「2023년 2학기 응용자료분석」
프로젝트 보고서

제 목	머신러닝을 활용한 주가예측

이 름	하서경	학 번	20202670

머신러닝을 활용한 주가예측

1. 머리말

가. 주제선정 이유

딥러닝, 머신러닝, 인공지능이 세상을 떠들썩하게 만든 이후 각 분야에서 해당 기술들을 활용한 예측을 시도하고 있으나 주식수익률 예측에는 활발하지 않다. 과거의 패턴을 학습하고 미래를 예측하는 기술인만큼 장이 열리는 날에만 데이터가 발생하며 과거의 추세가 이어지지 않으며 불규칙 변동이 있어 주식수익률을 예측할 경우 정확도가 낮다는 문제가 있기 때문이다. 또 예측에 반영되어야 하는 경제지표 및 기업 정보는 월간, 분기 말에 집계되어 시점 차이가 나 일간 주가 변화를 예측하기 어렵고 경제지표간 상관성이 높아 전처리를 해줘야 해 전문화된 연구자에게도 까다로운 주제이다.

나. 분석 필요성 및 분석방법 개요

머신러닝을 활용한 주식수익률 예측 결과에 대비되게 딥러닝을 활용할 경우 나은 성과를 얻을 수 있다. 머신러닝 활용을 고집하는 이유는 간편하게 사용할 수 있도록 발전되어 배포되고 있기 때문이다. 인공지능, 데이터분석에 발 담구고 있지 않은 재테크 목적의 개인투자자들도 손쉽게 따라할 수 있기에 머신러닝의 각 과정을 시계열 데이터에 맞춰 변경해보며 예측 성능을 높일 방법을 마련하는 연구가 필요하다고 생각했다.

성능을 보장하는 예측 모델을 생성할 경우 시장에서 수익을 벌고 손실을 줄일 수 있는 돌파구가 될 것이다. 시장은 점점 복잡해지고 개인화된 투자전략을 선보이는 투자자가 늘어 기존의 분석방법으론 Alpha를 발견해내지 못한다. 머신러닝의 기술로 복잡한 주식가격 변화를 이해할 수 있고 밝혀지지 않은 Alpha를 자동으로 구해내며 투자 의사결정 속도를 높일 수 있다. Feature Transformation, Feature Selection, Dimension decomposition, Model 선택 및 Hyperparameter Tunning 같은 머신러닝의 각 과정을 조정하며 주식 구분별 효과, 추가 연구방향 등을 알아내고자 한다.

2. 데이터 분석

가. 데이터 선정

주식 투자를 위해선 주가 데이터(시가, 종가 등), 경제지표, 재무정보(매출액성장액, PER/PBR 등)을 고려해야 한다. 주식데이터는 FinanceDataReader를 이용해 수집하며 TA-Lib으로 투자 분석지표를 산출해 변수로 사용한다. FinanceDataReader는 한국 및 미국의 주가, 지수, 환율, 암호화폐 등 다양한 금융 데이터를 제공하며 주로 주식 데이터를 부를 때 사용된다. TA-Lib은 다양한 기술적 분석 지표와 패턴을 계산하기 위한 함수를 제공하는 라이브러리이다.

본 연구는 주식 특성별 모델 구축 과정의 차이를 알아보기 위해 우량주, 성장주, 가치주, 배당주, 기술주로 구분하고 각 구분에 해당하는 주식으로 2013년 3분기부터 2023년 2분기까지 데이터를 확보할 수 있는 미국의 Microsoft, MasterCard, Intel, McDonald, Apple를 사용한다.

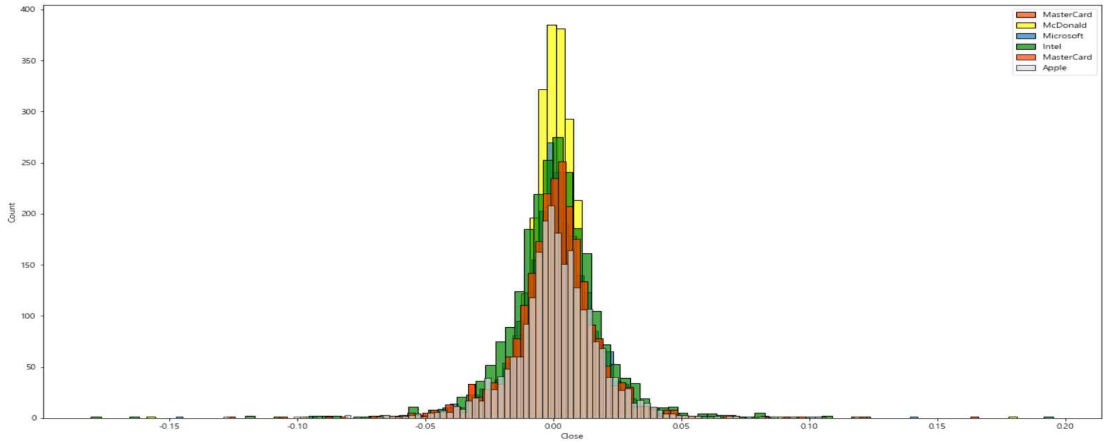
예측에 반영할 경제지표로 금리, GDP, 소비자물가지수(CPI), NASDAQ, S&P500, 반도체지수(SOX), 변동성지수(VIX), 실업률 등을 사용한다. 경제지표는 pandas_datareader와 FinanceDataReader, yfinance 패키지로 수집한다. pandas_datareader와 yfinance는 주식 가격 및 금융 데이터를 가져오기 위한 라이브러리로 Yahoo Finance에서 데이터를 불러온다.

주식을 예측하는데 기업의 정보도 포함되어야 한다. stockanalysis 사이트를 통해 각 기업의 10년 치 재무제표를 크롤링한다. 이후 성장률, ROA, 배당수익률, EPS 등을 구해 학습데이터로 사용한다. 결론적으로 종가로 구한 주가지표(이동평균(5일, 20일) 외 13개), 재무지표(매출액 외 14개), 경제지표(GDP 외 12개)을 수집했다.

나. 데이터 분석(분석 프로세스, 분석방법, 분석 내용 등)

1) Target 정의

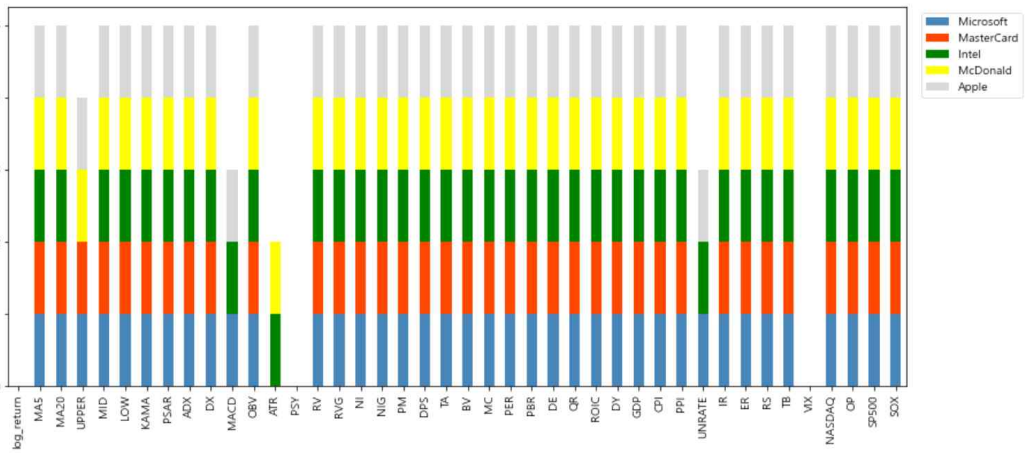
유사 연구를 조사했을 때 대부분의 연구가 주식의 방향성, 전날대비 상승하면 1, 동일하면 0, 하락하면 -1을 예측하는 모델 개발에 대해 이루어졌음을 알 수 있었다. 0.0001%와 10% 오른 주식을 1로 예측하는 것을 동일한 성과로 보면 안되고 수익을 더 키울 수 있는 10% 오름을 잘 예측하도록 해야 한다고 생각한다.



5개 회사의 주식수익률 분포를 보았을 때 모두 0을 중심으로 한 정규분포 형태를 띠며 IQR 상자가 [-0.3, 0.3]에 그려짐을 확인하였다. 데이터 불균형을 고려해 -0.2 이하는 -3, 0.2 이상은 3으로 이외의 데이터는 0.1 간격으로 -2부터 2까지의 라벨을 갖도록 처리하였다. 이로써 주식수익률 변화 방향과 상승/하락 정도를 예측하게 Target 변수를 생성했다.

2) Feature Selection

수집한 39개의 변수 중 예측에 유의미한 변수만을 선택한다. 불필요한 변수가 포함될 경우 주식데이터 특성상 타 변수간의 상관관계가 높아 일반화된 모형을 얻기 어려우며 모델 관리에 어려움이 따르기 때문이다. 전체 데이터를 9 : 1로 분할해 학습 데이터와 평가 데이터로 사용하며 학습 데이터에 한하여 Welch Anova를 사용해 Target 라벨별 차이를 보이는 변수만 선택한다. Welch Anova란 3개 이상의 그룹 평균을 비교하는 방법 중 그룹의 크기가 일정하지 않을 경우에 사용한다.



Welch Anova 결과에 따라 로그 수익률 외 6개 변수를 제외한 뒤 35개의 변수만을 사용한다.

3) Scaler

MinMaxScaler, RobustScaler, Normalizer, StandardScaler 중 성능이 뛰어난 Scaler를 선택해 데이터의 분포를 고르게 만들고자 한다. Microsoft로 기준 데이터로 하며 기준 모델을 Catboost로 삼고 동일한 전처리, IPCA를 적용하고 동일한 Hyperparameter를 설정한 뒤 모델 정확도를 비교했을 때 아래와 같았다.

MinMaxScaler	RobustScaler	Normalizer	StandardScaler
0.25	0.268	0.228	0.259

불규칙 변동이 많아 RobustScaler의 성능이 뛰어났을 것이라 해석하며 RobustScaler를 사용한다.

4) Reshape dataset & IPCA

다음 날의 주가를 예측하는 것은 전날의 데이터만으론 부족하다. 15~30일의 데이터 간 흐름을 파악해 예측해야 한다. 머신러닝 모델의 한계로 반영기간 내 데이터(2D)를 한 개의 행으로 축소해야 하는 문제가 있다.

여러 기법을 고려해본 끝에 IPCA기법으로 차원 축소해 학습데이터로 사용하기로 한다. IPCA기법이란 다차원 데이터의 차원을 축소하는 주성분분석의 일종으로 변수 간 상관관계를 없애며 데이터에 담긴 정보는 최대한 담는 방법이다. 기존 PCA 방법과 달리 새로운 데이터가 들어올 때마다 기존에 계산된 주성분을 업데이트하는 기법으로 불규칙 변동과 기간별 추세 차이가 뚜렷한 주가데이터에 유용할 것이라 생각했다.

Scaler를 선택한 것과 동일하게 한 시점 예측을 위한 반영기간, IPCA의 주성분을 업데이트할 횟수, IPCA의 차원 수를 지정해가며 최적의 조합을 찾아내야 한다. 실험 조건은 Scaler 선택 시와 동일하다.

IPCA 차원수	5개		10개	
Batch / 반영기간	64개	128개	64개	128개
15일	0.196	0.21	0.259	0.268
25일	0.233	0.228	0.192	0.201

매 예측시점의 전날부터 25일 전까지의 흐름을 학습하며 데이터 128개마다 IPCA의 주성분을 계산해 10개의 축소된 데이터를 얻어낸다.

5) Catboost & LightGBM

학습 속도가 빠르고 Tunning parameter 종류가 많아 다양한 시도를 해볼 수 있는 Catboost, LightGBM 모델을 연구에 사용하도록 한다. CatBoost는 그래디언트 부스팅 기반의 머신러닝 알고리즘 중 하나로 범주형 데이터를 처리할 수 있어 전처리 과정이 간소화된다는 장점이 있다. 그러나 본 연구에서는 학습횟수(iteration)를 늘리면 자동으로 모델이 데이터에 최적화되도록 Tunning되는 장점을 활용하고자 선택했다.

LightGBM 역시 그래디언트 부스팅 기반의 머신러닝 알고리즘이며 비선형적 패턴을 잘 학습해 시계열 데이터에 강점을 갖는다고 알려진 모델이다. 그러나 과적합 되지 않게 Hyperparameter Tunning을 세심히 해야 한다는 단점이 있다. 본 연구에선 Optuna라는 패키지를 사용해 최적화된 모델을 얻어내고자 한다.

다. 분석 결과 및 해석

Microsoft	Parameter	Score
Catboost	iterations=800, learning_rate=0.03	0.214
LightGBM	learning_rate=0.0398, max_depth=26, n_estimators=62	**0.237**

MasterCard	Parameter	Score
Catboost	iterations=800, learning_rate=0.03	**0.3**
LightGBM	learning_rate=0.035 max_depth=17, n_estimators=70	0.263

Intel	Parameter	Score
Catboost	iterations=2000, learning_rate=0.03	0.192
LightGBM	learning_rate=0.042, max_depth=30, n_estimators=57	**0.214**

McDonald	Parameter	Score
Catboost	iterations=2000, learning_rate=0.03	**0.402**
LightGBM	learning_rate=0.042, max_depth=15, n_estimators=77	0.397

Apple	Parameter	Score
Catboost	iterations=2000, learning_rate=0.03	0.214
LightGBM	learning_rate=0.033, max_depth=19, n_estimators=65	**0.281**

모델의 성능이 좋지 않아 paramter를 수정한 결과 Microsoft와 MasterCard는 학습 횟수(iterations)를 줄일수록 성능이 좋았다. 반대로 Intel, McDonald, Apple은 학습 횟수(iterations)를 늘릴수록 성능이 좋았다.

이러한 차이가 나타나는 이유로 데이터 내 변동이 아닌 불균형 문제, 특정 label을 갖는 데이터가 많아 모델의 학습 능력이 떨어지는 경우를 고려하며 Over Sampling해 학습시켜 보았으나 현저히 차이나는 것은 아니나 더 낮은 성능의 모델을 얻었다.

Microsoft	Parameter	Score
Catboost	iterations=800, learning_rate=0.03	0.161
LightGBM	learning_rate=0.039, max_depth=22, n_estimators=77	0.237

MasterCard	Parameter	Score
Catboost	iterations=800, learning_rate=0.03	0.21
LightGBM	learning_rate=0.046, max_depth=29, n_estimators=76	0.228

Intel	Parameter	Score
Catboost	iterations=2000, learning_rate=0.03	0.196
LightGBM	learning_rate=0.033, max_depth=22, n_estimators=76	0.227

McDonald	Parameter	Score
Catboost	iterations=2000, learning_rate=0.03	0.281
LightGBM	learning_rate=0.047, max_depth=21, n_estimators=75	0.353

Apple	Parameter	Score
Catboost	iterations=2000, learning_rate=0.03	0.174
LightGBM	learning_rate=0.0454, max_depth=24, n_estimators=80	0.205

개입 없이 건전한 주식 시장의 경우 수익률 변동이 0을 기준으로 0.1 가량 변화하기도 흔하지 않다. 자연적으로 발생하기 어려운 0.2, 0.3 이상의 변화 데이터의 수를 늘려 학습시켰을 때 오히려 Noise가 된 것이라 해석할 수 있다.

3. 분석 활용 전략

가. 기대 효과 및 방향 제시

무료로 기본 패키지와 Sklearn, Catboost, LightGBM 패키지만으로 예측 파이프라인을 구축했다는 의의가 있다. 그러나 주식 구분이 중첩되어 뚜렷한 비교 결과를 얻을 수 없었다. 추후 주식 특성보단 산업 분야별로 구분해 비교결과를 얻어 보고자 한다. 또 주식수익률 변화 정도를 반영해 예측 변수의 Label은 구성했으나 큰 변화를 예측했을 때 가중치를 준다거나 목적함수를 크게 해 최종 모델이 더 잘 구분해내도록 할 수 없었음이 한계로 남는다.

나. 본 논문의 기여점

전통적인 머신러닝 기법 중 시계열 데이터에 맞춰 사용하면 좋을 기법을 알아내고 적용해보아 성능을 확인해보았으며 지표별 예측 가중치를 알아낼 수 있는 차원 축소 방법을 마련해야 함을 알아냈다. 머신러닝 관점에서가 아닌 데이터 발생 관점을 고려해 Over Sampling하면 안 된다는 점 등을 고찰할 수 있었다.