2023

미래에셋증권

빅데이터 페스티벌 개발 기획서

팀 명 : 샤크의 모험(SSS)

팀 원 : 김홍식, 오하은, 이진표

목차 (Table of Contents)

코로나 발생 이후 개인투자자가 주식시장에 대거 유입되었습니다. 개인투자자는 기업의 가치와 상관없이 가격이 급등하는 종목을 거래하는 하이리스크 단기투자를 반복하고 그로 인한 손실은 매매 시 고려하지 않는 것 같은 모습을 보입니다. 이러한 특징으로 인해 기존에 연구되어온 투자론 혹은 현재도 활발히 개발되고 있는 종목 분석과 종목 추천 서비스, 종목의 가격 방향성 예측 등의 알고리즘 서비스는 이러한 개인투자자에게 도움이 되지 않습니다. 보다 극단적인 리스크의 존재를 알려주는 것이 실용적일 것입니다. 이를 위해 극단적 하이리스크를 세 가지로 분류한 뒤 이를 각각 학습하여 특정 기업이 극단적 하이리스크 상황이 발생할 확률을 알려주는 서비스를 제공하고자 합니다.

Ⅰ. 서론 (Introduction)

1.1 연구의 배경 및 목적

"주식하면 삼대(三代)가 망한다." 코로나 이전 주식과 관련된 가장 유명한 말이었습니다. 주식을 해야 한다고 말하면 얼굴을 찡그리고 눈빛이 변했던 사람들이 코로나 발생 이후 동학개미 운동을 시작으로 주식시장에 대거 유입되었습니다. 어디를 가도 주식에 관해 얘기하는 사람들이 있었고 TV, 신문, SNS 등 어떤 것을 보더라도 주식에 대한 얘기는 빠지지 않았습니다.

코로나 19 발생 이후 국내 개인투자자의 신규 유입, 순매수 대금과 거래대금은 역사상 유례없는 수준으로 많이 증가하였습니다(자본시장연구원). 기본적으로 개인투자자는 투자에 대한 역량과 지식, 재무 정보 해석과 같은 능력이 부족한 경우가 많으며 심리적 요인으로 인해 비합리적인 투자 행동을 보이는 것으로 알려져 있습니다.

자본시장연구원의 「코로나 19 국면의 개인투자자: 투자행태와 투자성과」에 따르면 개인투자자의 특징은 다음과 같습니다.

"... (중략) 개인투자자 약 20 만 명의 거래자료를 바탕으로 분석한 결과는 다음과 같다. 첫째, 개인투자자의 주식 포트폴리오는 중·소형주 및 특정 섹터의 비중이 높고 평균 보유종목 수가 적어 개인투자자는 높은 투자위험을 감수하고 있는 것으로 나타난다. 둘째, 개인투자자는 거래회전율, 일중거래 비중, 종목교체율이 매우 높은 투기적인 투자행태를 보인다. 이러한 행태는 신규투자자, 젊은 투자자, 남성, 소액투자자에서 현저하게 나타난다. 셋째, 개인투자자의 투자성과는 거래비용을 고려할 경우 시장수익률을 하회하며, 신규투자자 중 60%는 손실을 시현한 것으로 분석된다. 특히 소액투자자와 거래가 빈번한 투자자의 투자성과가 저조한데, 높은 거래비용과 낮은 분산투자 수준 외에 투자 대상 및 투자 시점 선택의 비효율성과도 연관된 것으로 추정된다."

요약하면 개인투자자는 수익률만을 목적으로 비효율적이고 리스크가 높은 올인 형태의 투자를 하며 이에따라 결국 손실을 주는 행위를 반복합니다. 일명 단타라고 불리는 투자행태를 보이는 개인투자자가 점점 늘어나고 있는 만큼, 투자에 대한 지식이 부족한 사람들의 투기적 형태의 매매가 늘어나고 있는 것이 사회적으로 익숙해지고 있습니다. 이러한 투기적 매매 방식을 하는 개인투자자는 급격한 가격 상승이 발생한 주식을 시세차익의 목적만으로 매매 하기에 기업이 상장폐지가 되거나 횡령·배임, 불성실공시법인 지정이 될 가능성 등을 알지 못하거나 과소평가합니다.

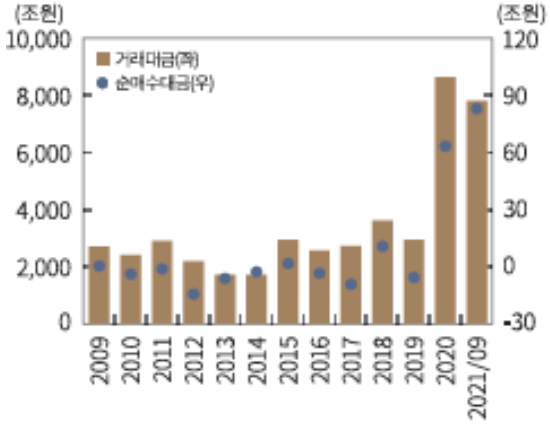
지금까지 주식투자에 대한 많은 논의와 연구가 있었고 이를 통해 다양한 투자 방법론이 등장했습니다. 이러한 발전 과정에서 주가를 예측하는 것은 불가능하다는 주장이 나오는 등 기존의 이론과 전혀 반대되는 관점 또한 대두되었습니다. 이에 따라 시장 평균 수익률을 따르는 지수 펀드나 인덱스 투자가 대두되었고 현재까지도 많은 기관투자자가 이러한 방식으로 포트폴리오를 구성하고 펀드를 운용하고 있습니다.

현재까지 투자에 대한 방법론은 시간이 지남에 따라, 투자에 대한 관점의 전환과 연구가 진행됨에 따라 투자 방식은 계속 바뀌고 진화하였습니다. 그러나 여태까지 논의되어 온 투자 방법론은 모두 장기적으로 시장은 우상향 한다는 가정과 경제적 주체인 투자자는 합리적 존재라는 가정 속에서만 의미를 가집니다. 기존의 경제학적 관점과 가정으로 설명할 수 없는 상이한 투자행태를 보이는 개인투자자를 위한 서비스를 제공하기 위해선 기존의 투자전략과 다른 방식으로 접근해야 할 필요가 있습니다. 현재까지 주식의 재무 정보를 이용하여 주식의 방향성을 예측하고자 하는 연구는 다수 존재합니다. 그러나 이러한 연구는 개인투자자에게 크게 도움이 되지 않습니다.

개인투자자는 이러한 연구의 가정과 관점을 깨고 반대로 행동하고 있기 때문입니다. 따라서 개인투자자에게 재무 정보를 통해 이 기업이 좋은 기업인지 나쁜 기업인지 알려주는 것은 효과적이지 않을 것입니다. 단기적이고 극단적인 수익률 추구를 위해 좋은 기업과 나쁜 기업을 따지지 않고 주식 가격이 급등한 종목에 투자하기 때문입니다.

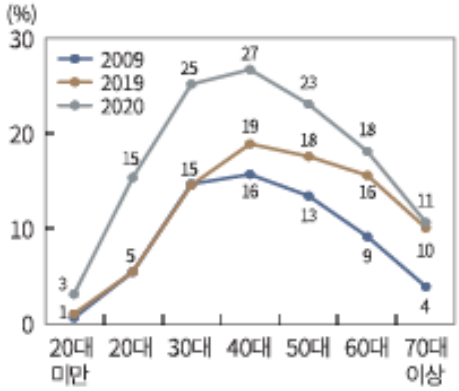
이미지 1)

<그림 Ⅰ-1>
연도별 개인투자자 거래실적 추이



자료: DataGuide (코스피, 코스닥 합산)

<그림 Ⅰ-2>
연령대별 개인투자자 비중 추이



자료: 예탁결제원, 통계청

이미지 출처: 김민기·김준석, 「국내 개인투자자의 행태적 편의와 거래행태」, 자본시장연구원, 2022

연령에 상관없이 전 세대에서 개인투자자가 늘어나고 있다는 결과를 위 자료에서 확인할 수 있습니다.

II. 문제의 정의 (Problem Statement)

2.1 구체적 정의

본 기획서는 기업을 평가하고 좋은 기업과 나쁜 기업을 구분하는 데에 목적이 있지 않습니다. 이 기획서의 목적은 기업의 가치와 비전에 관심이 없이 단기적으로 주가가 급등한 종목을 매수하는 개인투자자들을 대상으로 최소한의 방어 작용을 위한 서비스를 제공하는 것입니다. 이러한 개인투자자들은 어느 정도의 손실을 감수하는 경향이 있으므로, 이들을 위해 가장 위험한 수준의 리스크에 대한 정보를 제공해야 합니다. 이를 위해, 본 기획서는 다음 세 가지 경우를 극단적 하이리스크가 발생한 것으로 가정하겠습니다.

1. 경영 실패로 인한 상장폐지 기업과의 유사성
2. 횡령배임 혐의 발생 기업과의 유사성
3. 불성실공시법인과의 유사성

위 세 가지 중 하나 이상에 해당하는 기업을 AI 딥러닝을 통해 분석하여 서비스를 제공하고자 합니다.

첫 번째로, 경영 실패로 인한 상장폐지는 자본잠식, 감사 의견 부적정 등 기업 경영의 실패로 인해 상장 폐지된 기업이라 판단되는 기업들을 학습합니다. 경영 실패로 인한 상장폐지 가능성은 기업의 재무 정보와 비재무 정보를 AI 를 통해 학습합니다.

두 번째로, 횡령배임 혐의 발생 기업과의 유사성은 횡령이나 배임한 종목들을 파악하여 횡령 배임이 일어났던 기업들의 재무 정보와 비재무 정보를 AI 를 통해 학습합니다.

세 번째로, 불성실공시법인에 한 번 이상 지정된 기업들을 파악 후 해당 기업들의 재무 정보와 비재무 정보를 AI 를 통해 학습합니다.

학습된 결과를 통해 모든 기업에 대해 개별 기업이 경영실패로 상장 폐지될 가능성과 횡령배임이 발생할 가능성, 불성실공시법인에 지정될 가능성 세 가지 경우를 각각 수치로 알려주고자 합니다.

2.2 차별성

이 기획서와 유사한 서비스로는 '투자자 보호' 시스템이 있습니다. 투자자 보호 시스템은 관리종목, 거래정지 종목, 시장경보 종목으로 나누어 투자자들에게 종목의 위험성을 알려줍니다. 이 기획서의 첫 번째 제공 서비스인 상장폐지가 될 가능성과 유사합니다.

다만 관리종목으로 지정되지 않은 기업들에 대해서도 이 기업이 가진 상장폐지 가능성을 딥러닝으로 더 선제적이고 정확한 정보제공을 할 수 있을 것입니다. 그뿐만 아니라 재무제표를 올바르게 해석하지 못하는 투자자를 주 대상층으로 하며 재무제표를 해석할 능력이 있는 투자자 또한 재무제표 이외의 기타 변수로 인한 종목의 상장폐지 가능성을 확인할 수 있어 유의미한 서비스를 제공할 수 있습니다. 투자자 보호 시스템과 달리 이 서비스는 횡령배임과 불성실 공시에 대한 가능성도 확인할 수 있다는 차별점이 있습니다.

다른 유사한 서비스로 크래프트테크놀로지스 사의 '모자이크 AI'가 있습니다. 모자이크 AI 는 주식시장 및 연계 파생상품 시장에서 발생한 거래, 틱 데이터(tick data)와 대체 데이터를 자체 개발한 AI 모델입니다. 모자이크 AI 는 시장에서 이상 거래의 흔적을 실시간으로 탐지하고 이와 관련된 이벤트를 예측합니다. 이와 관련된 이벤트는 내부자 거래 정보와 스마트 머니 흔적 분석 등이 있으며, 앞선 테스트에서는 기업의 M&A 를 예측하기도 하였습니다. 모자이크 AI 를 비롯한 기존 시장에 존재하는 많은 종목 추천 등의 서비스는 투자 수익을 위한 주가의 방향성 예측 등이 주를 이루지만, 본 서비스는 투자의 수익보다는 극단적인 리스크에 대한 정보제공을 통해 하이리스크 매매를 반복하는 개인투자자에게 더욱 실용적인 서비스를 제공할 수 있다는 강점이 있습니다.

III. 연구 방법 (Methodology)

3.1 연구에 사용된 데이터 수집 방법

2000 년 1 월부터 2023 년 7 월까지의 모든 산업의 상장폐지, 횡령, 불성실 공시 기업을 선정하였고, 2000 년 1 월부터 2016 년 12 월까지를 학습데이터, 2017 년 1 월부터 2023 년 현재까지를 테스트데이터로 구분합니다. 또한 유가증권시장과 코스닥에 한정하여 기업을 선정합니다. 학습 과정에서는 상장된 계속기업을 무작위로 추출하여 위험기업과 1:1 비율을 맞추어 학습을 진행합니다.

가. 재무 정보: 일반적으로 기업경영분석과 선행연구에서 가장 많이 사용되는 재무비율 10 개를 변수로 두어 학습합니다.[1)] 재무 비율을 위한 상세 사항은 dart 오픈 api, 한국거래소의 정보 데이터시스템의 전 종목 시세에서 데이터들을 csv, txt 파일 등의 형식으로 얻습니다.

나. 비재무 정보: 비재무 데이터는 경영실패로 인한 상장폐지, 횡령배임 발생, 불성실공시법인 지정이 된 기업 리스트와 지배주주 지분율이 포함됩니다. 이에 해당하는 데이터는 kind, dart 에서 csv 파일의 형식으로 얻습니다.

데이터는 다음과 같은 유형을 사용합니다.

- 상장폐지 사유: 감사 의견 거절, 감사 의견 거절(감사범위 제한 및 계속기업 가정 불확실성), 외부감사인의 감사의견 의견거절, 감사의견 부적, 기업의 계속성 및 경영의 투명성 등을 종합적으로 고려하여 상장폐지기준에 해당한다고 결정 등
- 횡령 공시: 횡령, 배임 혐의의 발생, 회계 처리기준 위반에 따른 검찰 고발 등 조치 등
- 불성실 공시 유형 : 공시 번복, 공시 불이행 등

[1] 안동건·배기수,「산업별 상장폐지 예측모형 개발에 관한 연구 」, 상업교육연구, 한국상업교육학회, 30(2), p327, 2016

표 1) 사용 예정 데이터 명시

번호	인스턴스	상세사항	수집방법	샘플 예시
1	기업 규모	기업규모 = ln(시가총액), 사건 발생 전년도 연말	한국거래소정보데이터시스템 excel	ln(4,194,842,707,440) = 29.0648776
2	유동비율	유동자산/유동부채*100	dart api	756%
3	당좌비율	당좌자산/유동부채*100	dart api	159.2%
4	매출액순이익률	순이익/매출액*100	dart api	8.58%
5	총자산이익률	당기 순이익/총자산*100	dart api	10%
6	총자산회전율	매출액/총자산*100	dart api	138%
7	매출채권회전율	매출액/매출채권*100	dart api	442%
8	재고자산회전율	매출액/재고자산*100	dart api	9.9%
9	고정자산회전율	순 매출 / 평균 순 고정 자산*100	dart api	215.6%
10	매출액증가율	3 개년치 증가율 추이	dart api	8.6%
11	총자산증가율	3 개년치 증가율 추이	dart api	3.7%
12	지배주주지분율	최대주주, 특수관계인 및 계열회사임원 기말 보통주 총합	dart 사업보고서 주요정보조회	31.14%

3.2 설계 및 연구 절차

사용 예정 모델은 1) 랜덤 포레스트 2) LightGBM 3) CatBoost 4)AutoML 입니다. 저희가 제공하고 싶은 정보는 1. 경영 실패로 상장 폐지할 가능성 2. 횡령배임 할 가능성 3. 불성실공시법인에 속할 가능성으로 3 개의 클래스입니다. 학습 모델의 과적합을 방지하기 위해서 각 클래스에 다중 모델을 사용하지 않고, 하나의 모델을 사용하여 3 개의 클래스에 대해 예측을 동시에 수행하는 다중클래스 분류 모델로 학습시킵니다. 지도 학습 알고리즘을 사용할 예정이며, 앙상블 알고리즘과 AutoML 으로 총 4 개의 알고리즘을 활용해 예측합니다.

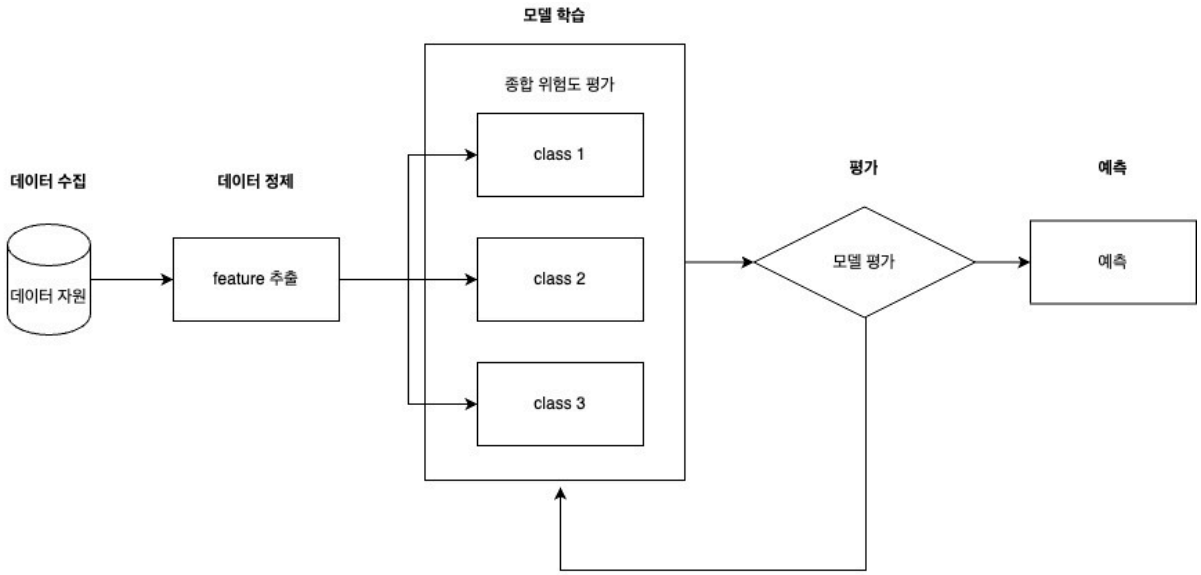
앙상블 알고리즘으로는 먼저 랜덤 포레스트(RandomForest)를 사용합니다. 랜덤 포레스트 알고리즘은 배깅의 확장으로 특성 무작위성과 모두 활용하여 상관관계가 없는 의사결정 트리를 만듭니다. 여러 개의 결정 트리를 사용하는 앙상블 기법이기 때문에 과적합 방지에 강점이 있습니다. 또한 앙상블 효과로 인하여 예측 성능을 향상할 수 있는 특징이 있습니다. 다음으로는 부스팅 알고리즘의 일종인 LightGBM 과 CatBoost 모델을 사용합니다. LightGBM 은 기존의 트리 기반으로 학습하는 그레디언트 부스팅 모델(GBM)입니다. 트리가 수평적으로 확장되는 다른 알고리즘과 다르게 수직적으로 확장되기 때문에 트리 분할 방식에서 발생하는 비대칭으로 인한 불균형을 최소화할 수 있는 특징이 있습니다. CatBoost 는 특성들을 모두 동일한 대칭적인 트리 구조를 형성합니다. 이러한 대칭 트리 형성 구조로 CatBoost 는 시계열 데이터를 효율적으로 처리합니다. 또한 기존 부스팅 계열 알고리즘보다 예측 시간을 감소시키므로 시간적인 장점도 있고, 다른 GBM 보다 과적합 방지에 유리합니다.

마지막으로는 AutoML 을 사용할 것입니다. AutoML 은 기계 스스로 적합한 학습 모델을 찾아 자동화합니다. 데이터의 특성을 분석하여 알고리즘을 선별해 관련 매개변수 값을 최적화하는 솔루션을 알아서 제공해 주는 장점이 있습니다.

3.3 사용된 도구나 기법 설명

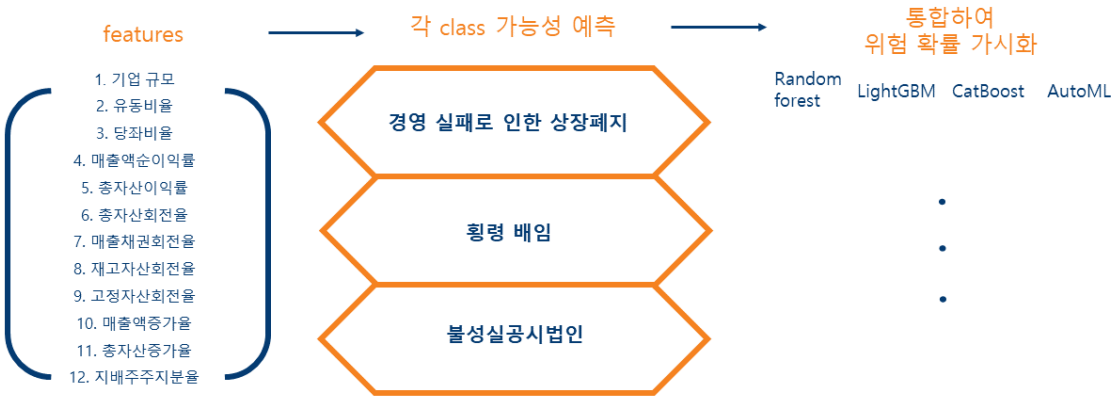
전체적인 학습 과정은 다음과 같습니다.

이미지 2)



메인 알고리즘의 흐름은 다음과 같습니다.

이미지 3)



데이터 자원으로부터 특성을 추출하기 위한 데이터들을 수집합니다. 일반적으로 기업 예측분석에서 많이 사용되는 11 개의 재무 요소와 비재무 요소로서 지배주주 지분율을 포함한 12 개의 특성을 사용합니다. 기간에 따라 나눈 학습데이터를 RandomForest, LightGBM, CatBoost, AutoML 로 학습시켜 기업경영 실패로 인한 상장폐지, 횡령배임 발생, 불성실공시법인 지정이 될 가능성을 예측하고 평가합니다. 이에 따라 서비스 이용자는 특정 기업이 보이지 않는 극단적 하이리스크의 잠재적 요소를 확인할 수 있습니다. 따라서 투기적 매매를 반복하는 개인투자자를 위해 유의미한 정보를 제공하는 서비스가 가능할 것입니다.