한국 아이돌의 유튜브 영상 정보를 이용한  
소속사 주가 예측 가능 여부 분석 프로젝트

-서강대학교 23-1 AI와 빅데이터를 활용한 경제정보처리(ECOG051-01) 학기말 개인 프로젝트-

작성자: 노유정(20191245)

최근 몇년 사이 투자에 관한 국민들의 관심이 높아지면서, 우리 주변에 주식 거래자가 많아졌음을 체감할 수 있게 되었습니다. 올해(2023년) 3월 한국 예탁결제원에서 발표한 자료에 따르면, 2022년 12월 기준 대한민국의 상장 법인 주식의 총 소유자가 약 1441만명으로 추산되었습니다. 이는 2023년 기준 생산연령인구(15-64세)인 3637만명의 40%에 달하는 숫자입니다. 하지만 이들 모두가 투자하는 회사의 내재가치를 분석하고 주식을 매수하지 않습니다. 지인에게 들은 정보로 거래하거나, 그냥 유명한 회사이기 때문에, 또는 많은 사람들이 그 회사의 주식을 샀기 때문에 거래하는 경우를 쉽게 찾아볼 수 있습니다.

따라서 회사의 호재를 찾아내거나 ROE, ROA 등의 주가지표를 분석하는 등의 복잡한 과정 없이도, 직관적으로 회사의 내재가치를 알아낼 수 있는 방법이 있을지 생각해보았습니다. 접근성이 좋고, 받아들이기 쉬우며, 가치를 포함하는 정보를 지금 이 순간에도 쌓고 있을 곳이 어디일지 생각하다보니 세계 최대 규모의 비디오 플랫폼인 유튜브를 떠올리게 되었습니다. 유튜브는 영상을 얼마나 많은 사람들이 보았는지, 그들이 영상을 좋아하고 싫어했는지에 대한 수치를 제공하며 그 영상의 댓글 또한 컨텐츠에 대한 수많은 평가를 내리고 있습니다. 저는 많은 회사들 중 k-pop 엔터 회사가 유튜브에서 얻을 수 있는 위 정보들로부터 내재가치를 파악하기 가장 적합한 회사라고 판단하였는데, 그 이유는 회사의 산물이 영상(방송, 뮤직비디오, 리얼리티 등)이며 회사의 가치 대부분이 이에 대한 '대중의 평가'에 비롯하기 때문입니다. 따라서 저는 한국 아이돌의 유튜브 내 정보들을 이용한 소속사 주가 예측을 주제로 선택하게 되었습니다.

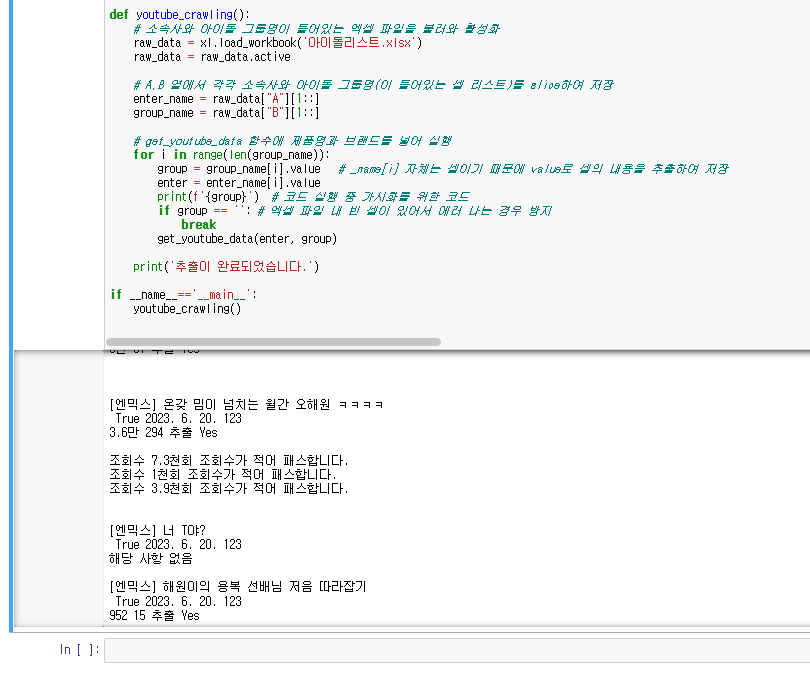
기존에 3대 엔터로 불리던 SM, YG, JYP를 분석하고자 하였으나, SM이 다른 곳으로 인수되는 등 대중의 평가만을 지표로 사용하기 어렵다고 판단하였습니다. 따라서 SM을 제외하고 방탄소년단 등이 소속되어있는 레이블 HYBE를 넣어 YG, JYP, HYBE의 3가지 엔터 회사를 분석하게 되었습니다. 세 회사의 총 열 네그룹의 아이돌 데이터를 사용하여, 각 아이돌의 유튜브 영상 조회수, 좋아요 수, 댓글 수를 크롤링하여 일자별로 합친 후, 그 날 올라온 영상의 갯수, 좋아요 수 / 조회수, 좋아요 수 / 댓글 수 데이터를 준비하였습니다. 즉 소속사에 소속된 아이돌이 n그룹일 때 각 그룹의 조회수, 좋아요/조회수, 좋아요/댓글수, 영상 수 총 4n개의 칼럼(X)을 가지고 그 날의 종가(y)를 예측하는 XGboost 활용 Regression 모델을 만들 것입니다. 또한 overfitting을 막기 위해 early\_stopping을 사용하고, 가장 좋은 성능의 모델을 찾기 위해 grid-search를 사용하여 모델을 만들고 이 결과를 분석할 것입니다.

소속사와 아이돌 그룹명, 그룹명의 영어 약칭, 소속사의 종목 코드 데이터를 가지고 있는 엑셀 데이터를 만들어서 준비하였습니다. 유튜브의 데이터는 파이썬의 selenium과 bs4 모듈을 이용하여 크롬 드라이버 바탕의 웹 크롤링 방식으로 가져올 예정입니다. 준비한 엑셀 데이터를 이용하여 예측하고자 하는 엔터 회사에 소속되어 있는 아이돌을 유튜브에 검색, 일정 기간(300개장일, 약 1년)내에 올라온 영상에 접속한 후 영상의 조회수, 영상의 좋아요 수와 댓글 수를 가져와 분석에 활용할 것입니다. 데이터의 일차적인 정제를 위해, 데이터 수집 과정에서 조회수가 1만 회보다 적은 영상의 경우는 제외하였습니다.

-준비한 엑셀 데이터



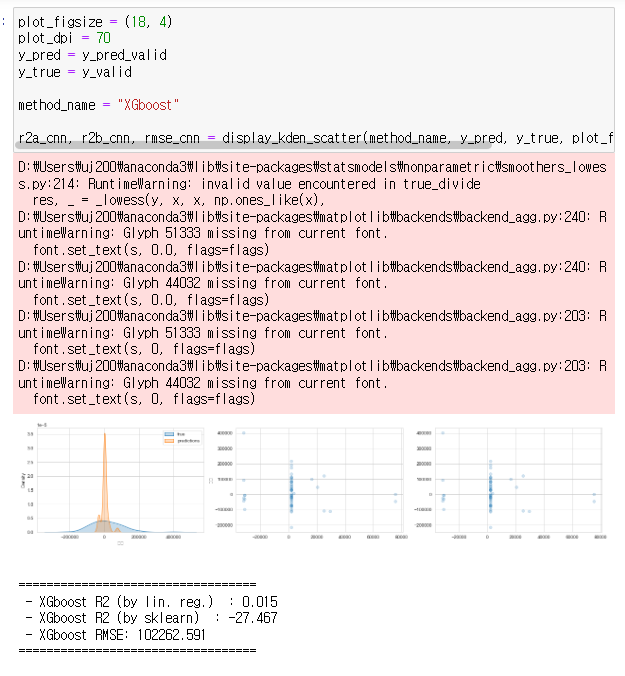
-데이터 크롤링 과정 (전체 코드가 길어, 자세한 설명은 코드에 직접 적어두었습니다.)



분석을 진행한 결과, 다음과 같은 결과를 얻을 수 있었습니다.

하이브

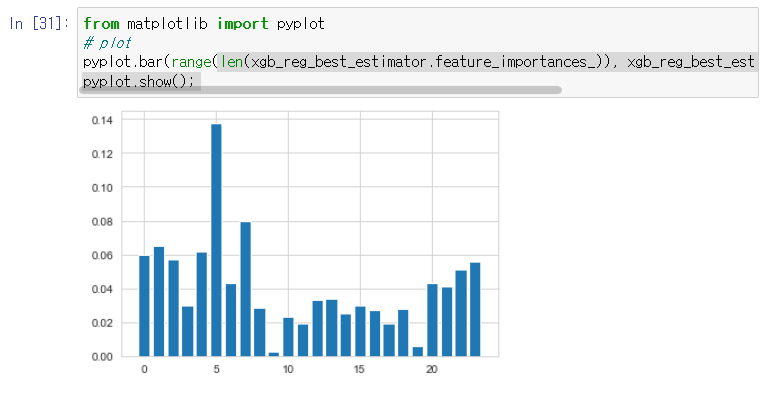
그리드서치 진행 전



진행 후

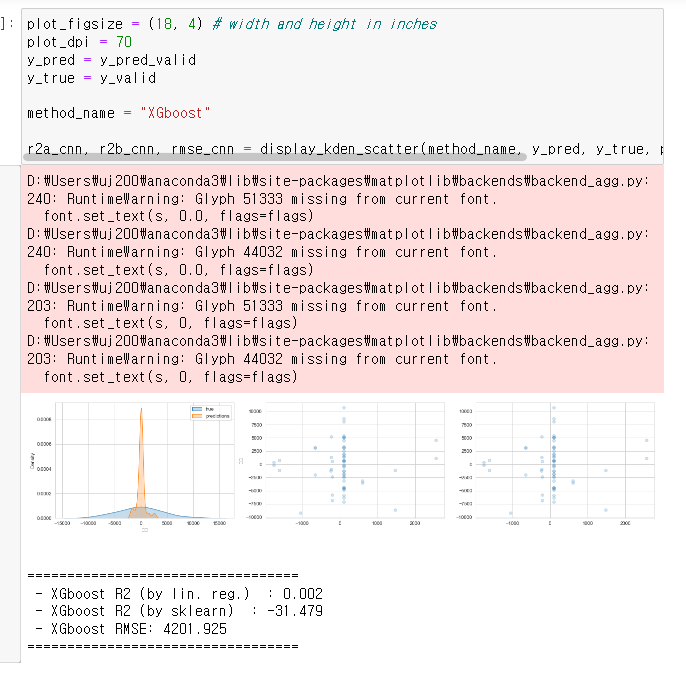


best\_feature



YG

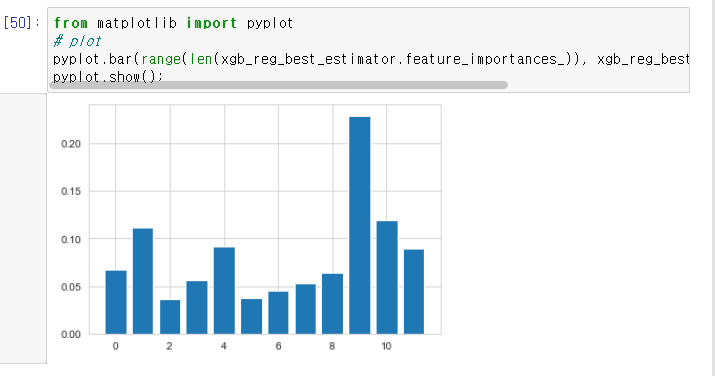
그리드서치 진행 전



진행 후

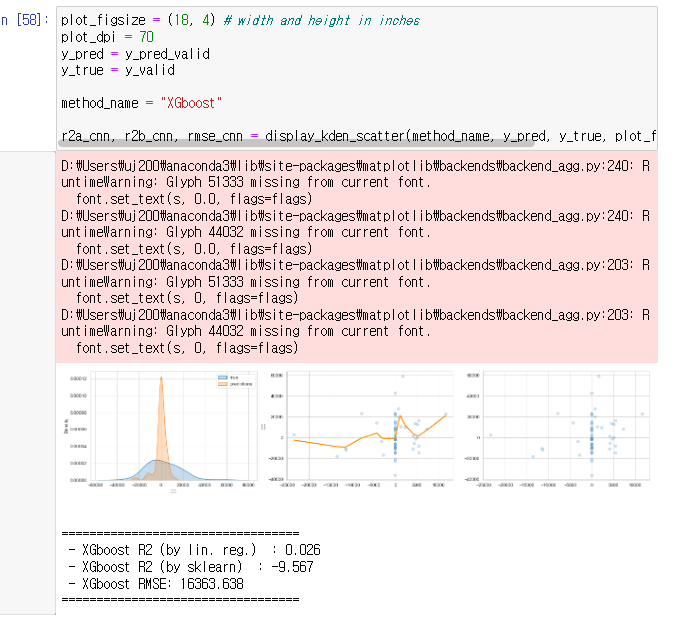


best\_feature



JYP

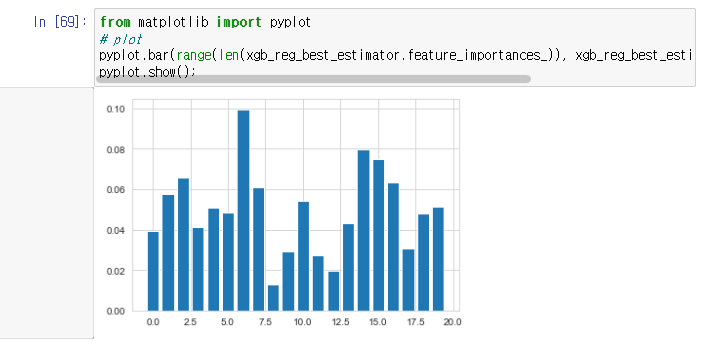
그리드 서치 진행 전



진행 후



best\_feature



각 분석 진행 결과, 그리드 서치 진행 전 RMSE값이 102262.6 / 4201.9 / 16363.6 , R2값이 0.015 / 0.002 / 0.026 진행 후 RMSE값이 98348.6 / 4157.5 / 16365.6 , R2값이 0.018 / 0.009 / 0.003 로 나왔습니다. RMSE는 예측 오차의 평균 크기를 나타내므로(데이터와 동일한 단위) 이가 작을수록 예측이 실제 값과 가깝다고 해석할 수 있는데, 분석 결과 값이 주가의 크기(최대 30만원)인 것에 비추어 볼 때 생성된 각 모델의 예측력이 거의 없다고 볼 수 있습니다. 또한 R2는 1에 가까울수록 모델이 종속변수의 변동성을 잘 설명한다고 볼 수 있는데, 이 값도 전부 0.03 미만으로 통상적으로 모델의 성능을 평가하는 척도인 0.7에 비해 한참 낮게 나왔습니다. 종합적으로 보았을 때, 각 모델이 가지고 있는 독립변수, 즉 유튜브의 영상 수, 조회수, 좋아요 수, 댓글 수(좋아요/조회수, 좋아요/댓글수)로는 주식의 종가를 예측하지 못했다는 것을 의미합니다.

이번 분석 자체는 좋은 결과를 얻지 못했지만, naive한 데이터가 가득한 유튜브에서 필요한 데이터를 가져다가 분석에 사용했다는 점에서 다음 분석에서 발전한 형태로 활용할 것을 기대할 수 있습니다. 또한 이번 분석에서 활용한 엔터 회사의 경우 대부분의 소비자가 한 플랫폼에 피드백을 주는 특수한 형태로 이루어져 있어 다른 주가 분석 프로젝트와는 달리 전수조사에 가깝게 데이터를 얻을 수 있었다는 장점이 있습니다.

분석에서 제한되었던, 개선할 점은 우선 아이돌 그룹 이름이 특징적이지 못한 경우에 원하지 않은 영상이 추출될 수 있다는 것입니다. 예를 들어 YG에는 보이그룹 '트레져'가 존재하는데, 크롤링 진행 시 '트레져헌터' 등의 다른 내용들이 검색되고 아이돌 영상은 육안으로 찾아볼 수 없을 정도로 적었기에 분석에서 제외하기도 하였습니다. 두번째는 데이터 자체의 오류입니다. 영상 업로드 일자 이후로 조회수 등의 데이터가 현재까지 쌓여있기 때문에, 이로 인해 발생하는 문제를 줄이기 위해 비율 데이터로 변환해 사용하였지만 그럼에도 불구하고 현재까지의 데이터가 포함되어있다는 문제가 여전히 남아있었습니다. 이를 더 확실히 줄이기 위해서는 댓글을 전부 크롤링하여 업로드 일자에 달린 댓글 비율을 구하고 이를 각각 조회수, 좋아요 수에 곱해주는 형태로 진행하는 것이 하나의 방법이 될 수 있습니다. 또한 댓글 자체를 크롤링하여 감정 평가 진행 후 이 데이터까지 사용하고 싶었으나, 이를 크롤링으로 가져오기에 환경이 받쳐주지 않아 실패하였습니다. 고성능의 GPU를 사용하게 된다면 이 문제를 해결할 수 있을 것입니다.

참고 자료

한국 예탁결제원 22년 12월 결산 상장법인 주식 소유자 현황   
<https://www.ksd.or.kr/ko/about-ksd/ksd-news/press-release/36531>

대한민국 인구상황판  
https://kosis.kr/visual/populationKorea/PopulationDashBoardMain.do