(인사)

안녕하십니까.

오늘 해커톤에서 저희 조가 발표할 주제는 '분류분석 모델을 활용한 sk하이닉스의 주가 방향성 예측'입니다.

<주제 선정배경>

<내일의 주가, 예측 가능한가>

과연 주식시장은 예측 가능할까요.

오늘까지의 데이터로 내일의 주가 등락을 예측할 수 있을까요.

어쩌면 누구나 한번쯤은 던져보았을 이 간단하면서도 도전적인 질문에서 저희 프로젝트는 시작합니다. 데이터와 머신러닝의 힘을 빌려 시장의 다음 움직임을 예측하고 시장의 변동성을 이해할 수 있다면 리스크는 낮아지고 수익은 높아질 것입니다.

<연구 목표>

이번 프로젝트에서는 기술적 지표를 활용하여 sk하이닉스의 주가가 상승할지, 하락할지 예측하는 분류분석 모델을 개발하고, 검증해보았습니다. 가격과 거래량은 주식시장에서 가장 기본적이고도 핵심적인 정보를 제공합니다. 주가는 시장의 불확실성과 기대를 반영하며, 거래량은 이러한 가격 변동에 대한 참여자들의 반응을 나타냅니다. 이 두 지표와 이들의 상호작용에서 파생되는 여러 기술적 지표를 활용하여 sk하이닉스의 주가 방향성을 예측해 보고자 합니다.

<선행 연구>

(마지막 선행연구) (한국소프트웨어종합 학술대회, 세종대학교) 연구에 따르면 트레이닝 기간이 길수록, 단기간의 주가 예측일수록 딥러닝 기반 주가 예측 모델의 정확도는 향상됩니다. 그러나 해당 연구는 최대 1년까지의 주가 데이터만을 학습하였다는 한계를 갖습니다. 저희 프로젝트는 한걸음 더 나아가 지난 10년간의 주가 데이터를 학습합니다. 긴 기간동안의 데이터는 다양한 경제 사이클과 시장 상황을 포괄하기에 더 높은 정확도와 신뢰성을 달성할 수 있을 것이라는 기대를 가지고 프로젝트를 진행하게 되었습니다.

<왜 기술적 지표인가>

일반적으로 주가예측모델에서는 재무 지표나 경제요인이 중요한 변수로 고려됩니다. 그러나 저희는 이러한 전통적인 접근 방식에서 벗어나, 재무지표와 거시경제 요인을 모델에서 의도적으로 제외하고, 기술적 지표만을 활용한 주가 예측모델을 구축하여 보다 범용성있는 모델을 구축하는 것을 목표로 하였습니다.

일단, 재무지표는 대개 연간 또는 분기별로 발표되기에, 다음날의 주가 up, down을 예측하는 본 연구의 피처로는 적절하지 않아 제외하였습니다.

외부 요인들이 하이닉스의 주가 예측에 중요한 역할을 할 수 있음을 알기에 많은 논의가 있었습니다. 예컨대, *~~엔비디아의 주가 혹은~~* 필라델피아 반도체 지수의 등락은 산업의 전반적인 동향과 기대를 반영하며 하이닉스의 주가에 직접적인 영향을 미치기에 해당 요인들을 피처로 넣은 모델의 정확도는 매우 높아질 것입니다. 그러나 저희는 보다 광범위한 적용성을 가지는 모델을 구축하는 것에 초점을 맞추기로 결론내렸습니다. 따라서 종목 특이적인 요인들을 직접 피처에 포함하지 않고 오로지 기술적 지표에만 기반하여 모델을 구축하고 학습시킴으로써 보다 넓은 범위의 시장 조건과 다양한 종목에 유연하게 활용할 수 있는 모델을 만들고자 하였습니다.

<데이터 수집 및 EDA>

<데이터수집>

모델에서 사용된 데이터는 2014년 3월부터 2024년 3월까지 총 10년간의 하이닉스 주식 데이터를 기반으로 합니다. 이중 Train data로는 2014년부터 2022년까지 8년, test로는 2022년부터 2024년까지 2년을 설정하였습니다. *~~장기간의 데이터를 분석하여 최근 10년간의 경제상황과 시장변화를 포괄적으로 반영하고자 하였습니다~~*. *~~주가예측의 모델의 정확도를 높이기 위해, 다양한 변수와 파생변수들을 활용하여 모델을 구축하였습니다.~~*

*~~중장기적 추세를 파악하는데 중요한 지표로 꼽히는 240일 이평선을 파생변수로서 모델에 포함하기 위해 2013년부터 데이터를 추출하였습니다~~*. pykrx라이브러리를 통해 시가, 고가, 종가, 등락률등의 기본적인 주가정보를 획득하였습니다.

그러나, 투자자별 거래량 정보는 Pykrx에서 제공하지 않기 때문에, 추가적인 데이터 수집 방법이 필요했습니다. 이를 위해 대신증권의 HTS(Home Trading System)를 이용하여 해당 데이터를 수집했습니다.

<파생변수 형성>

복합적인 파생변수의 생성과 활용을 통해 주식가격의 움직임을 보다 정교하게 모델링하고 예측할 수 있을 것입니다. 저희는 주식시장에서 널리 사용되는 5대 보조지표- 이동평균선, 상대강도지수(RSI), 이동평균수렴확산지수(MACD), 볼린저밴드, 그리고 스토캐스틱을 생성하는데 필요한 다양한 파생변수들을 모델에 추가하였습니다.

--------

시간 남으면..?

이러한 지표들은 흔히 주식가격의 움직임을 이해하고 예측하는데 있어 중요한 도구로 사용됩니다. 예를들어 이평선은 주가의 중장기적 추세를 파악하는데 사용됩니다. 또한 RSI는 특정주식이 과매수 또는 과매도 상태에 있는지를 판단하는데 사용되고 MACD는 가격의 모멘텀을 평가하는데 사용되고 있습니다. 볼린저밴드는 주가의 변동성을 측정하며, 마지막으로 스토캐스틱은 가격의 위치를 현재 추세내에서 평가합니다

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

<피처 총 100개>

총 100개의 피처를 사용하였습니다. 이중 85개는 가격 관련 피처로 언급한 5대 보조지표뿐 아니라 이들 지표를 계산하는데 필요한 다양한 파생변수들을 포함합니다. 나머지 15개는 거래량 관련 피처입니다.

<데이터 전처리>

Pykrx 라이브러리와 증권사 ‘hts’에서 가져온 데이터이기에 많은 전처리가 필요하지 않았습니다.

첫번째로, 대신증권 HTS에서 제공하는 데이터는 기간별 누적 값으로만 이루어져 있어 목적에 맞도록 주체의 거래일별 순매수량을 도출하는 과정을 거쳤습니다.

또한 이평선 도출을 위해 분석시점 이전의 데이터를 추가로 추출하였기에 파생변수 형성 후분석시점 이전의 불필요한 데이터를 제거하는 과정을 거쳤습니다.

<데이터 스케일링>

변수간의 척도를 유사하게 만들어 모델의 성능을 개선하고 학습시간을 단축하기 위해 데이터 스케일링을 진행하였습니다.

가격관련 모든 피처들에는 표준화 스케일링을 적용하였습니다. 하이닉스의 주가에는 극단적인 이상치가 존재하지 않고 범위가 제한되어 있기에 표준화 스케일링이 가장 적합하다고 판단하였습니다.

거래량 관련 피처들에는 로버스트 스케일링을 적용하였습니다. 거래량데이터에는 대규모 거래가 이루어진 날들이 존재하여 이상치가 존재하였으며 이상치에 강건한 로버스트스케일링을 활용하여 이러한 이상치의 영향을 줄이는 효과를 도모하고자 하였습니다.

<레이블 설정>

분류분석의 타겟 레이블은 다음 거래일의 종가 등락률을 기준으로 설정하였습니다. 구체적으로 해당일 대비 다음 거래일의 종가 등락률이 수수료율 0.25%보다 높은경우 레이블을 1로 설정하였고 그렇지 않으면 0으로 설정하였습니다. 이는 실제 투자에서 발생할 수 있는 수수로를 고려햐여 모델이 실질적인 수익을 예측할 수 있도록 하는 전략입니다.

*~~이러한 방법을 적용한 결과 트레인데이터에서 ‘1’의 레이블은 890번으로 전체의 45.3%을 차지하였고 ‘0’은 1075번으로 54.7%를 차지하였습니다. 테스트데이터에서는 ‘1’의 레이블이 201번으로 40.3% 0의 레이블이 294번으로 59.7%로 나타났습니다.~~*

레이블 설정 결과 0과 1의 비율이 상대적으로 균형을 이루고 있으며, 트레인데이터와 테스트데이터간에도 비슷한 비율을 유지하고 있음을 알 수 있습니다. 따라서 추가적인 샘플링 작업 없이도 데이터가 분류모델 학습에 적합하다고 결론내렸습니다.

<4. 피처선정>

<피처선택>

피처를 선택하기 위해 상관성분석, t-test 그리고 랜덤포레스트 기법을 활용한 중요도 분석을 진행하였습니다. 세가지 방법을 모두 적용한 뒤 상위 30위에 두번 이상 등장한 피처들 중에서 랜덤포레스트 중요도 분석을 기반으로 최종 피처를 선정하였습니다.

*~~랜덤포레스트는 다수의 결정트리를 조합하여 사용하기에 데이터의 비선형 관계와 구조를 효과적으로 포착할 수 있습니다.~~* 주가예측과 같은 금융 시계열 데이터는 비선형 관계를 가지고 있기에 랜덤포레스트 방법을 메인으로 선택하여 진행하였습니다.

<피처 개수 선정>

분류분석 모델의 precision\_score을 최대화하기 위한 최적의 피처 개수를 결정하기 위해 피처개수를 변수로 두고, 5개에서 15개 사이에서 가장 높은 precision\_Score을 주는 피처 개수를 도출하였습니다. 피처개수가 7개일때 네개의 모델에서 precision-score가 상위 3개 안에 들어, 최종적으로 피처개수를 7개로 결정하였습니다.

<5. 모델링 및 성능평가>

<모델링 및 성능평가 지표>

하이닉스의 주가 방향성을 예측하기 위해 서포트벡터 머신, 로지스틱회귀, 랜덤포레스트, 결정트리, 그리고 그래디언트 부스팅 다섯개의 분류분석모델을 적용하였습니다.

precision은 주가가 상승할 것으로 예측된 경우 중, 실제로 주가가 상승한 비율을 나타냅니다. recall, 즉 재현율은 실제로 주가가 상승한 경우 중 모델이 캐치한 비율을 나타냅니다. 저희의 분석은 precision\_score을 비교하는 방법에 초점을 두었습니다. 실제 투자 상황에서 잘못된 매수신호를 따라 투자할 경우 자본손실로 이어질 수 있기 때문입니다. 정밀도에 중점을 둔 모델을 이용하여 투자 결정시 발생할 수 있는 리스크를 감소하여 보다 신뢰성 있는 투자에 도움을 받을 수 있을 것입니다.

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

<분석과제>

하이닉스 주식 가격 변동 예측을 목적으로 네가지 중요한 분석 줄기를 통해 모델의 precision을 비교하는 과정을 수행하였습니다. 첫번째 분석목표는 스케일링한 데이터가 원본 데이터에 비해 모델의 성능을 어떻게 개선하는지 평가하는 것입니다. 두번째로, 타겟레이블을 두가지와 세가지로 설정했을 때 모델 성능을 비교합니다. 세번째 분석은 train test 데이터를 시간 순서에 따라 분할하는 것과 무작위로 분할하는 것이 모델 성능에 미치는 역할을 평가합니다. 마지막으로, 모델 하이퍼파라미터를 튜닝하여 성능 고도화를 시도합니다.

<Scaled data vs. Raw data>

*~~스케일링 처리된 데이터와 원본 데이터의 예측 모델 성능 비교를 통해 데이터 전처리 과정이 모델의 예측 정밀도에 미치는 영향을 평가할 수 있습니다~~.* 원본 데이터를 사용했을 때, 로지스틱 회귀모델이 가장 높은 정밀도인 0.55를 달성하였습니다. 스케일 처리된 데이터에서는 그래디언트 부스팅 모델이 0.53의 정밀도를 보여주었으나 이는 원본 데이터를 사용했을 때보다 다소 낮은 결과입니다. 특히 서포트벡터머신의 경우 정밀도가 0.27로 크게 떨어지는 현상을 보여줍니다. 반면 결정트리와 그래디언트 부스팅모델은 스케일링된 데이터에서 소폭의 성능개선을 보였습니다.

<2레이블 vs 3레이블>

두번째로 *~~타겟레이블을 2개와 3개로 설정했을 때 모델의 성능차이를 비교하여 레이블의 구분수가 모델의 예측 정밀도에 미치는 영향을 평가하였습니다~~*. 3레이블 설정은 타겟 레이블을 세가지로 보다 세분화하여 모든 날짜의 등락률중 상위 33%는 2, 중간 33%는 1 하위 33%은 0으로 설정하였습니다. 타겟 레이블을 3개로 설정하였을 때 모든 모델의 precision이 감소하는 결과를 보였습니다.

<time series split vs random split

세번째로, 시계열 데이터의 특성을 고려하여 시간기준 분할과 무작위 분할을 통한 훈련 및 테스트 데이터 세트 방식이 모델의 예측 정밀도에 어떠한 영항을 미치는지 비교하였습니다. *~~무작위분할의 경우 테스트 사이즈를 0.2에서 0.3으로 변화시키면서 randomstae를 1부터 100까지 변경하여 높은 precision score을 달성하는 설정을 찾고자 하였습니다~~*. Random split을 적용했을 때 svc모델이 randomstate 7에서 0.68의 높은 정밀도를 달성하였습니다. 또한 전반적으로 모든 결과에서 randomsplit을 통한 데이터 분할이 더 높은 정밀도를 달성했다는 점을 확인할 수 있습니다.

<하이퍼파라미터튜닝>

SVC와 로지스틱 회귀모델의 하이퍼파라미터를 다양하게 조정하여 최적의 성능을 도모하려 하였으나 튜닝후모델의 성능이 디폴트 설정보다 오히려 저하되는 결과를 보였습니다. 이 결과는 하이퍼파라미터 튜닝이 항상 유익한 결과를 보장하지는 않는 다는 것을 나타냅니다. 향후 프로젝트에서는 각 모델의 하이퍼파라미터가 가지는 의미와 영향력을 충분히 이해하여 보다 세심한 하이퍼 파라미터 조정을 통해 효과를 낼 수 있도록 노력하겠습니다.

<결론>

결론입니다.

피처 셀렉 결과, 선정된 7개의 피처는

'고가 대비 종가 등락률' '연기금 일 순매수량' '거래량 등락률' '사모펀드 ' '기타법인 ' '투신 ' '은행 '입니다. 투자주체별 순매수량이 가격이나, 가격에서 파생된 5대 보조지표보다 유용하다는 결과를 함의합니다. 거래량은 가격 변동의 전조로 해석 할 수 있을 것이며 각각의 투자 주체들이 시장에 미치는 영향력이 시장의 방향성을 이해하는 데 있어 가장 결정적인 요소가 되고 있음을 보여줍니다.,

또한 '연기금' '사모펀드' '기타법인' '은행'의 순매수량이 외국인이나 기관 투자자의 순매수량보다 더 강력한 피처로 작용한다는 결론입니다. 이러한 분석 결과는 투자전략을 수립할 때 흔히 가장 많이 보는 '기관'이나 '외국인' 투자자의 활동 이상으로 기타 다른 투자주체의 활동에 주목해야 함을 시사합니다.

<결론>

본 연구를 통해 하이닉스 주식가격변동예측 모델의 성능을 최적화하기 위해 다양한 접근방법을 탐색하였습니다. 분석과정에서 얻을 결과와 이에 대한 저희 팀의 해석을 요약하여 말씀드리겠습니다.

분석 결과, 원본 데이터를 사용한 예측 모델의 성능이 스케일링 처리된 데이터를 사용한 모델보다 더 높게 나타났습니다. 특히 거래량과 같은 변수들을 고려할 때, 로버스트 스케일링이 이상치를 어떻게 처리하는지가 모델의 성능 저하에 영향을 미쳤을 것으로 추측됩니다. 스케일링 과정에서 데이터의 중요한 정보가 손실되거나, 모델이 데이터의 본질적인 패턴을 학습하는 데 방해가 되었을 수 있다고 결론내렸습니다.

두번째로, 타겟 레이블을 3개로 설정했을 때보다 2개로 설정했을 때 precision score가 높았습니다. 이는 타겟 레이블의 수를 증가시키면 분류해야 할 클래스가 많아지고, 모델이 각 클래스를 정확히 구분하기 더 어려워짐을 의미합니다. 타겟 레이블의 수가 증가함에 따라 모델이 결정해야 하는 경계가 더 많아지고 복잡해지며, 더 세밀한 변동까지 예측해야 하기 때문입니다

세번째로, 시간 기준 분할 대신 무작위 분할 방식이 더 높은 정밀도를 달성했습니다. 시계열을 다룬 데이터의 경우 보통 시간기준 분할이 좋은 결과를 내는 것과 상이한 결과였습니다. 이는 10년이라는 장기간에 걸친 시계열 데이터에서 시장 상황의 변동으로 인해, 기존의 주가 이동 패턴과 다른 새로운 패턴이 등장했기 때문으로 해석됩니다. 특히, 코로나19 팬데믹 이후 주식 시장은 이전과 다른 양상을 보였으며, 이러한 상황 변화는 2022년을 기준으로 데이터를 분할할 때 예측 정밀도의 감소를 초래하는 요인이 되었다고 생각합니다.

<추가연구>

본래의 연구 의도는 범용성이 높은 금융 시장 예측 모델을 개발하는 것이었습니다. 그렇기에, 개별 기업인 하이닉스에 국한되지 않고, KOSPI 전체에 모델을 적용해 보는 방향으로 연구를 확장하였습니다.. 이번 연구에서는 HTS에서 가져온 투자 주체별 거래량 정보 없이, pykrx를 통해 접근 가능한 총 거래량, 가격 피처들, 그리고 가격에서 파생된 피처들만을 사용하여 분석을 진행하였습니다. 분석 기간을 보다 짧게 설정하여, 모델의 범용성을 다양한 시장 조건에서 테스트하였습니다. 포트 벡터 머신(SVM)의 precision score을 기준으로 내림차순으로 정렬한 결과, KOSPI 상장 종목 중 여러 종목들에서 높은 예측 정밀도를 달성했습니다.. 하이닉스의 경우보다 높은 성능을 보인 종목들이 다수 관찰되었습니다. 이는 하이닉스 같은 대형 주식이 가격과 거래량 외에도 다양한 외부 요인의 영향을 크게 받는 반면, 다른 종목들은 이러한 요인들의 영향을 상대적으로 덜 받을 수 있기 때문으로 추측됩니다.

.

<의의>

이 프로젝트를 통해 하이닉스 주가 데이터를 활용한 예측모델 개발에 있어서 데이터 전처리, 타겟레이블 설정, 데이터 분할 전략, 그리고 하이퍼파라미터 튜닝의 중요성을 종합적으로 조명하고자 하였습니다.   
  
다만 본 연구는 이러한 한계가 존재합니다.

첫 번째, 본 연구에서 사용된 분류분석 모델은 주가의 방향성 예측에 초점을 맞추었습니다. 즉, 주가가 상승할지 또는 하락할지에 대한 예측은 가능하지만, 실제 수익률을 예측하는 것과는 다릅니다. 이는 수익성을 직접적으로 보장하지 않는다는 것을 의미합니다. 따라서, 실제 투자 수익률을 예측하는 향후 연구가 추가적으로 필요합니다

두 번째, 이 프로젝트는 한정된 수의 피처를 사용하고 특정 기간 동안의 데이터에 근거하여 모델을 개발했습니다. 피처의 수를 늘리고, 다양한 기간에서의 연구를 수행하며, 다른 예측 모델을 사용한다면 예측 정확도를 더욱 향상시킬 수 있을 것입니다. 특히, 시장의 변화를 더 잘 반영할 수 있는 추가적인 피처의 도입과 다양한 모델링 기법의 적용은 모델의 성능을 개선하는 데 중요한 역할을 할 것입니다

오늘 저희 팀이 진행한 SK하이닉스 주가 예측 모델 개발 프로젝트에 대한 발표를 마치려 합니다. 저희가 유용한 정보를 제공했기를 바라며, 이상 발표를 마치겠습니다 감사합니다.