ChatGPT

# 모델 엔진 성과 및 데이터 정리

최근 10년치 주가 데이터를 활용한 **LightGBM 모델**을 개발하여 **종목 추천 및 수익률 예측**에 적용했습니다. 이 중 **최근 1년은 테스트**(검증) 데이터로 남겨두고, **이전 9년간의 데이터**로 모델을 훈련했습니다. 여러 시뮬레이션 중 약 20여 가지의 설정을 비교한 결과, **가장 높은 성과를 낸 모델 세팅**을 찾았습니다. 이 모델은 **5일간의 향후 수익 여부를 예측**하도록 구성되었으며, 해당 조건에서 **평균 약 3.2%의 수익률**을 기록했습니다. 참고로 동기간 **시장 평균 자연 상승률**은 약 **0.45%** 정도였으므로, 모델이 **시장 대비 우수한 성과**를 보였다고 볼 수 있습니다.

## 사용된 주요 지표 (12가지)

현재 모델에는 **총 12개의 기술적 지표**가 활용되고 있습니다. 이들은 **가격 추세와 모멘텀**을 파악하는 지표 8가지와, **변동성과 시장 대비 성과**를 보는 지표 4가지로 구성됩니다. 각 지표의 의미와 역할은 아래와 같습니다.

### 1. 추세 및 모멘텀 분석 지표 (8가지)

이 범주의 지표들은 **주가 움직임의 방향성**과 **과열/침체 여부**를 판단하는 데 사용됩니다.

- **SMA_20 (단순 이동평균 20)**: 최근 20일간의 주가 평균을 나타냅니다. **단기 추세**를 파악하여 주가의 단기 움직임이 상승세인지 하락세인지 확인합니다.
- **SMA_60 (단순 이동평균 60)**: 최근 60일간의 주가 평균으로, **중기 추세**를 보여줍니다. SMA_20과 함께 보면 **단기 vs 중기 추세 비교**가 가능하며, 두 평균선의 교차를 통해 추세 전환점을 짐작할 수 있습니다.
- **MACD (이동평균 수렴/확산 지표)**: 단기 이동평균과 장기 이동평균 간의 차이를 계산한 모멘텀 지표입니다. **추세의 강도**와 **추세 전환 가능성**을 판단하는 데 활용됩니다. MACD 값이 양수/음수로 변화하거나 증가/감소하는 추이를 통해 **상승 추세 지속 여부** 등을 평가합니다.
- **MACD 시그널 (MACD Signal)**: MACD의 9일 이동평균선입니다. 일반적으로 MACD 선과 이 시그널 선의 **교차 시점**을 매매 신호로 사용합니다 (예: MACD선이 시그널을 상향 돌파하면 매수 신호). 모델에서는 MACD와 함께 입력되어 **추세 전환 신호 포착**에 기여합니다.
- **RSI_14 (상대강도지수 14)**: 14일간의 상승 폭과 하락 폭을 비교하여 **주가의 과매수/과매도 상태**를 나타내는 모멘텀 지표입니다. RSI 값이 높으면(일반적으로 70 이상) **과매수**로 투자 심리가 과열되었음을, 낮으면(30 이하) **과매도**로 침체되었음을 의미합니다.
- **스토캐스틱 %K (STOCH_K)**: 최근 예를 들어 14일간의 **가격 범위 내에서 현재 종가의 위치**를 백분율로 나타낸 값입니다. %K 값이 높으면 최근 범위 중 **상단 근처 가격**, 낮으면 **하단 근처 가격**임을 뜻해 **단기 모멘텀**을 파악할 수 있습니다.
- **스토캐스틱 %D (STOCH_D)**: %K의 3일 이동평균선으로 **스토캐스틱 신호선**에 해당합니다. %K와 %D의 **교차**를 매매 결정에 활용하며, 단독 %K보다 **신뢰도를 높인 신호**를 제공합니다. (예: %K가 %D를 상향 돌파하면 매수 모멘텀 신호)
- **CCI_20 (상품채널지수 20)**: 20일간의 평균 가격 대비 현재 가격의 **편차를 표준화**한 지표입니다. +100 이상 높으면 **상승 추세의 과열** 또는 **강한 상승 모멘텀**, −100 이하로 낮으면 **과도한 하락 (침체)** 상태를 나타냅니다. 추세의 **과열 여부**나 **반전 가능성**을 판단하는 보조지표로 사용됩니다.

### 2. 변동성 및 시장 비교 지표 (4가지)

이 범주의 지표들은 **주가 변동성의 크기**와 **시장 대비 종목의 성과**를 측정하여 **위험 수준과 상대적 강도**를 평가합니다.

- **ATR_14 (평균 진폭 14)**: 최근 14일간의 **평균 실제 변동폭**을 나타냅니다. 하루하루의 고가~저가 변동 범위를 평균낸 지표로, 값이 크면 **주가가 크게 출렁이는 상태(높은 변동성)**임을 의미합니다. 이는 투자 시 **리스크 수준** 판단에 활용됩니다.

- **BB%_20 (볼린저 밴드 %B 20)**: 주가가 20일 볼린저 밴드 구간 내에서 **상단 밴드(1)와 하단 밴드(0) 중 어디쯤 위치하는지**를 백분율로 표현한 지표입니다. %B 값이 1에 가까우면 상단 밴드 근처 (상대적으로 **고점 영역**), 0에 가까우면 하단 밴드 근처 (상대적으로 **저점 영역**)에 있음을 뜻합니다. 이를 통해 **가격의 상대적 위치**와 **밴드 이탈 가능성**을 파악합니다.
- **거래량 SMA_20 (Volume SMA 20)**: 최근 20일간의 **거래량 평균**입니다. 주가와 함께 거래량 추이를 보면 **수급의 힘**을 알 수 있는데, 이 지표는 **거래량이 꾸준히 증가 추세인지 감소 추세인지**를 나타냅니다. 급등락 시 거래량이 평균 대비 많이 증가하면 **추세의 신뢰도**가 높다고 판단할 수 있습니다.
- **알파 SMA_20 (Alpha SMA 20)**: 개별 종목의 **20일간 평균 초과수익률**을 뜻합니다. 즉, 해당 종목의 수익률에서 같은 기간 **시장 지수(KOSPI)의 수익률을 뺀 값**을 20일 평균낸 것입니다. 양의 알파가 높으면 시장 대비 **초과성과**를 내고 있는 종목으로 **상대적 강세**를 보인다는 의미입니다. 반대로 음의 알파이면 최근 시장보다 **부진한 종목**임을 나타냅니다. 이 지표는 **종목의 진정한 실력**을 측정한다고 볼 수 있으며, 향후 **시장 대비 얼마나 더 오를 잠재력**이 있는지 판단하는데 도움을 줍니다.

## 추가 고려 중인 지표 (기본적 가치 및 수급, 4가지)

현재 모델에는 포함되지 않았지만 **추가 투입을 계획**하고 있는 지표들이 있습니다. **기업의 가치지표** 2가지와 **투자 주체의 수급지표** 2가지로, **종목의 내재가치와 자금 흐름**을 파악하기 위한 데이터입니다. 이들은 장기적인 **펀더멘털 강도**와 **수급 상황**을 반영하여 기술적 지표만으로는 포착하기 어려운 측면을 보완할 것으로 기대됩니다:

- **PBR (주가순자산비율)**: 주가를 주당순자산(BPS)으로 나눈 비율로, **저평가/고평가 여부**를 가늠하는 대표적인 가치 지표입니다. 일반적으로 1배 미만이면 **자산 대비 주가가 낮아 저평가**되었다고 보고, 지나치게 높으면 고평가로 판단합니다. PBR을 통해 **기업의 자산 가치 대비 현재 주가 수준**을 파악할 수 있습니다.
- **PER (주가수익비율)**: 주가를 주당순이익(EPS)으로 나눈 값으로, **기업 이익 대비 주가 수준**을 나타내는 지표입니다. PER이 낮으면 **이익 대비 주가가 저렴**하다는 뜻이고, 높으면 **이익 대비 주가가 높다(기대감 반영)**는 뜻입니다. PER 지표는 **성장주/가치주 판단**이나 업종 평균과의 비교를 통해 **투자 매력도**를 평가하는 데 활용됩니다.
- **외국인 순매수 (FOR_NET_BUY)**: **외국인 투자자들의 순매수량**입니다 (일정 기간 동안 외국인이 얼마나 순매수했는지). 외국인은 국내 주식 시장에서 **큰 손**으로 여겨지며, 이들의 매수세는 주가 상승에 기여하는 경향이 있습니다. 따라서 특정 종목에 외국인 순매수가 꾸준히 유입되면 **주가 상승 동력**으로 해석할 수 있습니다. 반대로 지속적인 순매도는 수급 측면에서 **악재**로 볼 수 있습니다.
- **기관 순매수 (INS_NET_BUY)**: **국내 기관 투자자들의 순매수량**입니다. 연기금, 자산운용사 등의 기관 투자자 동향을 나타내며, 기관의 매매는 비교적 **안정적이고 정보력에 기반**한 경우가 많습니다. 기관 순매수가 증가하는 종목은 **중장기적으로 긍정적 시그널**로 인식되며, 반대로 기관이 이탈하면 수요 감소로 **주가에 부담**이 될 수 있습니다.

  **현재 이 4가지 지표는 데이터 확보 문제로 모델 훈련에 포함되지 않았습니다.** 10년치의 **긴 기간 데이터를 수집**하기가 어려워서 일단 기술적 지표 12개만으로 모델을 구성했습니다. 다만 향후 데이터가 준비되는 대로 이들을 통합하여 모델을 **업그레이드**하거나, 또는 현재 **모델이 추천한 종목들을 평가**하는 **보조 지표**로 활용할 계획입니다. 펀더멘털 지표와 수급 지표를 함께 고려하면 **종목 선정의 신뢰도**를 높이고, 변동성 장세에서도 **안정적인 종목 발굴**에 도움이 될 것으로 기대하고 있습니다.

## 모델 설정 및 결과 요약

- **알고리즘**: LightGBM (Light Gradient Boosting Machine). 결정 트리를 기반으로 한 부스팅 모델로, **학습 속도가 빠르고 성능이 우수**하여 사용되었습니다.
- **학습 데이터**: 2015년부터 2023년까지 약 **9년간의 일별 데이터**로 모델을 학습시켰습니다. **최신 1년치(2024년)** 데이터는 모델이 보지 않은 채로 남겨 두어 **테스트 검증**에 사용했습니다.
- **예측 목표**: 향후 **5일간의 수익률 조건**을 예측하는 것에 중점을 두었습니다. 구체적으로, 60일간의 과거 데이터를 입력으로 사용하여 **5일 후에 주가가 상승할지 여부** 또는 **특정 수익률 이상 달성할지**를 분류/예측했습니다. 이러한 세팅에서 모델은 **평균 3.2%의 수익률**을 기록한 반면, 같은 조건에서 **시장 자체의 자연 상승률은 약 0.45%**에 그쳤습니다. 이를 통해 모델이 단순 보유 대비 **약 7배 이상의 초과 수익률**을 나타냈습니다. (물론 해당 수익률은 과거 데이터 테스트 결과이며, 실제 투자 환경에서는 달라질 수 있습니다.)

- **시뮬레이션 결과**: 다양한 피처 조합과 파라미터로 약 20개의 모델 케이스를 실험한 결과, **위의 기술적 지표 12개와 60일 기간**을 사용한 설정이 가장 우수한 성과를 냈습니다. LightGBM 모델의 **피처 중요도**를 살펴보면, 일부 모멘텀 지표(MACD, RSI 등)와 변동성 지표(ATR 등)가 상대적으로 높게 나타나 **단기 주가 방향을 결정하는 중요한 요인**임을 시사했습니다. 반면 **비슷한 역할을 하는 지표들**(예: %K와 %D, MACD와 MACD 시그널)은 상호 보완적으로 작용하여 교차 신호 등을 포착하는 데 도움을 준 것으로 판단됩니다.

## 추가 데이터 필요성 및 향후 방향

이번 모델에서 사용한 **12개의 기술적 지표**는 비교적 다양한 측면을 커버하지만, **유사한 속성을 가진 지표들 간의 중복**은 없는지 검토가 필요합니다. 예를 들어 **스토캐스틱 %K와 %D**, **MACD와 시그널선**은 각각 한 쌍의 지표로 서로 연관성이 높습니다. 다만 이러한 지표들은 교차 여부 등 **각기 활용되는 포인트가 다르므로 함께 포함**하는 것이 유의미하다고 보았습니다. 현재로서는 **특별히 의미 없는 지표는 없으며**, 모두 주가 예측에 나름의 정보를 제공하고 있습니다. 다만 **추세 관련 지표**(SMA 계열 등)나 **모멘텀 지표**(RSI, Stochastic 등)들은 일부 상관관계가 있을 수 있으므로, 추후 모델 정교화 단계에서 **피처 중요도 분석**이나 **상관분석**을 통해 **차원이 높은 지표**로 대체하거나 **불필요한 지표를 제거**하는 방안을 고려할 수 있습니다.

추가로, 향후 모델 성능을 높이기 위해 **다른 유형의 데이터**를 보강하는 것도 고려됩니다. 이번에 포함하지 못한 **펀더멘털(PER, PBR)** 지표와 **수급(외국인/기관 매수)** 지표는 그러한 노력의 일환입니다. 이 밖에도 고려해볼 수 있는 자료나 개선 방향은 다음과 같습니다:

- **기타 기술적 지표**: 현재 사용 중인 것 외에 **Williams %R**, **ADX(평균 방향성 지수)**, **모멘텀 지표(예: 12개월 모멘텀)**, **이격도** 등도 잠재적으로 추가 가능합니다. 새로운 지표를 추가할 때에는 기존 지표들과 **정보 중복도를 평가**하여, **추가적인 정보가치**가 높을 경우에만 포함하는 것이 바람직합니다.
- **패턴 및 가격구조 정보**: 캔들패턴(예: 연속 양봉/음봉 일수, 장대양봉 발생 여부)이나 **저항선/지지선 돌파 여부**, **갭 상승/하락 횟수** 등의 **가격 패턴 특징**도 성과에 영향을 줄 수 있습니다. 이러한 특성들은 현재의 지표들로는 완전히 포착되지 않는 **특정 상황에서의 단기 변동**을 설명해줄 수 있습니다.
- **거시 및 섹터 지표**: 금리, 환율, 원자재 가격 등의 **거시경제 지표**나, 해당 종목이 속한 **산업 지수의 동향**도 장기적으로는 고려할 수 있습니다. 특히 펀더멘털 지표는 거시 환경에 따라 해석이 달라질 수 있으므로, 필요한 경우 **시장 전반의 흐름 지표**를 추가해 **맥락을 반영**할 수도 있습니다.
- **데이터 기간 확장 및 업데이트**: 현재 10년치 데이터를 활용했지만, 더 많은 과거 데이터를 확보하거나 **실시간 최신 데이터로 지속 학습(온라인 러닝)**하는 것도 중요합니다. 특히 **펀더멘털/수급 데이터**는 시계열로 축적되면 **장기간의 패턴**을 볼 수 있으므로, 데이터가 쌓이는 대로 주기적으로 모델에 **재훈련 또는 전이학습**을 실시해 성능을 개선해야 합니다.

마지막으로, **모델의 궁극적인 목적은 실제 투자 시 수익률 극대화**입니다. 이를 위해서는 **백테스트**를 통한 전략 검증뿐만 아니라, 실제 시장 환경에서의 **리스크 관리**와 **거래비용 고려**도 필요합니다. 앞으로는 현재의 모델을 기반으로:

- 다양한 **모델 파라미터 튜닝**과 **피처 엔지니어링**을 지속하여 예측 정확도를 높이고,
- 신규 지표(PER/PBR 및 수급 데이터 등)를 통합해 **다각적인 관점**에서 종목을 평가하며,
- 필요시 **모델 앙상블**이나 다른 알고리즘(예: 딥러닝 모델과 혼합)도 검토하면서,

**지속적으로 수익률을 높이는 방향**으로 나아갈 계획입니다. 이러한 발전 과정을 통해 **시장 평균을 능가하는 알파**를 꾸준히 창출하는 **고도화된 투자 모델**을 구축하는 것이 목표입니다.

---