혁신성장 Project

|  |
| --- |
| **금융 관련 선물 및**  **옵션 시세 예측** |

**2019년 2월 25일**

**딥러닝 기반 핵심 산업별 빅데이터 분석 전문가 과정** A**반**

장진원

최형석

목 차

**1. 프로젝트 개요**  **1**

1.1 프로젝트 기획 배경 및 목표 1

1.2 구성원 및 역할 2

1.3 프로젝트 추진 일정 3

**2. 프로젝트 현황**  **4**

2.1 분석 목적 4

2.2 데이터 설명 4

**3. 프로젝트 개발 결과**  **9**

3.1 주요 기능 9

3.2 개발 분석서 13

3.3 개발 상세 설명 16

3.4 활용 방안 19

**4. 기대 효과**  **21**

4.1 향후 개선 사항 21

4.2 기대 효과 23

**5. 개발 후기**  **24**

**6. 강사 의견**  **25**

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경 및 목표

대한민국 시장경제가 불황으로 접어드는 가운데 금융시장의 악화로 시장 상황이 더욱 나빠지고 있다. 시스템트레이딩 기법은 미국의 월가에서 선물시장이 도입되면서 기술적 분석(technical analysis)을 중심으로 발전하였다. 시스템트레이딩(system trading)이란 미리 정해진 진입규칙과 청산규칙에 따라 자동으로 매매가 이루어지는 주식거래방법이다.

한국에서도 KOSPI 200 선물시장이 세계적인 선물시장으로 성장하면서 학계와 실무에서 시스템트레이딩에 대한 많은 연구가 이루어지고 있으며, 실제 투자에서도 많이 활용되고 있다.

외국인 투자자 비중이 50% 가량 차지하고 있어 외국인 투자자들의 영향을 받을 때마다 휘청거리는 금융시장의 현 상황을 타개하고자 객관적인 데이터 분석을 통해 자주적인 투자 방안을 모색한다.

1.2 구성원 및 역할

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 전공 | 역할 | 구현 부분 |
| 장진원 | 회계학과 | 팀장 | 모델링  데이터 전처리 |
| 최형석 | 컴퓨터공학과 | 팀원 | 데이터 자료 수집  데이터 전처리 |

1.3 프로젝트 추진 일정

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | *12/8(토) ~ 12/15(토)* | *팀 구성* |  |
| *12/15(토) ~*  *12/22(토)* | *프로젝트 주제 선정 및 기획서 완성* |  |
| *12/22 (토) ~*  *12/29(토)* | *프로젝트 데이터 수집*  *논의 및 수집* |  |
| PJT  수행  /  완료 | *1/2(수) ~ 1/19(토)* | *프로젝트 수행(EDA)* |  |
| *1/19(토) ~*  *2/2(토)* | *프로젝트 수행(모델링)* |  |
| *2/2(토) ~*  *2/16(토)* | *중간 산출물 논의 및 추가 데이터 수집* |  |
| *2/16(토) ~*  *2/23(토)* | *최종 산출물 논의* |  |
| *2/28(목)* | *팀별 최종 발표 (구축 완료 보고)* |  |

**2. 프로젝트 현황**

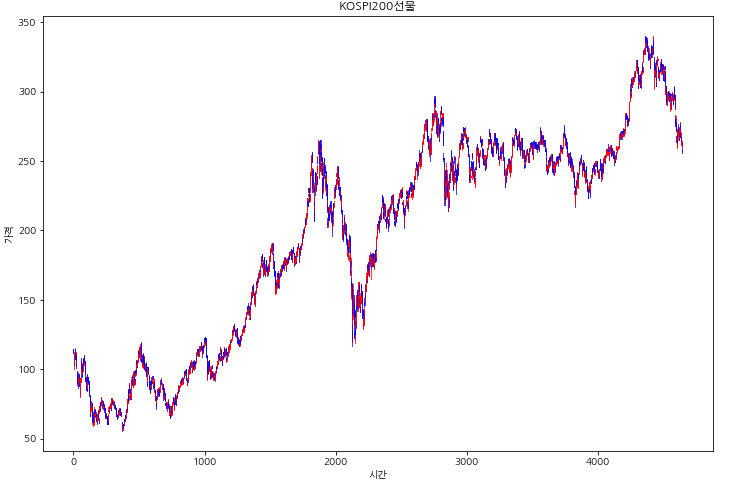
2.1 분석목적

시스템 트레이딩에서 수익에 영향을 미치는 요인에는 시장상황, 전략의 특징, 자금관리 등이 있다. 특히, 시장상황에 따라 주가 움직임이 미리 정해진 진입규칙이나 청산규칙에서 벗어나서 움직이면 시스템 트레이딩 전략은 큰 손실을 입을수도 있다. 따라서 대부분의 투자자들은 다양한 시스템 트레이딩 전략을 결합하는 시스템 포트폴리오를 이용하여 투자위험을 줄이고 안정적인 수익을 추구하고 있다. 시스템 트레이딩 전략들을 어떻게 결합하여 포트폴리오를 구성하느냐에 따라 투자수익과 위험에 큰 차이가 발생하지만, 아직까지 시스템 포트폴리오의 구성에 대한 연구는 이루어지지 않고있다. 기계학습 알고리즘을 이용하여 시장상황이 변할 때마다 포트폴리오를 적절히 재구성할 수 있다면 좋은 투자결과를 기대할 수 있다.

적절한 포트폴리오 재구성에 앞서 기계학습 알고리즘을 이용하여 시장상황을 사전에 예측하여 해당 전략들을 대응할 수 있다면, 포트폴리오 구성이 더욱 매끄러워질 수 있다. 본 프로젝트는 시스템 트레이딩 구축을 위해 시장상황을 상승장, 하락장, 안정적 변동 세가지 단계로 구분하는 것을 예측하여, 적절한 투자전략을 통해 안정적이고 높은 수익률을 보장할 수 있도록 한다.

2.2 데이터 설명

2000년 3월 10일부터 2018년 12월 27일까지의 KOSPI200 선물 일별자료를 사용하였다. 전체 자료 중 앞 기간 70%를 학습 데이터로 사용하고 뒤 기간 30%를 시험 데이터로 사용하였다. 모델링 입력 데이터로는 KRX 사이트에서 제공되는 기본적인 OHLC 가격과 현물 관련 변수들을 사용하였다. KOSPI 200 선물 일별데이터는 시가, 고가, 저가, 종가로 구성되는데 이 중 당일의 대표값 이라고 할 수 있는 종가의 변화값을 타겟변수로 사용하였다. KOSPI 200 선물가격은 <그림 1>에서 보듯이 연도에 따라 값의 크기가 크게 차이가 나므로 전일대비 변동가격을 사용하였다.



<그림1>

**3. 프로젝트 개발 결과**

3.1 주요 기능

KOSPI200 선물 데이터 내의 기존 변수들로 파생변수 생성(각종 기간별 거래량 변화, 시간대의 거래량 변화 등), 각 종 국제지수 변화 간의 관계 파악, 선물의 현재가격과 다른 변수들간의 상관관계 파악을 통하여 종가를 예측 한다

3.2 개발 분석서

KRX의 코스피200 선물 데이터를 가져와서 분석에 적합하게 정제한 뒤, 기술적 분석에 필요한 이동평균선, BollingerBand, ATR, Stochastic 등의 파생변수를 생성하여 모델링을 진행하였다.

모델링은 상관분석 후 다음날의 종가에 대하여 PCR 분석을 진행하였고, 종가 변화를 구간으로 나누어 Classification을 진행하였다. LinearSVC, KNN, SVC, Tree, Ensemble, LightGBM을 사용한 결과, SVC의 Score가 가장높게 나왔으며 LightGBM의 Confusion Matrix를 통해 각 구간별 f1-score를 확인하였다. 추가적으로 시계열 분석을 진행했는데, ARIMA 분석은 예측하기보다 결과값을 따라가는 형태를 나타났으며 LSTM 분석은 일별 데이터가 적었기 때문에 score가 낮게 나왔다.

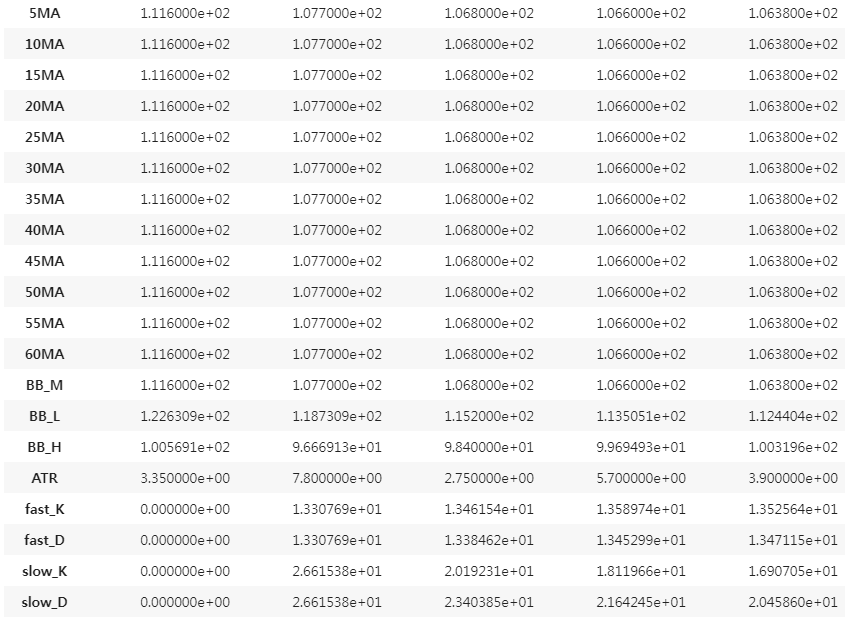
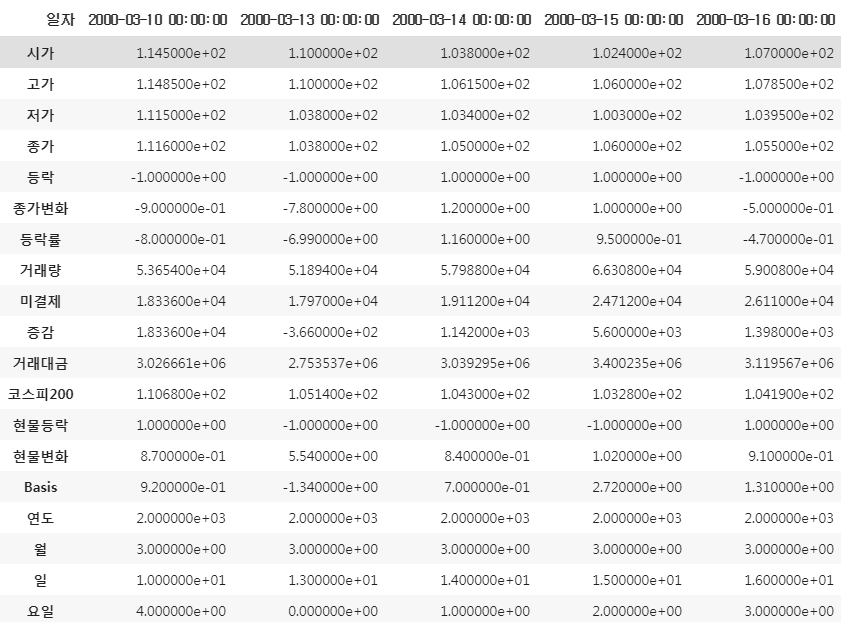
분류 모델을 보강하기 위해 기본적 분석을 진행하였고, 해외 투자자의 비중이 높은 것을 고려하여 ForexFactory 사이트의 경제 지표 및 사건의 데이터를 추가하였다.

종합적으로 분류에 영향을 끼치는 변수를 간단하게 확인하기 위해 DecisionTree 모델링을 진행하였고, 앞서 score값이 높게 나온 SVM모델과 LightGBM모델을 적용하였다.

3.3 개발 상세 설명



기본 데이터



데이터 전처리 및 파생변수 생성

이동평균(moving average) : 일정 기간 동안의 가격의 평균을 그래프로 그린 것이다. 이를테면, 오늘의 20일 이평값은 오늘의 가격(종가)부터 19일전의 종가까지를 평균한 값이다. 20일 이평선은 이런 이평값을 그래프로 이은 것이다. 이동평균은 추세를 계량적으로 표현하기 위해 개발된 기본적이고 강력한 도구이며, 기술적 분석 밖에서도 많이 채용된다.

Bollinger Band : 레이딩 밴드 개념에서 발전된 볼린저 밴드와 상대지표 그리고 대역폭은 이전 거래와의 상대적 가격의 높음과 낮음을 판단하는 데 사용된다. 볼린저 밴드는 켈트너 채널과 비슷한 변동성 지표이다. 볼린저 밴드는 주가가 이동평균선 중심으로 표준편차 범위 안에서 움직인다는 전제로 개발되었다

ATR : Average True Range 이며, TR이라는 변동성의 값을 평균화하여 선으로 표현하는 지표이다. ATR지표는 다른 지표와 다르게 매매 신호를 보여주는 것이 아닌, 변동성을 측정하여 주식의 위험도와, 투자적합구간이 적합한지의 여부를 투자자가 알기쉽게 보여주는 특징이 있다.

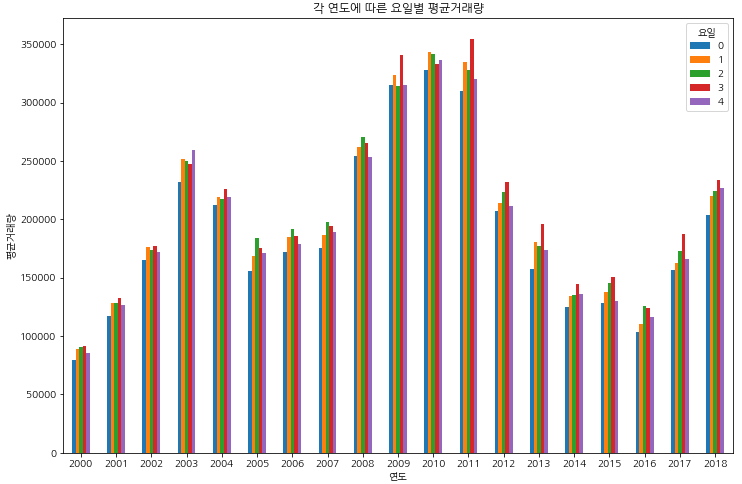
Stochastic : 주가가 과열 구간에 들어 서게 되면 조만간 하락할 것으로 반대로 주가가 침체 구간에 들어 서게 되면 조만간 반등할 걸로 예상해 볼 수 있는데 이러한 속성을 지표화 한 것이 바로 스토캐스틱이다. 조지 래인(George Lane)이란 사람이 만든 것으로 원래는 채권시장에서 사용되기 위해 만들어진 것으로 주식시장에서도 효과를 발휘하면서 유명해진 지표이다.

**EDA**



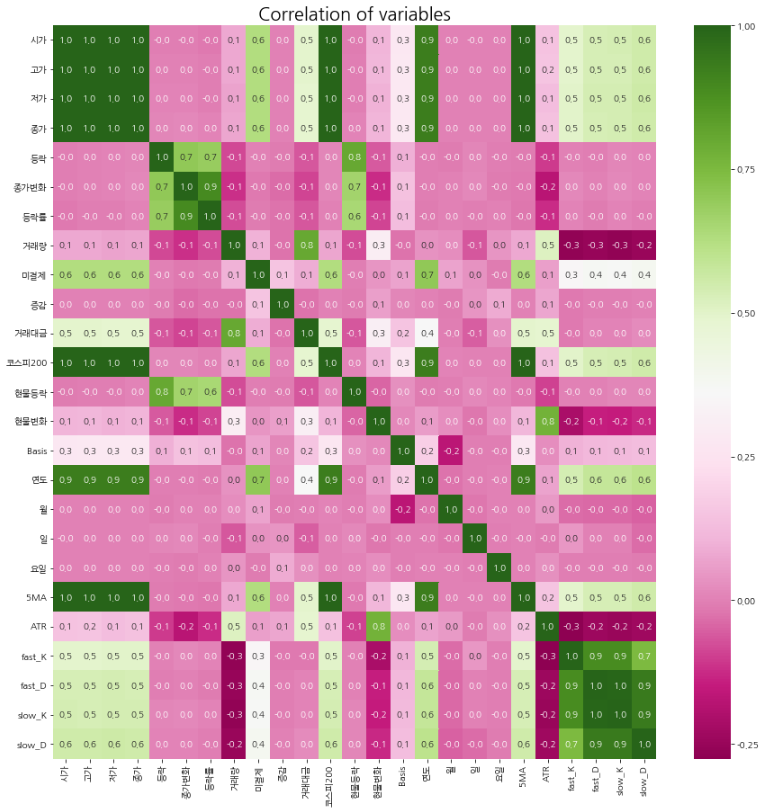
파생변수의 그래프

보라색 : ATR / 녹색 : 종가 / 빨간색 : BB\_M / 파란색 : BB\_L / 주황색 : BB\_H

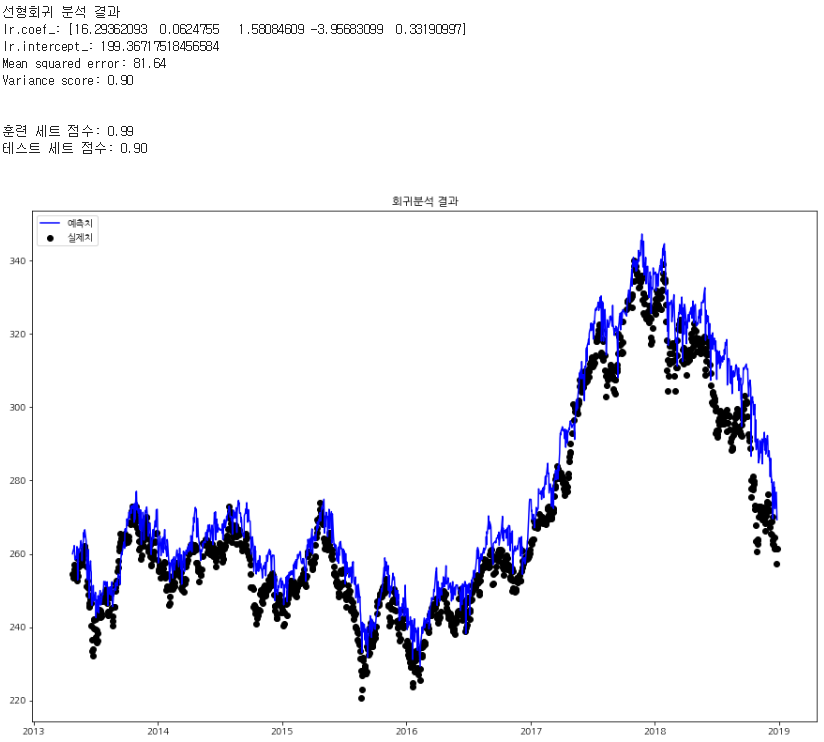


거래량이 높을 수록 가격의 변동이 크게 나타났으며, 무조건적으로 가격의 상승과 하락에 영향을 미치지는 않았다. 2007년 이후 거래량이 급격하게 증가하였는데 이는 금융위기로 인해 시장의 불안감을 나타내고 있다. 차이는 크지 않지만 상대적으로 월요일과 금요일에 거래량이 적은 것을 확인할 수 있는데, 투자자들이 개장과 파장의 여파를 피해서 투자하려는 경향을 보이는 모습을 유추할 수 있다.

**Regression**

종가의 증감을 예측하기에 앞서 독립변수들 중 어떤 변수들이 다음날의 종가에 큰 영향을 미치는지 확인하기 위해 상관분석을 실시하고, 단순히 선형 회귀분석으로 다음날 종가를 예측. 

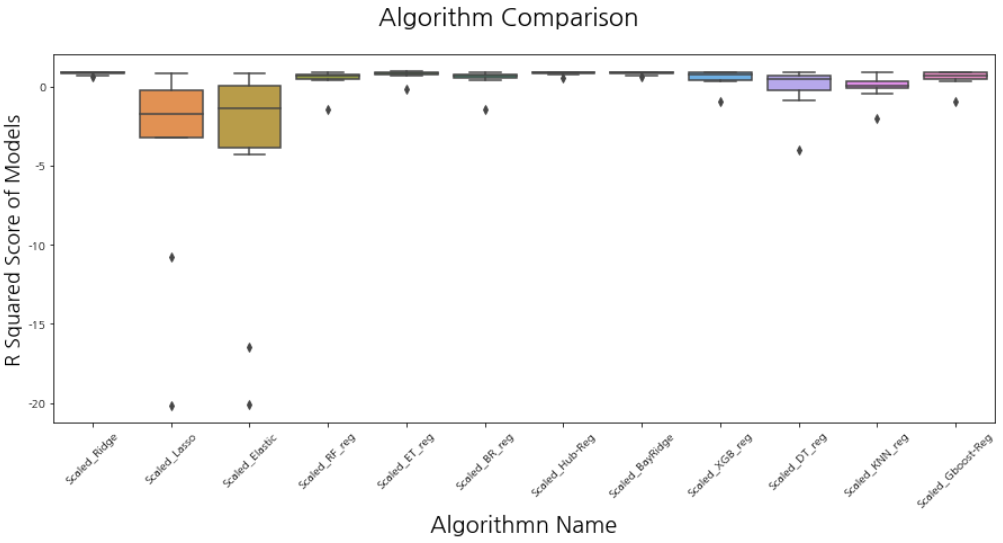
종가와 다른 변수들의 관계를 보았을 때 가격 관련 변수인 시가, 고가, 저가, 현물, 이동평균가격이 높은 것을 확인할 수 있다. 의외인 것은 연도가 가격과 어느정도 상관성을 보이는데 이는 연도 기준으로 보았을때 전체적으로 가격이 증가하는 추세를 보이고 있기 때문이라고 볼 수 있다. 구체적으로 살펴본다면 2007-2008 금융위기 이후 종가의 변화가 매우 탄력적으로 나타났던 것을 관측할 수 있다.



<그림2>

과최적화를 피하기 위해 Scaling과 PCA분석을 하고난 후 종가예측에 대해 Regression 분석을 실시하였다. 적절값인 5개의 주성분으로 분석을 하였고, 단순 선형회귀분석을 한 결과 훈련세트와 테스트세트의 정확도는 각각 99, 90점으로 높았다. 하지만 MSE가 81로 높게 나온 것을 확인할 수 있었다.

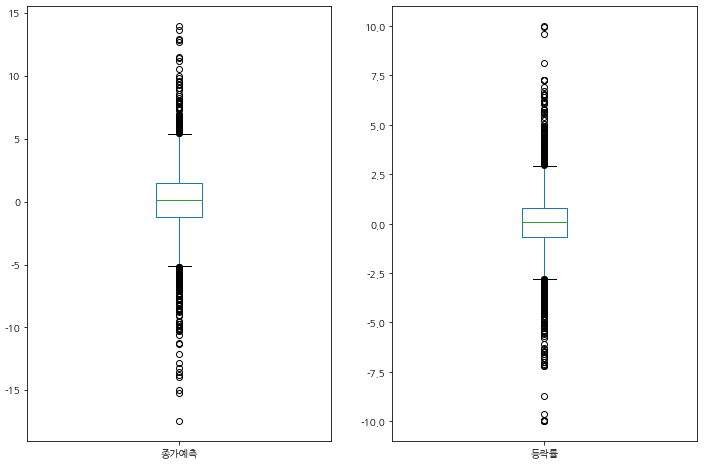
<그림2>를 살펴보면 과최적화를 피한 대신에 정확도가 상대적으로 많이 떨어졌는데, 갑작스러운 가격 변화에서 예측을 정확하게 하지 못했음을 확인할 수 있다.

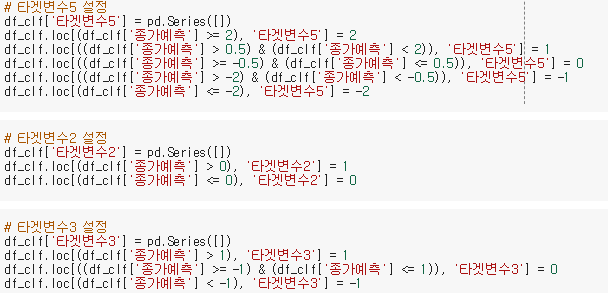


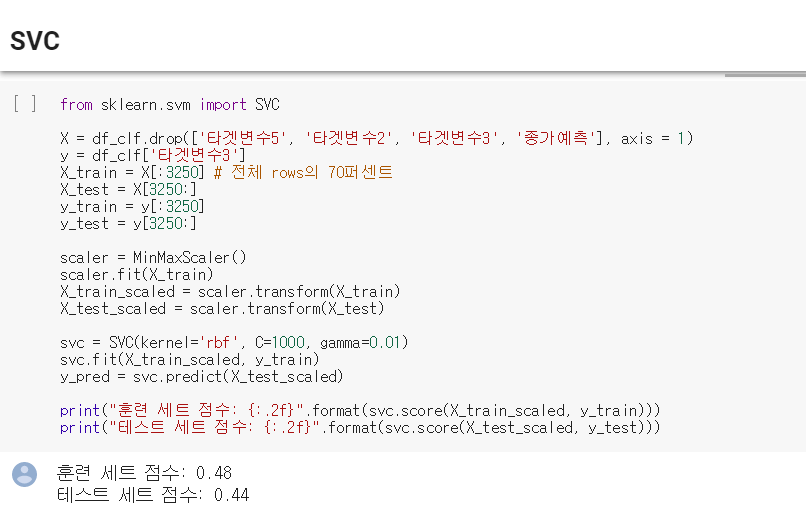
다른 회귀분석 알고리즘을 통해 CV 결과, MSG값이 0.844561로 나온 BayesianRidge가 가장 높게 나왔다.

회귀분석을 한 결과 변수들이 많아서 과최적화가 있었고, PCA분석을 통해 일반화를 시키더라도 높은 MSE수치를 봤을 때 정확하게 예측하지 못한다는 사실을 확인할 수 있었다.

**Classification**

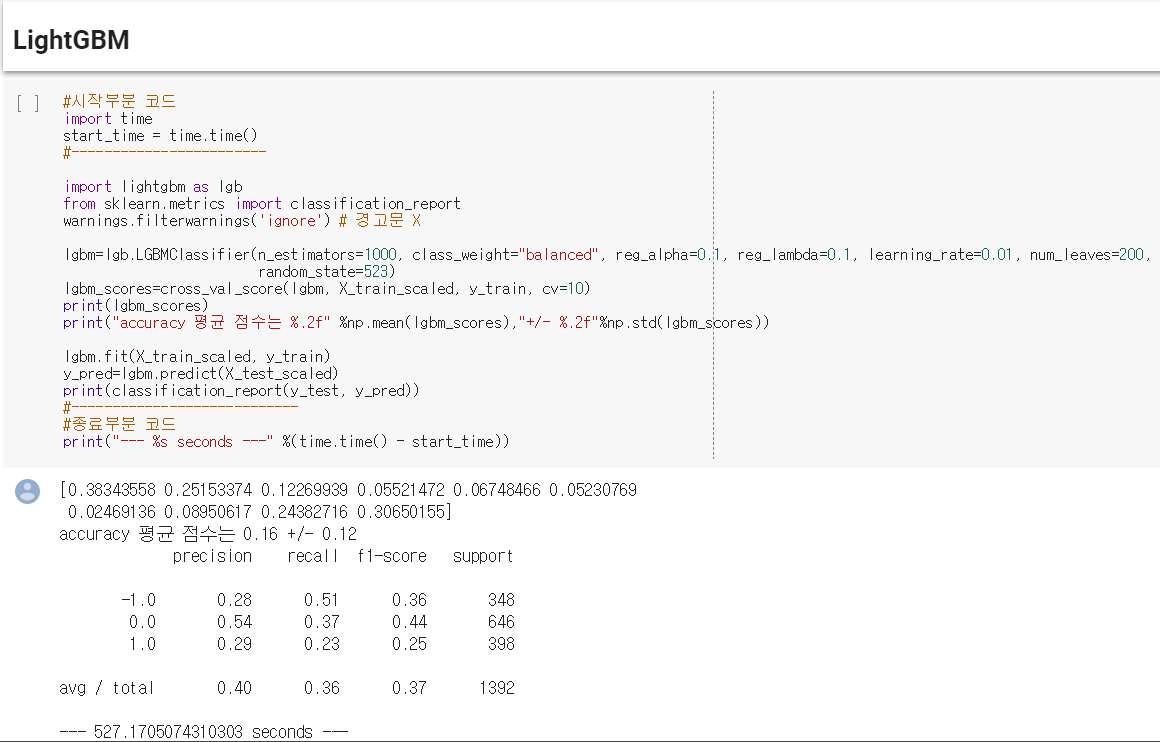
분류모델을 사용하기 위해 타겟변수의 편차를 알아보았다

위의 박스 플롯을 기반으로 타겟변수 설정



10만개 이하의 sample 데이터를 고려하여 SVM 알고리즘으로 모델링을 진행.

다른 여러 알고리즘 모델링도 진행했지만 결과값 또한 SVC가 가장 높게 나온 것을 확인할 수 있다. CV를 고려한 알고리즘 중에서는 RandomForest가 가장 높게 나왔지만, LGBM 모델링에서Classification Report를 통해 문제점을 발견할 수 있었다.



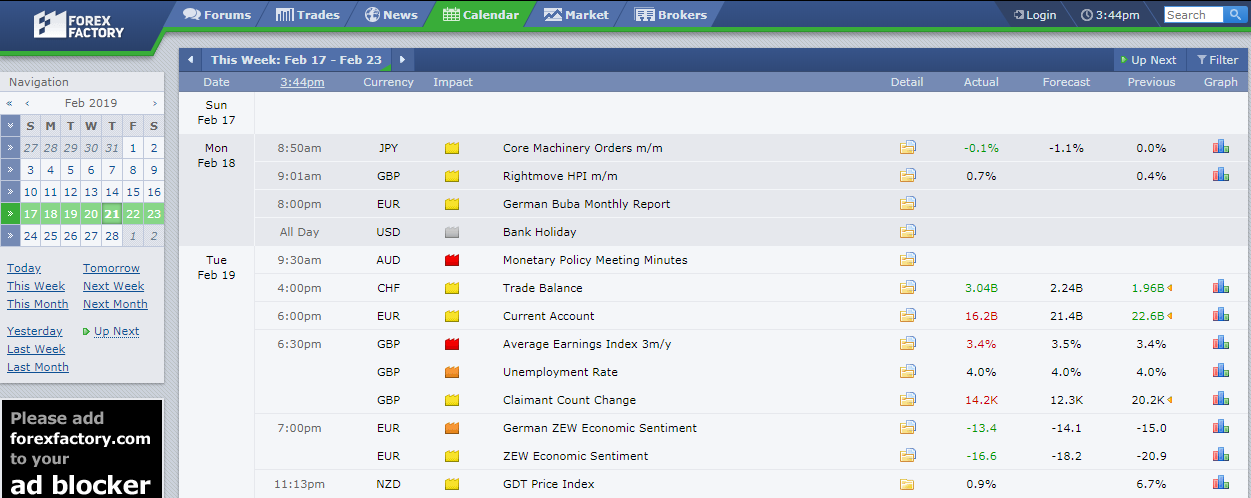
-1과 1에 대한 예측을 통해 매도, 매수포지션에서 전략으로 수익을 창출할 수 있는데 위의 Confusion Matrix을 살펴보면 f1-score가 0보다 상대적으로 낮게 나온 것을 확인할 수 있다.

결론적으로 기술적 분석을 통한 모델링으로는 한계가 있어서 기본적 분석을 더한 모델링을 통해 분석하는 것을 필요로 한다.

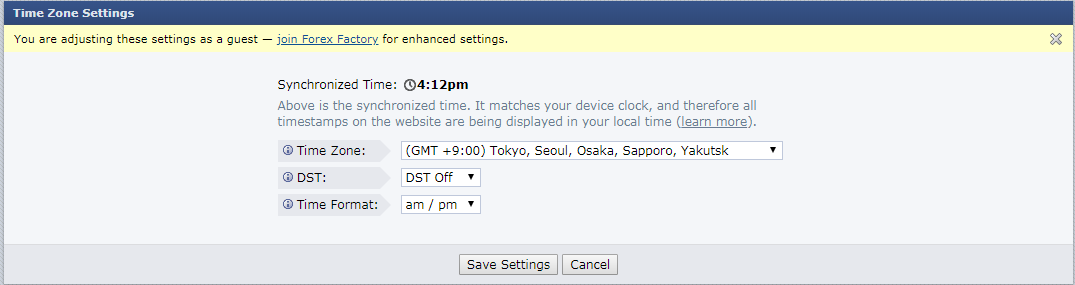
**기본적 분석을 위한 데이터 수집 및 전처리**



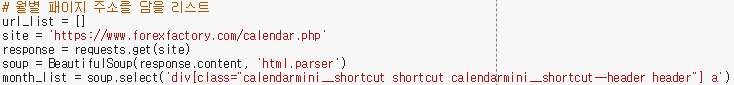
기술적 분석을 통한 모델링을 했을 때 만족스럽지 않은 결과를 보완하기 위해 기본적 분석을 보강하기로 결정. 외국인 투자자의 비중이 전체의 과반수 이상을 차지하고 있기 때문에 외국시장의 영향력을 고려하여 관련 사건이나 경제 지표들을 고려했다.



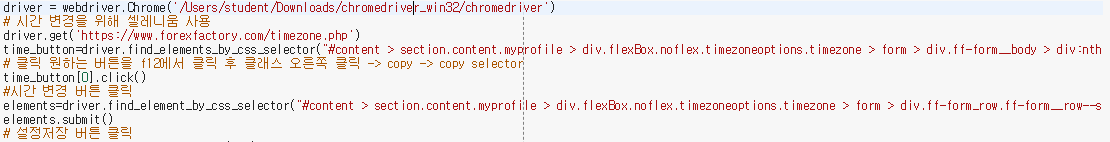
각 나라의 주요 경제지표 발표에 관한 내용을 참고하실 수 있는 포렉스 팩토리 홈페이지



포렉스 팩토리 시간을 변경 한 후 크롤링 시작

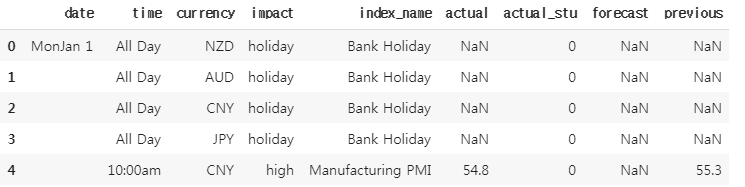


BeautifulSoup 모듈은 [HTML](http://en.wikipedia.org/wiki/HTML)과 [XML](http://en.wikipedia.org/wiki/XML)을 파싱하는 데에 사용되는 파이썬 라이브러리이다



일반적인 Beautiful soup로는 크롤링하기에 어려움을 겪어 셀레니움과 같이 사용 하여 크롤링

Selenium은 주로 웹앱을 테스트하는데 이용하는 프레임워크다. webdriver라는 API를 통해 운영체제에 설치된 Chrome등의 브라우저를 제어하게 하여 JavaScript를 이용해 비동기적으로 혹은 뒤늦게 불러와지는 컨텐츠들을 가져올 수 있다.



사이트에서 한국 시간으로 변경 후 크롤링 한 데이터

date : 날자

time : 해외 시간 기준을 한국 시간으로 변경 후 저장

currency : 나라

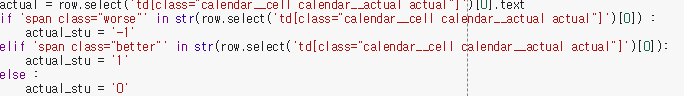
impact : high mid low non 종류로 중요도를 나타냄

index\_name : 경제지수 혹은 사건 이름

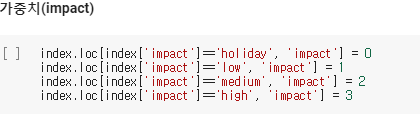
actual : 실제 수치

forecast : 예측 수치

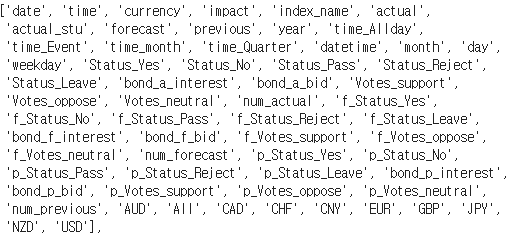
previous : 발표 전 수치



actual\_stu : 크롤링 단계에서 수치의 긍정, 부정 내용 긍정일 경우 1 부정일 경우 -1 영향없을 경우 0으로 변환



가중치는 0부터 3까지 지정해주었다

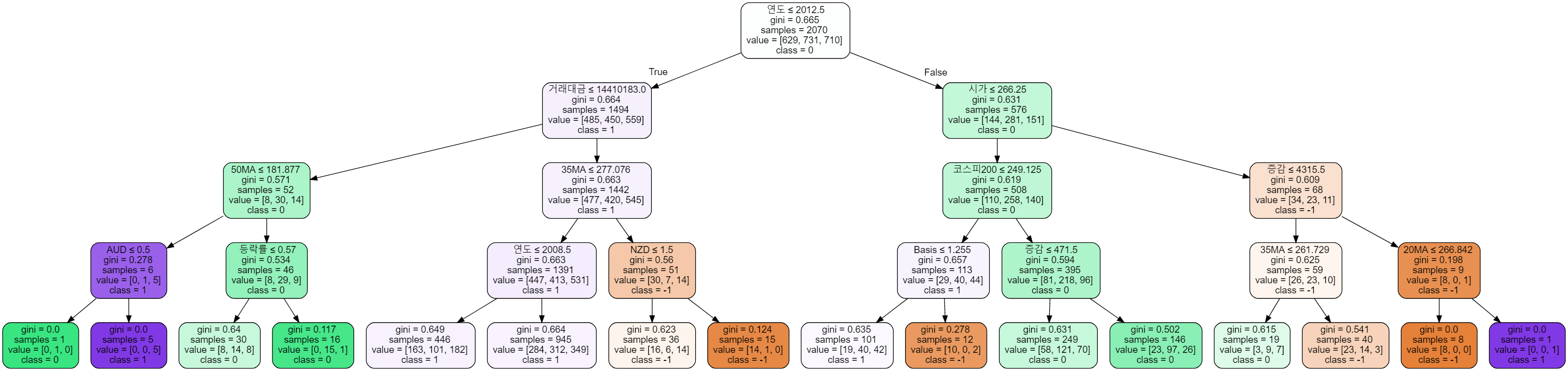


actual ,forecast ,previous, currency의 내용을 이용하여 추가변수 생성

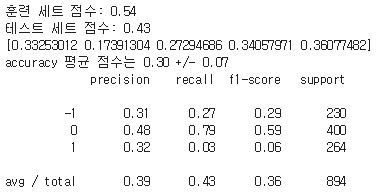


사건별로 정렬되있는 forexfactory 데이터를 일별 데이터인 코스피200지수 데이터와 합치기 위해, 예측치에서 실제치로 변경되었을 때 유의한 변수들의 변화율과 파생변수들을 고려하여 데이터를 가공.

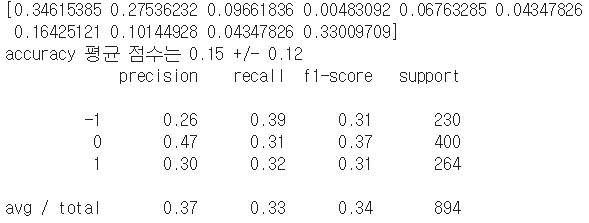
**최종 Modeling**



DecisionTree를 통해 모델링을 하였을 때 연도, 거래대금, 시가, 이동평균가격, 현물종가, 종가의 증감 등의 변수들이 구분하는데 영향을 끼치는 것을 알 수 있다.

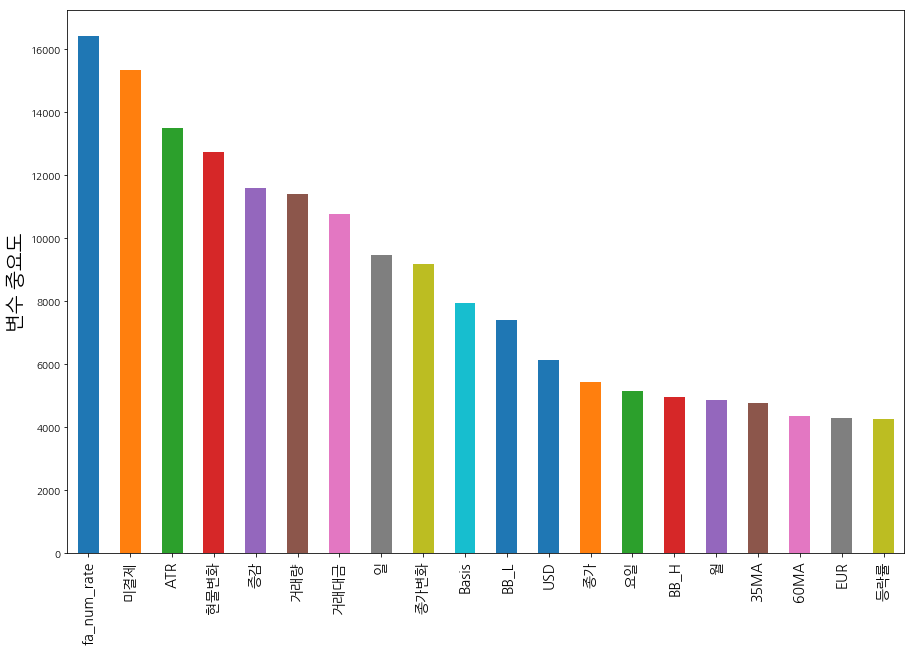


오히려 RandomForest, GradientBoosting, SVC 등 알고리즘의 score값이 전체적으로 줄어든 것을 확인할 수 있었다. Data를 merge하는 과정에서 기간이 짧아짐으로써 sample수가 줄어들었기 때문에 어느정도 score감소가 나타난 것으로 보인다.



LightGBM을 통해 모델링을 했을 때 score값은 마찬가지로 줄어들었지만, Confusion Matrix에서의 1과 -1 타겟의 f1-score값이 어느정도 소폭 상승한 것을 볼 수 있었다.

결론적으로 데이터 sample 수가 감소했음에도 사건이나 경제지표를 나타내는 변수를 추가함으로써 f1-score값이 증가하는 것을 보았을 때, 기본적 분석이 어느정도 효과가 있음을 확인할 수 있다.



LGBM 분석을 통해 변수 중요도를 확인해보면 ForexFactory 사이트의 경제지표지수 예측치와 실제치의 변화율을 나타내는 변수가 가장 높게 나온 것을 볼 수 있다. 모델링을 할 때에 있어서 기술적 분석보다 기본적 분석에 해당되는 변수의 중요도가 높게 측정된 것을 확인할 수 있다.

다음으로 미결제 변수의 중요도가 높게 나왔다. 미결제량은 선물의 매도 또는 매수 포지션을 취한 상태로 반대매매를 하지 않고 남아 있는 계약의 건수를 말한다. 청산하지 않게 되면, 미결제량이 늘어나게 되고 시장의 참여자도 늘어나게 된다. 강세장일 때에는 시장 가격이 상승하는 효과를, 약세장일 때에는 시장 가격이 하락하는 효과를 예상할 수 있다.

3.4 활용 방안

본 프로젝트는 매 거래일 장의 마감 시점 데이터를 활용하여 다음날의 종가를 예측하는 분석에 해당되기 때문에 실시간 크롤링을 통해 빠른 분석을 요구한다.

반복되지 않는 미래상황을 반영한 가격을 예측하는 것은 실제 어려운 일이기 때문에 투자로 인한 구조적 위험을 최소화하려면 다양한 전략을 수립하는 것이 중요하다.

모델링의 보완을 통해 정확도를 높이더라도 예상치 못한 사건이나 경제적 변화가 나타난다면 예측 모델의 정확도는 떨어지기 때문에, 매수 매도에 있어 다양한 전략을 고려한 포트폴리오를 통해 수익률을 보장하는 것이 요구된다.

**4. 기대 효과**

4.1 향후 개선 사항

실시간 크롤링을 적용하여 데이터를 수집하고 자동으로 분석을 하게끔하여 시스템을 구축하는 것이 우선시되어야 한다.

모델의 정확도를 상승시키기 위해 일별 데이터 뿐만 아니라 틱 데이터를 대상으로 분석한다. Sample의 수가 늘어남에 따라 딥러닝 같은 다양한 분석방법을 적용할 수 있고, ForexFactory 사이트의 데이터를 적극 활용할 수 있다. 시간대별 사건이나 수치 발표시 적극적으로 활용할 수 있고, 나아가 단순한 수치 데이터 분석뿐만 아니라 관련 사건이나 발표의 텍스트 및 영상 분석을 통해 가격에 미치는 영향을 적용할 수 있다. 대표적 예로는 우리나라의 박근혜 대통령 탄핵 당시 주가시장의 움직임이 탄핵인용 결정문의 내용에 따라 매우 탄력적으로 반응했던 것을 <그림2>를 통해 확인할 수 있다.

<그림2>

구체적으로 수익률을 보장하기 위해 다양한 전략을 수립한 포트폴리오를 반영한 트레이딩 시스템 구축을 마련해야 한다. 매수나 매도의 적절한 시기를 고려한 것과 손절 및 청산을 감안하는 방식의 투자 전략들을 고려하여 포트폴리오로 작성하여 수익률을 보장하게끔 한다.

4.2 기대 효과

앞에서 구축한 시스템과 포트폴리오를 통합하여 트레이딩 시스템을 구축한다면 다양한 환경을 고려하기 때문에 안정적인 수익률을 보장할 수 있다.

**5. 개발 후기**

|  |  |
| --- | --- |
| 성명 | 후기 |
| 장진원 | 금융 분석이 전공분야에 해당되지 않지만 개인적인 흥미와 멘토님의 추천으로 시작하였다. 배경 지식 공부 및 이론 공부를 하다보니 프로젝트 수행보다 지식 습득에 많은 시간을 투자하였고, 진행 속도가 더디게 되었다.  하지만 이론적으로 습득한 기술적 분석의 한계와 이를 극복하기 위해 트레이딩 시스템을 통해 수익률을 보장하는 방식을 알게 되었다. 해당 프로젝트를 보강하기 위해 기술적 분석뿐만 아니라 기본적 분석을 최대한 활용하는 방식으로 시스템을 구축하는 방향으로 진행을 해야겠다. |
| 최형석 | 빅데이터 라는 새로운 기술을 배우고 프로젝트 주제조차 평소에 접할 수 없던 선물이라는 주식 데이터를 주제로 진행하면서 새로움과 새로움의 조합이 시너지를 통해 데이터를 다루는 것과 새로운 것에 대한 이해도가 한단계 높은 수준으로 올라간 자신을 볼 수 있었고 결고 혼자서 할 수 없었기에 조장과 멘토분들에게 감사함을 느끼고 있다. |

**6. 강사 의견**

|  |  |  |
| --- | --- | --- |
| 평 가 요 소 | 배점 | 평 |
| 아이디어 :  유사한 서비스 존재 유무 및  체계성 | /20 |  |
| 2. 개발 : 실제 구현 정도 및 배포 유무,  코드의 무결성 및 난이도, 현업적용도, 실무기술 반영정도 | /30 |  |
| 3. PJT 수행력 : 일정관리 및 역할분담, 목표 일정 달성도, 팀내 참여도 등 | /30 |  |
| 4. 준비도 : 프리젠테이션 및 프로젝트 준비 정도 | /20 |  |
| 계 | /100 | 강사 의견 필수 |