**금융서비스마〮케팅 분야 분석 종합 프로젝트 기획안**

**2023년 5 월 18 일**

**과정명: 서비스 산업 데이터분석가 취업캠프(Python) 11회차**

|  |  |
| --- | --- |
| 프로젝트 주제 | 상관분석과 카이제곱검정을 이용한 주가 독립변수 선정 및  랜덤포레스트와 LSTM을 활요한 주식종목 선정과 주가 수익율 예측 |
| 팀 명 | 김종우 |
| 팀 원 | 김종우 |
| 프로젝트 내용  해결하고자 하는 문제  최종 산출물의 청사진 | 1. 주가수익율 관련 변수의 수치형 및 연속형 과거 데이터를 수집하고  분석하여 객관적 논리적으로 독립변수를 선택한다.  2. 선정된 독립변수를 기준으로 과거 데이터를 수집하고 데이터를 구성  하여 랜덤포레스트 모델로 주식종목 선정을 훈련시키고, 한국거래소의  약 2600개 상장종목을 전부 크롤링하여 수익 가능성 높은 종목을  선정한다. 신규상장과 상장폐지를 주기적으로 반영할 수 있다.  3. 선정된 종목의 과거 데이터를 수집하여 주식가격의 변동추세,  계절성 같은 입력시퀀스 패턴을 파악하여 미래에 적용하는  LSTM모델을 학습시켜서 미래의 주가수익율을 예측한다.  4. 주기적인 투자레터 및 텔레그램 발송 프로그램을 만든다.  5. 신문기사 및 종목리포트 같은 비정형데이터의 활용(미정) |
| 팀원간 역할  분담 및 일정 | 1. 데이터 크롤링: 김종우  2. 독립변수 상관분석 및 선정 : 김종우  3. 랜덤포레스트 학습 및 종목선정 : 김종우  4. LSTM 학습 및 미래수익율 예측 : 김종우  5 투자레터 및 텔레그램 발송 프로그램 : 김종우  6. 기획안 및 포트폴리오 작성 : 김종우 |
| 프로젝트 수행 방향  수행 방법/도구 | 1. 데이터 크롤링: 상장종목 크롤링은 ‘한국거래소’에서 진행,  독립변수(PER/PBR/ROE/DebtRatio/DividendRate/InvestGrade 등) 및  종속변수(주가시계열자료)는 ‘네이버증권’에서 수집  ☞ Tool: Scrapy, Requests, Pandas  2. 독립변수 상관분석 및 선정 : 연속형변수(예:투자의견등급평균)는  t-검정을 수행하고, 수치형변수(PER/PBR/ROE 등)는 피어슨  상관분석으로 비례/반비례 관계를 파악하여 독립변수로 선정  ☞ Tool: Scipy.stats, Numpy, Matplotlib.pyplot, Seaborn  3. 랜덤포레스트 학습 및 종목선정 : 독립변수와 종속변수를 기준으로  선정/비선정이 설정된 종목으로 학습을 진행시켜서 새로운  목록으로 투자대상 종목을 선정  ☞ Tool: Scikit-learn, Numpy, Pandas, Matplotlib  4. LSTM 학습 및 미래수익율 예측 : PER(주가수익비율), PBR(순자산비율),  ROE(자기자본이익률), Debt Ratio(부채비율), Dividend Rate(배당률),  Invest Grade(투자의견등급평균) 등 자료를 입력하고, Input gate로  LSTM Cell의 시그모이드함수, 편향 등 구성요소의 가중치를 조절하여  예측정보를 output gate에서 결정한다.  ☞ Tool: TensorFlow, Keras, PyTorch, Numpy, Pandas  5 투자레터 및 텔레그램 발송 프로그램 : 이메일본문에inline 기입된  투자종목과 정보, 예상수익율을 텔레그램과 함께 동시에 발송시킴  ☞ Smtplib, Email, Telegram, Asyncio, Requests |
| 프로젝트 범위  (상세 구현 기능) | 1. 데이터 크롤링:  1)종목: 2601개(5/18일기준) 코스피/코스닥 종목정보(종목명/종목코드)  ☞ url: https://kind.krx.co.kr/corpgeneral/corpList.do?method=loadInitPage  2)독립변수: PER/PBR/ROE/DebtRatio/DividendRate/InvestGrade 등  ☞ url: https://finance.naver.com/item/coinfo.naver?code={code}  2. 독립변수 상관분석 및 선정  1)투자의견등급평균과 주가수익율 귀무가설 t-검정  ☞ 귀무가설: 등급평균이 4점이상일 때 6개월 주가수익율이 10%이상이다.  2)독립변수와 주가수익율에 대한 피어슨 상관분석  ☞ 피어슨상관분석: 독립변수와 종속변수(주가수익율)간의 상관계수 도출  ※ 1에 가까우면 비례관계, -1에 가까우면 반비례 관계  3. 랜덤포레스트 학습 및 종목선정  1)학습데이터셋: 500개종목을 선정/비선정으로 설정  ☞ 학습데이터는 훈련에 사용하고, 검증데이터는 성능을 평가한다..  2)튜닝: 트리의 개수/깊이, 노드의 분할등을 조정하여 성능 제고  최적의 매개변수를 선택할 수 있도록 평가한다.  3)종목선정: 새로운 종목 2000개에 대한 종목선정을 수행한다.  4. LSTM 학습 및 미래수익율 예측  1)독립변수 전처리: 스케일 조정  ☞ 정규화(normalization), 표준화(standardization)  2)독립변수 입력  ☞ 수치형 독립변수들의 일정기간 데이터 일괄 입력  3)데이터분할 및 학습  ☞ 학습데이터셋으로 학습하고 테스트데이터셋으로 성능 검증  3)예측실행  ☞ 랜덤포레스트에서 선정된 종목의 미래 주가예측(테이블 및 그래프)  5 투자레터 및 텔레그램 발송 프로그램 |
| 프로젝트 관리 방안 | 1.종목 업데이트: 한국거래소 신규상장/상장폐지 종목 반영  2.분기별 재무지표 반영: 네이버증권에서 크롤링 데이터 업데이트  3.데이터 업데이트시 자동 이메일 및 텔레그램 발송 |
| 상세 일정 | ● 일정  5/10일 ~ 17일 : 주제선정  5/18일 : 기획안 확정 및 업로드(to구글드라이브)  5/20일 : 1차 멘토링 및 기획안 검토  5/23일 : 기획안 발표  5/25일 : 크롤링 및 데이터세팅 완료  6/02일 : 상관분석(독립변수vs종속변수) 완료  6/03일 : 2차 멘토링 및 중간발표(질의/응답)  6/10일 : 랜덤포레스트 완료  6/15일 : LSTM 완료  6/17일 : 3차 멘토링 및 중간발표(질의/응답)  6/20일 : 메일링 및 텔레그램 완료  6/27일 : 성능테스트 완료  6/29일 : 프로젝트실행(설계구현/테스트), 포트폴리오및프로젝트 완료  6/30일 : 프로젝트발표 |
| 기대 효과 | ▲수익모형(창업)으로 연결  1)단순한 변수설정으로 투자자의 직관적인 사용을 유도  2)매개변수설정 및 위험범위선택으로 투자성향별 퀀트투자 실현  3)쉬운설정으로 기존 증권사별 HTS/MTS를 대체 |