

|  |
| --- |
| 공모주의 시가와 고가 예측 |
| K-20110901-034369.png[제닉스] |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 학번 | 20191233 |  | 이름 | 김범수 |
| 학번 | 20201235 |  | 이름 | 백승민 |
| 학번 | 20221212 |  | 이름 | 이준석 |

[ 목차 ]

1. 서론…………………………………………………………………………………...3

A.공모주에 대하여

B.주제 선정 이유

2. 데이터…………………………………………………………………………...……4

A.데이터 종류

B.데이터 수집

C.데이터 전처리

3. 분류……………………………………………………………………………………9

A. Random Forest

B. SVM

C. Logistic Regression

4. 딥러닝………………………………………………………………………...……16

A. MLP

B. CNN

C. HYBRID

D. RNN + MPL

5. 클러스트링………………………………………………………………..………27

A. 클러스트링 목적

B. 클러스트링 알고리즘

C. 시행방법, 결과 및 분석

6. 주식 자동매매 시스템………………………………………………………...41

7. 결론………………………………………………………………………………....46

**1. 서론**

A) 공모주에 대하여

일반투자자들이 큰 수익을 기대할 수 있는 주식투자 방법 하나가 바로 공모주 투자입니다. 공모주는 공개모집주식의 약자입니다. 기업이 주식시장에 상장하는 단계에서 일반 투자자들에게 주식을 배분하는 것으로, 공모주 투자는 성장성이 큰 기업의 주식을 상장 전에 상대적으로 저렴한 가격에 매수하였다가 상장 후 매도를 통해 큰 수익을 볼 수 있어 투자자들에게 인기가 높습니다. **기업공개(IPO : Initial Public Offering)**는 기업을 일반 대중에게 공개하고 자본을 조달 받는 것을 의미합니다. 기업은 공개모집(공모) 절차를 통해 개인 및 기관투자자에게 투자자금을 받아 회사의 발행주식을 나누어 주고, 이후 이 주식은 주식시장에 상장되어 자유롭게 거래가 되는 것입니다. 2021년도에 공모주 투자에 대해 개미투자자들 또한 시도할 수 있게 균등 배분이란 제도가 생겼다. 이로 인해 많은 사람들이 공모주에 대하여 많은 관심을 갖게 되었다. 균등 배분이란 청약에 참여한 모두에게 차별 없이 공무주를 나누는 방법이다. 최소 청약금만 넣어둬도 참여자 모두 배정받을 수 있는 기회가 생기는 것이다. 원래의 비례배분은 청약증거금을 많이 넣을수록 더욱 더 많은 주식을 받을 수 있는 방법이다. 현재는 균등배분과 비례배분을 50대 50으로 맞추었다. 다음으로 우리가 공모주 청약을 해야 하는 이유에 대하여 적어보자면 첫째, 주식시장 참여의 기회를 얻을 수 있습니다. 공모주 청약을 통하여 주식을 처음으로 매수하고 매도해보는 경험을 얻을 수 있습니다. 이는 개인 투자자에게 주식시장에 참여해보고 기업의 성장과 함께 할 수 있는 기회를 얻을 수 있는 것만으로도 매우 좋은 일입니다. 또한 초저가로 주식을 구매할 수 있는 기회입니다. 공모주의 공모가는 일반적으로 낮게 책정이 되기 때문에 첫날 수익을 줄 확률이 매우 높습니다. 공모주를 청약해서 저렴한 가격으로 주식을 얻을 수 있고 높은 가격에 매도를 하여 이에 대한 수익을 얻을 수 있습니다. 또한 위의 설명한 균등배분을 통해 돈이 적든 많든 공평하게 주식을 획득할 수 있습니다. 그래서 비교적 돈이 적은 대학생들에게도 추천하는 투자입니다.

B) 주제 선정 이유

2021년도 이후의 많은 기사와 뉴스에 나오는 공모주들을 통하여 공모주에 대한 관심이 생겼고 주식을 투자하다 보면 절대 빼놓을 수 없는 IPO에 대하여 더욱 더 공부하고 싶었다. 이러한 공모주 투자는 22년도 초반 LG에너지솔루션의 상장 당시 우리나라 국민 총 442만명이 청약을 했을 정도로 엄청나게 많은 사람들의 관심 속에 있는 투자입니다. 주식 상장 전에 증권신고서를 통하여 공시하기 때문에 이 신고서에 있는 데이터가 상장 당시의 상황을 정확히 반영하기에 정확성이 매우 높고 또한 쉽게 온라인 사이트에서 데이트를 얻을 수 있기에 접근성이 높기 때문에 이러한 데이터를 이용하여 결과에 대한 예측 가능성이 높기 때문에 이러한 주제를 선정하였다. IPO를 하는 기업들에 대하여 머신러닝을 이용한 분석을 통하여 어떠한 공모주에 투자해야 돈을 벌 수 있는지에 대해 알아가기 위하여 정하였다. 최종적으로는 공모주를 통하여 꾸준히 돈을 벌고 싶다는 생각이 큽니다.

**2. 데이터**

A) 데이터 종류

먼저 데이터는 2023년도 6월이후에 상장한 기업들에 대하여 다루었습니다. 먼저 2021년도 이전의 데이터는 균등배분이 생기기 전이라 아예 필요가 없습니다. 2023년도 6월에 신규상장 주식들의 상장일에 한하여 상한가와 하한가의 한계를 300프로, -40프로로 늘렸다. 기존에는 100프로에서 -30프로 사이에 시가가 형성되고 이후의 상한가와 하한가는 일반 주식시장과 같은 30프로, -30프로입니다. 이는 엄청나게 큰 변화입니다. 공모주 청약에서 유명한 말인 따따블 상황시 전에는 수익률이 130프로인데 지금은 300프로 상승 인만큼 차이가 큽니다. 그만큼 데이터가 다르기에 2023년 6월 이후부터 2024년 9월에 상장한 회사 총 152개 회사의 IPO들의 데이터에 대하여 다루었습니다.

|  |  |  |
| --- | --- | --- |
| 라벨 | 독립변수 | 변수 설명 |
| X1 | 공모가 | 기업이 상장 주관사와 함께 적어낸 희망 가격과 기관투자자가 적어낸 희망 가격의 확정 난 가격 |
| X2 | 기관경쟁률 | 기관 투자자가 신청한 주식수를 해당 기관에 할당된 주식수로 나눈 것 |
| X3 | 의무보유확약 | 공모주를 상장이후 일정기간동안 팔지 않을 것을 약속한 비율 |
| X4 | 보호예수비율 | 은행, 증권회사 등 주식을 대량으로 구매하거나 회사의 많은 지분을 보유한 경우 의무 보유기간을 설정해 회사의 주식의 판매를 일정기간 동안 늦추는 옵션의 비율 |
| X5 | 나스닥 | 신규 상장 주식의 상장 전 날의 나스닥 지수 |
| X6 | S&P500 | 신규 상장 주식의 상장 전 날의 S&P500 지수 |
| X7 | 영업이익 | 순수하게 기업이 영업활동으로만 벌어들인 이익, 매출액에서 매출원가(구입가 또는 제조원가)와 판관비를 뺀 수치 |
| X8 | 매출액 | 회사가 만든 제품 또는 서비스를 판매한 금액 |
| X9 | 실제시가 | 상장일 신규 상장 주식의 시초가 가격 |
| X10 | 유통비율 | 주식 상장일에 실제로 유통되는 비율 |
| X11 | 공모밴드 상단 | 기업이 상장 주관사와 함께 적어낸 희망 가격과 기관투자자가 적어낸 희망 가격의 확정 난 가격의 밴드 |
| X12 | 공모밴드 하단 |
| X13 | 이전상장 여부 | 코넥스에서 상장한 주식이 IPO를 통하여 상장을 하려는 경우 |
| X14 | 주간사 규모 | 주간사의 크기와 이름값, 주간하는 횟수 등을 보고 3가지 클래스로 나누었다. |

표

공모가위치: 위의 공모밴드상단, 하단을 이용하여 공모가 위치라는 새로운 데이터를 만들었다. 공모밴드 상단과 하단을 이용하여 공모밴드 중간 가격을 구하고 공모가와 차이를 비교하여 공모가위치를 구하였다.

먼저 코넥스 시장은 코스닥 코스피에는 상장하기에 조건이 부족한 중소기업들을 상장시켜놓은 소규모 시장이다. 이전상장을 따로 체크한 이유는 이전상장의 경우 다른 신규 상장하는 경우와 달리 코넥스 시장에서 코스닥, 코스피 시장에 비해서는 소수이지만 거래가 지속되고 있었기 때문에 데이터가 다를 수 있기에 확인했다.

공모주 청약을 하다 보면 어떠한 증권사는 매달 청약을 하는 회사가 하나 이상씩 꼭 있고 어떠한 증권사들은 1년에 하나 있거나 만 것도 있다. 또한 kb증권이나 nh증권처럼 자주 들어보고 규모도 큰 증권사가 있는 반면 신영증권, 유진투자증권처럼 비인기 증권사들도 있다. 이러한 것들을 고려하여 클래스를 나누어 보았다.

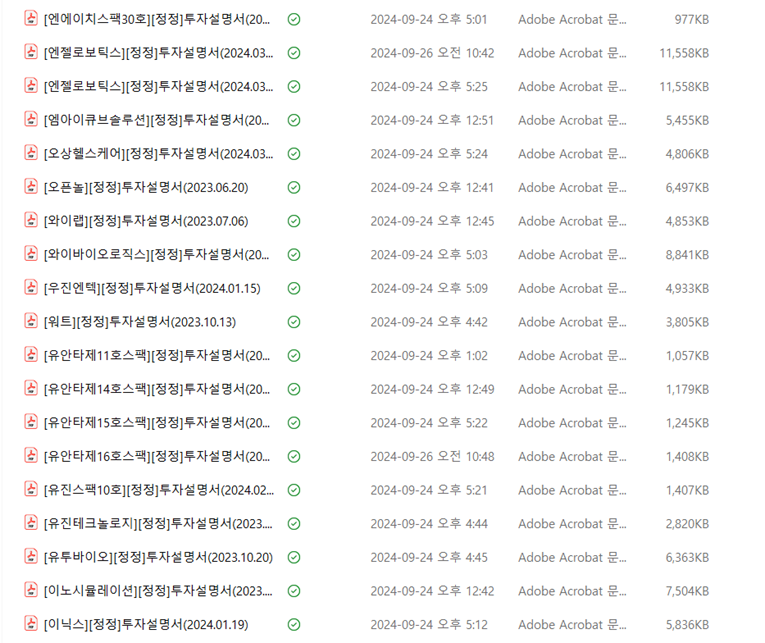


그림

시가에서 공모가를 뺀 값을 공모가로 나누어 시가 수익률이라는 데이터를 획득한 후 30프로 이상과 이하 두가지 경우로 클래스를 나누었다. 첫 번째 목표로 공모 청약을 해야 할 지 안 해야 할 지에 대한 판단을 내려야 하는데 이에 대한 기준을 30프로로 잡았다. 팔았을 때 최소한 30프로 이상의 수익을 얻어야 이 공모주를 청약하는 과정에 대한 최소한의 수치를 30프로로 판단하였다. 그리하여 시가 수익률이 30프로 이상일시 1 이하일시 0으로 클래스를 분류했다.

B) 데이터 수집

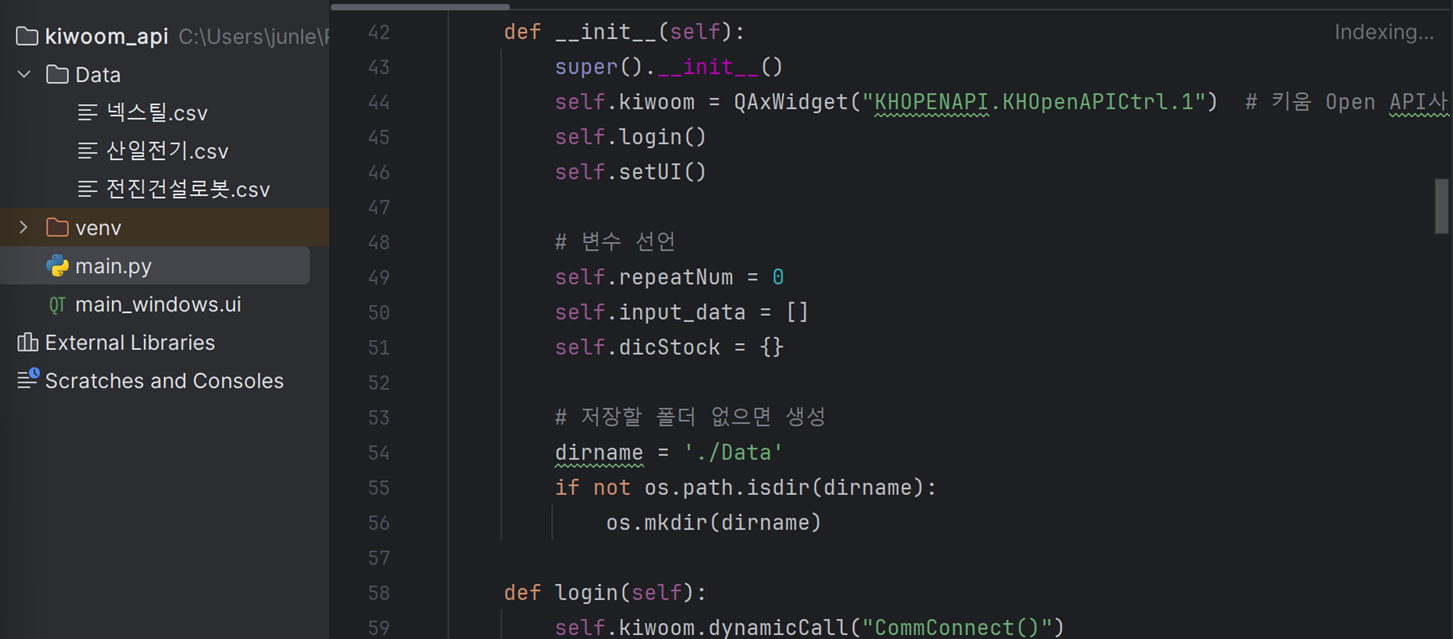
1. DART API:



그림

각 기업의 증권신고서를 다운로드 후 파이썬으로 원하는 데이터를 추출한다.

2. 키움 OPEN API:



그림

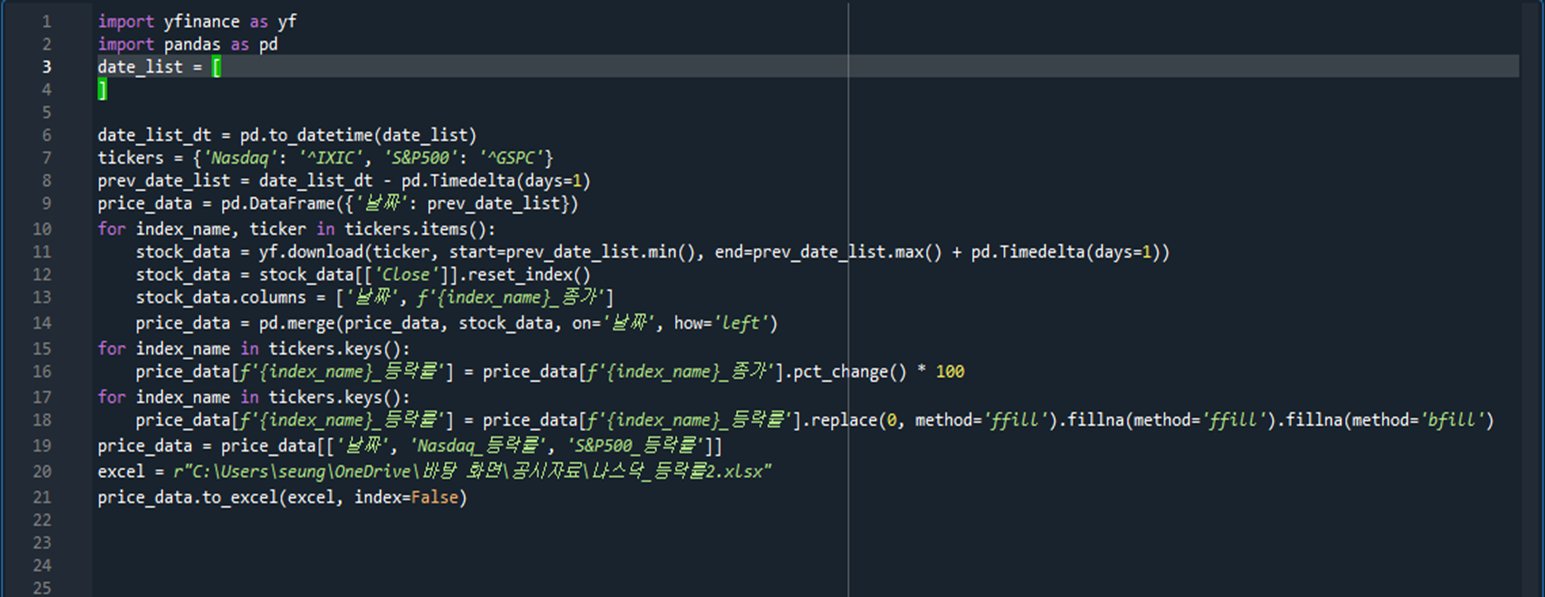
텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림

키움 OPEN API를 통하여 신규상장날의 시가, 고가 등의 정보를 추출한다.

3.나스닥, S&P500 지수:



그림

기존에 얻은 상장날짜 데이터와 Yfinance를 이용하여 신규상장 전일의 미국 지수를 추출

C) 데이터 전처리

로그처리: 이상치의 영향력을 감소시킨다. 최솟값과 최댓값의 차이가 매우 큰 경우 로그를 통해 차이를 줄였다. 예를 들면 시가총액의 차이가 어떠한 기업은 1조이상이고 어떠한 기업은 300억이하이므로 차이가 심한 경우에 로그를 처리해주었다.

인위적인 변형: 공모밴드 상단과 공모밴드 하단 그리고 공모가 가격을 통해 공모가 위치라는 새로운 데이터를 생성했다. 또한 시가수익률을 계산하여 30프로이상의 경우와 이하의 경우에 2개의 클라스로 구분하여 종속변수로 사용했다.

스모트: 두 분류간의 데이터 수 불균형을 해소시킨다. 시가수익률이 30프로 이하인 경우의 데이터가 부족하여서 데이터에 대한 불균형을 해결하기 위하여 도입했다.

스케일러: 각 변수의 영향력을 동일하게 조정해준다. 데이터 값들을 평균이 0, 표준편차를 1이 되도록 변형하는 방법이다.

상관성 측정: 이를 통하여 상관계수가 0.7이상인 공모가, 유통물량, 시가총액에 대하여 PCA와 kernal PCA를 적용했다. 또한 보호예수비율과 유통비율의 상관계수가 1이 나왔다. 이러한 정보를 통하여 둘의 데이터가 정확히 일치한다고 파악하였고 하나만 데이터로 적용했다.

Dimension reduction model: PCA를 통하여 원본 데이터의 분산을 최대한 보존하는 새로운 축을 찾아 데이터의 차원을 축소했다.

Kernal PCA를 통하여 비선형 데이터를 분석하여 더욱 복잡한 패턴을 파악할 수 있도록 새로운 축을 찾아 데이터 차원을 축소했다.

**3. 분류**

A) Random Forest

랜덤 포레스트 분석법은 분류와 회귀 모두에 사용 가능한 앙상블 학습 기법이다. 여러 개여 의사결정 나무(Decision Trees)를 훈련시켜 훈련 결과를 결합하여 최종 예측을 도출한다. 결과를 결합하는 방법에는 여러가지 방법이 있지만 이 논문에서는 개별 나무들의 예측 결과를 다수결 방식으로 취합하여 가장 득표수가 많은 클래스를 최종 예측 값으로 선책하였다. 랜덤 포레스트 분석법의 작동 원리는 아래와 같다.

-무작위 데이터 샘플링(Bootstrap Sampling): 원본 데이터에서 여러 개의 샘플을 무작위로 추출하여 각기 다른 결정 나무들을 훈련시킨다. 이 과정에서 복원추출이 사용되며 하나의 데이터 샘플이 여러 번 선택될 수 있다.

-다양한 트리 구성(Building Trees): 복원추출이 이루어지기 때문에 여러 샘플이 생성된다. 이런 과정에 의해 각 나무는 서로 다른 트리 구조를 가지게 된다. 이로서 예측의 다양성이 확보된다.

-다수결에 의한 결과 도출(Bootstrap aggregation): 각 트리의 결과를 취합하여 가장 많은 예측 결과를 가진 클래스를 최종 결과로 도출.

-특징 랜덤화: 각 의사결정 나무를 구성할 때, 분할에 사용되는 변수를 무작위로 선택한다. 일반적으로 나무의 각 분기점에서 사용 가능한 모든 변수 중 일부만 선택하여 최적의 분할 기준을 찾는다. 이로서 예측의 과적합을 방지한다.

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명-변수 중요도 측정: 각 트리에서 어떤 변수가 얼마나 자주 사용되는지를 계산하여 각 변수의 중요도를 계산합니다.

그림

랜덤 포레스트 분석 법을 이용하여 시가의 수익률을 +30%기준으로 두개의 클래스로 분류해보았다. 30%이상을 ‘A’ 30%이하를 ‘B’ 클래스로 분류하였다.

|  |  |
| --- | --- |
| Class | Range(시가수익률) |
| A | 시가수익률 30% |
| B | 시가수익률 30% |

표

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명전처리 과정을 통해서 최종적으로 선정한 7개의 변수에 대해서 모든 조합으로 분석을 진행하여 Accuracy를 구하였으며 가장 높은 Accuracy를 가진 5개의 변수 조합은 위와 같다. 같은 Accuracy를 가진 조합들이 여러 개 있으면 단순성의 원리에 의해서 가장 적은 변수 수를 가진 조합을 선택하였다. 가장 높은 Accuracy를 보이는 변수 조합은 [로그기관경쟁률, 의무보유확약, 보호예수비율] 이다. 이 변수 조합으로 분석을 해본 결과 아래와 같은 결과를 얻었다.

그림 8

그림

45개의 Test Set에서 33개의 데이터를 올바르게 분류하였다. 이번 프로젝트에서는 B의 Precision을 높이는 것이 중요하다. B로 분류한 상장 주식을 이용하여 다양한 매수 전략을 구성할 것이기 때문에 B로 분류한 데이터들 중 올바르게 분류한 데이터의 비율이 중요한 것이다. 이런 관점에서 봤을 때 B의 Precision이 0.91이라는 것은 나쁘지 않은 결과라고 보여진다.

B) SVM

Support Vector Machine(SVM)은 분류와 회귀 모두에서 사용할 수 있는 지도 학습 알고리즘이다. 주로 고차원의 데이터 즉 데이터의 개수가 많고 데이터들의 관계가 복잡할 때 효과적이다. 소수의 중요한 데이터(서포트 벡터)를 이용하여 최적의 결정 경계(초평면)을 찾는다. 이런 SVM의 작동 원리는 아래와 같다

-결정 경계: SVM은 주어진 데이터 포인트를 두 클래스로 나누기 위해 결정 경계를 찾는다. 이 결정 경계는 두 클래스 간의 마진을 최대로 만드는 선 혹은 평면을 의미한다. 이때 마진은 결정 경계와 각 클래스의 데이터 포인트 중 가장 가까운 데이터 포인트 사이의 거리이다. SVM은 이 마진을 최대화하는 방향으로 학습하여 최적의 경계를 찾는다.

-서포트 벡터: 결정 경계에 가장 가까운 데이터 포인트를 서포트 벡터라고 한다. SVM에서는 이 벡터만을 이용하여 모델이 학습된다. 따라서 클래스의 다른 데이터 포인트들은 결정 경계를 정의하는 것에 영향을 끼치지 못한다. 이런 원리에 따라서 다양하고 대규모의 데이터셋에서도 매우 효율적으로 학습할 수 있다.

-슬랙 변수: SVM은 데이터가 완벽하게 선형적을 구분되지 않을 경우를 고려하여 포인트가 결정 경계를 약간 넘어설 수 있도록 허용하는 슬랙 변수를 도입한다. 이 슬랙 변수로 인해서 데이터가 약간의 오류를 포함할 수 있지만 SVM의 목표는 최대 마진을 찾는 것이 목표이기에 선형적으로 구분되지 않는 데이터에 대한 유연성을 제공할 수 있다.

-커널 트릭: 데이터가 선형적으로 구분되지 않는 경우 SVM은 데이터를 더 높은 차원으로 변환하여 선형적으로 구분할 수 있게 만든다. 커널 함수로는 다항식 커널, RBF 커널, 선형 커널 등이 있다. 이 논문에서는 RBF커널을 사용했으며 이를 통해서 데이터를 실제로 고차원으로 변환하지 않아도 변환된 것처럼 계산할 수 있도록 했다.

도표, 종이접기, 패턴, 디자인이(가) 표시된 사진

자동 생성된 설명

그림

SVM 분석 법을 이용하여 시가의 수익률을 +30%기준으로 두개의 클래스로 분류해보았다. 30%이상을 ‘A’ 30%이하를 ‘B’ 클래스로 분류하였다.

|  |  |
| --- | --- |
| Class | Range(시가수익률) |
| A | 시가수익률 30% |
| B | 시가수익률 30% |

표

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

그림

전처리 과정을 통해서 최종적으로 선정한 7개의 변수에 대해서 모든 조합으로 분석을 진행하여 Accuracy를 구하였으며 가장 높은 Accuracy를 가진 5개의 변수 조합은 위와 같다. 같은 Accuracy를 가진 조합들이 여러 개 있으면 단순성의 원리에 의해서 가장 적은 변수 수를 가진 조합을 선택하였다. 가장 높은 Accuracy를 보이는 변수 조합은 [로그기관경쟁률, 의무보유확약] 이다. 이 변수 조합으로 분석을 해본 결과 아래와 같은 결과를 얻었다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림

45개의 Test Set에서 36개의 데이터를 올바르게 분류하였다. 또한 B의 Precision이 0.92로 Random forest 분석 보다는 0.01정도 높은 precision을 보인다.

C) Logistic Regression

로지스틱 회귀분석(Logistic Regression)은 종속 변수가 이진 분류(binary classification)인 문제를 해결하는 데 주로 사용되는 통계적 모델이다. 즉, 출력 값이 0 또는 1과 같은 두 가지 범주 중 하나로 분류될 때 사용된다. 회귀라는 용어를 사용하지만 실제로는 분류 문제를 해결하는 데 사용되며, 확률 값을 기반으로 예측한다. 로지스틱 회귀분석의 주요 작동 원리는 아래와 같다

-시그모이드 함수: 로지스틱 회귀는 시그모이드 함수를 사용하여, 예측된 값을 0과 1 사이의 확률로 변환합니다. 시그모이드 함수의 수식은 다음과 같다.

여기서 z는 독립 변수들의 가중합이며, 이 함수는 입력 값을 확률로 변환한다.

-결정 임계값: 시그모이드 함수를 이용하여 출력한 확률값을 이용하여 두가지 클래스로 구분하는 기준이 되는 임계값을 설정한다. 예를 들어 예측된 확률이 일정 기준(0.5)를 넘으면 1 그렇지 않으면 0으로 분류된다. 이 임계값을 이용하여 모델의 예측 성능과 정확도를 향상시킬 수 있다.

-로그 손실 함수: 로지스틱 회귀는 손실 함수로 로그 손실 함수(또는 이진 크로스 엔트로피)를 사용한다. 이 손실 함수는 예측된 확률 값과 실제 값을 비교하여 모델의 성능을 평가하는 데 사용한다. 로그 손실 함수는 다음과 같다.

****

-경사 하강법: 로지스틱 회귀는 최적의 가중치 파라미터를 찾기 위해 경사 하강법을 사용한다. 경사 하강법은 손실 함수를 최소화하는 방향으로 가중치를 반복적으로 업데이트하는 방식이다.

폰트, 타이포그래피, 디자인, 서예이(가) 표시된 사진

자동 생성된 설명-독립 변수와 종속 변수의 관계: 로지스틱 회귀는 독립 변수들이 종속 변수의 로그 오즈비(log-odds)에 선형 관계를 가진다고 가정한다. 여기서 오즈비(Odds ratio)는 특정 이벤트가 발생할 확률과 발생하지 않을 확률의 비율이다.

오즈비의 로그 값은 선형 관계를 형성한다, 로지스틱 회귀는 이 로그 오즈비를 모델링한다.

Logistic Regression 분석법을 이용하여 시가의 수익률을 +30%기준으로 두개의 클래스로 분류해보았다. 30%이상을 ‘A’ 30%이하를 ‘B’ 클래스로 분류하였다.

|  |  |
| --- | --- |
| Class | Range(시가수익률) |
| A | 시가수익률 30% |
| B | 시가수익률 30% |

표

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

그림

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명전처리 과정을 통해서 최종적으로 선정한 7개의 변수에 대해서 모든 조합으로 분석을 진행하여 Accuracy를 구하였으며 가장 높은 Accuracy를 가진 5개의 변수 조합은 위와 같다. 같은 Accuracy를 가진 조합들이 여러 개 있으면 단순성의 원리에 의해서 가장 적은 변수 수를 가진 조합을 선택하였다. 가장 높은 Accuracy를 보이는 변수 조합은 [로그기관경쟁률, X21, 보호예수비율, 미국지수] 이다. 이 변수 조합으로 분석을 해본 결과 아래와 같은 결과를 얻었다.

45개의 Test Set에서 36개의 데이터를 올바르게 분류하였다. 또한 B의 Precision이 0.92로 Random forest 분석 보다는 0.01정도 높고 SVM과는 동일한 precision을 보인다.

그림

지금까지 이용한 모델들은 Train Set을 일정하게 유지한 채 학습을 진행시켰다. 랜덤한 Train Set을 선정하여 학습을 진행하면 더 낮은 Accuracy가 나온다는 사실을 알았다. 따라서 이후 연구에서는 위 분석 모델들을 랜덤한 Train Set을 이용하여 학습시키며 그 Accuracy를 높일 수 있는 방법을 모색해야 한다.

**4. 딥러닝**

모든 딥러닝 머신에는 다음과 같은 분석 방법이 공통적으로 사용되었다

또한 수익계수를 예측하는 것이 아닌 시가수익률을 예측하는 모델을 설계해 보았다.

여기서 말하는 Acurracy는 예측 정확도로 Test set의 샘플들을 오차범위 내로 예측한 비율이다.

-K-fold: 데이터셋을 K개의 부분으로 나눈 후, K-1개의 부분을 학습 데이터로 사용하고 남은 1개의 부분을 검증 데이터로 사용하는 방법이다. 이 과정을 K번 반복하여 각 부분이 한 번씩 검증 데이터로 사용된다. K-fold 교차 검증은 모델의 일반화 성능을 평가하는 데 유용하며, 데이터셋의 모든 데이터를 최대한 활용할 수 있는 장점이 있다. 또한, 오버피팅을 방지하는 데 도움을 준다.

-Early-Stopping: 딥러닝 모델의 학습 과정에서 과적합(overfitting)을 방지하기 위해 사용되는 기법이다. 학습이 진행되는 동안 검증 데이터의 성능이 일정 기간 동안 개선되지 않으면 학습을 중단하는 방식이다. 이 방법은 훈련이 계속될수록 모델이 훈련 데이터에 과도하게 적합되는 것을 방지하고, 최적의 모델을 조기에 선택할 수 있도록 돕는다. 조기 중단은 훈련 시간을 단축시키고, 모델의 일반화 능력을 향상시키는 데 기여한다.

-Tolerance Factor: 수익률을 완벽하게 예측하는 것은 비효율적이기에 Test set에 대해서 학습을 진행할 때 종속변수에 오차범위를 도입하였다. 이로 인해 학습은 조금 덜 fit하게 이루어진다.

-역전파: 손실함수를 통해서 MLP의 예측값과 실제 값 사이의 차이를 계산하고 Loss를 구한다. 이때 자주 사용되는 함수로는 교차 엔트로피와 MSE가 있다. 이런 손실 함수로 구한 Loss를 최소화하기 위해서 가중치와 Bias를 조정하는 과정이 역전파 알고리즘이다. 각 뉴런의 출력에 대한 손실의 기울기를 계산하고 이를 바탕으로 경사 하강법을 통해 가중치와 Bias를 업데이트한다. 가중치의 변화량은 다음과 같은 식으로 계산한다.

폰트, 화이트, 디자인, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

은 손실 함수에 대한 가중치의 기울기이다.

해당 프로젝트에서는 교차 함수로 MSE를 사용하였다.

-Epoch: 전체 데이터셋을 모든 신경망에 통과시키는 과정을 Epoch이라고 한다. 한번의 학습에 하나의 Batch size에 해당하는 수의 데이터를 사용한다. 따라서 여러 번의 학습을 통해서 전체 데이터셋을 학습시킨다. Epoch수를 늘리게 되면 높은 예측 성능을 가질 수 있지만 과적합이 될 가능성이 높고 Epoch이 낮아지면 과적합을 예방할 수 있지만 예측 성능이 떨어질 수 있다.

A) MLP

MLP(Multilayer Perceptron)는 딥러닝의 기본적인 신경망 구조 중 하나로, 완전 연결 인공 신경망의 일종이다. MLP는 주로 분류, 회귀, 패턴 인식 등의 문제에 적용되며, 입력 데이터를 여러 계층에 걸쳐 처리하면서 복잡한 비선형 관계를 학습한다. MLP의 작동 원리는 다음과 같다.

폰트, 텍스트, 타이포그래피, 화이트이(가) 표시된 사진

자동 생성된 설명-MLP의 구조: 입력층과 은닉층 출력층으로 이루어져 있다. 입력층은 외부로부터 데이터를 받아드린다. 이때 각 입력 변수가 입력층의 하나의 뉴런에 해당한다. 입력층의 뉴런 수는 주어진 데이터의 차원에 따라 결정되며 이런 데이터를 그대로 은닉층에 전달한다. MLP는 하나 이상의 은닉층을 가질 수 있다. 이런 은닉층의 각 뉴런은 이전 층의 모든 뉴런과 연결되어 있으며 연결된 뉴런에서 여러 개의 입력을 받아 가중합을 계산한다. 가중합 계산 식은 아래와 같다.

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명출력층은 최종 예측값을 생성한다. 다양한 활성화 함수를 이용한다. 보통 ReLU, Linear가 널리 사용되며, 이진 분류 문제에서는 시그모이드 함수, 다중 클래스 분류 문제에서는 소프트 맥스 등의 함수를 사용한다. 출력층에서는 이런 함수를 이용하여 각 클래스에 속할 확률을 계산한다. 해당 프로젝트에서는 Linear를 사용하였다.

그래프

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림

그래프를 보면 Validation Loss가 매우 매끄럽게 감소하고 0.3 수준을 유지하고 있다. 또한 Validation Loss가 Train Loss 아래에 위치하고 있어 과적합이 이루어지지 않았다는 것을 알 수 있다. 하지만 표를 보면 R-2 score가 매우 낮게 나온 것을 보아 정확도가 매우 떨어진다는 것을 알 수 있다. 데이터가 선형적인 관계가 부족하고 복잡한 패턴이 있기 때문에 단순한 MLP모델로는 정확한 예측이 안되는 것으로 보인다.

B) CNN

텍스트, 그래프, 라인, 도표이(가) 표시된 사진

자동 생성된 설명CNN리 및 컴퓨터 비전 분야에서 주로 사용되는 딥러닝 모델이다. CNN은 Convolution layer, pooling layer, fully connected layer로 구성되어 있으며, 이미지의 공간적 특성을 효과적으로 학습할 수 있도록 설계되었다. 합성곱 계층은 이미지의 특징을 추출하고, 풀링 계층은 차원을 축소하여 계산 효율성을 높인다. 주로 이미지 분석에 사용되지만 이 프로젝트에서는 다양한 금융 데이터들을 2D로 변환하여 변환한 결과를 가지고 CNN 분석을 진행하였다.

그래프

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림

Validation Loss와 Train Loss는 0.25 수준에서 수렴하고 있지만 Early-Stopping으로 인해서 학습에 사용된 Epochs이 너무 적다는 문제가 발생한다. 이렇게 적은 Epoch은 특징 추출 성능의 감소와 데이터의 복잡성 반영 능력을 감소시킨다. 또한 표를 보면 R-2 score가 매우 낮게 나온 것을 보아 정확도가 매우 떨어진다는 것을 알 수 있다. CNN또한 좋은 성능을 보이지 않고 있다.

C) CNN + MLP

CNN과 MLP 하이브리드 모델은 이미지화 된 데이터와 같은 고차원 데이터를 처리하기 위해 설계된 강력한 딥러닝 구조이다. 이 모델은 CNN을 사용하여 이미지의 특징을 추출한 후, MLP를 통해 이러한 특징을 기반으로 최종 분류나 회귀 작업을 수행한다. CNN을 이용하여 복잡하고 국소적인 특징과 패턴을 추출하고 이를 반영하여 MLP를 통해 최종적인 모델 학습을 진행한다.

라인, 그래프, 스크린샷, 텍스트이(가) 표시된 사진

자동 생성된 설명CNN부분에서는 전체적인 특징부터 국소적인 패턴을 학습하기 위해서 filter를 64로 시작하여 128로 증가시켰고 이렇게 추출한 국소적인 패턴을 반영하기 위해서 MLP의 Layer Dense를 256부터 시작하여, 128, 64, 32, 16, 1 로 감소시키며 학습을 진행하였다. Layer Dense가 클수록 더 복잡한 패턴을 파악하는 것에 유리하기 때문이다. 이를 통해서 다음과 같은 결과를 얻었다.

그래프

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 16

CNN과 동일하게 Epoch의 수가 너무 적어 특징 추출에 효과적이지 못한 모델이다. 또한 정확도는 다른 모델보다는 높았지만 R-2 score가 더 낮아 예측성이 높다고 말할 수 없다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명또한 로지스틱회귀분석, SVM, Random Forest를 통해서 구한 수익계수를 변수로 하여 딥러닝 모델들을 학습시키는 방법도 있다. 로지스틱회귀분석과 SVM, Random Forest로 구한 수익계수가 정확하다는 가정하에 다음과 같은 결과를 얻을 수 있었다.

그림 17

0.3은 기존에 정한 임계값이고 0.81은 시가수익률의 평균값, 0.59는 중앙값이다. 수익계수를 0.81을 기준으로 나누었을 때 가장 높은 R-2스코어와 64.33%라는 예측 정확도를 보인다. 따라서 분류 모델들과 딥러닝 모델들의 성능을 모두 향상시키면 어느 정도 높은 예측 정확도를 확보할 수 있을 것이다.

**5. 고가예측**

고가의 수익률을 다음과 같은 기준으로 라벨링하고 이를 종속변수로 설정하여 랜덤포레스트와 LSTM + MLP, CNN + MLP 모델 분석을 실시하였다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 18

고가는 시가와는 다르게 시장의 변동성이 큰 영향을 미친다. 따라서 미국지수와 코스닥 등락률이 중요한 변수로 작용한다. 이를 위해서 시가를 예측하기 위한 다양한 모델보다는 LSTM모델과 MLP모델을 결합한 딥러닝 모델, CNN과 MLP모델을 결합한 딥러닝 모델을 사용하기로 했다. 또한 기존 시가 예측에서 가장 높은 정확도를 보인 RandomForest을 이용하여 분석을 진행해보겠다.

1. CNN + MLP

CNN과 LSTM을 결합한 모델은 다음과 같이 레이어를 설정했고 결과는 다음과 같다.

그림 19

그림 20

A클래스에서 약 0.52의 f1 score, B클래스에서 약 0.48의 f1 score를 나타낸다. 이는 성능이 매우 안좋은 모델이라는 의미이다. CNN과 MLP의 결합 모델은 여러 측면에서 우리가 가지고 있는 데이터를 효과적으로 분석하지 못한다. 그 이유는 아래와 같다.

일단 CNN은 복잡한 수치 데이터를 분석하지 못한다. 이 모델에서 사용한 데이터는 기관경쟁률, 의무보유확약, 주간사규모, 보호예수비율, 미국지수, 공모가위치, Kosdaq\_등락률을 이용했다. 데이터가 다 수치로 이루어져있고 복잡성이 높기 때문에 이를 이미지화 하여 분석하는 CNN 기법은 좋지 않은 기법이었다. 또한 데이터 수가 160여개로 매우 적기 때문에 단일 모델보다 더 복잡한 복합 모델은 우리의 데이터를 분석하는 것에 적합하지 않다는 결론에 도달했다.

1. LSTM + MLP모델

주가 예측에 가장 많이 쓰이는 딥러닝 모델은 LSTM이다. IPO기업들은 개별적인 기업이지만 하나의 ‘공모주 시장’에 속해 있기 때문에 특정 상장 주식의 변화에 직전 상장 주식의 주가 변화가 영향을 줄 수밖에 없다. 또한 같은 섹터에 속해 있는 주식들은 서로 영향을 끼치기 때문에 이를 반영할 수 있는 LSTM 모델이 더 적합하다는 결론을 내렸다.

다음과 같이 모델을 구성해보았다.

그림 21

LSTM + MLP 모델의 작동 방식은 다음과 같다. LSTM을 이용해 미국지수, Kosdaq\_등락률 을 분석하고 나머지 데이터를 MLP 알고리즘을 이용하여 분석한 후 두 경로의 출력값을 결합하여 최종 예측값을 산출한다. LSTM + MLP 모델의 성능은 다음과 같다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 22

A클래스의 f1 score는 0.67, B클래스의 f1 score는0.54정도로 좋은 예측 성능을 보이지는 않는다. 이 또한 적은 데이터수에 의한 결과라고 판단된다.

1. RandomForest

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명전처리 과정을 통해서 렌덤포레스트 분석 방법의 최적 변수를 다음과 같이 구하였다.

그림 23

가장 높은 정확도를 보이는 변수들로 분석을 해본 결과 다음과 같은 결과를 얻었다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 24

LSTM + MLP모델과 비교해 보았을 때 정확도가 약간 높은 것을 확인할 수 있다. 고가 예측에서도 랜덤포레스트 분석법이 가장 좋은 성능을 보인다.

다른 딥러닝 모델과 비교했을 때 랜덤포레스트 분석법이 높은 정확도를 보이는 이유는 복잡한 모델이 우리가 다루는 데이터에 적합하지 않다는 것이다. 적은 데이터수로 인해서 데이터의 복잡한 패턴을 파악하기 힘들다. 따라서 세세한 패턴을 파악하여 분석하는 딥러닝 모델들은 낮은 정확도를 보이고 비교적 매우 간단한 RandomForest 분석법은 비교적 높은 정확도를 보이는 것이다.

단순한 딥러닝 모델을 통해서는 정확도 높은 예측을 하기 어렵다는 결론에 도달했다. 일반적인 딥러닝 주가 예측 모델은 적게는 5000개의 데이터에서 많게는 10000개가 넘는 데이터를 활용한다. 하지만 현재 대한민국에서 공개된 의미 있는 공모주 데이터는 160개 정도이기 때문에 딥러닝을 통한 예측은 힘들다

따라서 공모주 시가/고가에 대한 정확한 예측보다는 공모주의 시가/고가에 대한 의미이는 영향을 끼치는 변수를 파악하는 것에 초점을 맞추어 이후 프로젝트를 진행하였다.

**5. 추가적인 데이터 전처리.**

A) 산포도 분석

기존에 진행했던 딥러닝 모델의 한계로서 데이터가 너무 단순하다는 점이 있었다. 따라서 제일 단순한 분석법을 이용하여 데이터의 특성을 알아보고자 했다. 보호예수비율을 이용하여 산포도 그래프를 그리면 다음과 같은 그래프를 얻을 수 있다.

그림 25

그래프를 보면 고가 수익률이 0.4가 넘는 모든 종목들은 보호예수비율이 50을 넘는다. 이를. 이 결과에 따라서 보호예수비율이 50넘는 데이터들을 이용해서 다음과 같은 산포도 분석을 계속했다.

작년매출액, 유통물량, 시가총액을 이용해서 다음과 같이 산포도 그래프를 그려보았다.

텍스트, 라인, 그래프, 번호이(가) 표시된 사진

자동 생성된 설명

그림 27

그림 26

**라인, 그래프, 텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명**

그림 28

시가총액, 유통물량, 작년매출을 기준으로 0에 가까울수록 고가 수익률이 높은 데이터가 분포한다. 이를 반영해서 세가지 변수를 스케일링하고 평균낸 값을 새로운 데이터 K로 설정하여 다시 산포도 그래프를 그려보았다.

**라인, 그래프, 텍스트, 도표이(가) 표시된 사진

자동 생성된 설명**

그림 29

이를 통해서 데이터를 분석해보면 고가 수익률이 0.5이상인 데이터 16개중 13개가 모두 K데이터 값 기준 음수에 분포한다는 사실을 얻게 되었다.

**6. 클러스트링**

A) 클러스터링목적

공모주를 분석하는데 있어서 산업군마다 혹은 기업의 특징마다 공모주의 변동성과 같은 특성에서 차이를 가질 것이라 생각했다. 그래서 공모주 사이의 관계를 알아내거나 서로 같은 특징을 같는 공모주끼리 분리하기 위하여 클러스터링을 통해 분석을 시도하였다.

해당 분류는 공모주 또는 해당 기업의 특징을 분석하기 위한 것이기 때문에 클러스터링에 사용한 데이터는 공모주의 특징을 잘 나타내는 것으로 보여지는 데이터들로 선정하였다. 이하의 4개의 데이터를 사용하여 클러스터링을 진행하였다.

'기관경쟁률', '의무보유확약', '시가총액', '보호예수비율’

이하의 3개의 데이터가 분석에 사용된 추가적인 데이터이다.

‘스팩주여부’, ‘공모가위치’, ‘공모여부’

스팩주: 공모주가 스팩주인지 아닌지 라벨링 한 것

(스팩란 특수목적인수회사의 약자로 이미 상장된회사가 특정 기업을 인수하기 위해 만들어진 회사이다. 스팩주 상장을 통해 얻은 자금을 통해 비상장 기업을 인수한다.)

공모가 위치: 공모가(X)와 공모밴드의 상대적인 거리를 나타내는 수치 공모밴드 중심으로 상단(A), 하단(B)과 얼마나 떨어져 있는지를 나타낸다.

(공모가 위치 = (X-(A+B)/2)/((A+B)/2) )

공모여부: 시가수익률이 30%가 넘을 경우 A, 그 이외에는 B

B) 클러스터링알고리즘

* StandScaler

데이터를 표준화하는 방법으로 데이터의 평균을 0으로 하며 표준편차가 1이 되게 변환하는 전처리 방식이다. 데이터의 각 항목 별로 크기 차이가 클 경우 특정 항목이 학습에 지배적일 수 있기 때문에 해당 과정을 통해 영향을 줄일 수 있다. 해당 과정은 데이터가 정규분포에 유사할 경우 효과적이다.

z= (x−μ)​ / σ

* MinMaxScaler

데이터 표준화 방법이며 데이터가 0이상 1이하가 되게 변환하는 방법이다. 이 방법도 데이터의 영향을 균일하게 하기 위한 과정이다. 해당 과정은 최대최소값에 따라 결정된다. 그래서 이상치가 존재할 경우 전체 데이터 범위가 쉽게 영향을 받아 민감하다.

* RobustScaler

데이터 전처리 기법 중 하나로 데이터를 중앙값과 사분위를 이용해 변환하는 방법이다. 이 방법은 데이터를 중앙값을 기준으로 1사분위수와 3사분위수의 차이로 나누는 방식이다. 이 방식은 standscaler나 minmaxscaler보다 이상치에 강하다.

* K-means

k-means 클러스터링은 데이터를 주어진 k개의 클러스터로 나누는 비지도 학습 알고리즘이다. 임의의 k개의 중심점을 데이터 공간에 배치한다. 이후 각 중심점과 가장 가까운 중심점을 골라 중심점의 클러스터에 할당한다. 이때 주로 L2-norm을 사용한다. 그 후 중심점을 클러스터 내 데이터의 평균 위치로 업데이트하고 위 과정을 중심점에 변화가 없을 때까지 반복한다. 이 과정을 통해 k-means 알고리즘이 데이터를 나눈다.

이 알고리즘은 k값을 지정해야하는 필요성이 있으며 거리를 사용하여 분류하기 때문에 구형으로 뭉쳐진 데이터가 아닌 경우 잘 작동하지 않는다.

k값을 결정할 때는 elbow method를 사용하여 결정한다. 각 k 값에 따른 클러스터링을 진행한 뒤 각 k에 대해 SSE를 계산한 뒤 SSE가 급격히 감소하다 완만해지는 지점의 k를 선택한다.

SSE: 클러스터 내 거리의 제곱합이다. 클러스터링을 평가할 경우 각 클러스터의 sse 점수 평균을 사용한다.

실루엣 점수: 군집 분석에 사용되는 지표로 각 포인트 I 에 대해 a\_i가 다른 포인트와 평균 거리이며 b\_i가 가장 가까운 군집의 포인트와의 평균거리일 때

전체 실루엣 점수는 S\_i의 평균이 된다.

* DBSCAN

DBSCAN은 k-means와는 다르게 밀도 기반 클러스터링 알고리즘으로 밀도가 높은 데이터를 클러스터링으로 묶고 밀도가 낮은 데이터를 노이즈로 여긴다. 이 알고리즘은 클러스터의 개수가 알고리즘을 통해 결정된다. 이 알고리즘은 epsilon과 minpts요소를 지정할 수 있다. Epsilon은 한 데이터의 반경을 의미하며 이 반경내 충분한 데이터가 존재하면 클러스터를 형성하며 한 클러스터를 형성하기 위한 최소 데이터 수를 minpts로 지정한다. Epsilon과 minpts로 클러스터의 중심, 경게를 구분짓고 이 밖에 있는 데이터를 노이즈로 간주한다. 밀도 방식이므로 k-means와 달리 구형 데이터가 아닌 형태도 구분할 수 있다.

* Affinity Propagation

이 알고리즘은 데이터 사이의 유사도를 기반으로 클러스터를 형성하는 밀도기반 알고리즘이다. 일반적으로 거리가 가까울 경우 유사도가 높다고 판단하고 이를 유사도 행렬로 저장한다. 유사도가 가장 높은 지점에서 각 데이터를 대표 포인트로 선택하고 한 데이터가 다른 데이터를 대표자로 선택할 가능성을 나타내는 책임도와 한 데이터가 대표가 될 가능성인 가용성 계산한다. 각 포인트의 책임도와 가용성을 반복적으로 계산하여 클러스터링을 수행한다.

* Ward’s linkage

병합식 계층적 클러스터링의 한 종류로 데이터의 분산을 최소화하는 방법으로 병합을 진행한다. 모든 데이터를 개별 클러스터로 본 뒤 두 클러스터링 병합했을 때 가장 분산 증가량이 적은 두 클러스터를 병합한다. 이를 반복하여 1개가 될때까지 병합한다. 이를 통해 얻은 클러스터는 다른 병합 방식보다 밀도가 고르게 유지된다. 그러나 분산 방식이기때문에 구형 데이터가 아닐 경우 성능이 떨어질 수 있다.

C) 시행방법

클러스터링을 진행할 때 데이터 사이의 간격이 클러스터링에 영향을 줄 것이라 생각하여 데이터와 로그처리한 데이터를 비교햐였으며 정규화 방법으로 StandScaler와 MinMaxScaler, RobustScaler를 사용한 것을 비교하였다.

우선 엘보우 방식과 실루엣 점수를 통해 군집화를 비교할 수 있는 k-means 알고리즘을 주된 알고리즘으로 분석하고 다른 알고리즘을 추가적으로 비교하는 것으로 하였다.

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 30

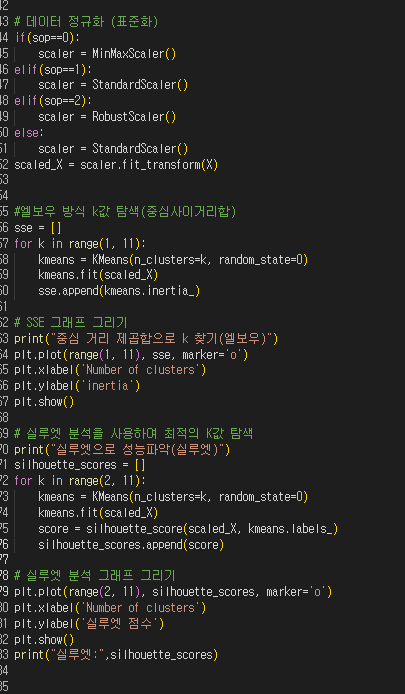


그림 31

위와 같은 코드로 k-means에 대한 데이터, 로그데이터를 3가지 스케일러로 처리한 것을 k에 따라 모두 실루엣을 조사한다. 그중 가장 실루엣점수가 높은 것을 골라 분석한다.

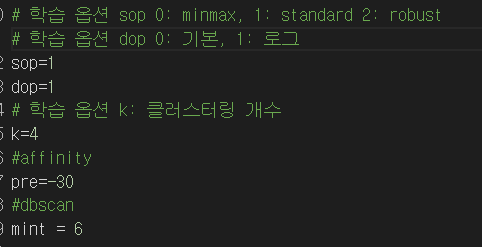


그림 32

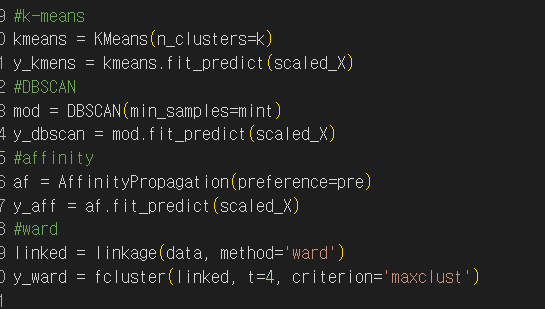


그림 33

그 후 다른 알고리즘과 비교하기 위해서 위와 같은 방법으로 모두 분류하였다.

D) 결과및분석

K-means 클러스터링 결과

Table k-means를 사용해 얻은 실루엣 표

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | k=2 | k=3 | k=4 | k=5 | k=6 | k=7 | k=8 | k=9 | k=10 |
| minmax | 0.610389 | 0.46606 | 0.48294 | 0.461968 | 0.46296 | 0.475735 | 0.45405 | 0.445376 | 0.428599 |
| minmax\_log | 0.655778 | 0.539173 | 0.409355 | 0.387355 | 0.378799 | 0.383222 | 0.408322 | 0.381927 | 0.387141 |
| standard | 0.488737 | 0.4404 | 0.47851 | 0.459247 | 0.461369 | 0.42194 | 0.477205 | 0.418193 | 0.409936 |
| standard\_log | 0.566384 | 0.600704 | 0.612872 | 0.608024 | 0.450761 | 0.458266 | 0.483315 | 0.416931 | 0.444181 |
| robust | 0.840235 | 0.382709 | 0.399564 | 0.395425 | 0.414397 | 0.445506 | 0.3947 | 0.394796 | 0.425252 |
| robust\_log | 0.815917 | 0.378114 | 0.392112 | 0.49442 | 0.523496 | 0.461938 | 0.442802 | 0.438411 | 0.393184 |

표

클러스터링을 k-means로 실행했을 때 데이터 처리 방법과 k값에 따른 실루엣 값이다.

이중 실루엣 값이 가장 높은 것은 robust 스케일로 전처리한 데이터이나 이는 군집의 개수가 2개일 경우만 높고 3개 이상일 경우부터는 급락하기 때문에 k=2일 경우만 의미가 있다.

텍스트, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 34

도표, 스크린샷, 텍스트, 디자인이(가) 표시된 사진

자동 생성된 설명

그림 35

하지만 위 군집화를 색으로 표시한 도표를 볼 경우 해당 분류는 큰 의미가 없다는 것을 알 수 있다. 150개 중 단 5,6,개만 분류하는 것은 군집 분석에 큰 의미가 없다고 여겨져 robust 스케일러는 고려하지 않았다.

라인, 도표, 텍스트, 그래프이(가) 표시된 사진

자동 생성된 설명

그림 36

라인, 도표, 텍스트, 그래프이(가) 표시된 사진

자동 생성된 설명

그림 37

도표, 스크린샷, 텍스트, 디자인이(가) 표시된 사진

자동 생성된 설명텍스트, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명

위도표, 스크린샷, 텍스트, 디자인이(가) 표시된 사진

자동 생성된 설명텍스트, 도표, 스크린샷이(가) 표시된 사진

자동 생성된 설명 엘보우 기법을 위한 그래프를 보아도 k =2 말고는 큰 의미가 없다는 것을 알 수 있다.

그래서 standard와 minmax중에서 가장 높은 실루엣 점수를 받은 방식은

A: standard k=2

B: standard logdata k=4

C: minmax k=2

가 된다. A인 경우 1개의 false true를 제외하고는 스팩주를 분리할 뿐이기 때문에 분석의 가치가 없다고 생각되며 C도 유사하였다. 그래서 B의 경우를 k-means 알고리즘의 적절한 파라미터라고 생각하였다.

Table A경우

텍스트, 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

그림 38

Table B경우

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

그림 39

Table C경우

텍스트, 도표, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명

그림 40

Table C의 엘보우 기법

도표, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 41

이하는 해당 분류 결과 4개의 군집으로 분류된 공모주를 관찰하여 얻은 특징이다

0번 클러스터: (15/150개)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 4/15 | 11/15 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 1/15 | 4/15(모두 스팩주) | 10/15 |

표

이 군집은 의무보유확율이 스팩주만큼 낮으며 대부분이 청약할 가치가 없는 낮은 수익성을 나타낸다.

1번 클러스터: (46/150개)

이 군집은 스팩주를 모았다. 46개 모두 스팩주이며 1번 군집으로 분류되지 않는 스팩주는 0번에 4개, 2번에 1개로 전체 스팩주 51개 중 46개를 분류하여 92%정도의 정확도로 스팩주를 분리하였다.

2번 클러스터: (86/150개)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 75/86 | 11/86 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 81/86 | 1/86 (모두 스팩주) | 4/86 |

표

이 군집은 대부분이 높은 수익률을 보이는 공모주로 0번 군집과 2번 군집을 통해 수익성이 있는 공모주와 없는 공모주로 나누어졌다고 볼 수 있다. 이미 라벨링된 공모여부를 보면 이 클러스터링으로 약 77%의 정확도로 공모주가 분리된 것을 알 수 있다.

3번 클러스터: (3/150개)

해당 클러스터는 표본의 크기가 작아 특징을 발견하기에는 신뢰성이 낮다. 발견된 특징에는 시총이 가장 큰 공모주들이며 시총이 큰 공모주 중에서 보호예수비율이 높은 편이다. 또한 3가지 모두 증권사가 3개 이상인 대형이다.

DBSACN

해당 알고리즘은 3개로 분류되었다.

0번 클러스터 (45/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 25 | 20 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 22 | 10 (모두 스팩주) | 13 |
|  | O | X |  |
| 스팩주 | 10 | 35 |  |

표

1번 클러스터 (41/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 37 | 4 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 0 | 41 (모두 스팩주) | 0 |
|  | O | X |  |
| 스팩주 | 41 | 0 |  |

표

2번 클러스터 (64/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 60 | 4 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 64 | 0 | 0 |
|  | O | X |  |
| 스팩주 | 0 | 64 |  |

표

위 결과에서 1번은 스팩주를, 2번은 수익성있는 공모주를 분리하는 것 같지만 k-means에 비하면 그리 뛰어나게 분리하지는 못하는 모습을 보인다.

Affinity propagation

해당 알고리즘의 경우 그룹이 5개로 분리되었다.

0번 알고리즘 (17/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 8 | 9 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 7 | 10 |  |
|  | O | X |  |
| 스팩주 | 0 | 17 |  |

표

1번 알고리즘 (72/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 67 | 5 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 71 | 1 | 0 |
|  | O | X |  |
| 스팩주 | 0 | 72 |  |

표

2번 알고리즘 (8/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 1 | 7 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 0 | 4(모두 스팩주) | 4 |
|  | O | X |  |
| 스팩주 | 4 | 4 |  |

표

3번 알고리즘 (46/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 41 | 5 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 0 | 46 | 0 |
|  | O | X |  |
| 스팩주 | 46 | 0 |  |

표

4번 알고리즘 (7/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 5 | 2 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 6 | 0 | 1 |
|  | O | X |  |
| 스팩주 | 0 | 7 |  |

표

이 경우 0번 1번이 각각 수익성이 낮은, 높은 공모주를 분리했다고 할 수 있으며 3번이 스팩주를 분리했다고 할 수 있지만 k-means에 비하면 잘 분리되었다고 할 수 없다.

Ward

해당 알고리즘은 그룹이 4개로 분리되었다.

0번 알고리즘 (21/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 15 | 6 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 18 | 0 | 3 |
|  | O | X |  |
| 스팩주 | 0 | 21 |  |

표

1번 알고리즘 (122/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 102 | 20 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 62 | 51 | 9 |
|  | O | X |  |
| 스팩주 | 51 | 71 |  |

표

2번 알고리즘 (2/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 2 | 0 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 2 | 0 | 0 |
|  | O | X |  |
| 스팩주 | 0 | 2 |  |

표

3번 알고리즘 (5/150)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 0 |  |
| 공모여부 | 3 | 2 |  |
|  | 양수 | 0 | 음수 |
| 공모가위치 | 4 | 0 | 1 |
|  | O | X |  |
| 스팩주 | 0 | 5 |  |

표

이 알고리즘이 가장 성능이 낮게 나왔다. 스팩주를 분리하지 못하며 수익성을 분리하지도 못했다.

위의 결과를 통해 공모주의 클러스터링은 k-means를 통한 것이 가장 잘 분리된 것이란 것을 확인할 수 있었다.

**7. 신규 주식 자동매매 시스템**

본 프로젝트는 초기 단계에서 신규주의 시가수익률 분석을 통해 공모 청약 여부를 판단하는 데 초점을 맞췄습니다. 이를 통해 신규주 공모 시점에서의 투자 가능성을 평가하였습니다.

그러나, 프로젝트를 진행하면서 신규주의 매매 관점에서도 지원할 수 있는 모형의 필요성을 느꼈습니다. 이에 따라 프로젝트의 초점을 고가수익률 분석으로 확장하였습니다.

또한, 새로운 데이터를 입력했을 때 해당 데이터가 매매의 관점에서 어떤 수익 가능성을 가지는지 판단할 수 있는 예측 모델을 개발하고자 하였습니다.

키움증권 API를 활용하여 특정 조건에 따라 자동으로 신규주를 매수 및 매도하는 시스템을 구현하였습니다. 아래는 시스템에서 적용된 매수/매도 조건과 실행 로직에 대한 상세 설명입니다.

매수 조건: 랜덤 포레스트 모형을 이용하여 새로운 데이터를 넣은 후 매수 신호가 뜬다면 신규주의 시가에 매수를 실행합니다

매도 조건: 종목의 현재 가격이 매수 가격 대비 15% 이상 상승했을 경우, 해당 종목을 매도합니다. 종목의 현재 가격이 매수 가격 대비 10% 이상 하락했을 경우, 해당 종목을 매도합니다.

**실제 구현 결과:**

텍스트, 스크린샷, 폰트, 디자인이(가) 표시된 사진

자동 생성된 설명

그림 42

Account.py : 계좌를 불러오는 페이지

Gui.ui: 실제 띄우는 인터페이스를 위한 pyqt5를 이용한 페이지

Kiwoom.py : 키움서버와 연결

KiwoomType.py: 키움주문의 상수와 이어주는 페이지

Randomforest.py: 랜덤포레스트를 돌려 새로운 데이터의 예측결과를 알려주는 페이지

Trade.py: 실제 매수 매도 해주는 페이지

Main.py: 모두 모아놓은 페이지

텍스트, 스크린샷, 디스플레이, 폰트이(가) 표시된 사진

자동 생성된 설명

그림 43

Main.py 실행시 첫 화면

텍스트, 스크린샷, 디스플레이, 폰트이(가) 표시된 사진

자동 생성된 설명

그림 44

계좌평가잔고 내역 확인 버튼을 누를시 계좌에 관한 정보를 불러온다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그림 45

실제 12월달에 상장하는 엠앤씨솔루션 신규주에 관한 정보를 입력

텍스트, 폰트, 스크린샷, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

그림 46

텍스트, 폰트, 스크린샷, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

그림 47

입력 후 randomforest 버튼을 누를시 모형의 정확도와 함께 예측결과를 나타냅니다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

그림 48

매수리스트에 들어갔다면 자동매매버튼을 누르게 되면 구매 후 5초에 한 번씩 현재가를 확인하며 매도 조건에 들어올 경우 매도합니다.

**8. 결론**

본 프로젝트는 공모주 시가수익률 분석을 통해 공모 투자의 의사결정을 지원하고, 이를 확장하여 자동매매 시스템을 구현하는 데 성공적으로 초점을 맞추었습니다. 프로젝트 진행 과정에서 2023년 6월 이후의 다양한 데이터를 활용하고, 여러 머신러닝 기법(회귀분석, 분류, 클러스터링, 딥러닝)을 적용하여 공모주 투자와 관련된 중요한 통찰을 도출하였습니다. 최종적으로, 새로운 데이터 상황에서도 활용 가능한 자동매매 시스템을 설계하였다.

아쉬운 한계점으로는 데이터의 범위와 양이 2023년 6월 이후로 제한되어 있어, 장기적인 패턴 분석에는 제약이 있었습니다. 자동매매 시스템이 실제 시장 환경에서 발생할 수 있는 현실적인 요소를 완전히 반영하지 못한 점이 존재합니다

앞으로 데이터 범위를 확대하여 장기적인 패턴 및 다양한 시장 상황에서의 모델 성능을 검증할 필요가 있습니다. 또한 외부 변수를 추가하여 모델의 예측력을 향상시킬 수 있습니다. 또한, 추가적인 강화학습 기법을 도입하여, 모델이 스스로 학습하며 점진적으로 성능을 개선할 수 있도록 설계할 수 있게 하고싶습니다