**머신러닝으로 코스피 주식 가격 예측하기**

CUAI 3기 Machine Learning Track A팀

공태웅(공공인재학부), 권예진(응용통계학과), 김원준(국제물류학과)

**[요약]**

코스피에 상장된 800개 기업 중 약 96%에 해당하는 766개사의 영업 실적과 전문가의 예측치, 부채 등의 지표를 데이터로써 수집했다. 가장 최근일 주가를 타겟 값으로 삼아 해당 데이터를 기반으로 기계학습을 진행하여 나름대로 주가를 예측하는 형태의 모델을 만들고자 하였다. 프로젝트를 기획하며 해당 모델을 바탕으로 팀원이 실제 관심 갖는 기업의 주가를 예측하는 형태로 발전할 수 있길 희망하였다.

**1. 서 론**

기실 '미래 예측'이란 비단 SF적이고 거창한 무엇이 아니라, 우리가 숨 쉬듯 행하는 일이기도 하다. 어제 오늘 몸이 으슬으슬 떨리는 것을 보아하니 내일이면 콧물이 날 것을 경험적으로 알고 있듯, 과거 상태를 진단해 가까운 미래의 상황을 내다보고 우리는 조치를 취하곤 한다. 그 '조치'로서 머신러닝 프로젝트를 어떻게 시도해 볼지 우리는 고민했다.

근래 매스컴에서 빠짐없이 오르내리는 것은 주식이었다. 친구 따라 묻지 마 매매를 하는 사람도 있다지만, 머신러닝을 공부한 사람으로서 우리 A팀은 본능적이고 원초적인 욕구에 굴복할 수밖에 없었다.

과거 기업 지표 데이터를 학습 피쳐로 사용해 특정 기업의 곧 다가올 주가를 어느 정도 예측할 수 있는가가 궁금했다. 물론 이러한 과정은 수많은 애널리스트와 투자사가 전문적으로 다루는 영역이나, 공부한 머신러닝 기법을 낯선 경연대회가 아닌 우리 핸드폰 속 주식 앱에 접목해보는 것 자체만으로 의미가 있다고 판단했다.

**2. 본 론**

**1) 기계학습 과정 기획**

주가에 반영되는 지표는 셀 수 없이 많으나 토이 프로젝트 사이즈의 프로젝트 진행을 위해 대표적인 지표를 선별하는 과정을 거쳐야 했다. 크게 세 가지 성격의 지표로 구분하여 피쳐로 쓸 데이터를 결정했다.

1. 기업의 영업 실적과 부채가 담긴 데이터  
   (분기별 영업/순이익률, 부채비율 등)
2. 투자 심리가 담긴 데이터  
   (공매도 비율, 외국인 보유 주식 수, 투자의견)
3. 기업의 규모  
   (시가총액, 매매단위, 상장주식수 등)

이러한 항목의 데이터를 추출하여 코스피 상장사를 데이터셋의 행(row)으로, 각 지표들을 열(column)로 배치하여 학습용 데이터셋으로 구축하고자 계획하였다.

그러나 통상적인 기계학습 프로젝트 경험에 비추어 본 프로젝트가 가질 한계점 역시 기획 단계에서부터 쉽게 예상할 수 있었다. 사용될 데이터 행의 수에 비해 코스피 상장사는 800개사에 국한되어 있고, 피쳐에 사용된 기업의 건전성에 대한 지표 (예를 들어 영업이익률과 EPS)는 대체로 각 지표 간의 상관관계가 존재할 가능성이 높았다.

이러한 데이터 부족 + 피쳐 간 다중공선성 문제가 우려되는 가운데서도 최대한 좋은 퍼포먼스를 내는 모델을 구축하고자 하였다. 더욱이 팀원 중에는 실제로 주식 투자를 적극적으로 하고 있는 사람도 있었고 경영경제대학 소속으로서 기업 지표들에 익숙한 팀원들도 있는 등 프로젝트 성사에 대한 의지가 강했다.

**2) 데이터 구축**

코스피 상장사의 피쳐 데이터는 기본적으로 회계규칙에 의해 공시된 데이터를 크롤러를 사용하여 수집하였다. 네이버 금융 ([www.finance.naver.com](http://www.finance.naver.com)) 사이트에는 기업 실적 데이터가 DART로부터 데이터를 받아 대리제공하고 있는데, 이를 크롤러를 사용해 데이터셋을 구축하였다. 이에 공매도 수량과 같이 DART에서 다루지 않는 데이터는 한국거래소(KRX)의 공매도 데이터를 수집해 병행하여 데이터셋을 보충하였다.

현재 기준으로 코스피 상장사는 총 800사이지만, 각종 우선주와 데이터가 이례적으로 결여된 기업들을 제외하여 총 766개 (약 96%) 의 레이블을 구축할 수 있었다. 이러한 레이블에는 삼성전자, 네이버와 같이 국내 대표 우량기업들 외에도 JW홀딩스, 도화엔지니어링 등 생소한 기업들도 많았다.

주가는 특성상 시계열적 성격을 갖게 되며, 이는 곧 업종 특성 혹은 국내 경제 사이클에 따라 일정한 패턴을 갖게 된다는 의미로 볼 수 있다. 이를 반영하여 타겟 값인 현재 주가를 예측하기 위해서는 분기를 기준으로 영업이익률, 매출액과 같은 피쳐들을 근 5개 분기 (19년 3,4분기, 20년 1,2,3분기) 별로 펼쳐 수집하였다.

또한 데이터셋이 수치형 데이터로만 구성된 것은 아니고, 기업의 업종과 같은 범주형의 데이터도 수집하였다. 미래산업으로 부각되는 특정 업종은 기업의 평가에도 긍정적으로 평가된다는 점을 착안하여 업종 데이터를 피쳐로 사용하였고 이는 원 핫 인코딩 처리하여 학습에 사용되었다.

**3) 모델 구축과 파라미터 튜닝**

앞선 과정을 거쳐 수집한 데이터를 바탕으로 현재 주가 예측을 위한 모델을 구축하였다. 모델로서는 대중적이면서도 우수한 성과를 자랑하는 랜덤포레스트 모델을 차용하였고 많은 피쳐들을 사용하는 환경에서 최적화된 파라미터를 얻기 위해 하이퍼 파라미터 튜닝을 진행하였다.

튜닝 기법으로 사용한 것은 베이지안 옵티마이저이며, 기반 지식이 없는 상황에서 튜닝에 대한 명확한 기준이 없었기 때문에 튜닝 작업을 최대한 자동화하기로 결정하였다. RFTuner 클래스를 사용하여 RMSE를 낮추는 방향으로 파라미터 튜닝을 진행했고, 55회의 이터레이션 끝에 파라미터를 결정할 수 있었다.

나아가 주가 학습 및 예측에 어떤 피쳐(기업 지표)가 가장 큰 영향을 미치는지 파악하였고 주당순자산가치인 BPS과 주가순자산비율인 PBR이 이에 손꼽혔다. 이러한 파라미터 세팅을 기반으로 다음과 같은 주가 예측이 가능하다고 정리할 수 있다.

삼성전자의 근 시일의 주가 방향성을 알고자 할 때 삼성전자를 제외한 700여 기업의 데이터를 학습한 모델을 구축하는 것이다. 이를 바탕으로 삼성전자의 피쳐들을 기반으로 모델이 뽑아낸 타겟값 (주가) 을 예측 결과로 삼는 것이다.

**3. 결 론**

서론에서도 언급하였듯 데이터와 인공지능에 기반한 주가 예측은 투자전문기관뿐만 아니라 수많은 전업 투자자들이 전문적인 지식을 바탕으로 도전하는 영역이다. 그와 동시에 수많은 변수들과 데이터 노이즈로 아직 정복되지 않은 분야이기도 하다.

그러한 성격의 프로젝트인 탓에 A팀 역시 데이터를 선별하는 과정과 알맞은 머신러닝 예측 기법을 선택하는 과정에서 부족함을 보인 것이 사실이다. 그러나 데이터를 크롤링한 다음, 의미 있는 파라미터를 뽑아내고, 실제 주식 투자 활동에 활용하는 과정에서 많은 인사이트를 얻어낼 수 있었다.

이러한 일련의 과정이 비단 투자자들에게만 의미 있다고 보진 않으며, 머신러닝의 기본적인 프로세스를 익히는 사람에게도 좋은 경험이 될 수 있으리라 생각한다.