특집 04

딥러닝 알고리즘의 금융분야 적용 방안

최연지 (이스트스프링자산운용), 이동원 ((주)크래프트테크놀로지스)

목 차

1. 서 론

대용량 연산이 가능해짐과 동시에 딥러닝 알고리즘이 비약적으로 발전함에 따라 최근 들어 학문적 영역을 넘어 다양한 분야에서 딥러닝 알고리즘을 활용한 서비스들이 출시되고 있다. 금융분야에서도 마찬가지로 딥러닝 알고리즘을 활용하여 투자전략을 수립하거나, 자연어 혹은 질적 데이터와 같은 비정형데이터를 객관화하고 수치화하는 연구가 활발히 진행되고 있으며 이에 기반한 상품이나 서비스들이 출시되고 있다. 본고에서는 이러한 동향에 대한 분석과, 나아가 금융분야에서 딥러닝 알고리즘과 빅데이터를 보다 효과적으로 활용할 수 있는 방안에 대해 논하고자 한다.

딥러닝 알고리즘과 빅데이터를 금융 분야에 활용함으로써 얻는 이점으로는 적시성과 효율성을 들 수 있다. 금융분야는 실시간으로 쏟아지는 수많은 데이터를 분석하여 최적의 의사결정을 올바르고 신속하게 내려야하는 특수성이 있다. 예를 들어, 금융자산의 경우 각 자산의 가격은 실시간 호가와 거래량을 통해 형성되어 여타 실물 재화에 비해 일중에도 가격 변동성이 크다. 때문에 의사결정에 수반되는 시간적 비용을 최소화하고 적시에 판단을 내리는 것이 중요하다. 뿐만 아니라 금융분야에서 내려지는 매매 의사결정은 비 가역적이기 때문에 의사결정의 정확성 또한 중요하다.

이러한 특수성으로 인해, 많은 데이터를 신속하고 정확하게 처리할 수 있는 알고리즘이 있다면 판단의 지연을 막고, 오류를 최소화하는데 도움이 될 것으로 보인다. 신속성에 관해서는, 금융시계열 빅데이터를 학습하여 특정 패턴을 인지, 포착한다면 실시간 데이터를 입력하여 즉각적인 예측결과를 얻을 수 있다. 정확성에 관해서는 딥러닝 알고리즘의 발전정도를 확인할 수 있는 대표적 이벤트인 LSVRC (ImageNet Large Scale Visual Recognition Challenge)에서 이미 2015년에 딥러닝 알고리즘 기반 이미지 분류모델이 사람의 정확도인 95%를 넘어선 사건은 잘 알려져 있다. 한편, 금융분야에서 분류, 또는 예측 정확도는 사람의 정확도를 기준으로 판단하기 어려우나 기존에 사용되고 있는 통계학적 방법론들에 비해 높은 정확도를 가지는 경향이 있다. 이는 특히 시계열 데이

터가 비선형성이 강하거나, 편향적일 때 등 통계학적 방법론에서 요구하는 가정들을 충족하지 못할 때 두드러진다.[1]

딥러닝 알고리즘을 금융분야에 활용함으로써 얻는 또 다른 이점으로는 양적인 데이터 뿐만 아니라 질적인 데이터를 활용하여 서비스 제공 범위를 확대할 수 있다는 점이다. 현대 금융 시장은 그 어느때보다 많은 이해관계가 얽혀있으며, 다양한 정보에 따라 요동치고 상호 영향을 준다. 금융은 항상 사회의 수많은 외력에 의해 움직이며, 때문에 금융 시장을 분석하는데 있어 지표화된 정형 데이터를 넘어 다양한 자료와 텍스트를 통하여 시대와 사회의 흐름을 읽는 것이 그 어느 때 보다 중요해졌다.

대표적인 예시로는 과거에 신용 평가, 또는 대출 심사 등에 있어 고객정보 중 수치형 자료가 주로 사용되었다면 최근에는 인구통계학적 자료, 온라인/모바일 소셜 활동 등 비금융, 비정형 데이터까지 활용한 평가모델들이 개발되고 있다. 또 다른 예시로는 대량의 뉴스 기사를 처리하여 매크로 지표를 개발하거나, 챗봇을 개발하여 고객 응대서비스에 이용하는 등 자연어처리를 활용하려는 시도들이 계속되고 있다.

따라서 딥러닝을 적절하게 활용했을 때 금융자산가치를 높이고 또한 의사결정을 내리는데 효과적인 지표를 생성할 수 있다고 보고, 금융분야에서의 딥러닝 활용 방안과 서비스의 발전 방향을 제시하고자 한다.

2. 금융AI 동향

인공지능은 지난 수년간 사회 전반에 걸쳐 큰 변화를 불러일으켰다. AlphaGo로부터 시작된 AI 열풍은 비단 인공지능 자체의 발달 뿐만이 아니라, 사회 전반에 걸쳐 큰 변화를 불러일으켰다. 사회 전반에 걸쳐, AI를 생활에 적용한다는 생각이 자연스럽게 녹아들었으며, 다양한 분야에서 AI를 적용하려는 시도가 활발하게 이루어졌다. 금융 AI 시장 또한 이러한 변화에 발맞추어 지난 수년간 지속적으로 성장해오고 있다. 금융 AI는 데이터의 분석을 통해 기존에 발굴되지 못했던 새로운 이윤을 창출하거나, 소비자의 행동 패턴과 사회적 정보를 바탕으로 소비자 개개인에게 최적화 된 금융 서비스를 제공하는 방식으로 나아가고 있다.

전자는 '알파'로 대표되는 시장 분석을 통한 신규 이윤 창출의 연장선에서, AI 모델을 통해 새로운 이익을 찾아낼 수 있는 팩터를 찾는 연구로 지속되고 있다. 현대 사회에서 개개인이 쏟아내는 금융 데이터의 양은 점점 더 방대해지고 있으며, 금융 시장이 보다 고도화되고 높은 수준의 전산화가 이루어짐에 따라 금융 데이터는 보다 거대해지고 있다. 방대해진 데이터는 인간의 인지 특이점을 넘어, 그 특징을 쉽게 파악할 수 없는 수준에 다다르고 있으며, 이에 대한 해법으로 딥러닝이 제시되고 있다. 다양한 팩터 간의 상호 관계성 분석, 그리고 나아가 성과에 대한 예측과 대안의 제시까지, 점차 그 폭을 넓혀가고 있다.

후자의 경우, 고도화된 산업화 시대가 빅데이터를 통해 개인의 취향과 성향을 파악하고 이를 통해 적합한 서비스를 추천하는 초개인화 시대로 변모함에 따라, 금융 서비스 또한 이전보다 개인의 취향에 맞추어 제공하는 방향으로 발달하고 있다. 소비자들이 금융 상품을 한 데 모아 판단하고 평가하는 것에서 나아가, 금융에 대한 전문적인 지식없이도, 개인의 성향을 반영한 상품을 직접 구성하고, 이를 토대로 개인의 금융투자를 수행할 수 있는 방향으로 나아가고 있다. 특히, 신용 대출과 자산관리 등, 개인의 성향과 특징이 크게 좌우하는 영역에서는 AI 상품의 도입이 적극적으로 검토되고 있다.

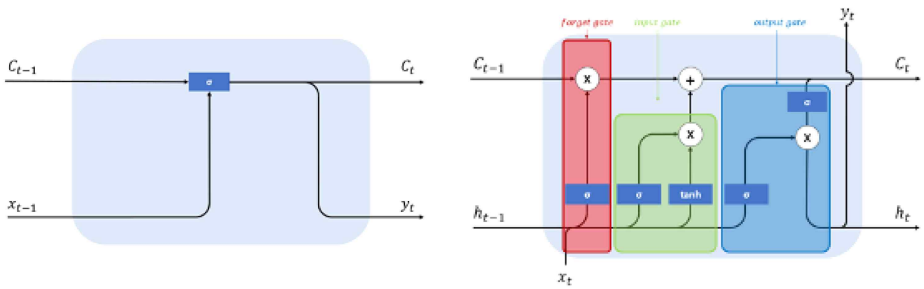
이외에도 시장 분석을 통해 시장의 안정성을 판단하거나, 시장 건전성을 확보하기 위해 이상 거래를 탐지하는 봇(bot) 프로그램을 도입하는 등, AI는 단순하게 추측 가능한 정답을 예측해주는 수준을 넘어, 인간의 인지를 확장시키고 다양한 방법으로 개인의 판단을 보조하고 있다. 그럼에도 현재 시장에서 AI 도입의 가장 큰 걸림돌은 AI 모델 도입에 대한 시장 규제와 인지 가능한 설명이 불가능한 AI 모델의 블랙박스성이다. 하지만 마이데이터 사업의 도입과 함께 시장 규제는 점차 완화되고 있으며, 후자의 경우 '설명가능한 인공지능'(XAI)으로 대표되는 인공지능의 발전 방향성 속에서 해결책을 모색할 수 있을 것으로 보인다.

3. 금융시계열 예측 모델

머신러닝과 딥러닝 등장 이전 시계열의 예측 모델 중 가장 전통적이며, 기초가 되는 모델은 다양한 회귀모델이다. 회귀 모델이란, 데이터의 잔차가 일정한 평균으로 회귀하는 경향을 보인다는 가정을 바탕으로 하는 모델로, 잔차와 여러가지 회귀 방정식을 바탕으로 미래를 예측하는 방법이다. 이러한 회귀 분석 모델 중, 가장 대표적으로는 ARIMA 모델이 있다. ARIMA 모델은 시계열을 분석함에 있어 일정 기간의 과거데이터를 이용하여 미래 시점을 예측하는 모델로, 미래값을 과거 값과 과거에 대한 예측 오차를 통해 설명하는 모델이다. 기존의 금융시계열 예측 연구는 Fama-French의 모델을 중심으로 ARIMA 모델을 수정 및 발전시키는 방향으로 이루어졌다.[2]

3.1 딥러닝을 이용한 시계열 분석 모델 변화

딥러닝의 등장은 이러한 전통적인 예측 모델의 패러다임을 크게 바꾸었다. Neural Network 기반 알고리즘 중 시퀀셜 데이터를 효과적으로 처리하기위해 등장한 RNN(Recurrent Neural Network)은, 입력과 출력을 일련의 입력 단위, 시퀀스로 처리하는 모델이다. 이러한 구조를 통해 각 셀은 이전의 값을 기억하는 메모리 셀의 역할을 하며, 모델은 자연스럽게 데이터의 선·후행 관계를 학습하게 된다. RNN은 장기 의존성에 의한 정보량 소실이 발생한다는 문제가 있었지만 LSTM(Long Short-Term Memory) 모델은 RNN 모델에 논리적인 게이트의 추가를 통해 이러한 문제를 극복하였다. LSTM 모델은 3개의 게이트를 통해 기억된 데이터의 삭제와 보존 여부를 결정하는 방식으로, 모델이 학습을 반복함에 따라 유의미한 정보량이 소실되는 것을 방지한다. 금융 시계열 데이터의 분석에서 가장 어려운 부분은 시계열의 다양한 특성 중에서 정보와 비정보를 식별하거나 구분하는 일이다. 주기성과 추세가 불분명한 금융 데이터의 이용에 있어, 학습 과정에서 유효한 정보량을 보전하는 것이 중요하게 여겨지기 때문에, 일반적으



(그림 1) RNN과 LSTM의 단위 셀 구조

로 RNN 모델에 비해 LSTM 모델이 보다 우수한 성능을 보인다. 아래 (그림 1)은 각각 RNN 모델과 LSTM 모델의 단위 셀의 구조[3]이다.

3.2 시계열 분석 모델 적용 방안

LSTM 모델을 이용한 금융 시계열 분석은 활발히 이루어 지고 있다 있다. “AI의 LSTM기법을 이용한 금융시계열 데이터 변동성 예측방법 연구”(송한진 외, 2018)[4]는 LSTM 모델을 이용하여 변동성을 바탕으로 다양한 지수 예측에 대해 연구했으며, “디노이징 필터와 LSTM을 활용한 KOSPI200 선물지수 예측”(이낙영 외, 2019)[5]은 LSTM 모델을 통해 데이터를 학습함에 있어 디노이징을 통해 모델의 성능을 향상시킬 수 있음을 보여주었다. 이외에도 다양한 연구에서 이러한 형태의 다양한 앙상블 모델을 제시하여 LSTM 모델을 통한 금융시계열 예측에 대한 방안을 제시하고 있다.

이렇듯 90년대 후반 등장하여 다양한 변형 및 개선 모델을 통해 연구가 이루어지고 있는 LSTM 모델 이후에도, 이를 더욱 개선한 GRU(Gate Recurrent Unit) 모델, 경쟁적 학습이라는 개념을 통해 새로운 학습 패러다임을 제시한 GAN(Generative Adversarial Networks) 모델 등, 보다 우수한 성능의 시계열 예측 딥러닝 모델들이 개발되고 있다. 이러한 시계열 예측 모델들에 금융 시계열 데이터를 입력하여 적절하게 학습하기 위해서는 다음과 같은 방안들이 시도될 수 있다. 알고리즘을 통해 분석하려는 금융 시계열 데이터의 특성을 고려하여 모델의 hyperparameter를 최적화하는 별도의 알고리즘을 합성한 앙상블 모델 개발, 또는 데이터셋에 따라 학습 데이터, 검증 데이터, 테스트 데이터의 비중을 최적화 하거나, 유사한 패러다임을 보이는 시기별로 데이터를 분리하여 학습하는 등 데이터셋 자체를 재구축하는 방안이 있을 것이다.

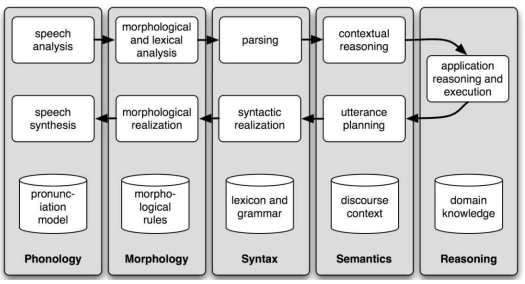
4. 자연어처리를 이용한 금융서비스

자연어 처리(natural language processing)란 인공지능의 한 분야로, 일상 생활에서 사용하는 언어인 자연어(natural language)를 컴퓨터가 처리하는 인터페이스 기능을 수행한다. 자연어 처리 기술에 의해 수행되는 작업의 예시로는 음성 인식, 내용 요약, 사용자의 감성 분석 등이 있다. 자연어 처리 기술이 비약적인 발전을 이룰 수 있었던 것은 딥러닝 알고리즘이 발전하면서 이다. 자연어 처리는 단어 간의 순서 및 상호 정보가 반영된 시퀀셜 데이터(sequential data)를 다루어야 한다는 점에서 머신러닝의 타 분야들에 비해 더딘 발전을 보였다.[6] 하지만 결국 어텐션 메커니즘 등 이러한 문제를 극복하는 딥러닝 알고리즘이 등장하면서 자연어 처리를 활용한 상품과 서비스들이 개발될 수 있었다. 금융분야에서도 이를 활용할 수 있는 방안에 대한 연구가 활발히 진행되고 있으며, 다만 한국어의 언어적 특수성과 경제×금융분야에서 오는 특수성으로 인한 모델의 성능 저하를 개선하는 과제가 남아있다.

4.1 자연어 처리 모델 변화

딥러닝 이전의 기존 자연어 처리의 구조는 (그림 2)과 같았으며 여러가지 모듈로 구성된 복잡한 모델이었기 때문에 구현이 어렵고 하나의 모듈이라도 완벽하게 동작하지 않으면 뒤에 오는 모듈의 오차들이 누적되는 문제가 있었다.[7]

한편, 자연어를 숫자의 나열인 벡터로 전환하여 vector space로 끼워넣는 임베딩을 통해서 단어를 연속적인 벡터를 통해 나타낼 수 있게 됐으며 이러한 벡터화된 데이터를 딥러닝 알고리즘에서 처리하여 end-to-end에 가까운 모델 구현이 가능해



(그림 2) Traditional NLP component stack[7]

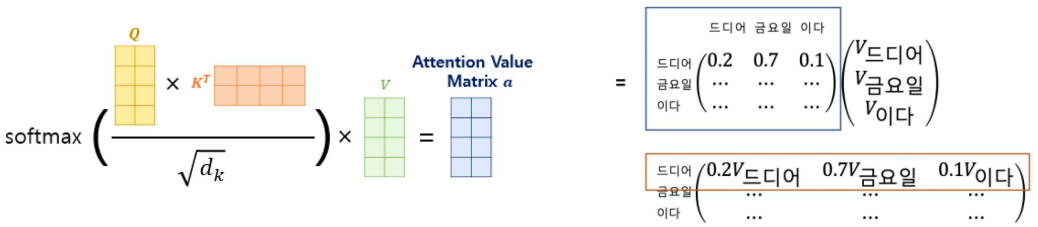
졌다. 앞에서 다뤘듯이 RNN의 단점을 보완한 LSTM이나 Attention Mechanism이 등장하면서 시퀀셜 데이터에 대한 학습이 용이해지면서 자연어 처리분야는 더욱 발전했다.

현재에 와서는 단어 수준의 임베딩에서 더 나아가 문장수준의 임베딩까지 가능해졌으며 문장수준의 임베딩을 수행하는 알고리즘 중 가장 고도화된 알고리즘으로 알려진 것은 Attention Mechanism을 이용한 Transformer Network이다. Transformer Network는 구글 연구팀이 2017년 NIPS (Neural Information Processing Systems) 컨퍼런스에서 공개한 딥러닝 아키텍처로, 이후 발표된 GPT, BERT 등 기법은 이를 기본 단위로 사용한다. 트랜스포머 기본 블록의 계산 과정과 작동 원리를 간단히 살펴보면, 문장을 임베딩할 때 쿼리(Q), 키(K), 값(V) 세 요소 사이의 관계들이 농축된 새로운 행렬을 만드는데서 단어들 사이의 의미적, 문법적 관계를 포착한다.[8] 예컨대 (그림 3)에서 상단의 행렬은 쿼리, 키의 내적을 scaling을 위한 특정 값인 로 나눈 뒤 softmax를 취한 결과이다. 이 행렬의 행은 쿼리단어들에 대응하며, 열은 키 단어들에 대응한다. 이 둘이 같은 구조를 self attention이라 한다. “드디어 금요일”의 값이 0.7로 가장 높은 것을 볼 수 있는데 이는 두 벡터의 내적 값이 크며, 벡터 공간 상에서 가까이 있을 가능성이 높다는 것을 뜻한다. 번역, 분류 등의 자연어 처리 태스크를 수행함에 있어 이처럼 단어들의 관계가 밀접할수록 가중치를 두는 방향으로 학습이 이루어 지는 것이 성능 개선에 중요하다.

4.2 자연어 처리 모델 적용 방안

앞서 소개한 Transformer Network는 이후 BERT, GPT3 등의 발전된 자연어 처리 모델을 구성하는 아키텍처가 되었다. 이들 모델에 금융분야의 말뭉치(corpus)를 input data로 입력하여 학습하려는 시도들이 계속되고 있다. 한편, 이들 모델은 금융과 같은 특수 분야, 혹은 전문 분야의 데이터를 통해 학습 및 검증한 모델이 아니므로 이러한 plain vanilla 모델들을 그대로 적용하여 금융 서비스를 개발하는 것은 적절하지 못할 수 있다. 따라서 도메인의 특성을 가진 말뭉치로 학습을 시킨 모델들이 개발되고 있으며, 의학분야에서는 미국에서 1,800만개의 의학 생명 논문을 학습해 개발한 BERT 기반의 BioBERT 모델이 vanilla BERT보다 좋은 성능을 보인 바 있다.

이후 금융 분야에서도 분야 특화된 모델을 개발하려는 노력이 있었고 그 중 대표적인 모델이



(그림 3) Scaled Dot-Product Attention 예시[8]

FinBERT이다. FinBERT는 기업 공시자료, 증권사 애널리스트 리포트, 금융 뉴스 데이터 등을 수집하여 학습한 모델로서, 금융관련 감성분석 2가지 태스크에 있어 SotA(state-of-the-art)를 달성하며[9] BioBERT와 마찬가지로 기존 BERT모델보다 높은 성능을 보였다. 이에 한국에서도 BERT모델이 가지는 메모리 문제를 극복하고자 한 ALBERT(A Lite BERT) 모델에 금융분야 한국어 말뭉치를 학습하여 챗봇에 도입하려는 시도 등이 이루어지고 있다.[10]

한편, 금융 분야에서 말뭉치를 수집하고 전처리함에 있어 한국어가 교착어로서 지니는 특성에 대해 고려하거나 도메인에 적합하게 전처리하는 기술의 고도화 정도가 향후 자연어 처리를 활용한 금융 서비스를 개발하는데 있어 주요 관건이다. 전처리가 완료된 공개 데이터는 주로 연구, 논문 작성을 위한 용도로 한정되어 있으며, 말뭉치 학습 데이터의 경우 일반 도메인의 말뭉치가 대다수를 차지한다. 이러한 제약사항들로 인해 기존의 성능 좋은 자연어 처리 모델들이 금융 분야의 한국어 말뭉치를 학습하는데 있어서는 좋은 성능을 보이지 않을 수 있다. 따라서 금융 분야에 자연어 처리를 활용하고자 한다면, 공개된 증권사 리포트 또는 기업 공시자료, 공공기관에서 공개하는 뉴스 데이터 등의 데이터를 수집하여 용도에 맞게 전처리할 수 있도록 기존의 존재하는 말뭉치 전처리 모델들을 수정 및 발전시키는 것이 향후 주요 과제일 것이다.

5. 결 론

금융 시장은 어느새 수많은 AI 모델이 치열하게 경쟁하는 무대가 되었다. 주가 변화 추이 예측, 포트폴리오 최적화, 리스크 관리, 신용 평가 등 다양한 금융 서비스에 머신러닝이 사용되고 있으며 특히 인공신경망 기반 딥러닝 알고리즘의 발전이 빠르게 이루어지고 있는 만큼 이를 금융 분야에 활용할 여지가 커지고 있다. 다만 금융 분야에 적용되는 법적, 제도적 특성과 더불어 금융과 경제 분야 데이터의 특성을 고려한 모델을 구축했을 때 앞에서 언급한 이점들을 극대화할 수 있으므로 모델 개발자들은 도메인의 특수성을 이해하고 이를 고려할 수 있어야할 것이다.

우리는 위의 사례들을 통해, 인공지능과 딥러닝 모델이 인간을 대신하여 금융 활동을 수행하는 미래가 머지 않았음을 예상할 수 있다. 그럼에도 불구하고 현재는 시장에 출시된 금융AI 모델들 중에는 여전히 기초적인 수준에 머물러 있거나, 고도화된 모델이라고 해도 고객에게 신뢰를 주지 못하는 경우도 있다. 소비자들을 추가적인 위험에 노출시키는 것을 방지하기 위해서는 제대로 된 학습과 검증이 필수적이다. 또한 인공지능과 딥러닝 알고리즘 모델의 안정성과 효과성을 고객에게 전달하기 위한 노력이 필요하다.

참 고 문 헌

[1] De Prado ML, Advances in financial machine learning, John Wiley & Sons, 2018.

[2] 박재환, “주식수익률 시계열의 ARIMA 모델 설정 및 분석”, 한국증권학회지, 제23권, 제1호, pp.187-210, 1998.

[3] Lee Dong Won, Oh Kyong Joo, “KOSPI200 Prediction through Low-Pass Filtered Long Short-Term Memory Algorithm”, Quantitative Bio-Science, 제39권, 제1호, pp.25-31, 2020.

[4] 송한진, 최흥식, 김선웅, 오수훈, “AI의 LSTM기법을 이용한 금융시계열 데이터 변동성 예측방법 연구”, 한국지식정보기술학

회지 논문지, 제14권, 제6호, pp.665-673, 2019.
[5] 이낙영, 오경주, “디노이징 필터와 LSTM을 활용한 KOSPI200 선물지수 예측”, 한국데이터정보과학회지, 제30권, 제3호, pp.645-654, 2019.
[6] 김기현, 자연어 처리 딥러닝 캠프(파이토치 편), 한빛미디어, 2019.
[7] Gao J, An introduction to deep learning for natural language processing, https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/dl-summer-school-2017.-Jianfeng-Gao.v2.pdf, 2017.
[8] 이기창, 한국어 임베딩, 에이콘, 2019.
[9] Dogu Araci, "Finbert: Financial Sentiment Analysis with Pre-trained Language Models," arXiv preprint arXiv:1908.10063, 2019.
[10] 정한영, “KB국민은행, 금융에 특화된 한글 자연어 학습 모델 'KB 알버트(ALBERT)' 개발했다”, 인공지능 신문, 2020년 6월 18일, https://www.aitimes.kr/news/articleView.html?idxno=16768.

저 자 약 력

최 연 지

이메일 : yeonji.choi@eastspring.com

- 2019년 연세대학교 정치외교학과 (학사)
- 2022년 연세대학교 산업공학과 (석사)
- 2022년~현재 이스트스프링자산운용 Quant Platform & Solution 본부
- 관심분야 : 퀀트 트레이딩, 금융AI, 머신러닝, 자연어처리

이 동 원

이메일 : dongwon.lee@qraftec.com

- 2018년 연세대학교 컴퓨터과학과(학사)
- 2020년 연세대학교 산업공학과(석사)
- 2020년~2021년 NICE P&I
- 2022년~현재 (주)크래프트테크놀로지스 데이터 엔지니어
- 관심분야 : 데이터 엔지니어링, 데이터 아키텍처, 퀀트 트레이딩