**비정형 데이터 활용 분석 프로젝트 기획안**

**2022년 4월 21일**

**과정명: [멀티잇]서비스 산업 데이터분석가 취업캠프(Python)**

|  |  |
| --- | --- |
| 프로젝트 주제 | Random Forest와 LSTM을 활용한 주식 종목 선정과 주가 예측 |
| 팀 명 | 김종우 |
| 팀 원 | 김종우 |
| 프로젝트 내용  해결하고자 하는 문제  최종 산출물의 청사진 | Random Forest와 LSTM모델을 활용하여 수익성 있는 주식 종목들을 선정하고, 그 종목들의 미래 예상주가를 그래프로 보여주는 프로젝트를 수행하여 투자자들에게 다음과 같은 편리를 제공하고자 합니다.  첫째, 데이터를 활용한 종목 선정으로 투자의 수익성 제고  둘째, 예상주가를 활용한 목표 수익율 설정  ☞ 사용자가 하이퍼파라미터 변경 기능 추가(미정) |
| 프로젝트 추진 일정 | ● 일정  4/20 : 주제협의 with 선생님(구현가능성 및 방법/절차 문의)  4/21(금) : 크롤링체크(dart.or.kr / finance.yahoo.com / finance.naver.com)  4/22(토) : 모델링 확인(Random Forest/LSTM, SVM/RNN)  4/24(월) ~ 25(화) : 기획안 작성/리뷰/발표  4/21(금) : 설계착수(구현방법/절차 등)  4/24(월) ~ 30(일) : 설계확정 및 구현  4/30(일) ~ 5/1(월) : 테스트  4/28(금) ~ 5/1(월) : 포트폴리오(파워포인트) 작성  5/2(화) : 검토 및 수정 with 선생님  5/3(수) : 최종검토 및 수정 with 선생님  5/4(목) : 포트폴리오 발표 및 포트폴리오 자료 업로드(to MLP) |
| 프로젝트 수행 방향  수행 방법/도구 | 프로젝트를 두단계로 나눴습니다.  첫번째 단계는 종목 선정입니다. 머신러닝의 Random Forest를 사용하여 수익 가능성 있는 주식 종목을 선정하는 것입니다. 종속변수는 선정or비선정이며 입력변수는 PER, 산업동향(기사내 특정 단어 빈도 순위), 환율, 금리입니다. 주식가격(일별종가), 환율, 금리는 네이버에서 뷰티풀수프 사용하여 크롤링하고, PER은 네이버에서 접근 불가할 경우 open DART의 재무정보를 받아서 직접 계산할 예정입니다. 산업동향은 네이버와 다음의 ‘경제’ 뉴스기사 크롤링하겠습니다.  2500개 주식종목에 대한 입력변수 칼럼과 종속변수 선정/비선정 칼럼을 별첨(semiPjt\_train.csv) 예시와 같이 만들고, 입력변수들의 조건을  전처리하여 주식종목별 선정(selection)여부를 선정/비선정 칼럼에 0과 1로 표시하여 target 데이터세트를 만들어 줍니다. 이러한 target데이터세트와 입력변수 칼럼으로 이루어진 훈련 데이터세트를 Random Forest모델에 넣어서 모델을 생성시킬수 있습니다.  ☞참조사이트 : https://finance.naver.com/item/coinfo.naver  [ 종목선정에 대한 머신러닝 ]  import numpy as np  import pandas as pd  from sklearn.ensemble import RandomForestClassifier  train\_df = pd.read\_csv("./semiPjt\_train.csv")  test\_df = pd.read\_csv("./semiPjt\_test.csv")  # 입력변수 데이터 생성  df = pd.read\_csv('semiPjt\_train.csv')  X = df[['환율', '금리', '배당수익율', '당기순이익율', 'PER', '핵심키워드종합순위']]  # 종속변수 라벨링  y = y.map({'선정': 1, '비선정': 0})  # 입력변수 조건 충족하는 종목 전처리 필터링  X\_selected = X[(X['환율'] >= 1250) & (X['금리'] <= 2.0%) & (X['배당수익율'] >= 4%) & (X['당기순이익율'] >= 10%) & (X['PER'] <= 12)& (X['핵심키워드종합순위'] <= 3)]  # Random Forest 모델 훈련  rf = RandomForestClassifier(n\_estimators=100)  rf.fit(X\_selected, y)  #정확도 계산  rf.score(X\_selected, y)  # 테스트 데이터에 대한 종속변수(선정/비선정) 예측  new\_X = pd.read\_csv('semiPjt\_test.csv')  y\_pred = rf.predict(new\_X)  # csv로 출력  submission = pd.DataFrame({  "종목명": test\_df["종목명"],  "선정/비선정": y\_pred  })  submission.to\_csv('selection.csv', index=False)  두번째 단계는 미래 1년간 일별 예상 주가입니다. 시계열 모델중에서 선정된 주식종목의 미래 예상주가를 LSTM방법으로 구하겠습니다. 종속변수는 주식가격입니다. 입력데이터는 과거의 주식가격, 거래량, PER, 배당수익율, 금리, 환율등 시계열 자료를 사용하겠습니다.  이상입니다. |