# ETF 분석

### 특성 선택 과정 및 결과

텍스트, 스크린샷, 라인, 번호이(가) 표시된 사진

자동 생성된 설명

XGBoost를 통한 변수 중요도 가로 막대 그래프이다.

우선 변수 중요도로 보았을 때 High와 Open이 중요도가 높게 나온다.

두 변수를 스케일링을 통해 값의 범위를 맞춰버리면, 원래 값의 크기에 의존하는 변수의 상대적 중요도가 모델에서 사라질 수 있기 때문이다.

그렇기에 두 변수를 사용한 표준화된 파생변수를 사용할 것이다.

또한, RSI가 중요도가 0으로 나오는 것은 이미 표준화 되어 있기에 영향력이 안나올수 있다.

그리고 RSI는 70을 넘으면 매도, 30 밑으로 떨어지면 매수 포지션을 취하기에 주식지표에서 중요하다고 생각하는 변수이기에 학습 시 RSI를 사용할 것이다.

Change(변동률)도 이전 거래일 종가 대비 당일 종가의 변화 비율이기에 전체적으로 변수 크기를 보았을 때 크지 않을 것이라고 판단이되고 어느정도 표준화가된 값이라 판단하기에 Change도 학습 변수로 선정

텍스트, 스크린샷, 메뉴, 폰트이(가) 표시된 사진

자동 생성된 설명

볼린저 밴드, MACD, Signal\_line, 스토오토캐스틱,ATR 변수를 파생하였고

표준화 되어있는 지표는 스토오토캐스틱이며 이 변수를 활용할 것이고 나머지 변수는 MACD와 Signal\_Line은 매수,매도 신호를 알 수 있는 중요한 지표라 판단하여 학습 시 사용 나머지 변수는 추가하 성능이 올라가면 판단하겠다.

* 파생변수 학습 및 중요도 파악
* 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  자동 생성된 설명
* 텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

  자동 생성된 설명

전체 파생 변수를 학습한 결과 1의 클래스가 적음에도 불고하고 0.5나온다.

또한 ATR, 볼린져 밴드, Signal\_Line의 변수 중요도는 낮기에 제외하고 학습

* Change, RSI, MACD, D, Close 변수로 학습한 지표

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

1의 클래스가 99개 인데도 precision이 0.83이 나오기에 우수하다고 판단하고 이 변수들로 모델링 변수 선정

1. 윈도우 길이 조정 과정 및 최적 값 선택 이유

### 윈도우 길이 조정 과정 및 최적 값 선택 이유

1. 각 파생변수의 윈도우 조정 이유

볼린저 밴드: 볼린저 밴드는 20일 이동평균과 표준편차를 기반으로 하므로, 주가의 평균적인 추세와 변동성을 파악하는 데 유용합니다. 이 윈도우 길이는 주가의 일반적인 패턴을 포착하는 데 적합합니다.

MACD: MACD는 단기와 장기의 지수 이동평균(EMA) 차이를 기반으로 하여 추세 반전을 예측합니다. 일반적으로 12일과 26일을 사용하는 것이 일반적이며, 이는 단기적인 가격 변화와 장기적인 추세를 비교하는 데 효과적입니다. signal\_window로 9일을 사용하여 MACD의 신호선을 생성하여 더 부드러운 신호를 제공합니다.

스토캐스틱 오실레이터: 스토캐스틱 오실레이터는 14일을 기준으로 하여 주가의 상대적인 위치를 평가합니다. 14일은 주간 시장 패턴을 이해하는 데 적합하며, 이를 통해 과매수 또는 과매도 상태를 확인할 수 있습니다.

ATR: Average True Range (ATR)는 변동성을 측정하는 지표로, 보통 14일의 이동 평균을 사용합니다. 14일은 시장의 변동성을 적절히 평가하는 데 필요한 과거 데이터를 제공합니다.

2. **최적 값 선택 이유**

성능 최적화, 50일, 100일, 150일로 테스트하여 precision 값을 비교하니

Windows =50 일때

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

100일 때

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

150일 때

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

Windows가 100과 150일 때 차이가 거의 없으므로 학습을 고려하여 windows를 100으로 선정

* 윈도우 길이 조정 이유

시계열 데이터의 특성: 주식 시장과 같은 금융 데이터는 연속적인 시간에 따른 패턴과 트렌드가 존재합니다. 따라서, 모델이 이전의 가격 행동을 학습할 수 있도록 충분한 양의 데이터를 제공해야 합니다. 윈도우 길이를 100으로 설정한 것은 과거 100일의 데이터를 바탕으로 다음 날의 가격 변화를 예측하는 데 필요한 충분한 정보가 포함되어 있다고 판단했기 때문입니다.

단기 및 중기 패턴 반영: 100일이라는 길이는 단기적인 변동성과 중기적인 트렌드를 동시에 반영할 수 있는 적절한 길이입니다. 이렇게 하면 특정한 시장 영향을 보다 잘 이해하고 모델에 반영할 수 있습니다.

### 모델 선택 및 하이퍼파라미터 조정

클래스가 불균형하여 XGBoost를 사용하여 선택

AutoML을 사용하여 판단하려 했으나 데이터 수 가 많아서 그런지 30%만 사용하여 돌려봤음에도 오래걸려 시간상 XGBoost로 선정

Optuna를 활용하여 최적의 하이퍼파라미터 값

Best parameters: {'max\_depth': 4, 'learning\_rate': 0.2160523826326754, 'n\_estimators': 837, 'min\_child\_weight': 2, 'gamma': 0.6788404324817511, 'subsample': 0.9242505996301495, 'colsample\_bytree': 0.876425335155132, 'reg\_alpha': 0.9069243453352921, 'reg\_lambda': 0.5586479973596679, 'random\_state': 857}

### 성능 비교 (조정 전후의 성능 차이 설명)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

조정 전 1의 precision이 0.71에서 0.83으로 상승

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

하이퍼 파라미터 조정후 recall 값도 상승

또한 n\_estimators 최적 값이 872였는데 오래걸려서 300으로 조정

### 최종 결론: 가장 좋은 성능을 보인 모델과 그 이유에 대한 설명

표준화된 파생변수를 활용하여 학습하고 최적의 windows와 하이퍼파라미터를 찾아 precision과 recall을 상승할 수 있었다.