문제점 파악

주식 매매에 데이터 분석을 활용하는 이유는 간단합니다. 단기 경험으로는 주가가 상승할 종목을 찾기가 상당히 어렵다는 것입니다. 일단 종목이 수 천개 입니다. 수천 개의 종목을 수작업으로 종목별 분석은 거의 불가능할 것 같습니다. 엑셀의 도움을 받을 수 도 있겠으나, 엑셀로 수천개의 종목과 몇 년치 데이터를 다루는 것도 어렵습니다. 따라서 파이썬을 이용해 데이터를 수집하고 분석하여 확률적으로 주가가 상승할 종목을 찾는 알고리즘을 만들고자 합니다. 주식매매 관련하여 데이터분석이 의사결정에 도움을 줄 수 있는 부분은 아래와 같습니다.

> 1. 매수 종목 선정
> 2. 매수 및 매도 시점 결정

이 책에서는 1번을 중점적으로 다루고,. 2 번 문제는 간단하게 논의해 보도록 하겠습니다.
이제 문제점이 파악되었으면 타겟(예측할) 변수의 결정이 가능합니다.

모델링에서 예측하고자하는 변수를 타겟변수(통계모델에서는 종속변수)라고 합니다. 단기매매를 위한 예측모델에서 우리의 타겟변수는 매수 후 1 주일(5 영업일) 동안의 주가변동입니다. 타겟변수를 어떻게 정의하는냐는 예측모델의 성공과 실패를 결정하는 중요한 요소입니다. 여기서 데이터 분석의 경험이 많은 분석가와 신입 분석가의 차이가 발생하게 됩니다.

모델의 선택에 대하여 시계열 모델을 우선적으로 생각해 볼 수 있습니다. 시계열 모델은 (T-n) .. (T-3), (T-2), (T-1) 주가 정보를 이용하여 T 시점의 주가 예측하는 방법입니다. 시계열 모델의 경우는 가격 자체를 예측하기 보다는 가격의 변동성(예를 들면 수익율) 을 시계열 변수로 해야 정상성(Stationary) 를 만족하게 됩니다. 정상성은 통계 시계열 모델에서 특히 중요합니다. 또한 머신러닝 및 딥러닝 시계열 모델 등도 학습이 잘 되려면 정상적을 만족해야 합니다. (시계열 모델에서 왜 정상성을 확보해야 하는 이유에 대하여 별도 기술). 하지만 시계열모델은 과거의 주가 흐름이 미래에도 반복한다는 기본 가정이 있습니다. 문제는 주가흐름은 이 가정을 만족하기가 어렵다는 것입니다. 따라서 시계열모델 접근보다는 과거 주가정보의 특징이 요약된 피쳐를 입력변수로 활용하고 매수 후 일 주일간(5 영업일)의 수익률을 예측하는 모델을 구현할 것입니다.

가상의 익절 수익율(예를 들어 익절 5%)을 설정하고 매도를 합니다. 매도 후, 수익이 발생한 경우는 1, 나머지는 0 인 타겟변수를 생성합니다. 타겟변수의 특이값(예를 들어 10% 이상 익절) 이 모두 1 로 치환되므로 모델의 입력변수가 타겟변수의 변화에 민감하게 반응하지 않게 됩니다. 일반적으로 이진 분류 모델의 경우 타겟변수에 입력변수가 민감에게 반응하지 않기 때문에 모델 적합이 잘됩니다. 이는 오버피팅(과적합)을 피하기 위한 한 방법이기도 합니다. 예측모델링을 공부하신 분들은 잘 아시겠지만, 이진 타겟변수에 대하여 다양한 모델을 활용할 수 있습니다. 이진 분류의 문제는 로지스틱 회귀분석이 대표적인 모델입니다. 요약하면 과거의 주가 및 거래량 정보를 요약하며 피쳐를 만든 후, 피쳐들을 이용하여 수익/손해를 추정하는 모델입니다.

By KHS