



# SI와 퀀트 투자

▶ 리서치센터 | 2024 11월 7일

Quant Analyst 김동영, CFA [dy76.kim@samsung.com](mailto:dy76.kim@samsung.com)

\* 본 자료는 고객의 투자에 참고가 될 수 있는 각종 정보제공을 목적으로 제작되었습니다.  
본 자료는 합리적인 정보를 바탕으로 작성된 것이지만, 투자권유의 적합성이나 완전성을 보장하지 못합니다.  
따라서 투자 판단의 최종 책임은 투자자 본인에게 있으며, 본 자료는 어떠한 경우라도 법적 책임소재의 증빙으로 사용될 수 없습니다.

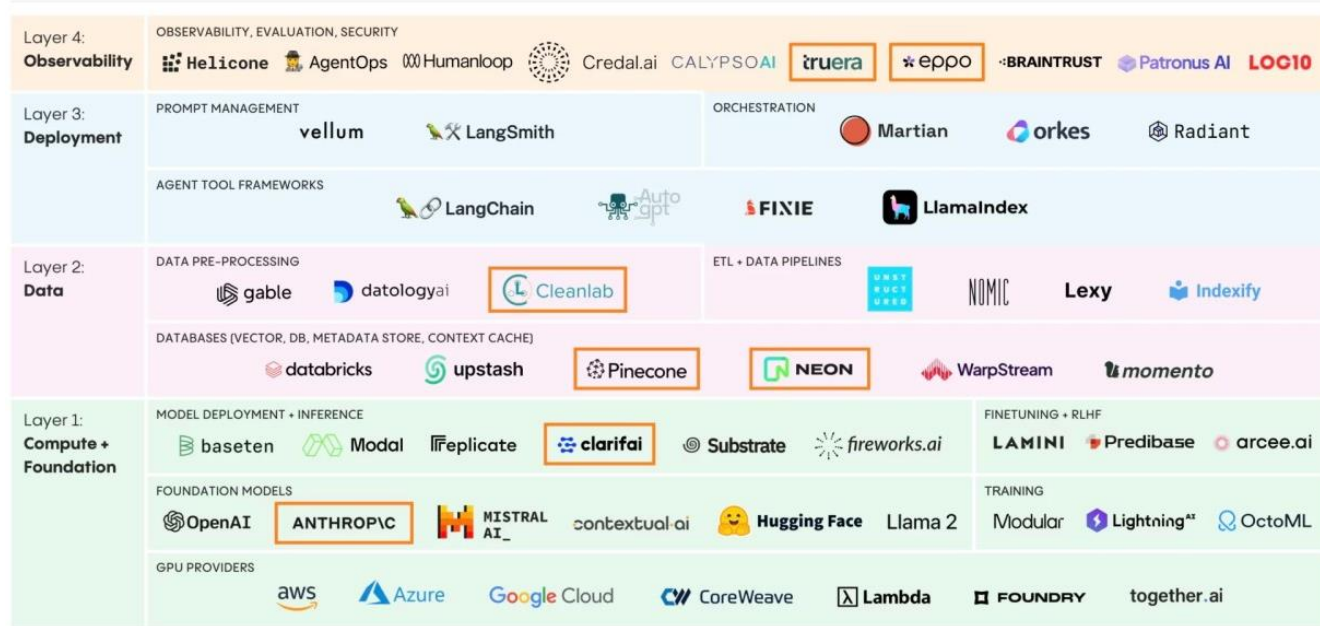
삼성증권

# AI와 퀀트 투자

## AI Impact

ChatGPT의 대성공 이후, LLM(large language model)과 생성형 AI의 서비스와 산업은 폭발적으로 늘어나는 중임. 기존 산업에서의 생산성 향상뿐 아니라, 새로운 산업 수요 또한 촉발되고 있음.

### Modern AI Stack: The Emerging Building Blocks for GenAI



© 2024 Menlo Ventures

Backed by Menlo Ventures

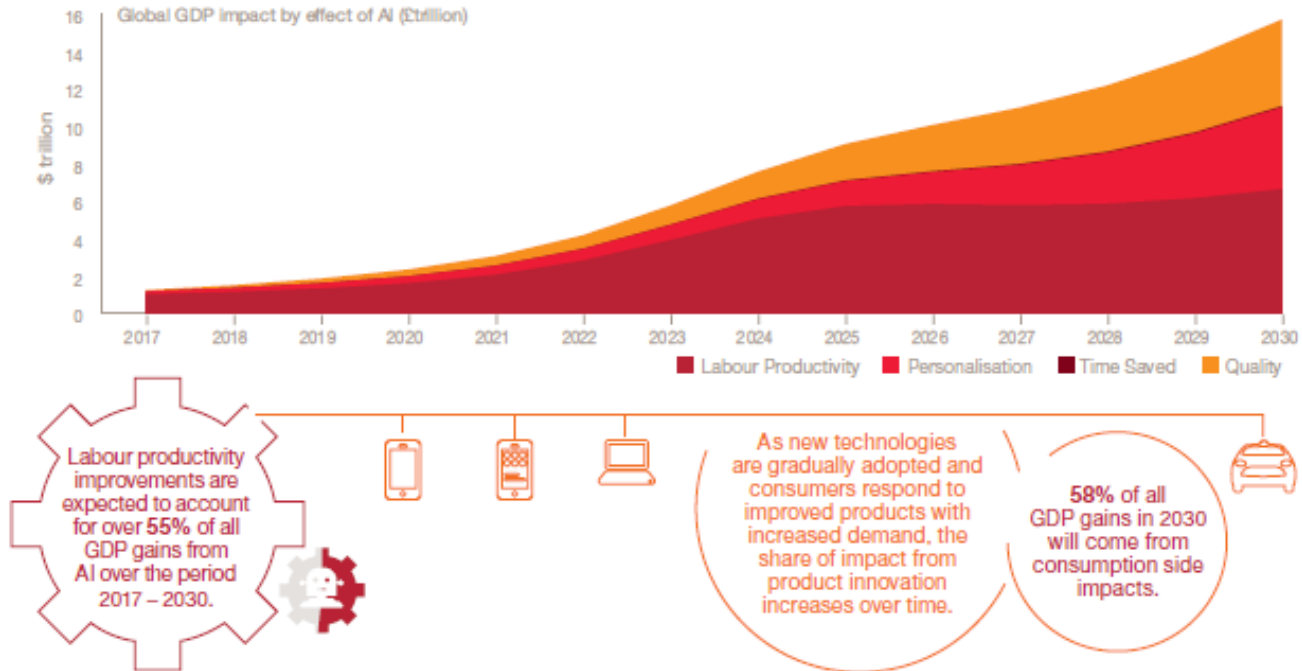
자료: Menlo Ventures

# AI와 퀀트 투자

## AI Impact

- PWC 리포트에 따르면, AI의 도입은 2030년까지 16조 달러의 추가 GDP를 창출할 것으로 전망.
- 이는 100조 달러 규모의 전 세계 GDP 대비 약 15%의 규모에 해당. 중국 규모만큼의 새로운 부가가치가 창출되는 것임.

### Where will the value gains come from with AI?



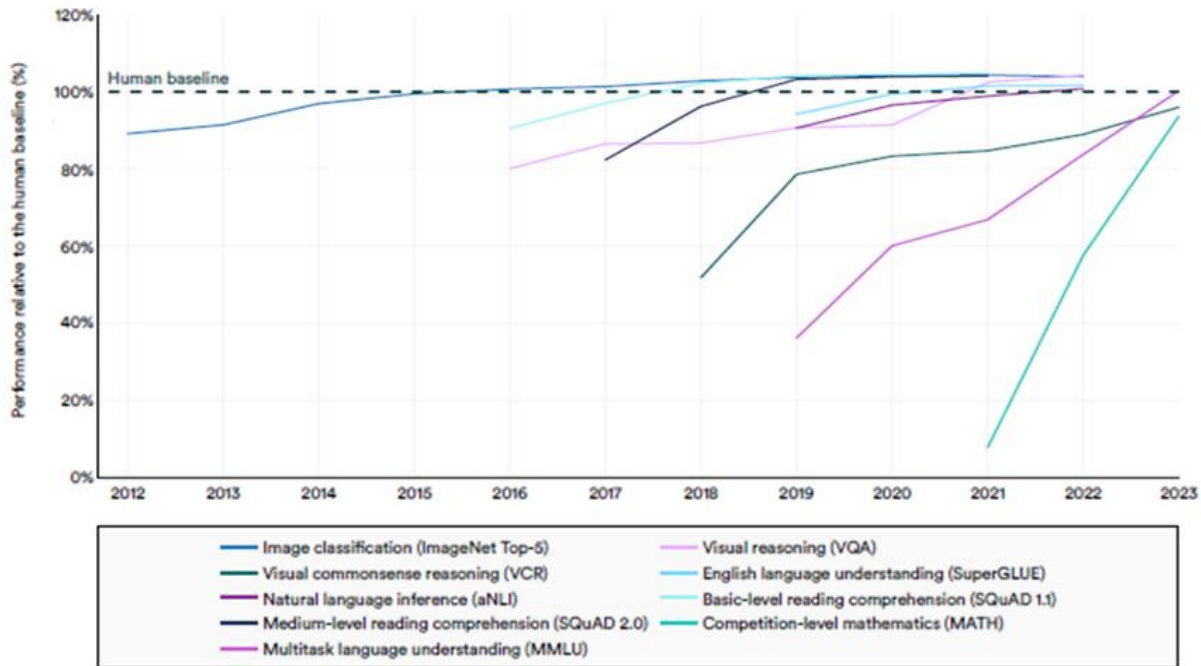
자료: PWC "Sizing the prize"

# AI와 퀀트 투자

## AI Impact

- 이런 AI 충격이 발생하게 된 이유는, 다양한 분야에서 AI가 인간의 능력을 뛰어넘는 시기가 본격적으로 도래했기 때문임.
- (아래 도표) AI는 Image classification(2015년), Basic-level reading comprehension(2017년), Visual reasoning(2020년)에서 인간의 능력을 추월함.
- 다른 고급 인지 능력에서도 AI의 능력치가 빠르게 향상되고 있음.

Select AI Index technical performance benchmarks vs. human performance



자료: Stanford University "AI Index Report 2024"

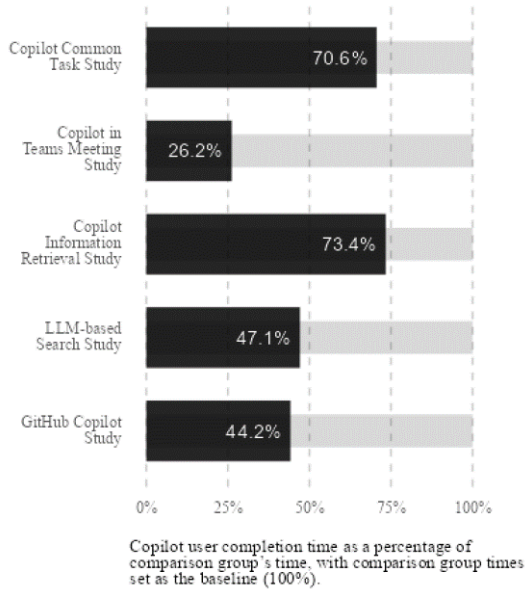
# AI와 퀀트 투자

## AI Impact

- Microsoft의 paper(아래 왼쪽 그래프)에 따르면, MS 코파일럿 및 깃허브 코파일럿을 사용하는 그룹의 작업 소요 시간은, 사용하지 않는 그룹의 작업 소요 시간 대비 26.2%에서 73.4% 수준의 낮은 비율을 기록함. 즉, AI의 보조를 받을 경우 작업 속도가 비약적으로 증가하는 것이 확인됨.
- AI Index Report(아래 오른쪽 그래프)에 따르면, 컨설턴트 대상 조사 결과 GPT-4 사용 시 업무의 질적(quality) 측면에서 40%의 개선이 있었다고 밝힘.

### Task completion speed of Copilot users versus baseline

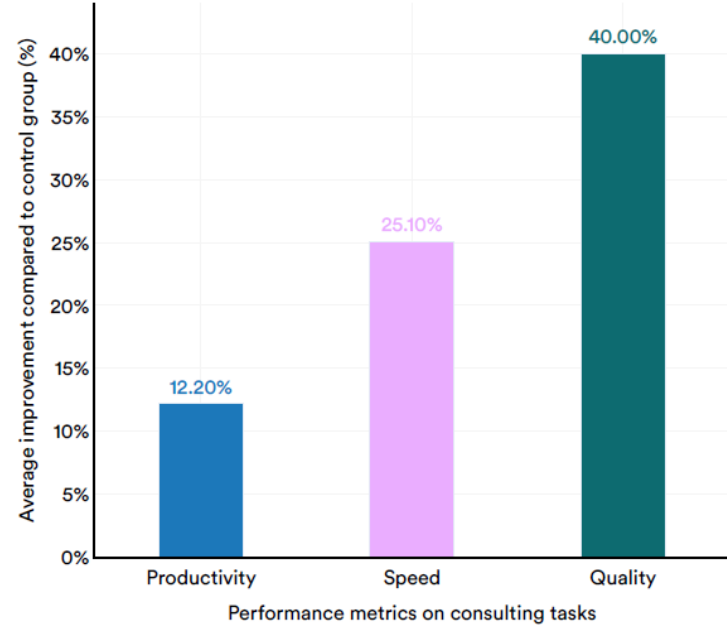
A cross-study comparison shows Copilot users consistently completed tasks more quickly



자료: "Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity" paper

### Effect of GPT-4 use on a group of consultants

Source: Dell'Acqua et al., 2023 | Chart: 2024 AI Index report



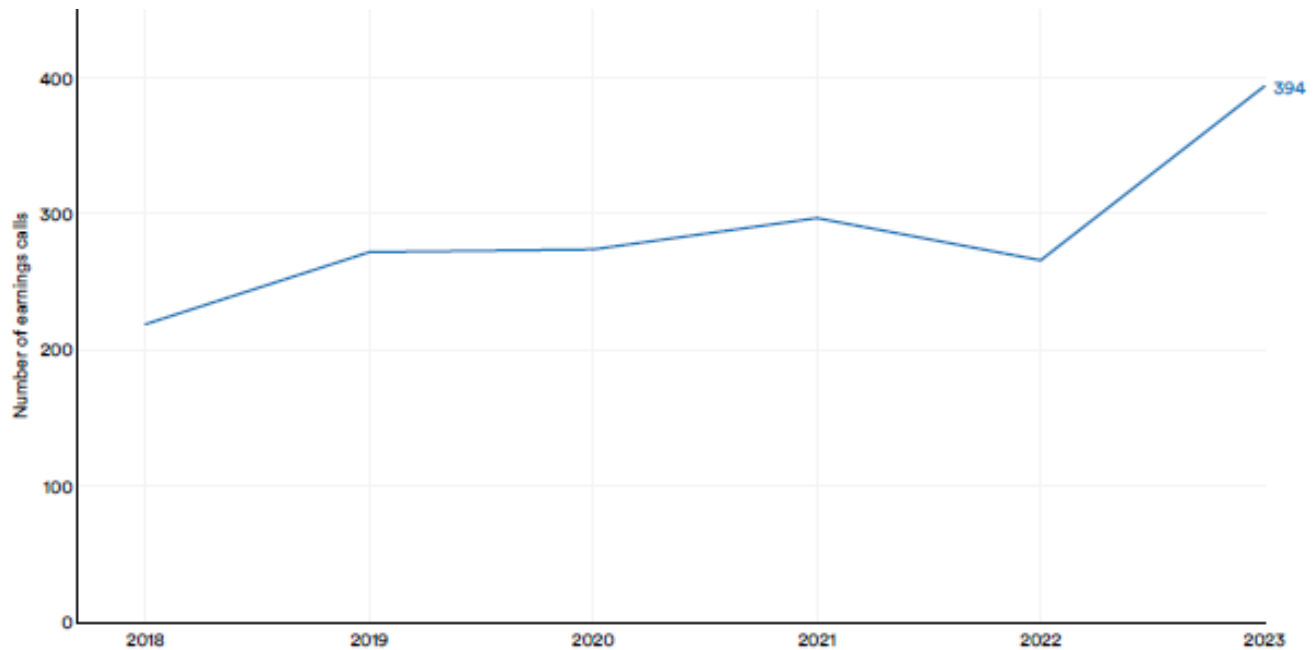
자료: Stanford University "AI Index Report 2024"

## AI와 퀀트 투자

# AI Impact

- 사회 각 분야에서 AI의 활용이 활발해지고 있음. 산업계도 마찬가지임.
- 미국 Fortune 500대 기업의 earnings call에서 “AI”을 언급한 회사의 수는, 2023년에 394개를 돌파함.
- 즉, 이제 대부분의 기업체가 업무에 AI를 접목하고, AI로부터 비즈니스 기회를 찾으려고 하는 상황임.

Number of Fortune 500 earnings calls mentioning AI, 2018-23



자료: Stanford University “AI Index Report 2024”

# AI 모델의 원리 (ChatGPT 안의 수학 및 알고리즘)

ChatGPT에서 사용되는 수학 공식 및 알고리즘에 대해서, 쉬운 수준부터 어려운 수준까지 순서대로 나열해줘

ChatGPT에서 사용되는 수학 공식과 알고리즘을 쉬운 수준에서 어려운 수준까지 나열하면 다음과 같습니다. 이들은 자연어 처리(NLP), 딥러닝 및 머신러닝 모델 학습에 적용되는 수학 및 알고리즘들입니다.

## 1. 기본 수학적 개념

- 덧셈, 뺄셈, 곱셈, 나눗셈: 기본적인 연산으로, 모든 계산의 기초가 되는 수학적 개념입니다.
- 확률 및 통계: 자연어 처리에서 자주 사용되는 개념으로, 특정 이벤트가 발생할 확률을 계산합니다.
  - 예: 사건 A가 일어날 확률  $P(A)$ .

## 2. 벡터 및 선형대수

- 벡터 연산: NLP에서 텍스트를 수치 벡터로 변환하여 계산합니다.
  - 예:  $\mathbf{v} = [v_1, v_2, \dots, v_n]$
- 내적 (Dot Product): 두 벡터 간 유사도를 계산합니다.
  - 예:  $\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i$
- 행렬 곱셈: 다차원 데이터를 처리하는 데 필수적입니다.
  - 예:  $C = AB$

## 3. 미분 및 기울기 계산

- 미분 (Differentiation): 머신러닝 모델에서 손실 함수의 최적화를 위해 사용됩니다.
  - 예: 함수  $f(x) = x^2$ 의 미분은  $f'(x) = 2x$
- 기울기 (Gradient): 기울기를 계산하여 파라미터 업데이트 방향을 결정합니다.
  - 예: 기울기 벡터  $\nabla f(x)$

초중고교 수학

## 4. 확률 이론 및 분포

- 베이즈 정리: NLP 모델이 단어 간의 확률 관계를 계산할 때 사용됩니다.
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- 가우시안 분포 (정규 분포): 많은 데이터가 따르는 분포로, 모델 학습에서 자주 사용됩니다.
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

고교 수학

## 5. 최적화 알고리즘

- 경사 하강법 (Gradient Descent): 모델의 손실을 최소화하기 위해 기울기를 따라 이동합니다.
  - 업데이트 규칙:  $\theta = \theta - \eta \nabla J(\theta)$
  - 여기서  $\eta$ 는 학습률(learning rate),  $\nabla J(\theta)$ 는 손실 함수의 기울기입니다.
- 스토캐스틱 경사 하강법 (Stochastic Gradient Descent, SGD): 데이터 샘플 하나씩 기울기를 계산하여 빠르게 학습합니다.

수치해석 이론

## 6. 신경망 및 딥러닝

- 역전파 알고리즘 (Backpropagation): 신경망에서 오류를 뒤로 전달하여 가중치를 업데이트하는 방법입니다.
  - 기본 아이디어: 각 계층에서의 기울기를 계산하여 가중치를 업데이트.
- 활성화 함수 (Activation Function): 뉴런이 활성화되는 방식을 결정합니다.
  - Sigmoid:  $\sigma(x) = \frac{1}{1+e^{-x}}$
  - ReLU:  $f(x) = \max(0, x)$

딥러닝 기본 알고리즘

## 7. 자연어 처리에서의 기본 모델

- TF-IDF (Term Frequency-Inverse Document Frequency): 텍스트의 중요도를 계산하는 방법입니다.
  - $TF-IDF(t, d) = TF(t, d) \times IDF(t)$
- Word2Vec: 단어를 벡터 공간에 임베딩하는 모델입니다.
  - 스킵그램 모델: 단어의 주변 단어들을 예측합니다.

자연어 처리의 기본 이론

# AI 모델의 원리 (ChatGPT 안의 수학 및 알고리즘)

## 8. 심층 신경망 및 트랜스포머

- 순환 신경망 (RNN): 순차 데이터를 처리하는 신경망입니다. 자연어 처리에 적합한 구조입니다.
  - 문제점: 장기 종속성 문제 (gradient vanishing problem)
- LSTM (Long Short-Term Memory): RNN의 장기 종속성 문제를 해결한 구조입니다.
- 트랜스포머 (Transformer): NLP 모델에서 주로 사용하는 모델로, 병렬화가 가능하여 빠른 학습이 가능합니다.
  - Self-Attention: 입력 문장에서 각 단어의 중요도를 계산하는 방법입니다.
  - 예:  $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$
- BERT (Bidirectional Encoder Representations from Transformers): 트랜스포머를 이용한 사전 학습된 모델로, 문맥의 양방향성을 학습합니다.

문장 생성을 위한, 순차 데이터용 순환 신경망

논리적 문맥 이해를 위한 자연어 알고리즘

## 9. 정교한 최적화 기법

- Adam Optimizer: 경사 하강법의 변형으로, 각 파라미터에 대한 학습률을 적응적으로 조정합니다.
  - 업데이트 규칙:  $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
  - $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
- 레이어 정규화 및 드롭아웃: 과적합을 방지하기 위해 사용되는 기법입니다.

딥러닝 기본 알고리즘

## 10. 최신 연구 및 고급 기법

- Attention Mechanism: 트랜스포머 모델에서 입력 시퀀스의 모든 단어가 서로 어떻게 관련되는지 계산합니다.
- GPT (Generative Pretrained Transformer): 거대한 파라미터를 가진 생성 모델로, 사전 훈련된 데이터를 통해 자연스러운 텍스트 생성을 가능하게 합니다.

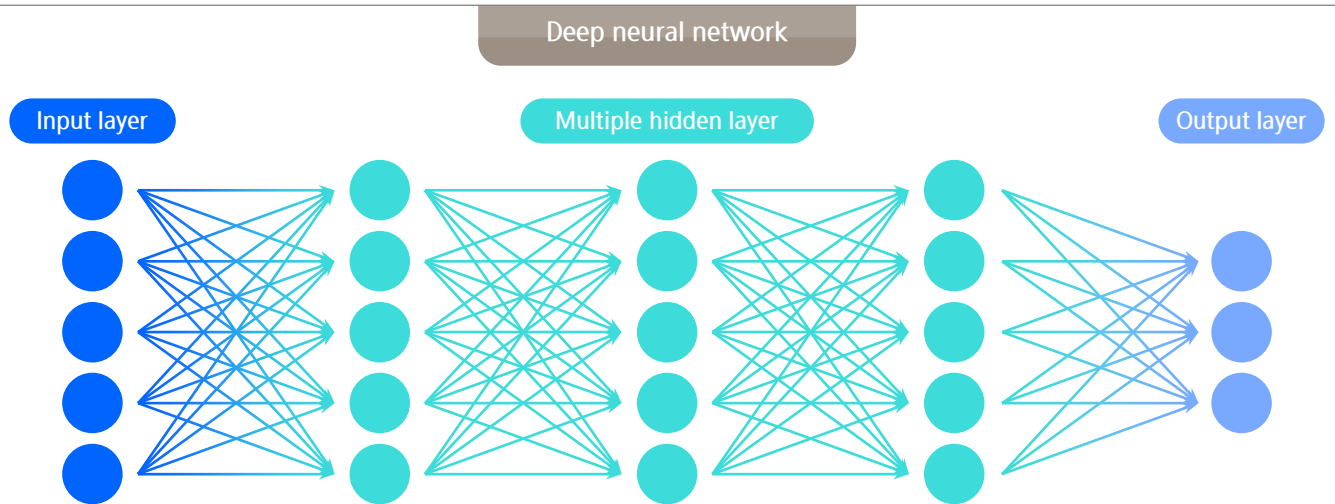
논리적 문맥 이해를 위한 자연어 알고리즘, 거대 파라미터 사용

이러한 수학적 개념과 알고리즘들이 점차 발전하면서 자연어 처리 모델이 더욱 정교해지고 있습니다. ChatGPT의 경우, 트랜스포머 기반의 모델을 사용하여 텍스트의 의미와 맥락을 이해하고 생성하는 데 탁월한 성능을 보여줍니다.

## 딥러닝 알고리즘의 이해 #1

- 딥러닝 혹은 신경망 모델이라고 하는 것은, 뇌에 있는 생물학적 뉴런 네트워크에서 영감을 받은 신경망 구조의 모델을 말함.
- 이 모델은 분류 혹은 예측 작업을 할 때, 입력 데이터가 여러 층을 통과하면서 변환되면서 최종적인 결론 층의 데이터로 변환시키는 일을 함.
- 예를 들어, 사진을 보고 개와 고양이를 구별하는 딥러닝 모델을 만든다고 하면,
  - 1개의 입력 데이터는 사진 한 장의 정보인 [가로 크기 \* 세로 크기 \* 각 점의 RGB 수치 정보]로 이루어짐. 이 크기가 Input layer의 차원 수가 됨.
  - 1개의 결과(출력) 데이터는 개에 해당하는지 고양이에 해당하는지의 정보가 있는 2차원 배열임.
  - 해당 입력 및 결과 데이터의 쌍을, 많은 수의 샘플들을 준비함. 이 샘플들을 사용해서, 인공 신경망 내부의 가중치를 학습시킴.
- 결국, 신경망 모델이 하는 일은, 입력층과 출력층 사이를 은닉층들로 연결한 다음, 그 사이의 행렬 계산에 사용되는 가중치를 학습으로 결정하고, 예측 시에는 그 고정된 가중치로 행렬 연산을 진행하는 것임.

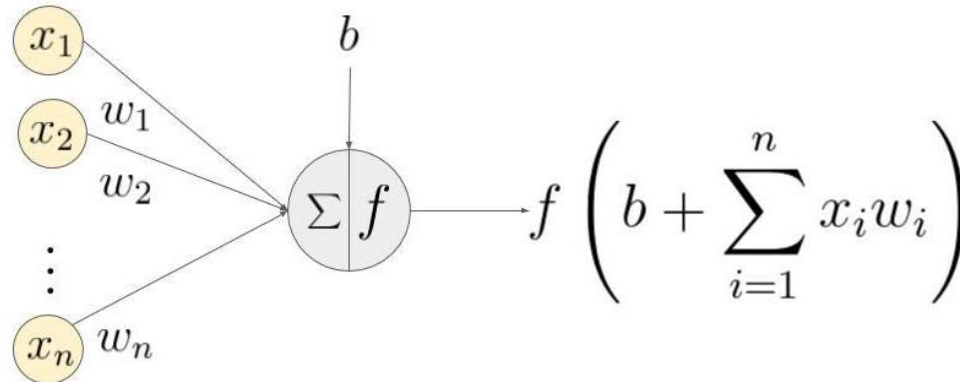
### 딥러닝 - 인공 신경망의 구조



## 딥러닝 알고리즘의 이해 #2

- 인공 신경망에서 한 개 계층 안의 한 개 노드는, 모두 아래의 수식으로 계산됨(수식 도표 참고).
- (완전 연결 신경망 가정 시) 중간 계층의 한 노드 수치는, 1) 직전 계층의 모든 노드값  $[x_1, x_2, \dots, x_n]$ 을 가중치  $[w_1, w_2, \dots, w_n]$ 로 가중 선형 결합한 다음 절편  $b$ 까지 더하고 2) 이 값에 activation 함수인  $f(\cdot)$ 를 한 번 더 적용하여 수치를 만들어 냄.
- 이 작업이 중간층 및 결과층의 모든 노드들에 대해서 모두 적용되는 형태임.
- 즉, 인공 신경망에 저장되는 정보는, 한 계층에서 다음 계층으로 계산 시 사용되는 각 노드별의  $[w_1, w_2, \dots, w_n]$ ,  $b$ 의 정보들의 집합임(행렬 정보).

### 인공 신경망의 수식



An example of a neuron showing the input ( $x_1 - x_n$ ), their corresponding weights ( $w_1 - w_n$ ), a bias ( $b$ ) and the activation function  $f$  applied to the weighted sum of the inputs.

## 딥러닝 알고리즘의 이해 #3

- 이렇게 여러 개의 층을 쌓는 것을 통해서, 현실 세계의 복잡한 문제들을 모두 해결할 수 있는 것이 딥러닝 모델의 장점임.
- 각기 다른 알고리즘과 접근법을 사용하는 단순 머신러닝 모델과 달리, 딥러닝 모델의 기본 구조는 모두 유사한 신경망 형태에서 출발함.  
(예로, 어떤 형태의 분류 문제라고 하더라도, 입력 정보는 입력층에 넣고, 분류 결과는 출력층에 넣고, 사이에는 적절한 은닉층을 넣음으로써 모든 문제를 해결할 수 있음)
- 역전파(back propagation) 알고리즘:
  - 딥러닝 모델이 하는 일은, 입력층의 수치를 행렬 연산을 통해서 은닉층들로 변환시키고 마지막 연산을 통해서 출력층의 수치를 만들어 내는 것임.
  - 딥러닝 모델을 학습시킬 때의 과정은, 미리 확보된 입력층 수치와 출력층 수치 정보를 차례대로 입력한 다음, 이들을 만족시키는 중간 은닉층의 연산 가중치를 반복적으로 찾아가는 과정임. 즉, 이는 단순 회귀 분석 식에서 "기울기와 절편 수치" 를 대입법을 통해 조금씩 찾아가는 것과 큰 틀에서 유사함.
  - 학습을 통해 은닉층의 가중치 결정이 완료되고 나면, 그다음부터는 새로운 입력층 정보를 넣었을 때 모델이 그에 맞는 출력층 정보를 예측하는 방식임.

매트릭스 영화 - 모든 세상을 수치 형태로 인식



## 딥러닝 알고리즘의 이해 #4

- 이때 중요한 것은, 딥러닝 모델 내부에는 항상 선형이 아닌 “비선형 변환”이 사용된다는 점임.  
(구체적으로는 각 노드에서 마지막에 적용되는 활성화 함수 activation function이 비선형 변환에 해당함. 예로, 시그모이드 함수, ReLU 함수 등이 사용됨)
- 만약 딥러닝 모델(심층 신경망)에서 아무리 층을 깊게 쌓더라도, 모든 연산이 “선형”으로만 이뤄진다면, 딥러닝이 잘 작동할 수 없음.
- 왜냐하면, 모든 선형 결합 작업은 계속 중첩을 시키더라도, 결국은 하나의 선형 결합으로 치환이 가능하기 때문임.
- 즉, 선형으로만 생각하는 구조라면, 모든 분석은  $y = b_1x_1 + b_2x_2 \dots + b_nx_n$ 의 선형 방정식을 벗어나지 못하게 됨. 따라서, 사물의 구별과 같은 복잡한 기능을 수행할 수 없음.
- 결국 딥러닝 모델이 고차원적인 일을 할 수 있는 것은, 세상을 선형으로만 보는 것이 아니라, 비선형의 세계로 보기 때문임. 비선형적 사고가 딥러닝 모델 안의 숨겨진 핵심 차별점임.

### 투자예의 함의:

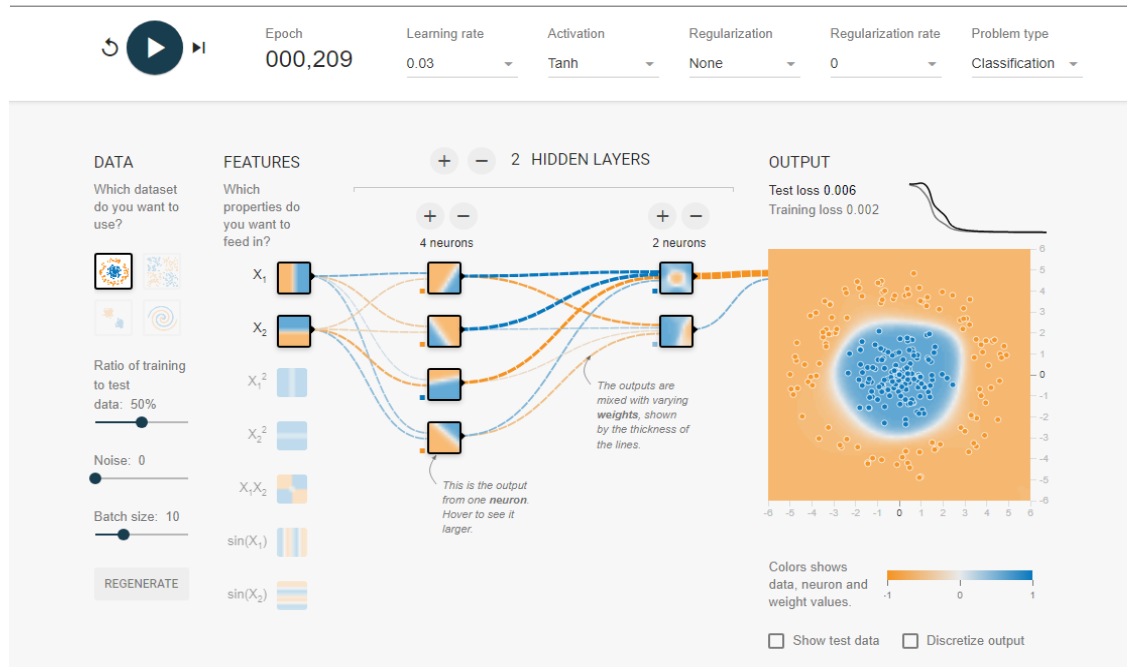
- 전통 경제학에서는 경제 변수의 움직임을 선형으로만 설명했었음. 가격이 내리면 수요가 많아진다, 금리를 올리면 통화가치가 올라간다, 통화가치가 오르면 수출이 감소한다 등의 모든 기존 사고 절차가 선형으로 연결되어 있음(가끔 반비례 함수가 사용되나, 선형 구조와 별반 차이가 없음).
- 이런 선형적인 사고는 복잡한 현대의 경제 구조를 설명하거나 예측하는데 유용하지 않음. 글로벌 경제의 움직임은 실상, 다양한 국면 전환의 트리거 포인트와 비선형 구조로 이루어져 있는 것에 가까움. 이를 단순한 “A가 발생하면 B가 발생하고, B가 발생하면 C가 발생한다”는 식의 3단 논법으로는 설명 및 예측을 할 수 없음.
- 결국, 가능성 높은 예측을 위해서는, 통섭(consilience)의 연결되는 복합적인 사고 및 판단이 필요함. 이때, 딥러닝 모델 같은 AI 기술이 도움을 줄 수 있음. 결국 금융 투자의 세계에서, 딥러닝 및 AI의 기술은 계속 그 중요성이 커질 것으로 예상함.

# 딥러닝 알고리즘의 이해 #5

## 예제:

- 이 예제는 [ $X_1$ ( $x$ 축에 표시),  $X_2$ ( $y$ 축에 표시)] 수치 정보를 가진 샘플들을, 좌표 상 가운데 점들의 집합인 “1번” 그룹과, 외곽 점들의 집합인 “-1번” 그룹으로 이항 분류하는 모델임. 여기서는 “2x4x2x1” 형태의 신경망 모델을 통해서 분류 작업에 성공함.
- 만약에, 신경망 내 계층 간 이동 시 (비선형 activation function이 없는) “선형 결합”만 사용된다면, 이런 복잡한 분류를 할 수 없음.
- 선으로 긋는 방식으로는 원 모양의 안쪽과 바깥쪽을 구별할 수 없기 때문임.

예제: 2x4x2x1 구조의 FNN(Feedforward Neural Network) 모델 (<https://playground.tensorflow.org/>)



## 딥러닝 알고리즘의 이해 #6

### 딥러닝 기법에서 다양성의 의미

- 딥러닝 알고리즘의 주요 주제 중 하나는 가중치 초기화 문제임. 만약에 신경망 은닉층의 가중치가 모두 동일한 0의 값 혹은 1의 값에서 출발한다고 하면, 해 찾기는 성공하지 못하고, 모델이 작동하지 않음.
- 만약 은닉층의 가중치가 모두 완전히 같은 0에서 출발하면, 어떤 사진(input 데이터)을 넣더라도 결과가 전부 0으로만 나오기 때문에, 역전파하더라도 모든 은닉층 노드의 가중치가 같은 값을 가지게 됨. 모든 노드가 같이 움직이기 때문에, 복잡한 규모의 신경망이 아니라, 단순히 1개 노드씩만 있는 직선 형태의 신경망으로 사실상 바뀜. 학습 진행 및 예측이 정상적으로 되지 않음.  
(학습 과정을 “역전파 알고리즘”이라고 함. 이 기법은 2차 함수가 xy 좌표상의 x축을 지나는 위치를 찾기 위해서 일일이 점을 대입에서 확인하는 방식의 “수치 해석” 접근 방법을, 행렬 버전으로 바꾼 정도의 개념에 해당함)
- 비유적으로 설명하면, 딥러닝 내부 은닉층의 각 노드의 역할을 사람이라고 칠 때 “동물의 배 모양만 구별할 수 있는 같은 역할을 하는 사람”만 100명을 데려왔기 때문에, 개와 고양이를 구별 못 하는 상황이라고 할 수 있음.
- 딥러닝에서 가중치 초기화 방식의 구체적인 알고리즘에는, 예를 들어 르쿤 초기화, He 초기화 등이 있음. 이런 초기화가 가지는 기본 원리는, 은닉층의 가중치가 적절히 랜덤하게 분산되어 있어서 향후에 있을 학습이 잘 되도록 준비한다는 개념임. 비유적으로 말하면, 딥러닝 내부에 동물의 눈을 구별하는 판단자, 동물의 코를 구별하는 판단자 등이 잘 성장할 수 있도록 초기 시작을 랜덤하게 만들어진 행렬에서 출발해야 한다는 뜻임.
- 시사점: 이런 개념은, 복잡한 사회 문제를 해결하려고 할 때도, 시사하는 바가 있음. 복잡한 문제를 해결하기 위해서는, 구성원 전체가 100% 일치하는 시각을 가지고 있어서는 그 문제가 해결되지 않음. 여러 구성원이 서로 다른 관점에서 문제를 바라보고, 이를 종합적으로 정리하는 과정이 있어야만, 진정한 문제 해결이 가능해짐.

# 금융, 과학, AI

- 금융 시장 및 금융 산업은, 항상 과학의 발전을 받아들이면서 성장해 왔음.
- AI의 발전이라는 변화를 통해서, 금융도 새롭게 변화할 것으로 기대함.

## 수학 및 과학자가 금융 시장 및 경제학의 발전을 주도 - 대표 사례

연구자	내용
로버트 머튼	컬럼비아 대학에서 Engineering Mathematics 학사. 캘리포니아 공대에서 Science 석사. 그다음에 MIT에서 Economics 박사 학위. 옵션 가격 결정 모델인 블랙-숄즈 모델에 대한 공동 기여로 1997년 노벨 경제학상 공동 수상
피셔 블랙	하버드 대학에서 물리학 전공. 하버드 대학에서 응용 수학 박사 학위. 1973년 마이런 솔츠와 같이 블랙-숄즈 모델의 논문 발표. 골드만 삭스에서 근무하며, 1990년에 자산 배분의 중요 모델인 Black-Litterman 모델을 개발하고 발표함. (95년 사망으로 노벨상 미수상)
케네스 프렌치	리하이 대학에서 기계 공학 학사. 그다음 로체스터 대학에서 MBA, 석사, 재무학 박사 취득. 노벨 수상자인 유진 파마 교수와 함께, CAPM에 이은 자본 시장의 주요 이론인 “Fama-French 3-factor model”을 개발, 발표.
존 내시	카네기 멜론 대학 수학 학사 및 수학 석사. 프린스턴 대학에서 박사 학위. 박사 학위 논문인 “내시 균형” 이론으로 인하여, 44년 후인 1994년에 노벨 경제학상 공동 수상
클라이브 그랜저	노팅엄 대학 수학 학사 및 통계학 박사. Granger causality 이론 개발. “ARCH를 이용한 경제 시계열 분석 방법”으로 로버트 앵글 교수와 함께 2003년에 노벨 경제학상 공동 수상
에드워드 프레스콧	스완스모어대학 수학 학사. 카네기멜론 대학 경제학 박사. Hodrick-Prescott 필터(HP 필터) 모델 개발. 경제학에 대한 공로로 2004년에 노벨 경제학상 공동 수상
제임스 사이먼	MIT 수학 학사. 버클리대 수학 박사. 스토니 브룩 대학 수학과 학과장 역임. 미국 수학 협회 Oswald Veblen Prize 수상자. 미국 국립과학원 회원. 르네상스 테크놀로지 헤지펀드 설립. 가장 성공적인 헤지 펀드 중의 하나로 성장. 2024년 사망 당시, 세계에서 51번째로 부유한 사람으로 추정됨.

## 퀀트 투자란

- 퀀트 투자(Quantitative Investment)는 수학적, 통계적 모델과 데이터를 기반으로 투자 의사 결정을 자동화하거나 체계화하는 투자 방식임. 퀀트(Quant)라는 단어는 quantitative의 줄임말임. 다양한 자산에 대한 투자 분석을 수학과 통계로 수행하는 전문가 또는 그들의 방식을 의미함.
- 퀀트 투자의 특징은 감정에 의존하지 않고, 철저하게 데이터와 분석 모델에 기반해 투자를 한다는 점임.


### World's Top 10 Hedge Funds

순위	헤지펀드	총 운용자산 (십억 달러)	Quant-base 여부
1	Citadel	397	퀀트
2	Bridgewater Associates	172	퀀트
3	AQR Capital Management	133	퀀트
4	D.E. Shaw	120	퀀트
5	Renaissance Technologies	89	퀀트
6	Two Sigma Investments	84	퀀트
7	Elliott Investment Management	70	
8	Farallon Capital Management	43	
9	Man Group Limited	36	
10	Ruffer Investment Company	27	

자료: <https://www.investopedia.com/articles/personal-finance/011515/worlds-top-10-hedge-fund-firms.asp> 2024/10/18 업데이트 기준

# 퀀트 투자에 머신러닝/AI를 접목하기

Quantitative Issue



**김동영, CPA**  
Analyst  
dy76.kim@samsung.com  
02 2020 7839

**목혜진**  
Research Associate  
hyejin.04@samsung.com  
02 2020 7795

삼성증권 SAMSUNG

---

2016. 1. 4

## 기계학습(Machine Learning)과 투자전략

---

빅 데이터 기법: Naive Bayes Classifier의 활용

---

**빅 데이터 그리고 기계학습**

빅 데이터는 최근 산업계의 가장 큰 화두다. 빅 데이터란 기존의 데이터를 넘어서는 대량의 정형/비정형 데이터 집합 및 이런 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다. 빅 데이터를 분석하는 작업은 인간의 처리 능력을 벗어나는 것이므로 기계의 도움이 필요하게 된다. 이 때 사용하는 도구가 바로 기계학습(Machine Learning) 알고리즘이다. 기계학습이란 컴퓨터가 데이터로부터 학습을 한 다음, 이를 바탕으로 데이터 분석 및 예측을 하는 알고리즘이다. 얼굴 자동인식 프로그램 혹은 스팸메일 자동분류 기능 등이 실생활에서 사용되는 기계학습의 사례에 해당한다. 미국 등의 선진국에서는 기계학습 기술을 활용하여 빅 데이터를 분석하는 Data Scientist의 수요가 폭발적으로 늘고 있는 상황이다.

기계학습(Machine Learning)은 주식투자 분야에서도 활용될 수 있다. 주식시장에는 수많은 데이터들이 존재한다. 매일매일 주식시장에 대한 뉴스 기사가 나오며, DART(전자공시시스템)에도 매순간 주식들의 공시정보가 올라온다. 각 주식의 여러 가지 펀더멘털 수치 데이터도 항상 조회할 수 있다. 이런 다양한 데이터들을 Machine Learning을 통해 자동 학습하여 패턴 인식 및 예측에 활용한다면, 자동화된 주식투자 모델을 만들 수 있다. 최근 증권업계에서 이야기되는 로보어드바이저 서비스에서도 기계학습 알고리즘이 일정부분의 역할을 하고 있다.

**Naive Bayes Classifier 소개 및 모델 제시**

나이브 베이즈 분류기는, 기계학습의 한 분야로서, 특정 자료가 여러 가지 속성을 가지고 있을 때 해당 자료를 어느 분류(Class)에 넣어야 할지를 베이즈 정리를 활용하여 판단해주는 알고리즘이다.

본고에서는 나이브 베이즈 분류기를 활용하는 W/L 모델과 BoW 모델을 제시한다. W/L 모델은 기존의 패턴분석 방법론과 유사하게 총괄될 다양한 펀더멘털 지표들 기반으로 주가상승/하락/주가하락예정 총괄을 구분하는 모델이다. BoW 모델은 'Bag of Words' 방법을 사용하여 텍스트 마이닝을 할 수 있는 모델이다.

---

**Naive Bayes Classifier 원리**

$$P(C_k|x_1, \dots, x_n) \propto P(C_k) \prod_{i=1}^n P(x_i|C_k)$$


---

자료: 삼성증권

### (발간된 퀀트 리서치 리포트를 통해서 사례 형태로 설명)

- 이세돌의 알파고 바둑 대결 이벤트(2016년 3월) 이전에 발표했던, 머신러닝 기법에 대한 소개 및 퀀트 투자 접목을 제시한 리포트.
- 전통 머신러닝 알고리즘 중의 하나인 나이브 베이저안 분류기의 원리를 설명.
- 나이브 베이저안 분류기를 팩터 모델에 적용하는 방법과, 텍스트의 감정에 사용하는 방법을 정리함.
- 자연어 처리의 가장 원시적인 방법인 Bag of Words(BoW) 모델을 소개.
- Python 코드를 보여주거나, sklearn 라이브러리를 쓰는 형태가 아닌 직접 함수를 만드는 방식 사용.

# 퀀트 투자에 머신러닝/AI를 접목하기

## Quantitative Issue



광동형, CFA  
Quant Analyst  
dl7@kpmg@samsung.com  
02.2020.7839

홍동훈  
Quant Analyst  
dl6.won@samsung.com  
02.2020.7982

신뢰에 가치로 답하다 삼성증권 SAMSUNG

2019. 11. 18

### 코스피 단기예측 AI 모델

#### 랜덤 포레스트 기법을 활용한 머신러닝 기반 모델

본 자료는 먼저 머신러닝 기법인 결정 트리와 랜덤 포레스트의 이론과 사용방법을 설명한다. 그리고, 랜덤 포레스트 기법을 활용하여 코스피를 단기 예측하는 AI 모델을 소개하고 이에 대한 장점을 제공한다.

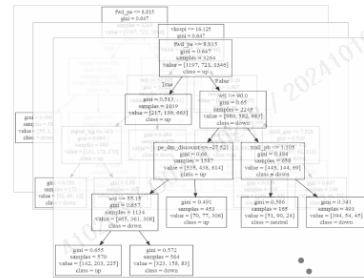
#### 결정 트리 기법

결정 트리(Decision Tree, 의사결정 나무라고도 함)는, 나무 형태의 판단 규칙 조합을 만들어서 기존 데이터를 학습하고 그 다음에 예측하는 알고리즘을 말한다. 쉽게 생각하면 어렸을 적의 '수고고개' 놀이와 같은 개념이다.

#### 랜덤 포레스트 기법

앞서 언급한 '결정 트리' 모델에 보강(bagging) 기법을 적용한 앙상블 모델을, 특별히 '랜덤 포레스트' 모델이라고 한다. 하나의 데이터 소스에서 랜덤하게 데이터를 만들어 여러 개의 학습된 결정 트리 모델을 만들고 이를 종합하는 방식이다.

#### 랜덤 포레스트 개념도



참고: 랜덤 포레스트는 수십 혹은 수백 개의 결정 트리 모델을 만든 다음, 이를 종합해서 판단함  
자료: 삼성증권

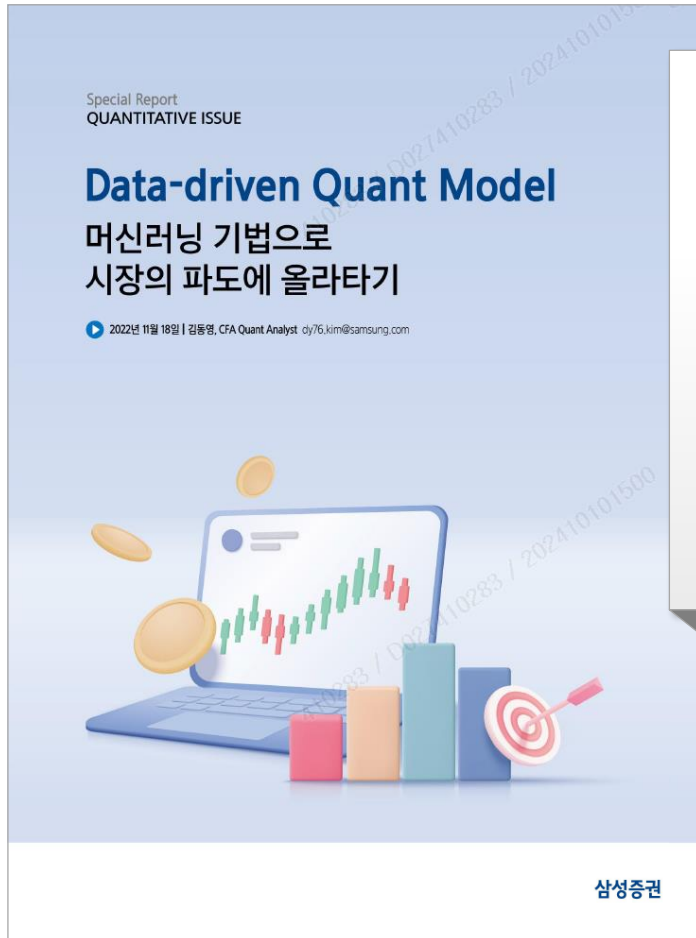
#### 랜덤 포레스트 기반 단기예측 모델

랜덤 포레스트 기법을 사용하여 코스피 시장을 단기예측하는 모델을 개발하였고, 모델 코드를 제시한다. 일반 시계열 데이터와 정교한 모델을 통해, 기존 예측 대비 정확도가 크게 향상된 것으로 확인된다.

• 본 자료는 Python 언어와 sklearn 라이브러리를 기준으로 설명함

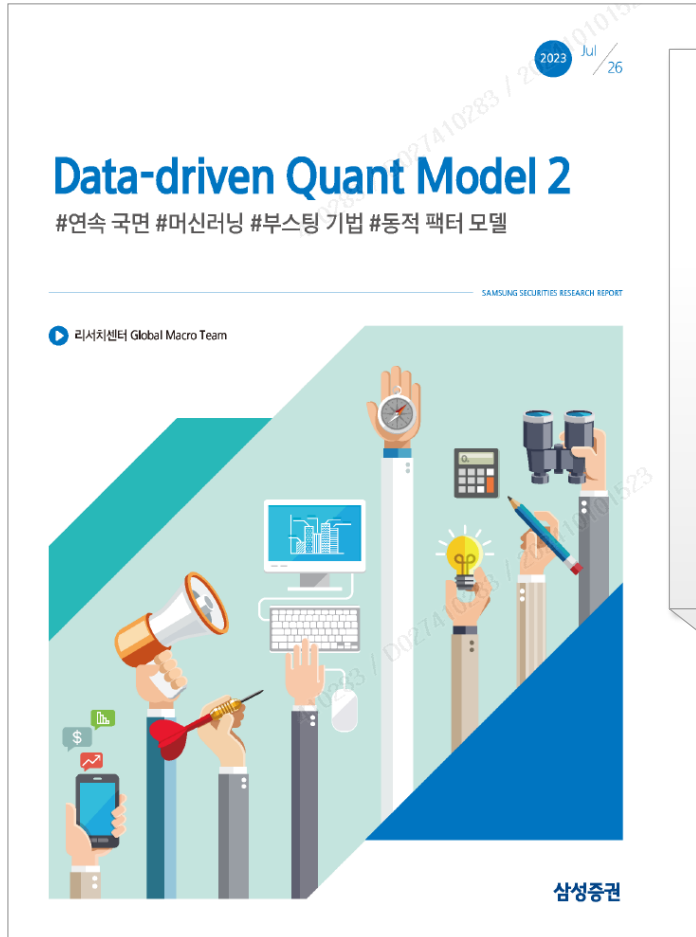
- 전통 머신러닝 알고리즘인 결정 트리 모델과 랜덤 포레스트 알고리즘을 소개하고, 코스피 지수를 단기 예측하는 모델의 예시를 제시함.
- Overfitting, 혼동 행렬, cross-validation 기법 등의 기본적인 머신러닝 방법론을 정리.
- Validation을 진행할 때, 시계열 데이터의 특성을 고려하지 않고 단순하게 접근한 부분은 자료의 취약점에 해당함. 주의 필요.
- Python의 sklearn 표준 라이브러리를 사용한 예측 모델 소스 코드를 전체 공개한 것이 특징임.

# 퀀트 투자에 머신러닝/AI를 접목하기



- 팩터 로테이션 모델에, 머신러닝 기법을 본격적으로 채용하여 재정립한 모델을 제시.
- 모델 프로세스의 모든 단계에, 객관적이고 효율적인 ‘데이터 기반의 접근법’을 사용함.  
(시장 판단을 위한 feature 데이터 선정, 추가 가공을 통한 파생 변수 생성, 동적 전략의 팩터 구성 요소의 선택법 등등)
- 사용하는 머신러닝 기법들에 대한 간략한 내용 소개 - K-Means 클러스터링, Elbow method, Random Forest Classifier 등.
- 단편적인 머신러닝 알고리즘 설명보다는, 머신러닝 기반 퀀트 모델의 전체 구성에 대한 설명을 강조함.

# 퀀트 투자에 머신러닝/AI를 접목하기



- DDQM 모델의 업그레이드 버전. 효율성을 더 높은 DDQM2 모델 소개.
  - 경제/시장 변수 feature를 가지고 부스팅 알고리즘을 통해, 동적 팩터 전략을 결정하는 모델 방식임.
  - 배경지식으로 앙상블 기법, 부스팅, 그래디언트 부스팅, LightGBM 등의 내용을 소개.
  - 시계열 데이터에서 쓸 수 있는 TimeSeriesSplit에 대한 힌트 제공.
  - 각 팩터별로 수익률 예측을 위한 머신러닝 학습기를 각각 운영하는 것이 특징.
  - 당사 리서치에서 사용하고 있는, 가장 최근 기준의 머신러닝 기반 모델임.
- \* 본격적인 딥러닝 / 신경망 기반의 리서치 발간을 아직 못한 것이 아쉬운 점임.

## 손쉬운 출발점: ChatGPT 사용하기

---

### [ 해외 금융기관의 ChatGPT 활용 사례 ]

- Morgan Stanley는 월가 투자은행 최초로 GPT-4 기반의 맞춤형 AI 모델인 AI@Morgan Stanley Assistant를 공개함.
  - 해당 모델을 통해 약 10만 개의 연구보고서 및 관련 문서에 빠르게 액세스하여 최적의 고객 자문 서비스를 제공할 계획
  - 또한 시장 동향과 투자 정보, 내부 프로세스 등 고객의 질문을 AI가 대신 소통함에 따라 고객 서비스 시간을 단축할 것으로 기대
- JP Morgan은 GPT-4를 활용해 연방준비은행의 정책 방향을 예측하는 Hawk-Dove Score를 자체 개발함.
  - 지난 25년간 발표된 연준 성명과 총재 연설 등을 학습시켜 완화부터 억제까지 등급을 매김

### [ 필자의 사용 사례 in ChatGPT / Perplexity ]

“미국 대통령 선거와 주식시장 움직임의 관계에 대한 대표적인 속설을 알려줘.”

“pd.merge로 df1과 df2를 병합할 때, df1에는 s\_date 열이 있고 df2에는 s\_date란 인덱스가 있을 때, 이 s\_date를 기준으로 병합하는 코드를 알려줘.”

“팩터 투자 백테스팅을 가능하게 하는 가장 쉬운 파이썬 라이브러리는?”

“생물 보안법이 뭐야?”

“역대 미국 대통령 선거일을 표로 만들어줘.”

“「트로이 메라이」를 한글로 번역해 줘.”

→ 매일 사용 중에 있음. 자료 검색의 출발점이 “구글 검색”에서 ChatGPT로 옮겨가는 추세임.

# Compliance Notice

---

- 본 조사자료는 당사의 저작물로서 모든 저작권은 당사에게 있습니다.
- 본 조사자료는 당사의 동의 없이 어떠한 경우에도 어떠한 형태로든 복제, 배포, 전송, 변형, 대여할 수 없습니다.
- 본 조사자료에 수록된 내용은 당사 리서치센터가 신뢰할 만한 자료 및 정보로부터 얻어진 것이나, 당사는 그 정확성이나 완전성을 보장할 수 없습니다. 따라서 어떠한 경우에도 본 자료는 고객의 주식투자의 결과에 대한 법적 책임소재에 대한 증빙자료로 사용될 수 없습니다.
- 본 자료에는 외부의 부당한 압력이나 간섭 없이 애널리스트의 의견이 정확히 반영되었습니다.

## 삼성증권

삼성증권주식회사

서울특별시 서초구 서초대로74길 11(삼성전자빌딩)

Tel: 02 2020 8000 / [www.samsungpop.com](http://www.samsungpop.com)

삼성증권 Family Center: 1588 2323

고객 불편사항 접수: 080 911 0900



Member of  
**Dow Jones  
Sustainability Indices**  
Powered by the S&P Global CSA