, 조해원님, 그리고 저로 구성되어 있습니다.

각각의 역할 분담의 경우

조장이신 김윤지님은 통계 도매인을 담당해주시고

조해원님은 기업 분석 및 경제 도메인 영역을 담당해 주셨습니다.

남슬아님은 PPT 준비 및 통계 도메인 영역을 담당해 주셨고

저는 코드 구현 파트를 맡았습니다.

#3

목차는 다음과 같은 순서대로 진행될 예정입니다.

#4

본격적인 발표 진행에 앞서서 지난번 중간발표 당시에 제대로 된 답변을 드리지 못한 것으로 생각되어 그 부분에 대한 피드백을 먼저 진행하고자 합니다. 질문 사항은 다음과 같습니다.

#5

첫번째, 무위험 이자율을 3년 국채, 미국채 10년 같이 장기로 가져온 이유입니다. 먼저 무위험 이자율에 대한 명확한 정의가 없습니다. 그렇기 때문에 조사 당시에 미국 재무성 채권 중 은행 같은 기관 투자자들이 가장 선호하는 미국채 10년물과 한국 기관에서 벤치마크 무위험 이자율로 사용하고 있는 국고채 3년물로 무위험 이자율로 정의 내렸습니다.

#6

두번째는 주식 분석에 있어서 PCR을 어떻게 고려할 것인지에 대해서 입니다. 중간 발표 이후에 회의 결과 저희는 해당 부분을 고려하지 않기로 했습니다. 그 이유는 주가 현금흐름의 정의에서 발생한 문제입니다.

먼저 PCR, 주가 현금흐름은 주가 나누기 현금흐름인데 이 현금흐름에 대한 정의는 기업에서 공시한 재무재표의 현금흐름표와 은행에서 분석한 현금수지분석표가 있습니다. 그런데 이 둘 사이의 차이가 4조에 이른 상태입니다. 이런 상황에서 현금흐름에 대한 명확한 추정이 불가능한 상태이기 때문에 PCR에 대한 실질적 의미에 대한 의문이 든 상태입니다.

#7

세 번째는 금을 독립변수로 추가한 이유입니다. 저희는 최초 분석 당시에 삼성을 반도체 기업으로 생각했어 그에 대한 원자재 성격으로 금을 독립변수로 생각하여 추가하였고 여기에 부가적으로 금과 주가가 양의 상관관계가 나온 이유를 질문하셨는데 기본적으로 안전자산인 금과 위험자산인 주식은 음의 상관관계를 갖는 이유는 주가의 높은 변동성에 대한 시장의 불안감이 금에 대한 투자로 이어지기 때문에 나타나는 것이라고 생각되는데 해당 시기에 주가가 전반적으로 크게 상승하는 시기였기 때문에 주가 상승과 동반하여 금값도 같이 상승한 것으로 생각됩니다.

#8

이어서 저희 프로젝트 목표에 대한 설명을 드리겠습니다.

저희의 프로젝트 목표는 이전에 말씀 드린 것처럼 이번 강의에서 배운 내용을 최대한 써먹어 보는 것이었고

주제는 지표를 이용한 삼성전자 주가 예측 및 상관관계 분석으로 잡았습니다.

프로젝트 목표와는 별개로 저희의 자료 분석 목표는 다음과 같습니다.

1. 삼성전자 주가와 상관관계가 높은 지표를 파악하고
2. 상관관계가 높은 변수들을 독립변수로 회귀 모형을 적용하여 삼성전자 주가 예측에 유의한 모형인지 검증한다는 것입니다. 그런데 유의미한 결과를 도출하지 못해서
3. 시계열 모형을 적용하여 삼성전자 주가 예측에 유의한 모형인지를 검증한다.라는 분석 목표가 추가되었습니다.

#9

본격적인 발표에 앞서 발표가 지루해질 수 있기 때문에 결과부터 말씀 드리면 다음과 같습니다.

외부 변수를 독립변수로 둔 회귀 분석 주가 예측은 유의미한 결과를 도출하지 못했고 시계열 예측의 경우 장기 예측에 어렵고 단기 예측에 적합하다는 결과입니다.

#10

그러면 지금부터 저희의 프로젝트 진행과정에 대한 설명을 이어서 하겠습니다.

이전에 발표한 내용을 다시 한번 말씀 드리면 저희는 주식이 다른 자산과의 상관관계가 크기 때문에 이것들을 기반으로 회귀분석을 할 수 있지 않을까? 라는 생각에 기반하여 주식, 채권, 외환, 원자재의 4개의 섹터로 나눠서 독립변수를 선정하였습니다. 이 과정에서 데이터의 전처리를 진행하고 로그 스케일 자료가 가장 분석에 적합하다는 결과를 도출하게 되었습니다.

그래서 이 독립변수와 삼성전자와의 상관관계를 히트맵 형식으로 도출하고 그에 대한 분석 결과는 다음과 같았습니다.

먼저 종합적으로 금과 항셍지수를 제외한 자산의 섹터 내의 상관관계는 대체로 높게 나타나는 것을 확인 할 수 있었습니다.

독립변수인 삼성전자 주가와 종속변수와의 상관관계를 확인해보면

항셍을 제외한 주가지수는 강한 양의 상관관계

채권(미국채10년, 한국채3년)은 음의 상관관계

금은 삼성 주가와 강한 양의 상관관계

원유는 양의 상관관계

환율(달러,엔화)와는 음의 상관관계를 확인 할 수 있었습니다.

중요한 것은 독립변수 간의 상관관계였는데,

앞서 말씀 드린 것과 같이 자산 섹터 내에서 강한 양의 상관관계를 확인 할 수 있었습니다.

자세한 내용은 ppt에서 보이는 것 과 같은 상관관계를 확인 할 수 있습니다.

상관관계 분석 결과 이대로 회귀 분석을 진행할 경우 독립변수간 높은 상관관계로 다중공선성 문제가 예상되고 삼성전자 주가와 상관관계가 낮은 지표를 제외하는 과정이 필요하기 때문에 다음에 소개드릴 독립변수 선정과정 feature selection 파트로 넘어가게 되었습니다.

#11

Feature selection을 위해 먼저 Lasso, Ridge, RandomForest에 별도 파라미터 조정 없이 넣어서 중요도에 따른 독립변수 선정을 진행해 봤습니다.

이 시기에서 13개의 독립변수 중 중요도 순서로 5개를 뽑아서 독립변수로 선정하였습니다. 그 결과 모델에 따른 feature 선택은 ppt에 나온 것과 같이 선정이 되었고 모델을 통한 feature selection을 마치게 되었습니다.

#12

그 다음 차원 축소를 활용한 독립변수 추출도 진행해 보았습니다. 이 과정에서 자산군별로 축소하고, 상관관계 낮은 부분을 제외하고 차원축소하고, 전체 독립변수를 축소하는 등 수많은 방법을 사용했는데 물리적 시간과 지면상의 한계로 일부만 소개를 하자면

1. 전체 섹터가 4개(주식,채권,외환,원자재)이기 때문에 component를 4로 지정하여 차원축소를 진행하는 경우
2. 금의 경우 원자재와의 상관관계가 낮았기에 원자재와 별도로 취급하여 강한 양의 상관관계가 나타난 주식과 같은 분류를 해 차원축소를 진행하는 경우
3. 금과 강한 음의 상관관계가 나타난 채권을 같은 분류를 한 후 이에 대해 차원축소를 진행하는 경우
4. 상관관계의 절대값이 0.3이하인 지표를 제외하고 자산군별로 차원축소를 진행하는 경우

입니다.

#13

그 결과 전체 자산에 대해 component를 4로 적용한 PCA1을 제외하고는 변수간의 높은 상관관계가 해결되지 않았습니다. 프로젝트 목표 상, 배운 내용을 다 사용하기 위해 지면상 소개되지 않은 부분을 포함하여 전부 회귀 분석을 진행하였습니다만 여기에서는 다중공선성 문제가 해결된 PCA\_1에 대해서만 소개할 예정입니다.

#14

먼저 이전에 회귀 모델 별 별도 파라미터 조정 없이 중요도에 따라 선정된 독립변수를 기반으로 개별 회귀 분석을 진행하였습니다.

분석 결과 mse, rmse의 값만 보면 상당히 잘 학습되었다고 생각하실 수 있으나 R square값이 처참한 값이 나왔습니다. 또한 잔차 검증 결과 독립성, 등분산성, 정규성 조건을 만족하지 못하였습니다.

#15

차원축소를 진행한 독립변수는 라쏘, 릿지, 랜덤포레스트, 일반 선형회귀 모두를 진행했습니다.

먼저 전체 독립변수에 대해서 component를 4로 적용한 pca1의 경우에는 MSE와 RMSE 값은 다음과 같이 도출 되었습니다만 Rsqure값이 과적합과 의미 없음이 나왔습니다.

결과적으로 다중 선형회귀 모델은 주가를 설명하는데 적절하지 못하다고 판단하였습니다.

#16

그래서 이번에는 차원축소를 진행한 독립변수에 대해서 다항 회귀를 진행했고 이때, degree는 2,3으로 설정하고 분석을 진행을 진행하였습니다.

그 결과는 다음과 같이 나왔습니다.

그 결과 degree 2에서는 과적합, degree 3에서는 R^2값이 음수가 나왔습니다.

.

이런 결과로 저희는 외부지표를 활용한 회귀분석이 주가 예측에 적절하지 못하다는 결론을 도출하고 시계열 분석을 시작했습니다.

#17

시계열 분석에서 저희는 삼성전자 전체시기에 대한 주가 데이터와 최근 3개년 삼성전자 주가를 기반으로 한 예측 모델을 각각 만들고 2022년 1분기 삼성전자 주가를 예측하여 실제 데이터와의 비교를 통해서 예측력을 검정하고자 하였습니다. 여기서 저희는 22년 1분기를 예측 대상으로 했기 때문에 3개년 주가 데이터를 가져왔고 해당 시기가 코로나 영향을 크게 받았기 때문에 이에 대한 비교군 역할로 상장 이후 전체 주가 데이터도 가져와 모델로 만들어 보았습니다.

#18

Log스케일을 진행한 주가 데이터에 대해서 ACF 그래프와 PACF 그래프를 확인하였습니다. 해당 자료를 확인하면 모든 데이터에서 PACF가 1기 이후 낮아졌기 때문에 저희는 해당 데이터에 1차 차분한 데이터가 적합하다고 생각했고 이에 대한 검증을 위해서 실제 차분한 데이터에 대해서 ACF 그래프와 PACF 그래프를 출력했습니다.

#19-1~19-2

그리고 해당 데이터를 기반으로 추가적인 검증을 위해서 KPSS 검정을 진행했습니다.

해당 검정에는 귀무가설이 ‘데이터가 정상성을 만족한다.’는 것이고 이에 대해 p값이 0.1로 나와 유의수준 0.05에서 귀무가설을 기각할 수 없기 때문에 데이터가 정상성을 만족한다는 것을 알 수 있었습니다.

#20

그리고 해당 데이터를 기반으로 가장 적합한 ARIMA 모형을 찾기 위해서 auto\_arima를 실행하였고 그 결과 상장 전체 시기에 대한 데이터는 ARIMA(1,1,2),

#21

최근 3개년 데이터는 ARIMA(0,1,0)이 가장 적합한 예측 모델이라는 결과가 도출되어 해당 결과를 기반으로 예측을 진행하고자 하였습니다.

--------------------------------------------------------------------------------------------------------------------

그래프;

-standardized residual( )

-> 백색잡음의 특성(평균 0을 중심으로 무작위하게 움직임)

-correlogram(ACF) : 시차 0과 시차 10 제외하고 모두 파란색 구간 안에 있음

-> 시차 0과 10을 제외한 각 시차의 자기상관계수값이

5% 유의수준 내에서 0과 다르다고 할 수 없다(자기상관관계가 없음!

-히스토그램

-> 정규분포 N(0,1)과 비슷한 평균을 가짐

-> but, 첨도가 더 뾰족하게 솟음

-QQ plot

-> 정규성을 만족하려면 데이터가 빨간 직선 위에 위치해야함

-> 양 끝에서 약간 벗어나는 모습을 보임

검정통계량;

-Ljung-Box (Q)(융-박스)

-> 잔차가 백색소음을 따르는지 판단하는 것

-> p-value : 0.49 >0.05 ->잔차가 백색소음을 따른다

-> 시계열 모형이 잘 적합되었고, 남은 잔차는 더이상 자기상관을 가지지 않는 백색잡음임을 확인함

-skew(왜도) : 0.37 , kurtosis(첨도) : 5.72

-> 잔차가 정규분포를 따른다면 왜도는 0에, 첨도는 3에 가까워야함

-Heteroskedast (이분산검정) : 0.88

-> 잔차가 이분산성을 띄지 않는지 검정

-> p-value : 0.31 > 0.05

-> 잔차가 이분산성을 보이지 않음(잔차가 등분산성을 만족함)

- Jarque-Bera(쟈크베라검정) (JB) : 245.9

-> p-value : 0 < 0.05

-> 잔차가 정규성을 띄지 않는다

결과적으로 잔차는 백색잡음이지만 정규성을 따르진 않는다.

시계열 모델이 유의하지 않다..?

--------------------------------------------------------------------------------------------------------------------

#22

저희의 시계열 예측은 예측 방법에 따라서 왼쪽에서부터 1~4번으로 분류하고 각각 시계열 예측을 진행하였습니다.

1~3번의 경우 우리가 현실에서 미래 시점에 대한 주식 가격 정보를 알 수 없기 때문에 이런 상황을 적용한 주가 예측 방식이었고 4번의 경우 실제 가격을 학습시키는 방식으로 모델이 잘 짜여졌는지 검증하기 위해 실제 테스트 값을 학습하는 방식으로 예측을 진행하였습니다.

#23

그 결과 1의 경우 <한 템포 쉼>

#24

2의 경우 <한 템포 쉼>

#25

3의 경우 다음과 같은 주가 예측에서 상승의 경향성을 확인할 수 있었으나 실제 주가를 잘 예측하지 못하였고

#26

이에 모델이 잘 짜였는지 확인한 결과 모델 자체에 대해서는 이상이 없는 것을 확인할 수 있었습니다.

#27

그래서 저희는 해커톤에서 초반에 말씀 드린 것과 같이 외부지표를 독립변수로 한 회귀 모형은 삼성전자 주가 예측에 적합하지 않았고 시계열 분석 역시 단기 예측에는 적합하지만 장기 예측이 어렵다는 결론을 도출하였습니다.

#28

그렇다면 왜 이런 결론이 도출되었는지에 저희 조의 해커톤 진행에 대한 한계점을 설명하도록 하겠습니다.

먼저 프로젝트 초기 단계에서의 한계점입니다.

첫 번째는 회귀분석의 독립변수 선정 과정에서 자료 조사를 충분히 하지 못한 채, 삼성전자와의 상관관계를 주관적으로 고려하여 선정하였습니다.

실제 삼성전자의 dart 공시 자료를 확인하면 환 위험에 대한 영향이 기재되어 있는데 해당 부분에서 달러, 유로, 인도 루피의 영향이 크다고 기재되어 있지만 실재 저희의 모델에서는 달러와 엔화를 적용하여 현실과 모델의 차이가 있었습니다. 또한 종속변수인 삼성전자를 반도체 기업이라는 가정 하에 금을 원자재로 생각하고 독립변수로 추가했으나 실제 삼성전자의 매출구조를 확인하면 반도체 부문은 40%로 기타 다른 부분이 더 컸기 때문에 반도체 기업으로 단정하기 어려웠습니다.

두번째는 통계적으로 선형 회귀 분석의 기본 가정 검증 결과 잔차의 등분산성, 독립성, 정규성 조건을 만족하지 못했습니다.

즉 실질적으로 자료 자체가 회귀 모형에 적합하지 않는 상태였습니다. 앞서 말씀 드린 문제도 해당 문제에 영향을 미쳤을 것으로 생각됩니다.

마지막으로 주식의 가격이 시간에 따라 변화하는 자료지만 그에 더하여 다양한 요인을 반영하고 있는 지수로 단순한 시계열 예측만을 기반으로 장기간의 미래 가격을 정확하게 예측하는 것은 어렵다는 것이 있습니다.

#29

그러나 이러한 결과에도 저희는 어떤 의미를 도출 할 수 있었습니다.

통계적 기법을 적용하는 방법적인 측면에서 배운 내용을 최대한 적용하자는 프로젝트의 목표를 달성하였습니다.

또한 분석 과정에서 효과적인 모델을 발견하지 못했으나 이를 통해서

회귀 분석 모형의 경우 데이터 선정 과정에서 통계적 검증 및 공신력 있는 자료를 통한 사전 조사가 필수적이라는 것을 몸으로 깨달았습니다.

그리고 주가 예측에 대해서 타 금융 자산을 기반으로 회귀분석을 하는 것이 어렵다는 것을 알았다는 부분인데, 만일 이번 해커톤 경험이 없었다면 주가 예측에 기타 자산을 독립변수로 추가하여 다중 공선성 문제 등이 발생한 것도 모르고 분석을 진행했을 것이라 생각됩니다.

그리고 시계열 예측의 경우 장기 예측에서 현실과 거리가 있는 결과가 나왔지만 단기 예측에서 유의미한 성과를 얻을 수 있었습니다.

#30

이것으로 발표를 마치도록 하겠습니다. 감사합니다.

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

예상 질문

회귀분석 변수 선택에서 인과관계를 왜 고려하지 않았나? :

회귀분석의 목적 자체가 인과관계를 밝히는 것이다.

Pca1에서 component를 4로 지정한 이유:

분석 자산을 4개의 분류로 고려하였기 때문이다.

회귀분석은 왜 의미가 없냐?

회귀분석을 통한 주가 예측은 실제로 의미 있다. 다만 본 해커톤 발표에서는 저희가 선정한 지표 선정이 문제가 되었고 실제 포항공대 연구 논문인 ‘회귀분석을 이용한 주식 가격 예측’에서 내부 지표를 활용한 주가의 회귀분석을 진행했는데 여기서는 유의미한 모델이 도출되었습니다.

코로나 시기는 이상치 아니냐? 왜 분석했냐?

그렇기 때문에 전체 데이터를 비교 집단으로 같이 분석을 진행했다. 결과적으로 둘의 정도 차이는 있었지만 실제로 분석 결과의 방향성이 같았다.

또한 해당 수치의 제거에 대한 본질적인 의문이 들었다.

1. 코로나가 경제에 큰 충격을 준 상황이기에 해당 수치를 이상치로 제거해야한다면 그 이전에 발생한 리먼 사태, imf사태, 미중 무역전쟁, 닷컴버블, 러시아-우크라이나 전쟁, 911사태 등의 예상치 못한 사건의 영향력을 전부 제거해야 한다. 그렇기에는 능력이 부족하기도 하고 그 과정에서 데이터를 수정하는 작업에서 개인의 주관이 너무 개입된다 생각한다.

독립변수의 사전 검증은 안했냐?

* 중간 발표 이후 사전검증을 진행했고 만족하지 못했다. 하지만 확인한 시점 상 데이터를 새로 선정하기에는 시간이 부족했고 프로젝트의 목표를 우선으로 하여 그대로 진행했다.

외부변수가 왜 의미가 없냐고 하냐?

* 외부 영향력 전체가 의미가 없다는 의미가 아니라 저희가 선정한 변수에 대한 의미이다.

단기 예측 결과는 어떤가?

전체 시기에 대한 mse,R^2

MSE : 0.00022 R^2\_score : 0.856

3개년에 대한 mse,R^2

MSE : 0.00022 R^2\_score : 0.857