2022-2 Financial Analytics
Final Research Paper

Prediction of Stock Performance Using Magic Formula and Machine Learning

December 15, 2022

Chan Gyu Lee
21700587
Department of AI Convergence & Entrepreneurship
Handong Global University

Abstract

다양한 데이터가 개방되고 새로운 투자전략이 지속적으로 연구되고 있음에도 불구하고 주식 시장 움직임에 대한 예측은 높은 변동성과 다양한 변수로 인해 예측에 어려움이 있다. 수많은 종목들이 경제적 및 정치적 요인과 같은 불규칙한 변동에 큰 영향을 받아 급격한 상승과 하락이 발생하게 되고, 많은 투자자들이 이로 인해 어려움을 겪고 있다. 이 문제를 해결하기 위하여 많은 학자들이 데이터 분석에 기반한 주식 성과 예측을 시도하고 있다. 정확한 예측을 수행해야 하는 투자자 입장에서 데이터 분석 기반의 주식 성과 예측은 투자 리스크의 범위를 줄이는 매우 효과적이고 필연적인 수단이다.
본 연구는 마법공식이라는 전통적인 투자전략 방법에 머신러닝 기술을 결합한 주식 성과 예측모델을 제안한다. 총 여섯 개의 알고리즘을 기반으로 대한민국의 82 개 주식 종목의 10 년치 데이터를 학습하여 비교분석 실험을 진행했다. 실험 결과, 과거 선행연구에서 주목받지 못하던 Gaussian process classification 모델이 가장 높은 성능을 기록하며 예측 연구의 효과적인 도구임을 증명하였고, 이는 주식 성과 예측 연구의 시야를 넓힌 점에서 의의를 갖는다.

Key Words: Stock Performance, Magic Formula, Machine Learning, Gaussian Process Classification

1) Introduction

대규모의 금융 데이터를 생산하고 다루는 금융 기관 및 관련 회사는 투자 결정을 위해 주식 및 시장에 대한 유용한 정보를 발견할 수 있는 효율적인 방법에 대해 지속적으로 연구한다. 또한, 다양한 데이터가 개방되면서 주식 정보와 데이터에 쉽게 접근할 수 있게 되어 주식 시장에서 금전적 기회를 극대화하고자 하는 전 세계 많은 개인 투자자들이 주가 예측에 대한 관심이 늘어나고 있다(Hargreaves and Yi Hao, 2013). 반면, 주식성과 예측은 과거부터 끊임없이 연구되고 있는 주제임에도 불구하고(Yoon and Swales, 1991; Hu et al., 2013; Waqar et al., 2017), 몇 가지 명확한 한계점을 띄고 있어 예측에 어려움이 있다. 먼저 주가의 변화는 글로벌 시장의 정치, 경제 관련 사건에 많은 영향을 받기 때문에 예측에

어려움이 있다(Zhang et al., 2018). 또한 시장에는 매우 많은 종류의 주식이 있기 때문에 처리해야 하는 주식 정보의 양이 너무 많고 이 또한 일반 신문, 매거진, 라디오, 텔레비전, 등 다양한 매체를 통해 전달되기 때문에 신뢰할 수 있는 정보를 가려내는 데 한계가 있다(Rodolfo et al., 2016). 이러한 이유로 주식 시장의 움직임을 완벽하게 예측하는 것은 불가능하지만, 과거 데이터 분석을 기반으로 주식 시장의 움직임을 예측하면 잘못된 시점에 주식을 매도하여 발생하는 손실과 그 영향을 크게 줄일 수 있다(Sakhare and Imambi, 2019).

항상 현명하게 주식 정보를 사용하여 정확한 예측을 수행해야 하는 투자자 입장에서 데이터 분석 기반의 주식 성과 예측은 투자 리스크의 범위를 줄이는 매우 효과적이고 필연적인 수단이다. 본 연구는 데이터 과학적으로 투자 위험을 줄일 수 있도록 머신러닝 기법을 활용하여 주식 성과를 예측하고, 실험 결과를 토대로 투자자의 의사결정을 지원하는 투자 방법론을 제시한다.

2) Literature Review

2.1) Stock Performance Prediction

주식 성과를 예측하여 확률적으로 손해를 줄이고 이윤을 극대화하는 연구는 과거부터 꾸준히 연구되고 있는 주제이다. Carol and Hao(2013)는 통계적 분류(Classification) 모형기반 주식 선택 방법론을 활용하여 높은 수익을 창출하는 주식을 가려내고 해당 방법론이 호주 시장 평균을 의미하는 Australian All-Ordinaries Index 를 능가하는 지 실험하였다. 실험 결과 호주 시장 평균보다 높은 이윤을 창출하는 주식 묶음을 선별하였고, 이를 통해 주식 선택 및 거래 전략을 구축하였다.

인도의 National Institute of Technology 의 연구자, Dutta et al(2012) 또한 다양한 재무 비율을 독립변수로 사용하여 기본적인 통계 회귀 모형 중 하나인, 로지스틱 회귀(Logistic Regression) 모델을 개발하였고, 이를 사용하여 인도 주식 시장에서 활발하게 거래되는 주식의 성과를 예측하는 연구를 진행하였다. 해당 연구에서는 주식의 성과에 유의미한 영향을 미치는 지표를 함께 조사하며 연구의 잠재성을 높였다.

이와 더불어, Sehal and Song(2008)은 웹 사용자의 감정을 기반으로 주식 시장 예측을 수행하는 새로운 방법을 소개했다. 해당 방법론은 금융 게시판을 스캔하여 개별 저자가 표현한 감정을 추출하고 감정과 주식 가치 간의 상관관계를 학습하여 모델을 구축했다. 해당 모델을 통해 주식 성과와 최신 웹 감정분석 정보가 밀접하게 연관되어 있음을 알 수 있었다. 위 사례들과 같이 다양한 연구자들이 각자의 새로운 도구를 활용하여 주식 성과를 예측하는 시도가 끊임없이 이어지고 있으며, 수많은 변수가 존재하는 연구 주제 특성상 앞으로도 꾸준히 연구되어야 하는 주제임을 시사한다.
본 연구는 위 사례를 참고하되 예측 연구에 특장점을 지니고 최근 많은 발전이 이루어진 머신러닝 분야를 접목시켜 보다 정확한 주식 성과 예측 연구를 진행한다.

2.2) Machine Learning

Samuel(1950)은 머신러닝을 인간의 반복적 인 프로그래밍 없이 데이터라는 경험을 통해 명시하지 않은 작업을 스스로 실행하는 기술 이라고 정의했다. 머신러닝의 활용범위가 다양해진 근래에는 머신러닝을 활용형태의 관 점에서 예측, 분류, 군집화 또는 그룹화를 수행하는 알고리즘으로 정의한다(Athey, 2019). 머신러닝에 의해 학습된 모델은 데이터 집합에서 정보와 지식에 대한 판단 또는 예측을 통해 추론에 가치를 더하여 연구의 완성도를 높인다. 전통적 관점에서 추론의 과정은 오직 기존 이론들을 토대로 인간이 생각해냈다면, 머신러닝 관점에서는 추론의 과정을 완전히 뒤집어 방대한 양의 데이터와 인간의 상상력을 통해 컴퓨터가 흥미로운 추론의 결과들을 전달하게 된다(Dhar, 2013). 머신러닝을 학습 방법의 대표적인 예시인 지도학습(Supervised learning) 기법은 결괏 값이 주어진 훈련 데이터를 사용하여 각각의 변수와 해당 결괏값의 패턴을 학습하고 이를 토대로 머신러닝 모델을 생성한다 (Cunningham et al, 2008). 학습을 통해 구현 한 이 모델에 새로운 입력값을 넣으면 예측 값을 도출하고 이 예측능력을 연구에 활용한 다. 이 연구방식은 기존 단순 통계에 의존하던 연구와 비교할 때 다양한 이점들이 존재한다. 첫째, 데이터의 복잡한 비선형 관계에 있어서 더 정확하게 패턴을 찾아낼 수 있다. 기존 회 귀적 접근 방식에서는 단순히 변수 간 상관 관계에만 의존하기 때문에 복잡하고 규모가 있는 데이터에 대해서 유의미한 정보와 지식을 찾기에는 한계가 존재한다. 비선형 관계에대한

패턴 분석과 예측능력에 우수성을 지닌머신러닝 기법은 현재 의학, 심리, 경제, 등다양한 산업군에서 활용되며 그 가치를 드러내고 있다. 둘째, 머신러닝은 자가 성능 개선 기능을통해 주어진 데이터와 알고리즘 내에서 스스로 꾸준히 학습하고 성장한다. 기존 회귀 및통계 모델에서는 이 모든 과정을 사람의 논리와 세상의 이론을 토대로 수동 작업을 해야 했다면, 머신러닝은 이론적 고찰에서 더나아가서 스스로 세상에 없던 새로운 인사이트를 창출하기 위해 학습을 시도한다. 이를통해, 인간이 직관적으로 고찰하기 어려운 변수 간의 관계에 대해서 새로운 통찰을 제시하여 혁신적 연구접근에 용이하다. 현재 이러한 예측 관련 이점들을 활용하여 주식 성과와 같은 금융 데이터에 대한 연구를 진행하는 사례가 다수 존재한다. 최근 NUS Fintech 학회에서 싱가포르의 주식 성과를 예측하기 위해 다양한 머신러닝 알고리즘을 활용한 사례가 있다(Cunrong et al., 2022). 해당 연구는 J.Greenblatt(2005)의 책에서 소개된 마법공식(Magic Formula)을 차용하여 해당 식에서 사용되는 ROC(Return On Capital)과 EY(Earnings Yield) 값을 독립변수 학습한 예측모델을 개발했다. 이는 86%의 정확도(Prediction Accuracy)로 높은 수익율의 주식을 선별하며, 마법공식과 머신러닝의 시너지를 증명하는 기반을 다졌다.
본 연구는 NUS Fintech 의 마법공식 기반 주식성과 예측모델 연구를 기초로, 다양한 알고리즘을 사용하여 예측모델을 만들고 성과를 비교 분석하는 연구를 진행한다.

2.3) Magic Formula

수 많은 투자 전략들이 소개되고 있었음에도 여전히 주식 시장에 대한 예측은 한계점이 많았던 중, Joel Greenblatt 라는 미국의 투자자가 발간한 ‘The Little Book that Beats the Market’에서 최적의 주식 포트폴리오를 구성하기 위한 마법공식(Magic Formula)을 소개하였다. Greenblatt 가 소개한 공식은 자본 수익률(ROC)와 이익 수익률(EY) 두개의 지표만을 고려하여 이윤을 극대화하는 상위 20~30 개의 주식을 선별하였다. 자본 수익률이란 영업이익을 투자자본으로 나누어 계산한 값이다. 이렇게 도출한 값을 활용하여 동일한 비용이 투입되었을 때 더 큰 수익이 나는 기업을 측정할 수 있다. 이익 수익률은 동일한 기업 구매가에 대해서 더 큰 수익이 나는 기업을 측정할 수 있다. 따라서, 이 두 가지 투자기준을 고려하여 높은 자본 수익률을 내고 높은 이익 수익률을 내는 상위

주식을 선별하는 것이 마법공식의 핵심이다. Greenblatt 의 마법공식은 17 년 동안 평균 30.8%의 수익률을 기록하며 S&P 500 시장 지수를 능가하는 성과를 보였다.
마법공식 방법론의 주요 이점은 단순성이다. 효과적으로 투자하기 위해 훈련된 투자 전문가의 도움이 필요하지 않으며, 신뢰할 수 있는 투자 포트폴리오를 구성하는데 단순한 몇 가지 지표를 활용할 수 있다. 이는 감정적이거나 비합리적인 의사결정을 줄이는 데 효과적이다. 그러나, 다른 투자 전략과 마찬가지로 시장의 영향을 많이 받는 주식 가격의 불완전한 특성상 마법공식 또한 항상 최선의 전략일 수 없다. 즉, 주식 시장의 움직임을 정밀하게 예측할 수 있는 최고의 알고리즘은 없기 때문에 머신러닝 알고리즘을 도입하여 주식 움직임을 최적의 정확도로 예측하여 주식 시장에 투자하는 투자자의 재정적 손실을 최소화하는 것을 목표로 분석을 수행해야 한다.
본 연구는 마법공식에서 활용하는 지표들을 머신러닝 기법을 활용하여 인간의 주관을 보다 제거하고 NUS Fintech 에서 진행한 연구에서 더 나아가 실험하지 않은 다양한 알고리즘으로 모델을 생성하여 예측 성과를 실험한다.

3) System Overview

이 연구는 마법공식에서 활용하는 ROC 와 EY 값을 활용하여 미래의 주식 성과를 예측하는 머신러닝 예측모델을 개발하고, 이를 활용하여 투자자의 위험 부담을 줄이고 이윤을 극대화하는 최적의 포트폴리오를 구하는 것을 연구의 목적으로 한다. 본 연구는 크게 6 개의 연구 과정으로 나눌 수 있으며 이는 다음과 같다.
첫째, 머신러닝 예측모델의 기반을 구축하기 위하여 과거의 Historical 데이터를 수집한다. 데이터는 연도별 주식 가격 종가 데이터와 EBIT, EV, EY, 등의 주식 정보를 포함한 두 개의 데이터로 나눌 수 있다. 두 번째로, 위 데이터를 연도별, 그리고 주식 종목 별로 통합하여 예측 모델의 재료가 되는 데이터를 완성한다. 이때 예측모델의 독립변수로 활용되는 ROC 와 EY 를 계산하여 데이터 셋에 포함시킨다. 셋째, 모델 개발 단계에서는 앞서 완성한 통합 데이터를 모델 학습에 사용하여 SVM(Support Vector Machine), KNN(K-Nearest Neighbors), XG Boost, GPC(Gaussian Process Classification), Random Forest, ANN(Artificial Neural Network)를 포함한 총 여섯 개의 단일 머신러닝 모델을 생성한다.

이렇게 개발한 여섯 개의 머신러닝 예측모델을 활용하여 네 번째 단계에서는 주식 성과를 예측한다. 다섯 번째 단계에서는 공통된 예측오차 지표를 활용하여 각 모델의 예측 정확도를 비교 분석하고 마지막으로 모델 최적화 기법(Feature Engineering 및 하이퍼 파라미터 튜닝)을 통해 모델의 성능을 개선한다. 이 모든 과정은 Appendix 의 <그림 1> System Overview 에서 확인할 수 있다.

4) Research Methodology

4.1) ROC & EY

ROC 와 EY 를 계산하는 공식은 비교적 간단하다. 우선 ROC 는 배당금에서 순이익을 빼고, 부채와 자기자본을 더한 후, 순이익과 배당금을 부채와 자기자본으로 나누어 계산할 수 있다. 즉, ROC 에 대한 계산 공식은 다음과 같다.

$$Return\ on\ Capital = \frac{(Net\ Income - Dividends)}{(Debt + Equity)}$$

EY 는 법인세를 계산하기 전의 이익 관련 이자비용을 기업의 시가총액과 순부채를 합한 기업가치 값으로 나눈 값을 의미한다. EY 는 아래 공식으로 표현할 수 있다.

$$Earnings\ Yield = \frac{Enterprise\ Value}{Earnings\ Before\ Interest\ rate\ Tax}$$

4.2) Machine Learning Algorithms

4.2.1) Support Vector Machine

서포트벡터머신은 Boser et al.,(1992)가 비선형 관계의 데이터를 결정 경계선을 통해 분석하는 방법을 제안하면서 시작되었다. 결정 경계선이란 데이터를 속성별로 나누는 선을 의미하고 이 선을 기준으로 비선형 관계의 데이터를 임의로 선형의 분석 가능한 데이터 집단으로 구성한다. 속성에 따라 부분집합으로 나누어진 데이터 중에 결정 경계선과 가장

가까운 데이터 포인트를 서포트 벡터(Support Vector)라고 하며 이 데이터 포인트가 모델 학습에 핵심이 된다.

4.2.2) K-Nearest Neighbor

K-Nearest Neighbor(KNN) 알고리즘은 간단하지만, 효과적인 비모수적(Non-parametric) 분류 모델이다(Guo et al., 2003). KNN 은 데이터 포인트를 k 개의 이웃 군집으로 분류하고 이를 유클리드 거리 기반으로 가중치를 두어 각 데이터 포인트를 군집에 속하게 하여 분류를 진행한다. 적절한 k 값을 찾는 것이 알고리즘의 핵심이며, k 값에 따라 성능에 변화가 생긴다. k 값을 선택하는 방법은 여러가지가 있으나, 대표적으로는 그리드 서치(Grid Search)를 사용한다. 그리드 서치 기법은 임의로 선정한 다수의 그리드를 사용하여 모델의 정확도를 실험하고 비교 분석하여 가장 성능이 높은 최적의 k 값을 도출한다.

4.2.3) XG Boost

XG 부스트는 그래디언트 부스팅(Gradient Boosting)이라는 기존 부스팅 모델의 예측 정확도와 같은 장점을 유지하고 느린 처리 속도와 과적합 규제와 같은 한계점들을 병렬 및 분산 처리 방식의 학습법과 학습 손실(Training Loss)의 최소화를 통해 극복한 알고리즘이다. 이렇게 개선된 XG 부스트 기법은 기존 그래디언트 부스팅보다 10 배 이상의 빠른 실행 속도와 더불어 모델링 최적화를 위해 사용자가 세밀하게 매개변수(Parameter)를 조절할 수 있기 때문에 모델링 최적화가 요구되는 다양한 연구 분야에서 활용되고 있다(Chen and Guestrin, 2016). XG 부스트의 단점으로는 다른 부스팅 모델과 마찬가지로 계산 과정의 추적이 어려워서 모델을 해석하는 데 한계가 존재한다.

4.2.4) Gaussian Process Classification

현실의 많은 문제는 복잡하면서도 아직 알려지지 않은 설명 변수와 반응 변수 간의 관계에 대해서 다루어야 한다. Gaussian Process Classification(GPC)는 복잡한 데이터를 잘 설명할 수 있는 동시에 정확히 예측할 수 있으므로 이러한 상황에서 추론을 수행하는 데 유용하게 사용될 수 있다(Schulz et al., 2018). GPC 는 비모수 베이지안 방식으로 회귀 문제에 접근한다.

따라서 설명 변수와 반응 변수 간의 관계의 포착 및 설명을 위해 방대한 수의 매개변수를 활용한다. 또한 GPC 은 베이지안 추론을 통해 복잡성 수준을 결정한다(Gershman and Blei, 2012; Williams, 1998). GPC 는 확률론적 예측 분포에 예측의 불확실성에 대한 추정치를 포함하여 제공하며 커널에 대한 파라미터를 학습하여 높은 성능을 낼 수 있다는 장점을 가지고 있다(Quinonero-Candela et al., 2007).

4.2.5) Random Forest

랜덤 포레스트는 수많은 결정트리가 울창한 숲의 모양으로 모인 형태를 의미하는 대표적인 앙상블(Ensemble) 기법이다. 앙상블이란 학습 데이터를 토대로 서로 다른 예측모델을 다수 생성하고 이 모델들의 결과를 결합하여 하나의 최적화된 모델을 생성하는 학습법을 의미한다. 즉 랜덤 포레스트는 하나의 거대한 결정트리가 아닌 여러 개의 작은 결정트리들을 만들고 이들을 통합하여 안정적인 예측값을 도출하는 모델링 방식이다(Brieman, 2001).

4.2.6) Artificial Neural Network

Artificial Neural Network(ANN) 모델은 뇌의 Sensory Processing 모델에서 영감을 받아 SC Wang(2003)가 컴퓨터에서 뉴런의 네트워크 모델을 시뮬레이션하며 고안한 인공 신경망 모델이다. 실제 뉴런의 수행 과정을 모방하는 알고리즘을 적용하며 다양한 유형의 문제를 학습하도록 만든 대표적인 모델이다(Anders Krogh, 2008). ANN 은 패턴 인식, 예측, 등에서 활발히 사용되고 있으며 기본적으로 입력값에 따라 부여되는 가중치를 겹겹이 네트워크 형태로 나열하여 구성한다. 할당된 가중치는 뉴런의 활성화를 결정하는 수학 함수에 의해 계산되며 출력 정보를 활용하여 예측값을 결정한다(N. Gupta, 2013)

4.3) Evaluation Metrics

본 연구는 머신러닝 모델의 성과를 모델의 분류 평가지표를 중심으로 비교 분석한다. 이와 같은 평가 방식은 Factor 형태의 변수를 예측하는 문제에서 자주 사용된다. 분류 평가지표는 실제 결과를 예측한 결과로 비교하는 표이며, 이는 <표 1> 과 같이 표현된다. 예측 오차

비교를 위한 공통된 지표로서는 정확도(Accuracy)를 가장 먼저 활용하고 이는 아래 수식으로 나타낼 수 있다.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

분류 정확도를 비교한 후, 더 정확한 분석을 위해 정밀도(Precision), 재현율(Recall), F1 Score 를 활용한다. 정밀도는 모델이 긍정적으로 분류한 데이터 포인트 중 실제로 긍정적이었던 포인트의 비율이다. 정밀도의 공식은 다음과 같다.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

이와 더불어, 재현율은 실제값이 긍정적인 것 중 예측값도 긍정적으로 나타난 비율을 의미한다.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

정밀도와 재현율은 서로 줄다리기의 관계를 가지고 있기 때문에 이 둘에 대한 적절한 비율을 맞추어 계산하는 지표가 필요한데 그것이 바로 F1 Score 이다. F1 Score 는 다음과 같이 표현된다.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

분류 정확도를 포함한 위 4 개의 평가 지표를 활용하여 각 모델의 성능을 비교분석하고 가장 높은 수치를 나타내는 최적의 모델을 선정하여 주식성과를 예측한다.

5) Case Study

5.1) Data (Korea Financial Data)

본 연구는 대한민국의 주식 시장 데이터를 적용하여 주식 성과 예측모델을 실험한다.
데이터 수집을 위해 DataGuide 에서 제공하는 대한민국 상장 주식 데이터를 활용하였고, 총 82 개의 가용가능한 주식 종목을 2002 년~2022 년에 걸쳐 추출하였다. 모든 데이터는 연율화 수치로 전처리 하였고, 각 종목별로 해당되는 연도에 대한 ROC 와 EY 를 계산하여 변수를 생성하였다.

이후 성과척도가 되는 레이블 변수를 추가하기 위하여 주식 종목별 3 년 후 가격을 시장 지수와 비교하여 높으면 1, 낮으면 0 을 부여하여 변수를 생성하였다. 이 과정에서 시장 지수는 KRX 정보데이터 시스템에서 동일하게 2022 년부터 2022 년 ETF 지수를 추출하여 활용하였다. 이 과정에서 2004~2006 년 데이터를 이상치로 판단하여 제거하였고, 만들어진 레이블 변수는 예측모델의 종속변수로서 활용했다. 이렇게 구성한 데이터는 총 1068 개의 데이터 포인트로 이루어져 있으며, 그 형태는 Appendix 의 <그림 2>에서 확인할 수 있다.

5.2) Model Construction & Optimization

본 연구는 모델 개발을 위해 R 프로그램의 함수와 패키지들을 사용하여 SVM, KNN, XG Boost, GPC, Random Forest, ANN 을 순차적으로 활용하고, 이를 검증한다. 모델링에 앞서, 모델 학습과 테스트를 위해 데이터를 9:1 비율로 추출하여 학습 데이터와 테스트 데이터를 만든다. 이후 모델별로 <표 2> 의 파라미터를 활용하여 모델 학습을 진행하였다. 이후 모델 생성 과정에서 각 모델의 성능 최적화를 위해 그리드 서치를 사용하여 성능을 극대화하는 파라미터(Hyperparameter)를 적용하여 모델을 개선하였다. 또한 Feature Engineering 을 통해 모델 성능을 극대화하는 파생변수를 생성하였고, 이는 기존 변수(ROC, EY)에 제곱한 값으로 설정하여 <그림 3>와 같이 데이터를 재구성하였다.

5.3) Experimental Result

최적의 파라미터를 적용한 각 모델의 예측 정확도(Accuracy)를 살펴보면 <표 3>와 같다. 표에 나타났듯이 가장 높은 성능을 보인 모델은 0.645 의 예측 정확도를 기록한 GPC 모델이다. 반면, RandomForest 모델과 ANN 모델은 0.51 의 정확도를 나타내며 무의미한 예측성과를 기록하였다. 보다 명확한 예측 성능 비교를 위해 정밀도와 재현율, 그리고 F1 Score 를 비교한 결과는 <표 4>과 같다. 역시 GPC 모델이 정밀도와 재현율의 조화 평균을 의미하는 F1 Score 에서 가장 높은 0.667 을 기록하며 상대적으로 뛰어난 성능을 기록하는 것을 확인할 수 있다. GPC 의 분류평가지표를 나타내는 <표 5>를 보면 예측 성과를 가시적으로 확인할 수 있다.

6) Conclusions

본 연구는 주식의 미래 성과를 예측하여 투자자의 의사결정을 돕기 위해 NUS Fintech 학회의 마법공식과 머신러닝의 관계 연구를 기반으로 다양한 예측모델을 만들어 실험을 진행하였다. 실험 결과 GPC 가 가장 높은 예측성과를 기록하는 것을 알 수 있었지만, NUS Fintech 가 선행연구로 진행했던 결과, 0.86 의 예측 정확도에 비해 다소 성능이 떨어지는 것을 확인할 수 있었다.

그럼에도 본 연구는 몇 가지 의의를 지닌다. 첫째, 시계열 예측에 큰 강점을 지녔으나, 주식 성과 분석에 사용된 적 없던 Gaussian Process Classification 이 효과적인 도구임을 증명하는 점에서 큰 의미를 지닌다. 주목받지 못하던 모델을 사용하여 예측 연구의 성과를 높이고, 주식 성과 예측 연구의 시야를 넓힌 점에서 의의를 갖는다.

또한, 이 연구는 기존 NUS Fintech 의 연구에 비해 정확도가 비교적 낮지만, 246 개의 종목으로 진행했던 선행연구 환경을 비교하였을 때, 82 개의 훨씬 적은 수의 종목으로 성과를 냈다는 점에서 중요한 의의를 갖는다. 추후 더 많은 데이터를 확보하여 모델링을 진행하면 예측 성능을 획기적으로 개선할 수 있음을 의미하며 이는 본 연구의 잠재력을 증명한다.

이 연구는 몇 가지 한계점도 갖고 있다. 우선, 시장 지수에 ETF 데이터를 대입하여 예측모델의 종속변수가 되는 레이블을 계산하였으나, 보다 적합한 지표가 존재할 수 있다. 모델에 알맞은 시장 지수 데이터를 수집한다면, 연구의 결과가 개선될 수 있을 것이다.

또한, 국내 주식에서 모델에 악영향을 주는 상당수의 이상치가 존재하였다. 이상치를 보다 효과적으로 처리하여 데이터의 양과 질을 개선한다면 연구의 결과가 고도화될 것이다.

Reference

Ashish Sharma, Dinesh Bhuriya, Upendra Singh, “Survey of Stock Market Prediction Using Machine Learning Approach”, *IEEE ICECAT 2017 proceedings,* pp.506-510.

Athey, S. (2019). 21. The Impact of Machine Learning on Economics. In *The economics of artificial intelligence* (pp. 507-552). University of Chicago Press.

Breiman, L. (2001). Random forests. *Machine learning, 45(1),* 5-32.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings *of the fifth annual workshop on Computational learning theory* (pp. 144-152).

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of *the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In *Machine learning techniques* for *multimedia* (pp. 21-49). Springer, Berlin, Heidelberg

Cunrong, K. A., Kumar, V. J., Tandjung, H. N., & Widjaja, N. (2022, June 12). *Beating the Singapore Stock Market with the 'magic formula'. Medium. Retrieved December 14, 2022.*

Dhar, V. (2013). Data science and prediction. Communications of the *ACM, 56(12),* 64-73.

Dutta, A., Bandopadhyay, G., & Sengupta, S. (2012). Prediction of stock performance in the Indian stock market using logistic regression. *International Journal of Business and Information*, *7*(1), 105.

Gershman, S. J. and Blei, D. M.(2012), “A tutorial on Bayesian nonparametric models”, *Journal of Mathematical Psychology*, 56(1), 1–12.

Greenblatt, J., & Tobias, A. (2005). *The Little Book That Beats the Market* (1st ed). Wiley.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.

Gupta, N. (2013). Artificial neural network. *Network and Complex Systems*, *3*(1), 24-28.

Hargreaves, C., & Hao, Y. (2013). Prediction of stock performance using analytical techniques. *Journal of Emerging Technologies in Web Intelligence*, *5*(2), 136-142.

Hu, Z., Zhu, J., & Tse, K. (2013, November). Stocks market prediction using support vector machine. In *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering* (Vol. 2, pp. 115-118). IEEE.

Jigar Patel, Sahil Shah, Priyank Thakkar, K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques", ScienceDirect, Expert Systems with Applications, 2014.

Jing Zhang, Shicheng Cui, Yan Xu, Qianmu Li, Tao Li, "A novel data-driven stock price trend prediction system", ScienceDirect, *Expert Systems With Applications 97 ,2018,* pp.60–69

Quinonero-Candela, J. and Rasmussen, C. E.(2005), "A unifying view of sparse approximate Gaussian process regression", *The Journal of Machine Learning Research*, 6: 1939-1959.

Rodolfo C. Cavalcante, Rodrigo C. Brasileiro, Victor L.F. Souza, Jarley P. Nobrega, Adriano L.I. Oliveira, "Computational Intelligence and Financial Markets: A Survey and Future Directions", *ScienceDirect*, Expert Systems With Applications", 2016, pp. 1-12.

Sakhare, N. N., & Imambi, S. S. (2019). Performance analysis of regression based machine learning techniques for prediction of stock market movement. *International Journal of Recent Technology and Engineering*, *7*(6), 655-662.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3), 210-229.

Schulz, E., Speekenbrink, M. and Krause, A.(2018), "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions", *Journal of Mathematical Psychology*, 85: 1-16.

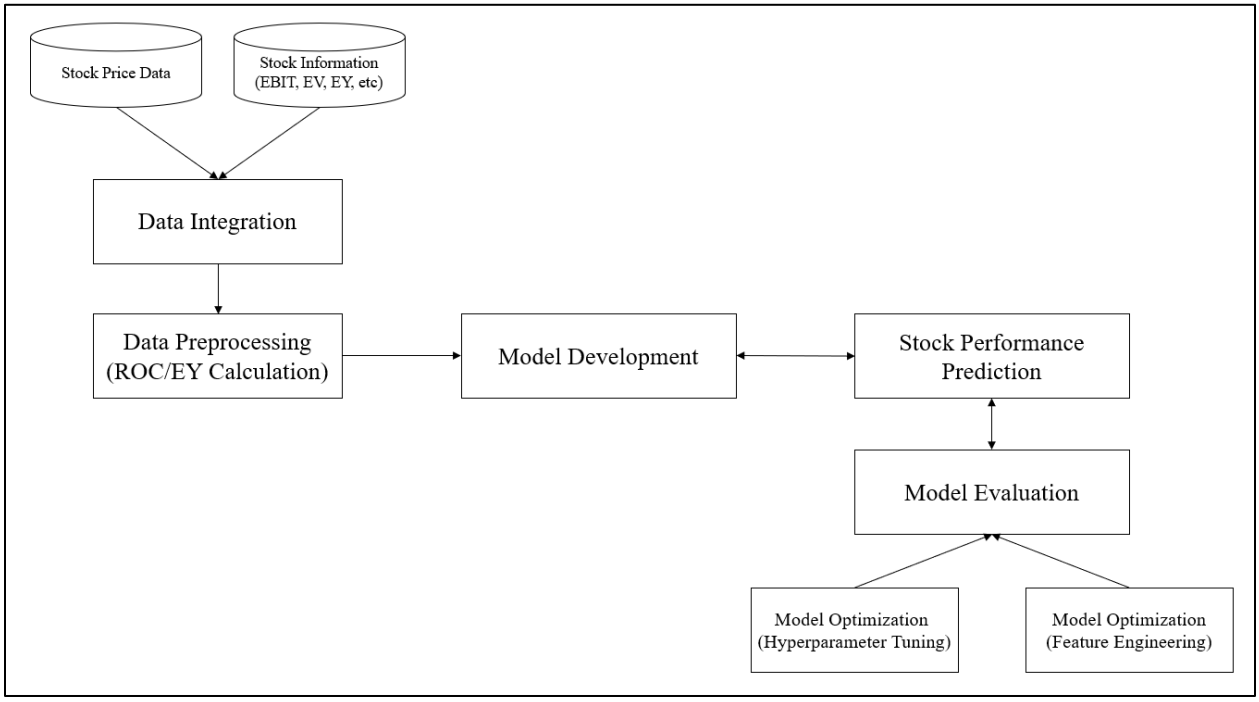
Sehgal, V., & Song, C. (2007, October). Sops: stock prediction using web sentiment. In *Seventh IEEE international conference on data mining workshops (ICDMW 2007)* (pp. 21-26). IEEE.

Waqar, M., Dawood, H., Guo, P., Shahnawaz, M. B., & Ghazanfar, M. A. (2017, December). Prediction of stock market by principal component analysis. In *2017 13th International Conference on Computational Intelligence and Security (CIS)* (pp. 599-602). IEEE.

Yoon, Y., & Swales, G. (1991, January). Predicting stock price performance: A neural network approach. In *Proceedings of the twenty-fourth annual Hawaii international conference on system sciences* (Vol. 4, pp. 156-162). IEEE.

Zahid Iqbal, R.Ilyas, W. Shahzad, Z.Mahmood, J.Anjum, “Efficient Machine Learning Techniques for Stock Market Prediction”, *IJERA*, Vol.3, issue 6, Nov-2013, pp.855-867.

Appendix



<그림 1> System Overview

<표 1> 분류평가지표(Confusion Matrix)

Confusion Matrix		**Predicted**	
		Negative (0)	**Positive (1)**
Actual	**Negative (0)**	True Negative	False Positive
	Positive (1)	False Negative	True Positive

	Year	종목	roc	ey	EBIT	EV	label
134	2017	1.삼성전자	0.147095719	0.816348645	30403215000	37242929443	0
228	2002	1.삼성전자	0.109636109	0.103883042	3106192090	29900858056	0
257	2003	1.삼성전자	0.199743223	0.190601317	7101016000	37255860004	0
446	2015	1.삼성전자	0.155077265	0.603908245	27563175000	45641329175	0
513	2022	1.삼성전자	0.156356535	0.594274328	52688168000	88659673663	0
552	2008	1.삼성전자	0.117989049	0.235644033	7474287000	31718549878	1
596	2019	1.삼성전자	0.225449067	0.926614861	60705978000	65513710758	1
633	2009	1.삼성전자	0.075873542	0.163927839	5525904000	33709368846	1
726	2014	1.삼성전자	0.232490000	0.786926244	37710730000	47921555928	1
803	2011	1.삼성전자	0.201942469	0.537777888	18981166000	35295549401	1
844	2010	1.삼성전자	0.154565852	0.412009361	12002644000	29131969158	1
899	2012	1.삼성전자	0.151336936	0.420996453	16815520000	39942189261	1
951	2007	1.삼성전자	0.143684500	0.272419270	7995948000	29351624095	1
983	2013	1.삼성전자	0.218817400	0.714458780	29255107000	40947228607	1

<그림 2> 삼성전자 데이터 스키마 샘플

	Year	종목	roc	ey	EBIT	EV	roc2	ey2	label
228	2002	1.삼성전자	0.109636109	0.103883042	3106192090	29900858056	1.202008e-02	1.079169e-02	0
257	2003	1.삼성전자	0.199743223	0.190601317	7101016000	37255860004	3.989736e-02	3.632886e-02	0
951	2007	1.삼성전자	0.143684500	0.272419270	7995948000	29351624095	2.064524e-02	7.421226e-02	1
552	2008	1.삼성전자	0.117989049	0.235644033	7474287000	31718549878	1.392142e-02	5.552811e-02	1
633	2009	1.삼성전자	0.075873542	0.163927839	5525904000	33709368846	5.756794e-03	2.687234e-02	1
844	2010	1.삼성전자	0.154565852	0.412009361	12002644000	29131969158	2.389060e-02	1.697517e-01	1
803	2011	1.삼성전자	0.201942469	0.537777888	18981166000	35295549401	4.078076e-02	2.892051e-01	1
899	2012	1.삼성전자	0.151336936	0.420996453	16815520000	39942189261	2.290287e-02	1.772380e-01	1
983	2013	1.삼성전자	0.218817400	0.714458780	29255107000	40947228607	4.788105e-02	5.104513e-01	1
726	2014	1.삼성전자	0.232490000	0.786926244	37710730000	47921555928	5.405160e-02	6.192529e-01	1
446	2015	1.삼성전자	0.155077265	0.603908245	27563175000	45641329175	2.404896e-02	3.647052e-01	0
134	2017	1.삼성전자	0.147095719	0.816348645	30403215000	37242929443	2.163715e-02	6.664251e-01	0
596	2019	1.삼성전자	0.225449067	0.926614861	60705978000	65513710758	5.082728e-02	8.586151e-01	1
513	2022	1.삼성전자	0.156356535	0.594274328	52688168000	88659673663	2.444737e-02	3.531620e-01	0

<그림 3> 삼성전자 데이터 파생변수 생성

<표 2> 모델별 적용 파라미터

Model	Package	Parameters
SVM	e1071	- type = C-classification - kernel = sigmoid
KNN	Class	- optimal k = 33
XG Boost	xgboost: gbtree	- eta = 0.3 - max_depth = 3 - gamma = 3 - nrounds = 100 - nfold = 5
GPC	kernlab	- kernel = polydot
RF	randomForest	- ntree = 20
ANN	neuralnet	- act.fit = logistic - hidden = 3

<표 3> 모델 예측 정확도 결과 및 성능 비교

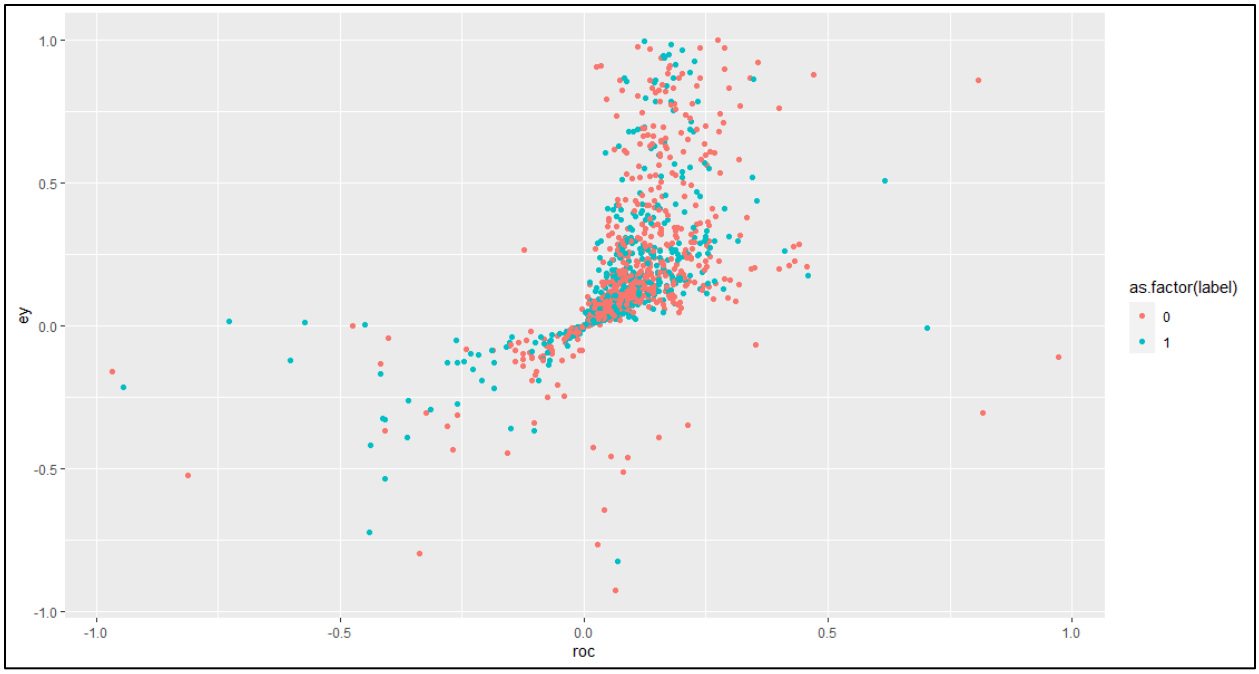
Model	SVM	KNN	XGB	GPC	RF	ANN
Accuracy	0.542	0.636	0.580	**0.645**	0.514	0.510

<표 4> 예측모델 종합 성능 비교 분석

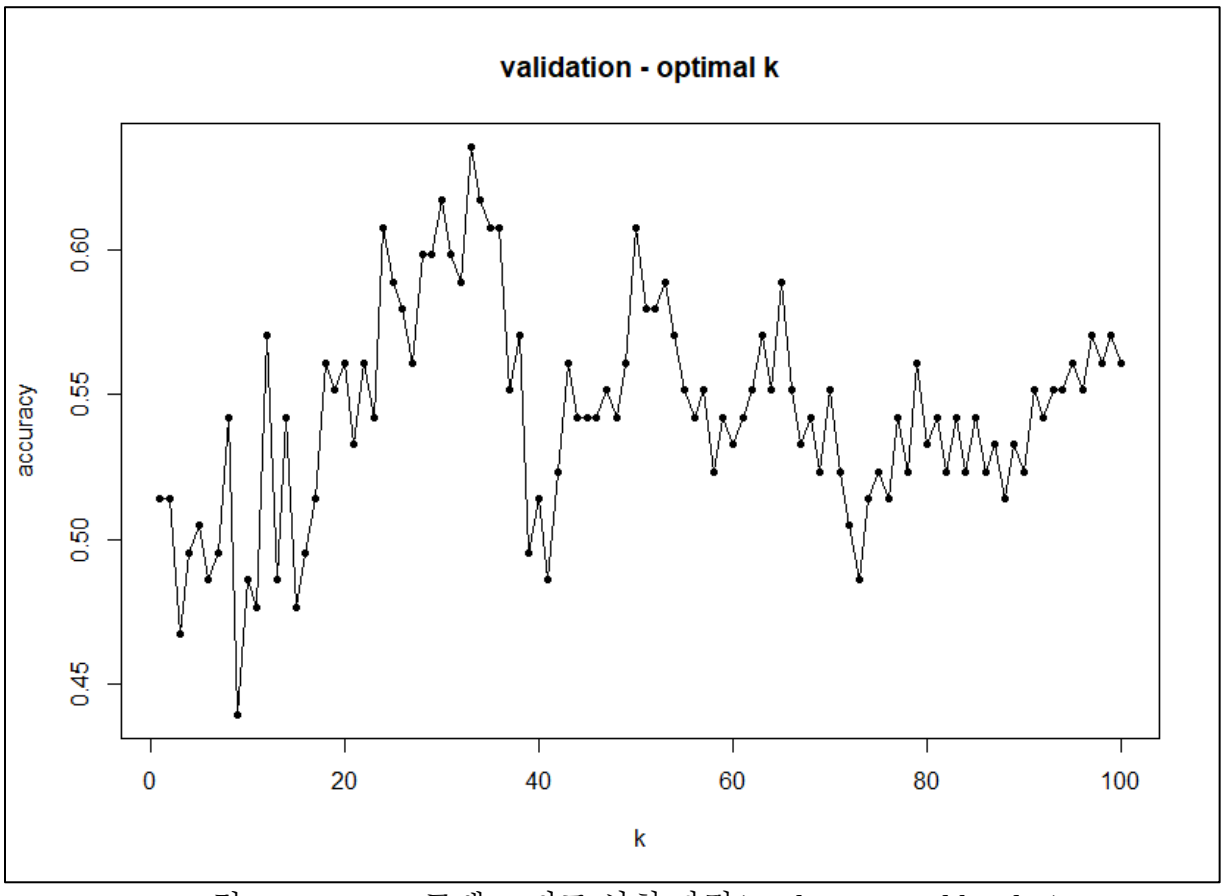
Model	SVM	KNN	XGB	GPC	RF	ANN
Precision	0.523	0.622	0.588	0.613	0.50	0.50
Recall	0.654	0.635	0.538	0.731	0.481	0.942
F1 Score	0.587	0.629	0.560	**0.667**	0.490	0.653

<표 5> GPC 모델 분류평가지표

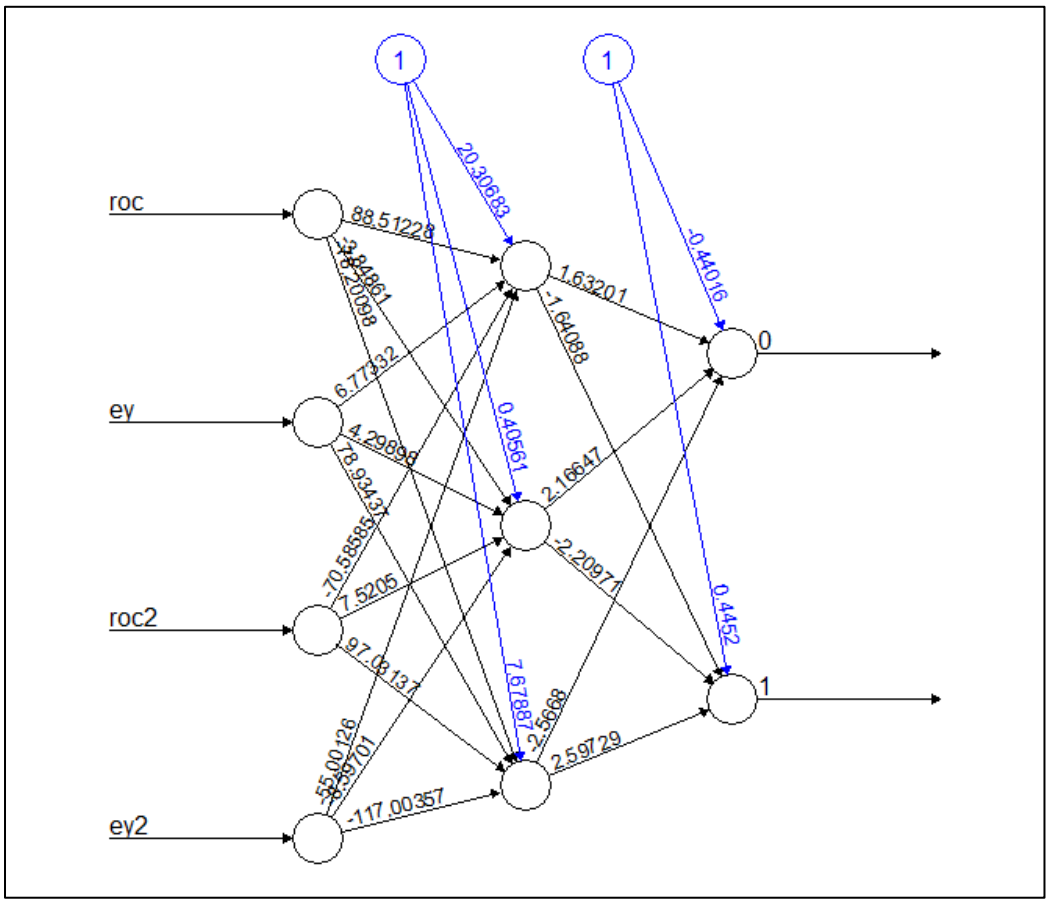
GPC Confusion Matrix		**Predicted**	
		Negative (0)	**Positive (1)**
Actual	**Negative (0)**	31	24
	Positive (1)	14	38



<그림 Extra> ROC / EY 산점도 (Non-linear Data 를 의미)



<그림 Extra> KNN 모델 그리드 서치 과정(Finding optimal k value)



<그림 Extra> ANN 모델 은닉층(Hidden Layers) 학습 과정