졸업논문 최종보고서

딥러닝 기법을 활용한 한정판 스니커즈 리셀 가능 여부 및 리셀 가격 예측

2022년 06월

경희대학교 공과대학

산업경영공학과

조 희 수, 이 수 진, 노 희 지, 김 은 비, 문 선 웅

목 차

국문초록

주요어 : 리셀 시장, 스니커즈, 스니커테크, 심층신경망(DNN), LSTM, 리셀 가능 여부, 리셀 가격 예측

학 번 : 2019100907 조 희 수
2019100889 이 수 진
2019100866 노 희 지
2019100858 김 은 비
2019100869 문 선 웅

최근 스니커즈를 중심으로 한 리셀 시장의 급격한 성장이 이루어지고 있다. 스니커즈는 그 브랜드, 라인, 콜라보 브랜드, 디자이너, 색상 등 다양한 변수에 의해 그 가치와 가격이 결정되며 그 변동 폭이 크다는 특징이 있다. 따라서, 리셀 시장에서의 수익 창출을 위해서는 신발을 구매하기 전 그 신발에 대한 가치 및 리셀 가능 여부 더 나아가 가격 변동을 예측하는 것이 중요하다. 하지만 현재 리셀 시장에 대한 분석 및 리셀 수익성 예측에 관한 연구는 미비한 실정이다. 따라서 본 연구는 딥러닝 기법을 이용한 스니커즈 리셀 가능 여부 예측 및 가격 예측 모델 개발을 제안한다.

구체적인 연구 목표 첫 번째는 스니커즈의 개별적인 특성(브랜드, 라인, 색상, 사이즈 등)을 입력으로 하여 리셀 가능 여부(고수익 가능/수익 창출 가능/ 수익 창출 불가능)를 예측하는 DNN 모델을 구현하는 것이다. 두 번째는 리셀 시장에서의 거래 데이터와 포털 사이트 검색량 데이터, 기온 데이터, 물가지수 데이터, 주가 데이터 등을 결합하여 기존에 시장에서 거래되고 있는 상품의 추후 가격 추이를 예측하du 딥러닝 시계열 예측 모델을 구현하는 것이다. 시장에 처음 출시되는 신제품의 경우 리셀 가능 여부 예측이, 기존 거래 상품의 경우 미래 가격 예측이 소비자들의 리셀 시장 진입 및 수익 창출 기회 마련에 도움을 줄 것이라 기대한다. 또한, 기존 연구에서 사용되지 않았던 딥러닝 기법을 활용함으로써 더 다양한 상품에 대한 예측 및 정확도 향상을 이뤄낼 수 있을 것으로 예상된다.

그 림 목 차

표 목 차

제 1 장 서론

제 1 절 연구 동기와 배경

1.1.1 스니커즈 리셀 시장의 급성장

최근 MZ세대 사이에서 리셀이 새롭게 떠오르는 재테크 방법으로 주목받으면서 스니커즈를 중심으로 한 리셀 시장이 매년 엄청난 성장을 이루고 있다. 이제는 MZ세대는 물론 유명인들까지도 한정판 스니커즈를 구매하기 위해 열을 올리며 리셀은 유행을 넘어 하나의 문화로 자리 잡고 있다.

리셀(resell)이란 명품이나 한정판 제품 등 희소성 있고 인기 있는 상품을 구매한 뒤 다시 비싸게 되파는 행위로, 특히나 패션 분야에서 리셀에 대한 관심이 급증하고 있다. 그중 명품보다 접근성이 높고 마니아 층이 탄탄한 스니커즈 리셀 시장은 그 규모가 더욱 급성장할 것으로 예상된다.[1] 미국 투자 은행 코웬앤드컴퍼니에 따르면 전 세계 스니커즈 리셀 시장은 2019년 20억 달러에서 2025년 약 60억 달러에 이를 전망이다.[2] 이에 따라 국내 대기업들도 스니커즈 리셀 시장에 뛰어들며 시장 규모를 키우는 추세다. 네이버 자회사 스노우가 선보인 '크림'의 경우 2020년 출시 이후 매월 거래액이 평균 121% 성장하고 있으며 1년 만에 누적 거래액 2700억 원, 누적 회원 수 190만 명을 돌파했다. 무신사가 선보인 리셀 플랫폼 '솔드아웃' 또한 출시 2개월 만에 월평균 120%가 넘는 높은 성장세를 기록하였다.

1.1.2 스니커즈 리셀 시장의 높은 수익성

사람들은 왜 이렇게 리셀에 열광할까. 스니커즈 리셀 시장은 주식 시장보다도 높은 변동성을 보이고 있으며, 소비자들은 발매가 10만~20만 원대 스니커즈를 재판매해 적게는 수십만 원, 많게는 몇백만 원까지도 이익을 보는 등 적은 자본으로 고수익을 기대할 수 있다는 점에서 청년층의 관심이 집중되고 있다.[3] 특히 셀럽이 착용하거나 유명 예술가와의 콜라보레이션 혹은 뜻밖의 사건에 의해서도 상품값이 몇십 배 치솟곤 하는데, 그 예로 2019년 출시된 조던과 미국 힙합 가수 트래비스 스캇의

1) 이윤화 기자, "[리셀의 세계]①한정판 되판다고 다 돈이 되는 건 아닙니다", 이데일리, 2020.05.22.
2)
3) 신은빈 기자, "'한정판'딱지 붙으면 몇 배씩 뛴다...MZ세대의 짭짤한 제테크 '리셀'", 매일경제, 2021.12.17.

콜라보 제품 ‘조던 1 레트로 하이 OG SP 모카’ 모델은 발매가 239,000원 대비 약 10배에 달하는 200만 원대의 시세로 현재까지도 활발하게 거래되고 있다. 또한, 불과 몇 달 전 브랜드 오프 화이트의 CEO로 스트릿 패션 신을 이끈 버질아블로가 세상을 떠나면서 그가 제작했던 스니커즈의 판매가가 폭등하는 일도 있었다. 사망 전 약 600만 원대에 거래되던 ‘조던 1 X 오프 화이트 레트로하이 시카고 더텐’ 제품은 사망 소식이 전해진 직후 1000만 원대로 가격이 폭등하며 발매가 약 22만 원의 50배 수준으로 거래되었다.

1.1.3 스니커즈 리셀 시장에 관한 연구의 필요성

이처럼 빠른 성장세를 보이는 스니커즈 리셀 시장이지만, 아직 해당 산업과 관련한 연구는 많이 이루어지고 있지 않다. 특히나 높은 변동성을 보이는 시장인 만큼 소비자들에게 어떤 제품이 높은 리셀가를 형성할 수 있을지, 또 어떤 시점에 리셀가가 높게 형성될지는 매우 중요한 정보일 것이다. 이에 본 연구에서는 신발을 이루고 있는 특성 데이터들을 활용한 제품의 리셀 가능 여부 예측과 기존 시장 거래가, 포털 검색량, 소비자 물가 데이터 등 각종 시계열 데이터를 활용한 스니커즈의 가격 예측 모델을 구현함으로써 소비자들의 리셀 시장 진입을 돕고 수익 창출의 기회 마련에 도움을 주고자 한다. 리셀 가능 여부 예측 모델은 기존 거래 데이터가 존재하지 않는 신제품의 상품 특성을 활용한 리셀 가능성 예측, 리셀 가격 예측 모델은 기존에 리셀 시장에서 활발히 거래되는 기존 상품의 수익성 예측에 높은 활용성을 기대할 수 있다. 더불어 리셀 시장 분석에 기존 연구에서 활용되지 않았던 딥러닝 기법을 도입해 분석을 진행했다는 점에서 탐색적, 탐구적 의의가 있다.

제 2 절 연구 목표

본 연구에서는 딥러닝 모델을 활용한 한정판 상품의 리셀 가능 여부를 예측을 수행하여 마니아층 위주로 이루어져 있는 리셀 시장에 일반 소비자들이 진입 장벽을 낮춰 시장 활성화에 기여하고 리셀 가능 여부 진단 및 가격 예측을 제시함으로써 효율적인 구매 및 투자를 돕는다.

기존 연구에서 주로 활용된 머신러닝 기법이 아닌 딥러닝 기법을 활용하여 예측을 수행하며 스니커즈의 특성에 대한 분석과 주요 변수들의 선별을 통해 예측의 정확도를 더욱 향상시키는 것을 목적으로 한다.

1.2.1 리셀 가능 여부 예측 모델

첫 번째 모델은 발매 예정이거나, 발매된 지 얼마 되지 않아 체결 가격 추이가 존재하지 않는 신제품의 특성 데이터를 사용한 '리셀 가능 여부 예측' 모델이다. 사용 데이터는 브랜드, 콜라보 유무 및 콜라보 브랜드의 유명도, 라인, 색상, 발매가, 발매 일자 등을 사용한다. 먼저 브랜드는 리셀 시장에서 대표적으로 거래되고 있는 나이키, 조던부터 뉴발, 아디다스, 컨버스까지 총 5개의 브랜드 제품이 포함되며, 콜라보 브랜드의 종류는 off-white, Travis Scott, Sacai, Undercover, Supreme, Stussy, Fear of God, Peaceminusone 등을 비롯해 총 50가지이다. 다양한 상품 중 소비자가 느끼기에 리셀 여부의 필요성이 두드러지는 상품군의 예측을 목표로 한다.

본 연구의 수행 단계는 크게 3단계로 구분된다. 먼저 KREAM 플랫폼의 상품 데이터를 크롤링을 통해 수집한 후, 전처리를 진행한다. 두 번째 단계로 회귀모델을 활용하여 다양한 상품 특성이 예측에 미치는 영향력을 분석 및 파악함으로써 입력 변수 선정 지표를 마련한다. 마지막으로 평가 지표를 선정하고 이를 기준으로 예측 모델의 성능 평가하며 지속적인 실험과 개선을 통해 모델의 성능을 개선해나가며 최적의 모델을 구현해내는 것이 본 연구의 최종 목적이다.

1.2.2 리셀 가격 예측 모델

두 번째 모델은 이미 출시되어 시장에 거래되고 있는 제품의 시계열 데이터를 사용한 '가격 예측' 모델이다. 사용 데이터는 해당 제품의 거래일자별 거래가, 포털 검색량, 소비자 물가 지수 등 제품 자체 혹은 사회·경제적 지표에 해당하는 시계열 데이터이다. 시계열 데이터는 일정한 시간 동안 수집된 일련의 순차적으로 정해진 데이터셋의 집합으로 추세 변동, 계절 변동, 순환 변동, 불규칙 변동 등 변화가 빈번하게 일어나 예측에 사용할 때 특별히 주의해야 한다. 다양한 변동들이 존재하기 때문에 결과값에 영향을 미치는 알맞은 변수들을 신중하게 선별하여 예측에 활용하는 것이 필요하다. 따라서 하루 뒤의 리셀 가격을 예측하기 위해 수집한 데이터 중 어떤 변수들이 리셀 가격 예측 유의미한 영향을 미치는지, 어떤 딥러닝 기법을 활용했을 가장 예측력이 뛰어난지를 비교 분석하여 최적의 방법론과 활용 변수를 선택하고자 한다. 우선 지난 가격 데이터의 추세나 변동을 분석하여 하루 뒤 리셀 가격을 예측하는 단일변량 모델 비교를 통해 리셀 가격 예측에 가장 알맞은 모델링 기법을 결정한다. 이후 포털 검색량, 주식 종가, 소비자 물가 데이터 등 다양한 수집 데이터 중 어떤 변수 조합이 가장 뛰어난 예측 성능을 보이는지 비교하여 최적의 입력 변수 조합을 선택한다. 위 과정을 통해 최종적으로 최적의 딥러닝 모델과 활용 변수를 확정하고 결과를 분석하는 것이 본 연구의 목표이다.

제 3 절 기존연구와 그 한계

리셀 시장 특히 스니커즈 리셀 시장이 최근 들어 급성장하기 시작한 만큼 한정판 스니커즈의 리셀 및 리셀 가격 예측에 관한 연구는 아직 많지 않으며 딥러닝 모델을 활용한 연구는 아직 진행된 적이 없다. 일정한 가격이 정해져서 판매되는 다른 상품들과는 달리 리셀 시장은 개인 간 1:1 거래 혹은 중개 플랫폼을 통한 거래가 이루어지며 판매자와 구매자가 제시한 가격이 일치하는 지점에서 거래 가격이 형성된다는 특징 때문에 가격 변동성을 가지고 있다. 따라서 리셀 시장에 관한 기존 연구는 가장 큰 시장을 형성하고 있는 한정판 스니커즈에 대한 시장분석, 리셀 가격과 가격 결정요인 탐색, 수익률 예측 연구가 주로 이루어지고 있다.

1.3.1 스니커즈 리셀(Resell) 현상 분석(2021)[4)]

이 논문에서는 기존 세대와는 다른 MZ세대의 특징이 복합적으로 작용하여 나타난 결과라고 할 수 있는 리셀 시장에 대한 분석을 진행하였다. 스니커즈 리셀은 이제 단순한 판매와 구매 행위를 넘어서 하나의 사회적 현상으로 자리 잡고 있다. 시장 분석을 통해 도출된 특징으로는 첫 번째, 거래되고 있고 다양한 모델 중 스포츠 브랜드를 중심으로 특정 브랜드와 모델이 선호되고 있으며, 유명 브랜드나 유명인과 콜라보한 모델이 큰 인기를 끌고 있다. 두 번째, 리셀 시장에서 거래되고 있는 스니커즈의 상당수가 한정판 모델로서 다양한 방식으로 판매되고 있고, 전체 상품의 평균 가격뿐만 아니라 상당한 비중의 상품이 높은 프리미엄 가격대를 형성하고 있다. 해당 연구는 리셀 시장에 대한 정량적 분석보다는 리셀 문화와 시장 특성에 대한 정성적 분석이 이루어졌다.

1.3.2 머신러닝 기법을 활용한 한정판 운동화 리셀 여부 예측 및 수익성 평가(2020)[5)]

이 논문은 글로벌 리셀 플랫폼 StockX에서 수집한 운동화 특성 데이터를 입력변수로 하여 발매 직후 30일 동안의 해당 운동화 리셀 가능 여부 예측을 핵심 과제로 연구를 진행하였다. 리셀 가능 여부는 리셀 가격에서 StockX의 검수 수수료(리셀 가격의 10%)를 뺀 후의 가격이 그 제품의 발매가보다 크거나 같으면 가능(1)으로

4) 이재영, 「스니커즈 리셀(Resale) 현상 분석(제1보)」, 『패션과 니트』, pp. 67-80(14), 한국니트디자인학회, 2021.
5) 김누리, 「머신 러닝 기법을 활용한 한정판 운동화 리셀 여부 예측 및 수익성 평가」, 서울대학교 석사학위 논문, 2020.

판단 아니면 불가(0)로 판단하였다. 예측을 위해 사용된 입력변수는 오직 신발 자체에 대한 정보이며 브랜드, 라인, 모델명, 콜라보 브랜드, 모델 타입, 사이즈, 주요 색상, 사용된 색상의 개수, 색상, 신발 갑피 재질, 미드솔 재질, 발매일, 발매 계절, 발매가, 빛 반사, 야광, 페인트 오프가 포함되었다. 구체적으로 사용된 방법은 머신러닝 기법 10가지로 로지스틱 회귀분석(Logistic Regression), 릿지(Ridge) 분류, K-인접(K-Nearest Neighbors), 서포트 벡터 머신(Support Vector Machine), 의사결정나무(Decision Tree), 다층 퍼셉트론(Multi-Layer Perceptron)과 앙상블 기법인 배깅(Bagging), 에이다부스트(AdaBoost), 엑스트라나무들(ExtraTrees), 랜덤 포레스트(Random Forest)가 포함된다. 각 모델의 결과를 정확도와 F1 score를 활용해 비교했고 가장 좋은 성능을 보인 모델은 의사결정나무와 앙상블 계열 기법이었으며 그 정확도는 약 80% 수준이다. 그 후 좋은 성능을 보인 모델을 대상으로 예측한 한정판 운동화를 실제 구입했을 경우의 수익률을 분석해보는 것까지 진행된 연구이다.

해당 논문은 다양한 머신러닝 기법을 사용하여 한정판 운동화의 리셀 가능 여부를 예측하고 그 예상 수익률까지 분석하며 의미 있는 결과를 도출하였다. 하지만 절대적인 수집 데이터 양의 부족을 이유로 딥러닝 기법을 활용하지 못했으며, 30일이라는 짧은 기한의 예측만을 진행하였다는 한계점이 있다. 또한, 스니커즈 자체의 특징 변수들이 예측에 미치는 영향력은 따로 고려하지 않고 모두 소거 없이 활용하여 예측을 진행했다는 점도 아쉬운 점이라고 할 수 있다.

1.3.3 XGBoost모형을 활용한 가격 상승 요인 탐색 및 예측을 통한 리셀 시장 진입장벽 해소에 관한 연구(2021)[6)]

이 논문은 리셀 시장의 법적인 문제점이나 행동학적 관점에서 진행된 기존 연구들과는 다르게 리셀 시장의 상품성과 리셀이 되는 이유 등에 초점을 맞춰 수행된 연구이다. 해당 연구 역시 글로벌 리셀 플랫폼인 StockX로부터 스니커즈별 품목명(브랜드, 라인, 제품명), 판매 횟수, 출시 가격, 사이즈, 색상, 콜라보 유뮤, 거래 일시, 거래 가격 데이터를 수집했고, SNS(instagram, Bidu)에서 해당 상품에 대한 게시글과 검색량을 크롤링하여 추가하였다. 수집된 데이터들을 전처리한 후 XGBoost 알고리즘을 활용하여 스니커즈의 리셀 가격이 출시가보다 높아지도록 만드는 요인 분석을 진행하였다. 그 결과 라인의 경우 Jordan1, Air Max, YeezyBoost 순, 색상은 무채색, 비비드 컬러, 블루 순으로 강한 영향력을 발휘한다는 결론이 도출되었다. 실제로 높은 영향력을 가진 속성으로 구성된 'Jordan1 Retro High Dark Mocha'와 'Yeezy Boost 350 V2 Beluga'가 출시가와 비교해 173%, 244%의 프리미엄 리셀가

6) 윤현섭, 강주영, 「XGBoost 모형을 활용한 가격 상승 요인 탐색 및 예측을 통한 리셀 시장 진입 장벽 해소에 관한 연구」, 『한국 전자거래학회지』, pp. 155-174(20), 한국전자거래학회, 2021.

를 형성하고 있는 것을 확인함으로써 해당 분석 결과가 의미 있다는 것을 검증하였다. 그 후 시계열 모형의 일종인 Prophet 모형을 통해 프리미엄 가격을 형성하고 있는 두 가지 상품의 거래 데이터를 활용하여 가격 추이 및 해당 연구를 시행한 2020년 이후인 2023년까지의 가격 예측을 시도하였고 해당 모델의 정확도는 80%에 못 미치는 것으로 나타났다.

해당 연구는 빠르게 성장하고 있는 스니커즈 리셀 시장을 머신러닝 기법을 활용해 분석하여 상품의 가격 상승 요인 탐색과 가격 예측을 진행함으로써 리셀 상품의 정보를 제공하고 소비자 가이드라인을 제시하여 시장 진입 장벽을 낮추는 데 도움을 주었다. 하지만 약 150만 건의 거래 데이터를 활용하면서 너무 많은 요인과 다양한 상품들로 인해 분석의 정확도가 다소 떨어진다는 점과 프리미엄 가격이 형성된 단 2가지 상품에 대한 가격 예측만을 진행하여 모델의 정확도를 평가했다는 점, 2023년까지의 장기간 예측에 대한 정확도는 평가할 수 없었다는 한계점이 존재한다. 또한, 스니커즈의 가격 상승에 영향을 미치는 요인 탐색을 수행했지만, 이는 인기 모델의 공통적 특징 정도의 의미만을 가질 뿐이며 실제 가격 추이 분석 및 예측에는 오직 리셀 거래 데이터만을 사용했다는 점도 한계점이라고 할 수 있다.

1.3.4 한국인 해외관광수요의 예측력 비교: 다변량 모형과 단일변량 모형(2003)[7)]

해당 논문은 1998년과 1999년에 경제위기 극복 과정을 거치면서 내국인 해외여행자 감소가 1인당 지출액 감소 등으로 해외 관광 흑자를 기록한 국내 기록이 감소함에 따라 그 원인을 분석하고자 진행한 연구이다. 국가 경제를 안정적으로 관리하기 위해서는 경상수지가 바람직한 수준으로 유지되어야 하며, 이것은 관광수지의 관리와 예측 없이는 불가능하기 때문에 해당 연구가 진행되었다. 기존의 논문들은 내국인의 관광 수요에 영향을 미치는 대표적인 변수로 명목환율만을 이용한 단일변량 모델링을 진행하였는데, 해당 논문에서는 기존의 연구들이 간과하고 있는 경제변수 변동에 대한 관광 수요의 동태적 반응도 고찰하여 연구를 진행하였다. 관광 수요를 예측하기 위해 결과값에 어떤 변수들이 영향을 미치는지 분산 분해를 통해 분석하였다.

논문에서는 여러 변수를 사용하는 다변량기법이 단일변량기법보다 더 우수한 예측 성과를 나타낼 것처럼 보이지만, 적지 않은 경우 단일변량기법의 예측력이 다변량기법보다 우수한 것으로 나타나고 있다고 말하였다. Somanath(1986), Meese and Rogoff(1983), MacDonald and Taylor(1994) 등은 많은 변수들로 구성된 다변량모형

7)

이 단일변량모형보다 성능이 우수하지 못함을 밝혔다.

하지만 예측 모델은 설명력보다는 예측 능력에 비중을 두어야 한다. 그러나 많은 경우에 있어 해당 논문에서 밝힌 바와 같이 단일변량 모형의 설명력이 우수하나 이것이 예측력의 우수함을 보장하지는 않는다고 말했다. 논문의 본론 부분에서는 단일변량모형과 다변량으로 된 구조적 모형의 예측력을 비교하여, 다변량 모형은 단변량 모형에 비해 낮은 예측오류 뿐만 아니라 예측 편의도 보이지 않음을 보였다. 이에 비해 단변량 모형은 높은 예측오류와 심각한 수준의 예측편의를 보여 비구조적 모형에 의한 예측은 의미가 없음을 증명했다.

이 논문은 단변량 모델링에 대해서는ARIMA, RW를 사용하고, 다변량 모델링에 대해서는 회귀분석 모델을 사용하여 진행하였다. 단변량과 다변량에 대해 모델링을 진행할 때 같은 모델링 기법에 대해서 변수만 제어하며 진행한 것이 아니라, 각각의 변수에 따라 다른 모델을 진행하여 정확한 단변량, 다변량에 대한 비교가 불가능하다는 한계점이 있다.

1.3.5 Stock Prediction Based on Optimized LSTM and GRU Models(2021)[8)]

위 논문은 변동성이 매우 높은 주식 시장에 대해 주가를 예측을 수행하였다. 주식 시장은 고수익, 고위험의 특징을 가지고 있다. 따라서 주식 시장 예측은 금융권의 주요 연구 주제이자 투자자들의 주목을 받고 있는 주제이다. 주식 시장에서 주가의 상승과 하락에 영향을 미치는 요인은 복잡하고 다양하다. 가격 지표, 유통 지표, 활동 정도, 경제적 불확실성이나 거래자의 기대, 거래자의 심리적 요인, 정치적 환경 같은 비경제적 요인 등이 그 예이다.

주가는 일반적으로 변동성이 크고 비모수적이어서 주식 시장에서 비선형 및 비정상 특성을 초래한다. 따라서 흔히 사용하는 ARIMA 통합 이동 평균 모델은 선형 모형이기에 주식 시장 분석에 적합하지 않다. 다음으로 활용할 수 있는 기법은 인공신경망으로 MP 신경망과 역전파 신경망이 있다. 하지만 인공 신경망의 모델은 구조가 너무 복잡하고 몇 가지 문제점이 존재한다. 대표적으로 과적합은 모델 일반화의 약한 일반화로 이어지고, 국부 극값은 모델의 예측 능력 저하로 이어질 수 있다. 또한 최적화 과정에서 뉴런과 과도한 가중치로 인해 그래디언트 소실이나 폭발 문제가 발생하는 경우도 있다. 따라서 해당 논문에서는 딥러닝 기법 중 LSTM, GRU로 주가 예측을 수행하였다.

그러나 해당 연구 과정에서 LSTM을 활용한 결과 바로 직전 날의 주가와 똑같이

8)

예측하여, 마치 지난 데이터가 한 칸씩 밀린 듯한 Shifted 현상이 발생하였다. 이런 현상은 변수가 결과값에 큰 영향을 주지 않는, 즉 종속변수와 무관한 독립변수로만 분석을 진행했을 때 발생한다. 주가 예측에서는 대표적으로 가격 변수가 그렇다. 주가 예측은 아직 종속변수에 막대한 영향을 미치는 독립변수가 존재하지 않기 때문에, 변수 몇 개에 의존해서 먼 기간까지의 예측 분석을 진행하는 것은 불가능한 일이다. 논문에서는 변수와 결과값 사이의 설명력을 다시 분석하여 유의미한 영향력을 가지는 변수들을 활용한 단기간의 예측을 수행하며 연구를 마무리하였다. 해당 연구에서와 같은 해결 방법을 활용하기 위해서는 종속변수에 영향을 주는 독립변수들이 존재해야만 한다.

제 4 절 본 연구의 필요성

스니커즈 리셀 시장이 날이 갈수록 그 시장의 규모가 커지고 있다. 이에 따라 스에 상응하는 연구 및 논문 역시 증가하고 있으며, 일명 '리셀 재테크'라는 말이 붙을 정도로 리셀 시장은 확장 중이다. 재테크라 하면 기존에 여겨지는 시장은 주식, 펀드, 경매 같은 종목들이 먼저 생각나던 과거와 달리 고가의 신발을 통해서도 재테크를 성공시킬 수 있는 시대가 된 것이다. 재테크의 핵심은 '되팔수 있을 것인지 유무'와 '되팔수 있다면 얼마나 수익성이 있을가'이다. 그런데 기존 논문에서는 가격 변화 추이 예측을 통한 수익성 분석만을 진행하거나 리셀 가능 여부에 대해서만 예측만을 수행하였다. 따라서, 본 연구가 추구하는 방향성은 첫 번째로 신제품을 겨냥한 리셀 가능 여부 예측 모델과 기존 시장 제품을 겨냥한 리셀 가격 예측 모델을 모두 구성하는 것이다.

현재 소비자들이 리셀을 할 경우 일반적인 상황을 살펴보면, 먼저 신발은 나이키 공식 스토어를 통해 진행된 추첨에 당첨되면 발매가를 기준으로 신발을 구매하게 된다. 구매한 스니커즈를 리셀하기 위해서는 대표 리셀 플랫폼인 stockX나 Kream 같은 리셀 사이트를 들어가서 본인이 직접 가격 추세를 확인하며 신발 상태를 검정받은 후 판매하는 시스템이다. 향후 해당 신발이 가격이 얼마나 오를 것이고 떨어질 것인지에 대한 정보는 주로 구전을 통해 듣는 소문이나 SNS, 뉴스 기사 등을 확인하여 소비자가 주관적으로 판단하게 된다. 따라서 이러한 주관적 판단에 의존해 신발의 구매/판매 여부를 결정해야 한다는 점이 큰 어려움이자 진입장벽으로 작용한다. 따라서 본 연구를 통해 신발의 특성 정보를 활용한 리셀 가능 여부 예측과 포털 검색량, 거래내역, 주식 종가와 같은 시계열 데이터를 활용한 스니커즈 리셀 가격을 예측 분석을 진행하여 앞서 설명한 소비자가 겪는 어려움에 대한 해결책을

제시한다.

또한, 기존의 논문에서는 데이터의 규모의 문제 등을 이유로 딥러닝 모델을 적용하지 못하거나 너무 많은 요인을 고려하여 상품의 다양성 때문에 분석의 정확도가 떨어졌다는 한계점이 존재하였다. 따라서 해당 연구에서는 이러한 한계점을 극복하고 예측의 정확도를 높이기 위해 KREAM 플랫폼에서 분석 조건에 해당하는 상품 데이터를 최대한 많이 수집하여 딥러닝 모델 학습에 활용하였다. 더불어 회귀 모델을 활용한 변수 분석을 통한 특성 소거 과정도 추가로 진행한다.

리셀 가격의 정확한 예측을 위해서는 기존 연구들에서 리셀 시장 분석에 활용해보지 않았던 다양한 접근 방식과 기법을 활용한다. 특히 본 연구에서는 종속변수를 설명하는 독립변수들의 조합과 변수들 사이의 상관관계가 예측력에 주요한 영향을 미치는 만큼, 변수 제어를 통한 실험을 수행하고 비교 분석하여 가장 알맞은 모델을 채택한다. 그 수행과정을 자세히 설명하면 먼저 거래 가격만을 사용한 단일변량 모델을 3가지 딥러닝 기법을 적용하여 구현하고 비교를 통해 분석에 가장 알맞은 모델을 결정한다. 이후 다양한 변수들을 여러 조합으로 구성하여 모델에 적용해봄으로써 종속변수에 대한 높은 설명력을 갖는 변수 조합들을 선별한다.

이러한 연구 수행을 통해 본 연구는 딥러닝 기법 적용, 분석의 정확성 및 예측력 향상을 도모하고 결과적으로 리셀 시장 분석에 새로운 기법 도입, 기존 연구들보다 더 발전된 형태의 리셀 가능 여부 및 가격 예측 모델을 구현이라는 의의를 지닌다. 더불어 기존 연구에서 진행하지 않았던 단일변량 모델링과 다변량 모델링에 대한 성능 비교를 통해 새로운 리셀 가격 예측 방식을 제안한다.

제 5 절 기대효과

1.5.1 스니커 테크의 주 소비자에게 편리성 제공

스니커 테크란 희소한 중고 신발을 고가에 사고팔며 이윤을 남기는 사람들을 일컫는다. 최근 국내 스니커즈 리셀 시장의 급성장과 함께 나타난 새로운 현상이다. 스니커즈 중고 거래를 성장시키는 데 가장 크게 기여한 사람들은 바로 MZ세대 소비자이다. 자신만의 개성을 표현하는 것을 선호하는 젊은 소비자층에게 스니커즈 구매는 일종의 경험 소비였다. 또한 인하대 이은희 소비자학과 교수 말에 의하면 리셀로 거래되는 스니커즈는 “과시 소비의 기본 요소인 ‘비싼 가격’과 ‘돈이 있어도 못 산다’는 요소를 모두 담고 있어 젊은 층의 소셜 네트워크 서비스 인정 욕구를 충족하기도 좋은 아이템”이라고 말했다. 이렇듯 스니커즈 리셀 시장의 주 소비자층 연령대는 매우 어리며, 심지어 학생들까지 포함하고 있다.

MZ세대는 최초의 글로벌 세대이자 인터넷 세대로 인터넷, 모바일 장치 및 소셜 미디어의 사용 증가와 친숙함을 특징으로 들 수 있다. 그들은 개별적인 쇼핑몰 사이트를 방문해서 온라인 구매를 하는 것보단 통합 온라인 쇼핑몰 앱을 이용하고, 배달 음식점을 검색해서 전화로 주문하는 것보단 통합된 배달 플랫폼을 이용하는 것이 더 익숙하다. 따라서 본 연구는 스니커즈 시장의 주 고객층인 MZ세대에게 많은 편리함을 제공해 줄 것이다. 개인이 노력하여 리셀 가능 여부에 대한 조언을 얻고, 리셀 가격이 얼마나 될지 검색하러 다니지 않아도, 데이터를 분석하여 소비자들이 원하는 정보인 리셀 가능 여부, 리셀가 예측의 정보를 제공하는 것은 빠르게 성장하는 스니커 테크 소비자에게 특히 커다란 편의성을 제공할 것이기 때문이다.

1.5.2 마케팅 정보로의 활용

스니커 테크 시장에서 제조사들은 주로 한정판 마케팅의 태도를 취하고 있다. 한정판 스니커즈의 대부분은 일명 뽑기라고 불리는 Raffle (추첨제) 방식으로 시장에 입고되고 있다. 스니커즈를 제작하는 브랜드들은 최근 이 한정판 마케팅을 더욱 강화시키고 있다. 대표적인 브랜드인 나이키, 아디다스, 뉴발란스 등의 글로벌 스포츠 브랜드에서도 스니커즈를 제작하고 한정판으로 상품을 내놓는 방식을 선호하고 있다. 래플에 참여하는 고객 수와 관여도가 점차 높아지는 만큼 제조사들은 이 같은 마케팅을 관두긴 어려울 것이다. 마케팅 효과를 노리는 제조사와 소액으로 단기 차익을 누리려는 소비자들 사이의 이해관계가 성립되면서 스니커 테크가 더욱 호황을 누리고 있다.

스니커 테크의 성장을 이룬 가장 큰 주축은 한정판 판매가 아닌 소비자들의 리셀이다. 소비자들은 본인이 착용하기 위해 신발을 구매하기도 하지만, 많은 부분이 되파는 것에 초점을 두고 있다. 스니커즈가 단기성 차익을 누리는 재태크 수단 중 하나로 급부상한 것이다. 가격대, 구매 경로 등 진입 장벽이 낮은 스니커 테크는 10만 ~ 20만 원대 스니커즈를 사서 수십, 수백만 원까지도 차익을 남길 수 있다. 따라서 소비자들이 래플에 참여하고, 스니커즈에 열광을 하는 주요한 이유로 리셀이 빠질 수 없다. 소비자들은 스니커즈를 되팔고 차익을 얻기 위해 브랜드의 상품들을 구매한다. 따라서 적절한 래플과 한정판 마케팅을 위해서는 현재 리셀 시장에서 소비자들에게 가장 높은 차익을 제공해 줄 것이라고 예상되는 상품은 무엇인지, 단기간에 그 가치가 급격하게 성장한 제품은 무엇인지에 대한 정보가 필요하다. 본 프로젝트에서 결과값으로 출력하는 리셀 가능 유무, 특정 상품의 판매가 예측을 통해 제조사 입장에서도 스니커즈 제품의 추이를 파악하는 데 도움이 될 것이다.

1.5.3 학문적 기여

국내 스니커즈 리셀 시장은 급성장을 이루고 있다. 글로벌 리서치 전문기업 코웬 엔코에 따르면 전세계 스니커즈 리셀 시장은 해마다 20%씩 성장해 2030년에는 약 25조원 규모에 이를 것으로 예측되었다.[9] 국내 시장에서도 네이버의 자회사인 스노우와 무신사가 선보인 스니커즈 리셀 플랫폼 '솔드아웃'이 출시되며 대기업들 역시 스니커즈 시장에 뛰어드는 모습을 확인할 수 있었다. 롯데쇼핑몰도 2020년 7월 '아웃오브스탁'과 업무 협약을 맺고 매장에 오프라인 리셀 매장을 열었으며, 서울 여의도 더현대서울은 개점 시점부터 오프라인 리셀 매장 '브그즈트랩'을 열었다.

빠른 성장세를 보이며 호황을 누리는 시장 규모와는 다르게 아직 스니커즈 리셀 분야에 대한 데이터 분석적인 연구가 활발하게 이루어지지 않고 있다. 스니커즈의 가격을 결정하는 중요 요인을 분석하는 것 역시 데이터 분석의 관점으로 모델을 통해 도출된 결과가 아닌, 개인에 의거한 간편 추론 수준에 머물고 있다. 따라서 본 연구는 활발한 거래가 이루어지며 소비자층이 두터운 시장의 성장세와 발맞춰 실제 데이터를 사용한 리셀 가능 유무 결과와 리셀 예측가를 도출해 낼 수 있다는 점에서 의의를 갖는다. 또한, 리셀 예측가 분석에서는 기존 연구와 다르게 단변량, 다변량 변수 모델링에 관해 비교하는 과정을 동일한 모델에 대해 진행해봄으로써 단변량, 다변량 변수 모델링에 대한 정확한 비교 분석의 한 가지 사례를 남긴다는 의의도 갖는다.

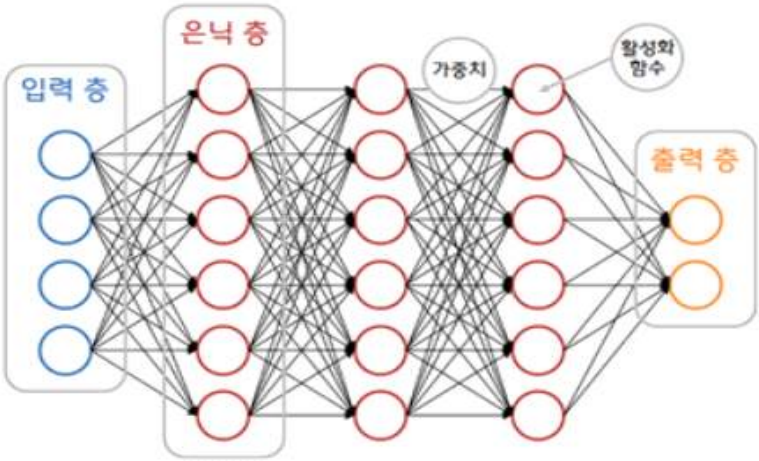
제 6 절 사용한 연구 방법론

1.6.1 리셀 가능 여부 예측 - DNN

DNN은 Deep Neural Network의 약자로 심층 신경망을 의미한다. 심층 신경망은 일반적인 인공 신경망(Artificial Neural Network)에서 2개 이상의 다중 은닉층을 사용하여 더욱 고도화된 학습을 가능하게 하는 가장 기본적인 딥러닝 기법이다.

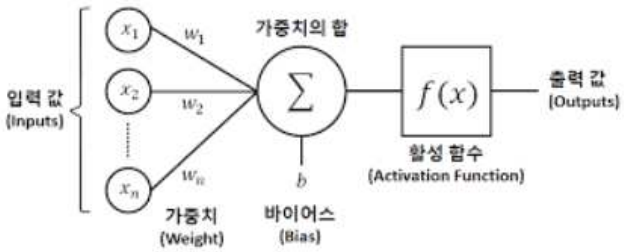
인공 신경망은 학습에 필요한 데이터를 입력받기 위한 입력층과 입력 데이터에 대한 학습을 수행하는 다중의 은닉층, 최종적인 결과값을 결정하는 출력층으로 이루어진다. 각 층은 입력으로 들어온 값을 처리하기 위한 노드(유닛)로 구성되며 각각의 노드들은 활성화 함수(Activation function)을 통해서 입력된 값을 다음 층으로 전달할 것인지 그 여부를 결정한다. 인공 신경망의 구조를 다음과 같은 그림으로 표현할 수 있다.

9)



[그림 1-1] DNN 모델 구조

여기서 은닉층들과 출력층 사이의 존재하는 가중치 의해 전달되는 값이 달라지고 최종적으로 가중치 값에 영향을 받은 출력층의 각 노드들의 값의 상대적인 크기를 바탕으로 최종값이 결정된다. 즉, 다음 그림과 같이 입력 값은 각각의 가중치와 곱해진 후 바이어스 값과 더해져 활성함수를 통과하고 이 과정을 통해 출력 값이 정해지는 것이다.



[그림 1-2] DNN 모델의 출력값 계산 과정

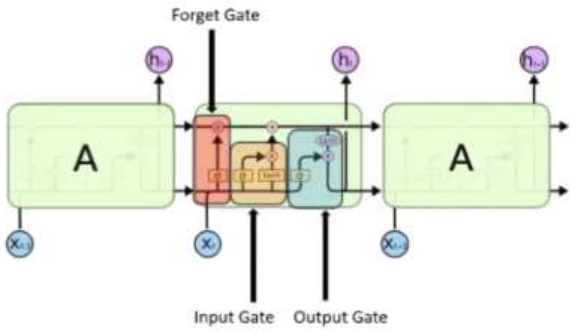
따라서, 인공 신경망을 대상 데이터에 최적화하고 그 성능을 높이기 위해서는 각 층 사이에 존재하는 수많은 가중치의 값을 적절하게 추정하는 것이 매우 중요하다. 이때 사용되는 방법이 오류 역전파(Backpropagation)이다. 이 방법은 최종 출력층에서 측정된 오차를 기반으로 각 층 사이에 있는 가중치의 값을 경사 하강법(Gradient Descent)을 사용해 조정한다. 전체적인 절차를 설명하면, 초기 가중치의 값을 무작위로 설정한 뒤 입력층으로부터 데이터를 입력받는다. 그 후 설계된 여러 개의 은닉층을 통과하며 가중치와 활성화 함수로부터 영향을 받은 추정값이 도출된다. 이 추정값과 실제 결과값의 오차를 측정하여 출력층으로부터 입력층의 방향으로 되짚어 가면서 각 가중치 값을 수정한다. 최종적으로 추정치를 바탕으로 구해진

오차값이 일정범위 내에 들어올 때까지 위 과정을 반복적으로 수행하며 가중치를 갱신한다. 심층 신경망은 이러한 인공 신경망에서 2개 이상의 은닉층을 포함한 깊어진 구조로 복잡한 비선형 문제도 효과적으로 해결할 수 있어 다양한 분야에서 널리 활용되고 있다. 구체적인 사례로는 음악 흥행 예측 모델 연구(2020), 'League of Legends' 게임 승패 예측 연구(2021), 일일 전력수요 모델 개발 및 예측 연구(2021), 서울시 도로 링크별 교통 혼잡도 예측(2019) 등의 연구가 수행된 적이 있다.

우리는 한정판 스니커즈의 특성 변수(브랜드, 라인, 모델명, 색상, 사이즈, 콜라보 정보 등)을 입력 데이터로 하여 은닉층의 개수 및 노드 수, 활성 함수 등 다양한 하이퍼 파라미터 조정을 통한 실험을 진행할 것이다. 최종적으로는 한정판 스니커즈의 리셀 가능 여부를 가장 정확히 예측할 수 있는 심층 신경망 모델을 도출하는 것이 이 과제의 첫 번째 목적이다.

1.6.2 리셀 가격 예측 - LSTM

시계열 모델은 이전에 관측된 값을 기반으로 미래의 값을 예측할 수 있는 모델이다. 연속적으로 나타나는 시계열 데이터를 가지고 미래의 값을 예측해야 하기 때문에 일정 기간 동안의 값을 반영할 수 있는 LSTM 모델을 사용하고자 한다. LSTM은 long short-term memory로 기존 RNN의 한 종류로 긴 의존 기간을 필요로 하는 학습을 필요로 하는 경우 사용할 수 있다. RNN과 비슷하게 neural network를 반복시키는 순환구조의 형태를 띠고 있지만 각 모듈에는 4개의 상호작용 하는 층이 들어있다. 기존 RNN의 기울기 소실 문제를 해결하도록 개발되어 있다. LSTM 셀에서는 상태가 크게 두개의 벡터로 나누어진다. Ht(Short-term state),ct(long-term state)가 존재한다. LSTM의 구조는 다음 사진과 같다.



[그림 1-3] LSTM 모델 구조

1) cell state

cell state는 정보가 바뀌지 않고 그대로 흐르도록 하는 역할을 한다.

2) forget gate

forget gate는 cell state에서 sigmoid layer를 거쳐 어떤 정보를 버릴 것인지 정한다. 현시점의 정보와 과거의 은닉층의 값에 각각 가중합을 구한 후 sigmoid 함수를 적용해 그 출력값의 직전 시점의 cell에 곱해준다. Sigmoid에 의해 0과 1 사이의 값을 갖는다. 1에 가깝다면 과거 정보를 많이 활용하고, 0에 가까우면 과거 정보를 많이 버리게 된다. 수식은 다음과 같다.

$$f^{(t)} = \sigma\left(U_f x^{(t)} + W_f h^{(t-1)}\right)$$

3) input gate

input gate는 앞으로 들어오는 정보 중 어떤 것을 셀 state에 저장할 것인지 정한다. Sigmoid를 거쳐 어떤 값을 갱신할 것인지 정하고 tanh layer에서 새로운 입력 후보 벡터를 만든다. 현시점이 실제로 가지고 있는 정보가 얼마나 중요한지를 반영해 셀에 기록한다. 수식은 다음과 같다.

$$i^{(t)} = \sigma\left(U_{in} x^{(t)} + W_{in} h^{(t-1)}\right)$$
$$\tilde{C}^{(t)} = \tau\left(U_c x^{(t)} + W_c h^{(t-1)}\right)$$

4) memory cell 계산

과거의 정보를 forget gate에서 계산된 만큼 버리고, 현 시점의 정보 후보에 입력 게이트의 중요도를 곱해준 것을 더해 현시점 기준 memory cell을 계산한다. 수식은 다음과 같다.

$$C^{(t)} = f^{(t)} * C^{(t-1)} + i^{(t)} * \tilde{C}^{(t)}$$

여기서 *는 pointwise operation이다.

5) output gate

현시점의 은닉 값으로 출력할 양을 결정하는 출력게이트이다. Sigmoid 층에 입력값을 넣어 output 정보를 정한 다음 cell state를 tanh 층에 넣어 sigmoid 층의 출력값과 곱해 output으로 내보낸다. 수식은 다음과 같다.

$$o^{(t)} = \sigma\left(U_o x^{(t)} + W_o h^{(t-1)}\right)$$
$$h^{(t)} = o_t * \tau(C^{(t)})$$

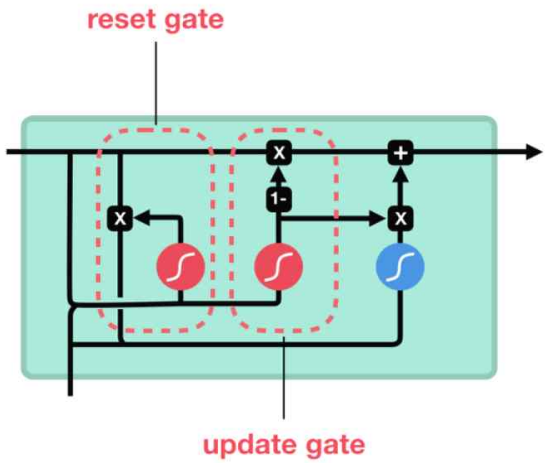
6) 출력층

출력층의 수식은 다음과 같다. RNN의 방식과 동일하다.

$$\hat{y}^{(t)} = softmax\left(Vh^{(t)}\right)$$

1.6.3 리셀 가격 예측 - GRU

GRU는 Gated Reccurent Unit의 약자로 LSTM과 비슷한 이유로 만들어졌지만, LSTM을 구성하는 Time-Step의 셀을 조금 더 간소화한 버전이다. GRU는 LSTM의 구조를 간단하게 개선했기 때문에 보다 빠른 학습 속도를 보인다고 알려져 있지만, 여러 평가에서 GRU는 LSTM과 비슷한 성능을 보인다고 평가된다. 하지만 데이터의 양이 적을 때는 매개변수의 양이 적은 GRU가 유리하고 데이터의 양이 많으면 LSTM이 낫다는 것이 보편적 인식이다. GRU의 구조는 다음과 같다.



[그림 1-4] GRU 모델 구조

LSTM과의 주요한 차이점을 뽑자면, 첫 번째로 GRU는 LSTM과 다르게 게이트가 2개이며 Reset 게이트와 Update 게이트로 이루어져 있다. 여기서 Reset 게이트는 이전 상태를 얼마나 반영할 것인지 나타내며, Update 게이트는 이전 상태와 현재

상태를 얼마만큼의 비율로 반영할 것인지를 의미한다. 또한, LSTM에서의 cell state와 hidden state가 hidden state로 통합되고 update gate가 LSTM에서의 forget, input gate를 제어한다. 마지막으로 GRU에는 output gate가 존재하지 않는다.

GRU의 핵심 내부 구조로는 Reset gate와 Update gate가 있다. Reset gate는 이전 상태의 hidden state와 현재 상태의 x를 받아 sigmoid 처리한다. 즉, 위에서 설명했듯 이전 hidden state의 값을 얼마나 활용할 것인지에 대한 정보를 담고 있다. Update gate는 LSTM의 forget gate, input gate와 비슷한 역할을 하며 이전 정보와 현재 정보를 얼마나 반영할지에 대한 비율을 구하는 것이 핵심이다. 즉, update gate의 계산 한 번으로 LSTM의 forget gate + input gate의 역할을 대신할 수 있다. 최종 결과는 다음 상태의 hidden state로 보내지게 된다.

1.6.4 리셀 가격 예측 - AdaBoost-GRU 앙상블 모델

1)Adaboost 알고리즘

Fruend와 Shapire가 처음으로 제안한 알고리즘으로 예측력이 약한 분류기를 결합해 강한 분류기를 만드는 알고리즘이다. AdaBoost 알고리즘은 분류가 잘못된 데이터는 추출 확률을 증가시키고, 잘 된 데이터는 추출 확률을 감소시키면서 추출 확률을 조정한다. 처음 데이터가 추출될 확률은 동일하고 계속해 반복 학습을 진행하며 강한 분류기를 만들어낸다.

데이터가 추출될 확률을 w_1, w_2, … w_n이라고 하고, 표본의 수를 M, 각 붓스트랩 표본에서 구한 분류기를 C_1, C_2, … C_M이라고 하자. 분류기 C_m의 오분류율은 이와 같이 계산한다.

$$\epsilon_m = \frac{\sum_{i=1}^{n} w_i^{(m)} I(C_m(\boldsymbol{x_i}) \neq y_i)}{\sum_{i=1}^{n} w_i}, \; m = 1, ..., M,$$

[그림1-]

여기서 오분류율은 C_m에 의해 각 데이터의 오분류에 대해 데이터가 추출된 확률을 가중 값으로 한다. 분류기의 신뢰도는 다음과 같다.

$$\alpha_m = \frac{1}{2} \ln \frac{1 - \epsilon_m}{\epsilon_m}, \; m = 1, ..., M,$$

[그림1-]

(m+1)번째 붓스트랩 표본의 추출 확률은 다음과 같다. 여기서 Zm은 정규화 상수이다.

$$w_i^{(m+1)} = \frac{w_i^{(m)} \exp(-\alpha_m y_i C_m(x_i))}{Z_m},$$

[그림1-]

따라서 M개의 분류기를 합쳐 만들어진 최종 분류기는 다음과 같다.

$$C^*(\boldsymbol{x'}) = sign(\sum_{m=1}^{M} \alpha_m C_m(\boldsymbol{x'})),$$

[그림1-]

최종 분류기는 각 분류기에 대한 중요도만큼의 가중치를 반영한 가중 결합이다.

2)AdaBoost-GRU 앙상블 모델

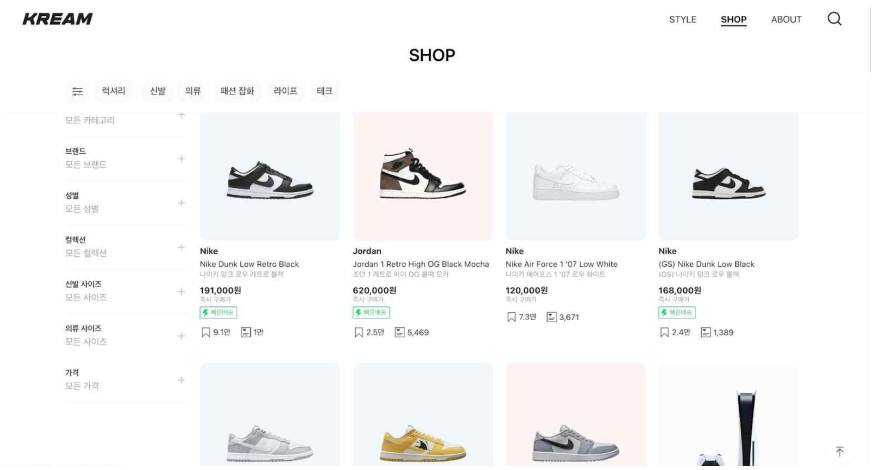
AdaBoost-GRU 앙상블 앙상블 모델은 GRU모델(예측기)를 여러 개 결합해 더 강한 예측기를 만들어낸다. 시계열 자료를 사용해 모델을 학습시키며, t 시점으로부터 하루 뒤의 결과(리셀가)를 예측한다.

제 2 장 본론

제 1 절 데이터 수집

2.1.1 스니커즈 특성 및 거래 데이터 수집

본 연구에 활용한 데이터셋은 국내 스니커즈 리셀 업계 1위를 차지하고 있는 한정판 스니커즈 거래 플랫폼 'KREAM'에서 수집하였다. 다음은 KREAM 웹사이트의 모습이다.



[그림 2-1] KREAM 웹사이트 구성

사이트 좌측에 브랜드, 컬렉션, 신발 사이즈 등의 기준에 따라 검색 결과를 필터링 할 수 있는 기능이 있다. 해당 연구에서는 이 기능을 사용하여 브랜드는 나이키, 조던, 뉴발란스, 컨버스, 아디다스 제품군으로 한정하였고, 컬렉션 또한 앞서 말한 브랜드에 해당하는 컬렉션만을 선별하여 데이터 수집을 진행하였다. 스니커즈 리셀에서는 신발 사이즈에 따라서도 가격 변동이 생기기 때문에 스니커즈 시장에서 가장 수요가 높고 거래가 활발한 270 사이즈에 해당하는 데이터로 한정하였다.

또한, 예측이 리셀 가능과 리셀 수익 창출 불가능 2가지로 이루어지기 때문에 학습에 사용되는 라벨별 데이터 불균형 문제로 인해 발생할 수 있는 문제를 최대한 예방하고자 수집과정에서 웹페이지 오른쪽 상단에 위치한 정렬 기준 옵션을 사용하였다. 인기순, 프리미엄순, 즉시 구매가순 등의 나열 기준을 통해 데이터를 나열한 후, 각각의 신발 특성 데이터와 발매 직후부터 약 3개월 간의 일별 리셀 평균 거래 가격들을 수집하였다. 수집한 신발들의 특성 데이터로는 모델명, 브랜드, 라인, 콜라보 브랜드, 발매일, 발매가, 색상으로 변수의 상세 정의는 [표2-]로 정리하였다. 해당 변수들을 고려해 총 477개 종류의 스니커즈 데이터를 수집하였고 각 브랜드별 비율은 표[2-]와 같다. 상대적으로 인기가 많고 한정판, 콜라보 상품 출시 및 거래가 활발한 나이키, 조던, 뉴발란스 제품의 제품이 차지하는 비율이 높다.

제품명	알파벳과 숫자로 구성된 제품 고유 정보
브랜드	신발을 생산 및 판매한 업체(나이키, 조던, 뉴발란스, 아디다스, 컨버스)
라인	브랜드의 세부 라인(예시: 나이키 - 덩크, 에어맥스1)
콜라보 브랜드	신발 제작을 위해 협업한 브랜드나 아티스트
발매일	신발이 발매된 날짜(연도, 월, 일)
발매가	발매 당시 가격
색상	신발을 이루는 다양한 색상

[표 2-1] 스니커즈 특성 설명

나이키	조던	뉴발란스	아디다스	컨버스	합계
178	99	100	59	41	477

[표2-]브랜드별 수집 데이터 수

477개의 개별 스니커즈의 발매 직후 거래 가격 데이터에 대해서도 수집을 진행하였다. 각 제품마다 발매 시기의 차이가 존재하기 때문에 정해진 날짜에 해당하는 데이터가 아닌 제품별로 발매일 기준 3개월에 해당하는 일별 리셀 거래 평균 가격 데이터를 추출하였다. 따라서 발매일로부터 아직 3개월이 경과하지 않았거나 KREAM 플랫폼의 창립 이전에 발매되어 발매 직후 데이터를 확보할 수 없는 제품은 분석 대상에서 배제되었다. 거래 데이터는 사이트에 공개된 일자별 평균 거래 금액 수집으로 진행되었으며, 거래가 매일 발생하지 않은 경우 전날의 거래 가격을 채워 넣어 결측치를 보정하였다.

2.1.2 스니커즈 가격 추이 데이터 수집

스니커스 거래 가격 예측에 활용하기 위한 제품 선정 기준은 가격의 변동이 작지 않고, 충분한 거래 데이터가 존재하며 'KREAM' 이 창립된 2020년 1월 1일 이후에 발매된 제품으로 선정하였다. 이러한 기준에 따라 선정된 제품은 조던 1 레트로 하이 OG 블랙 모카, 나이키 x 사카이 LD와플 블루 멀티이다. 이후 해당 제품별로 신발 사이즈 270에 해당하는 일별 평균 거래 가격 정보를 발매일 이후부터 2022년 4월 22일까지 수집하였다. 수집된 데이터 일부 예시는 다음과 같다.

제품명	날짜	가격(거래가)	발매가
조던 1 레트로 하이 OG 블랙 모카	2020-11-13	429000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-14	439000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-15	391000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-16	370000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-17	365000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-18	335000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-19	318000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-20	369000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-21	355000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-22	375000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-23	390000	199,000원
조던 1 레트로 하이 OG 블랙 모카	2020-11-24	396000	199,000원

[그림 2-2] 수집한 제품 거래 가격 데이터 일부

2.1.3 소비자 물가지수 데이터 수집

소비자 물가 지수가 리셀 가격에 영향을 미칠 수 있다는 가정하에 통계청 e-나라 지표 국정 모니터링 지표 서비스를 통해서 소비자 물가지수 데이터를 수집하였다. 소비자 물가 지수란 도시가계가 일상생활을 영위하기 위해 구입하는 상품가격과 서비스 요금의 변동을 종합적으로 측정하기 위해 작성하는 지수로 해당 데이터는 2015년을 기준(=100)으로 가계소비지출에서 차지하는 비중으로 1,000분비로 산출, 품목 460개를 대상으로 작성되었다. 분석 기간에 해당하는 2020년 1월부터 2022년 3월까지의 데이터를 수집하였다.

2.1.4 나이키 주식 종가 데이터 수집

선정된 제품이 모두 나이키 제품이므로 나이키의 주식 가격이 리셀 가격에 영향을 미칠 수 있을 것이라 예상하여 나이키 주식 종가를 데이터 수집을 진행하였다. 주식의 종가는 당일 해당 주식의 가장 마지막 거래가 이루어진 가격을 말한다. 나이키는 외국계 기업이기 때문에 가격 데이터의 기준이 달러로 설정되어 있다.

2.1.5. 네이버 검색량 데이터 수집

해당 제품에 대한 사람들을 보편적인 관심도와 화제성이 포털 검색량에 반영될 수 있으며, 리셀 가격에도 영향을 미칠 수 있다고 판단하여 제품이 발매 직후부터 2022년 4월 22일까지의 네이버 검색량을 네이버 API를 발급받아 수집하였다. 네이버에서 제공하는 검색량 데이터는 가장 검색량이 많은 시점의 양을 100으로 하여 상대적인 수치로 표현된다.

스니커즈 리셀 가격 예측에 활용한 모든 데이터 수집을 완료한 후 시계열 데이터 형식으로 결합한 결과는 다음과 같다.

날짜	가격	발매가	소비자물가더	검색량	종가
2020-02-18	720,000	179,000원	0.9	0	100.0548
2020-02-19	720,000	179,000원	0.9	0	100.506
2020-02-20	720,000	179,000원	0.9	0	100.5747
2020-02-21	720,000	179,000원	0.9	0	98.3381
2020-02-22	720,000	179,000원	0.9	0	98.3381
2020-02-23	720,000	179,000원	0.9	0	98.3381
2020-02-24	720,000	179,000원	0.9	0	94.0809

[그림 2-3] 가격 예측 모델 수집 데이터 일부

제 2 절 데이터 전처리

2.2.1 스니커스 리셀 가능 여부 모델 데이터 전처리

크롤링을 통해 수집된 데이터의 경우 제품의 색상이 하나의 컬럼에 슬래시로 구분되어 있기 때문에 입력 데이터로 활용할 수 없다. 따라서 pandas 라이브러리의 데이터 프레임 형식으로 바꾸고 내장 함수를 활용하여 슬래시의 개수로 사용된 색상의 수를 파악하여 새로운 컬럼(col_num)을 생성하였다. 더불어 제품당 최소 1개에서 최대 5개까지 사용된 색상들에 대해서도 각각의 색상을 구분하여 'col1'부터 'col5'까지의 컬럼을 생성하였다. 또한, 같은 색상이더라도 브랜드마다 표현하는 명칭이 상이하기 때문에 유사한 색상을 같은 계열로 묶어내는 분류 작업을 진행하였다. 예를 들어 라이트 그레이, 레인 클라우드, 울프 그레이, 스틸 그레이 등은 모두 그레이 계열에 포함된다. 이렇게 모든 색상에 대해 분류 작업을 진행한 뒤 블랙 및 화이트 계열(1), 베이지 계열(2), 레드 계열(3), 블루 계열(4), 형광 계열(5), 옐로우 계열(6), 그린 계열(7), 브라운 계열(8), 그레이 계열(9), 그 외 색상(10)으로 구분하여 값을 부여하였다.

다음으로 각 제품에 대한 3개월간 거래 가격 데이터를 통해 리셀 가능 여부에 대한 라벨링을 진행하였다. 라벨링 기준은 거래가가 제상품 발매가 이하일 경우 0(리셀 수익 창출 불가능) 발매가를 초과하는 경우 1(리셀 수익 창출 가능)로 분류하였다.

발매일 정보는 상세한 날짜 정보보다는 발매 시기를 반영하는 것이 적절하다고 판단하여 발매 월(month)정보만 추출하여 month 컬럼으로 추가하였고, 발매일로부터의 경과일수를 day 컬럼으로 추가하여 각 모델 학습에 활용할 데이터를 구분하는 과정에서 활용하였다.

또한, 스니커즈의 브랜드에 해당하는 나이키, 조던, 뉴발란스, 아디다스, 컨버스는 각각을 하나의 컬럼으로 생성하여 해당 브랜드에 제품이면 1 아니면 0으로 값을 부여하였다. 브랜드의 하위분류인 라인의 경우 브랜드와 같은 방식으로 컬럼을 생성하기에는 그 종류가 매우 많기 때문에 총 30개의 라인에 대해 1~30까지의 번호를 할당하는 방식을 채택하였다.

마지막으로 콜라보의 경우 우선 콜라보 유무 컬럼을 생성하여 콜라보 브랜드가 존재하는 경우 1 그 외에는 0의 값을 부여하였다. 이후 콜라보 컬럼을 생성하여 콜라

보 브랜드의 유명도를 반영하였다. 데이터셋 내에 존재하는 50가지의 콜라보 브랜드에 대한 유명도를 고려하여 유명한 경우는 1 아니면 0으로 값을 부여하였다.

발매가	라인	콜라보	month	color_num	col1	col2	col3	col4	col5	day	nike	jordan	newbalance	adidas	converse	콜라보유무	type
259000	21	0	4	1	9	0	0	0	0	0	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	1	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	2	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	3	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	4	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	5	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	6	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	7	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	8	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	9	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	10	0	0	1	0	0	0	1
259000	21	0	4	1	9	0	0	0	0	11	0	0	1	0	0	0	1

[그림 2-4] 리셀 여부 예측 모델 최종 활용 데이터 일부

변수명	전처리 방식
nike	나이키 상품에 해당하면 1, 아니면 0
jordan	조던 상품에 해당하면 1, 아니면 0
newbalance	뉴발란스 상품에 해당하면 1, 아니면 0
adidas	아디다스 상품에 해당하면1, 아니면 0
converse	컨버스 상품에 해당하면1, 아니면 0
라인	라인별 부여한 번호에 따라 1~30
콜라보유무	타사 브랜드와 콜라보가 이루어진 경우 1, 아니면 0
콜라보	콜라보 브랜드의 유명도에 따라 1 또는 0
color_num	상품에 사용된 색상의 수에 따라 1~5
col1	색상 계열에 따라 1~10
col2	색상 계열에 따라 1~10, 없다면 0
col3	색상 계열에 따라 1~10, 없다면 0
col4	색상 계열에 따라 1~10, 없다면 0
col5	색상 계열에 따라 1~10, 없다면 0
발매가	제품의 발매가
month	제품이 발매된 시기(월 기준)(1~12)
day	발매일로부터 경과일수(0~89)
type	리셀 가능 여부(0/1)

[표 2-] 변수별 전처리 방식과 값의 범위

2.2.2 스니커즈 가격 예측 모델 데이터 전처리

다른 데이터들의 경우 2022년 4월 22일까지의 수집이 모두 완료되었지만, 소비자 물가 데이터가 2022년 3월까지밖에 산출되지 않았으므로 학습 및 테스트에 사용할 모든 데이터의 범위를 3월로 축소하였다.

또한, 모든 변수들은 각자 다른 단위를 가지고 있기 때문에 이러한 단위 차이가

몇몇 회귀 모형이나 머신러닝 기법에서 문제를 일으킬 수 있다고 판단하였다. 따라서 입력값들마다 min-max 스케일링을 통해서 0-1 사이의 값으로 다시 매핑해주는 작업을 수행하였다.

마지막으로 나이키 주식 종가 데이터의 경우 주말이나 공휴일에는 주식장이 열리지 않아 결측치가 발생한다. 이를 해결하기 위해 기존 선행연구에서 주로 사용하는 방법인 선형보간법을 활용하여 결측치를 보정하였다. 다음은 결측치 보정을 수행한 결과이다.

종가	종가
100.0548	100.054800
100.5060	100.506000
100.5747	100.574700
98.3381	98.338100
NaN	96.919033
결측치 보정 전	결측치 보정 후

[표 2-2] 주가데이터 결측지 보정 결과

제 3 절 DNN을 활용한 리셀 가능 여부 예측 모델

2.3.1 리셀 가능 여부 예측 모델 실험 설계

예측 모델을 학습시키기 위해 사용된 전체 데이터는 477개 모델의 3개월 간 거래 내역에 해당하는 총 43,953개의 데이터이다. 한 가지 스니커즈 모델에 대한 데이터는 발매일로부터 발매 후 89일 차까지의 90일 치에 대한 데이터로 구성된다. 각각의 예측 모델은 발매일로부터의 경과일수를 기준으로 구분된 데이터를 학습에 활용하기 때문에 하나의 모델은 경과일수가 같은 477개의 데이터만을 학습하므로 최종적으로 90일 치에 대한 90개의 모델을 구현하였다. 모델마다 477개의 데이터 중 70%에 해당하는 333개는 train data, 30%에 해당하는 144개는 test data으로 사용되었다. 해당 연구의 핵심 과제는 각 스니커즈만의 특성을 입력 변수로 하여 90일 동안의 리셀 가능 여부를 예측하는 것이다.

2.3.2 DNN 모델 설계

모델 학습 및 평가에 활용되는 전체 데이터는 약 4만 4천 개 정도로 큰 규모이지만, 각 모델에 사용되는 데이터 수는 500개 이하로 그 양이 적다. 기존 논문을 검토해보았을 때 데이터의 양이 적다면 모델 설계가 복잡해질수록 신경망 과적합의 위험성이 높아지기에 은닉층의 노드 수와 레이어 수를 적게 설정하는 것이 더 높은 성능의 모델을 도출할 수 있는 방향이라고 판단하였다.[10] 또한, 아직 딥러닝 기법을 리셀 시장분석에 활용한 사례가 거의 없어 탐색적 수준에서 예측 모델을 설계를 시도하는 단계에 해당하기에 본 연구에서는 은닉층의 수를 3~9개 사이에서 조정하며 시도해본 후 가장 좋은 성능을 보인 모델의 설계를 채택하였다. 최종 모델의 구조는 변수를 입력받는 입력층, 5개의 은닉층, 마지막 분류 결과 출력을 위해 2개의 노드로 구성된 출력층으로 총 7개의 층으로 구성된 모델을 수립하였다. 은닉층은 16개~32개의 은닉 노드로 이루어져 있으며 활성함수로 'Relu', 과적합 방지를 위해 dropout = 0.01을 적용하였다. model compile을 과정에서는 옵티마이저로 'adam', loss함수는 'binary_crossentropy', 성능 평가 척도는 'accuracy'로 지정하였다.

10)

```
#모델 설계
model = tf.keras.Sequential()
#input layer
model.add(layers.Dense(18, input_shape=(9,)))
model.add(layers.Activation('relu'))
model.add(layers.Dropout(0.01))

#hidden layer
model.add(layers.Dense(18))
model.add(layers.Activation('relu'))
model.add(layers.Dropout(0.01))

model.add(layers.Dense(32))
model.add(layers.Activation('relu'))
model.add(layers.Dropout(0.01))

model.add(layers.Dense(16))
model.add(layers.Activation('relu'))
model.add(layers.Dropout(0.01))

model.add(layers.Dense(32))
model.add(layers.Activation('relu'))
model.add(layers.Dropout(0.01))

model.add(layers.Dense(16))
model.add(layers.Activation('relu'))
model.add(layers.Dropout(0.01))

#output layer
model.add(layers.Dense(2))
model.add(layers.Activation('softmax'))

# 모델 컴파일
model.compile(
    loss= 'binary_crossentropy',
    optimizer="adam",
    metrics=['accuracy'])
```

[그림 2-5] DNN 모델 설계 구현 코드

2.3.3. 로지스틱 회귀 모델을 활용한 모델 입력 변수 선별

데이터 전처리을 통해 생성, 정리된 데이터 feature(특성)는 총 16개이다. 변별 없이 모든 특성을 모델의 입력 변수로 활용할 경우 너무 많은 feature의 사용 즉, 고차원의 입력 데이터로 인해 모델의 성능이 떨어질 가능성이 존재한다. 따라서 이러한 '차원의 저주' 문제를 해결하고 예측의 정확도를 높이기 위해서는 데이터의 특징과 의미를 제대로 표현하고 있는 변수를 선별하여 모델에 반영하는 것이 중요하다.

이를 위해 16가지의 feature를 모두 입력 변수로 활용한 로지스틱 회귀 예측 모델

을 구현 및 실행 후 coefficient(계수)를 도출하여 예측 결과에 각 변수가 미치는 영향력을 확인하였다.

```
model.coef_[0]
```

```
array([ 0.17886373,  0.06963908, -0.04027577, -0.22340804,  0.05039864,
       -0.06245553,  0.18543126,  0.34718591,  0.20789797, -0.01530044,
       -0.02774572, -0.17652195, -0.03857877, -0.27740858,  0.01814583,
        0.01200784])
```

[그림2-]

특성	coefficient(순위)	특성	coefficient(순위)
발매가	0.018(14)	newbalance	0.04(11)
라인	0.34(1)	adidas	0.06(9)
콜라보 유무	0.178(6)	converse	0.22(3)
콜라보(유명도)	0.21(4)	color1	0.015(15)
month(발매월)	0.012(16)	color2	0.027(13)
col_num(색상수)	0.07(8)	color3	0.176(7)
nike	0.05(10)	color4	0.038(12)
jordan	0.18(5)	color5	0.28(2)

[표2-]

그 결과 라인 > color5 > converse > 콜라보(유명도) > jordan > 콜라보 유무 > color3 > 사용 color 수 > adidas > nike > newbalance > color4 > color2 > 발매가 > color1 > 발매 시기(월) 순으로 큰 coefficient 값을 가졌다.

2.3.4. DNN 모델의 성능 개선

예측 모델의 성능을 향상을 모델 학습에 영향을 미치는 batch_size, epochs, validation_set의 비율, 과적합을 방지하기 위한 earlyStopping 설정, 입력 데이터로 사용되는 feature의 종류 등 다양한 구성요소와 파라미터를 변경하며 test 정확도를 확인하였다. 입력 변수의 경우 로지스틱 회귀모델을 통해 도출한 coefficient 값과 순위를 기반으로 다양하게 조합하여 실험하였다.

2.3.4. DNN 모델의 성능 평가 방법

가장 핵심이 되는 성능 척도는 정확도와 F1 score이다. 정확도는 전체 데이터 중 올바르게 예측한 데이터의 비율로 가장 직관적으로 모델의 성능을 나타낼 수 있는 평가 지표이다. 정확도와 더불어 F1 score를 성능 평가에 기준에 포함한 이유는 실제로 리셀이 가능한 상품을 얼마나 잘 예측해내는지가 해당 연구의 핵심이기 때문이다. 따라서 얼마나 정확히 예측했을지를 나타내는 정확도와 정밀도, 재현율을 모두 반영하는 F1 score를 성능 평가의 기준이 된다. 여기서 정밀도는 True라고 예측한 값 중 실제 True인 데이터의 비율을 나타내며, 재현율은 실제 True인 데이터 중에서 모델이 True라고 예측해낸 데이터의 비율이다. F1 score는 예측의 신뢰성에 중점을 둔 정밀도와 데이터 발굴에 중점을 둔 재현율에 같은 비중을 두고 조화평균을 이용하여 구한 값이다.

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * (precision * recall)}{precision + recall}$$

[그림2-]

또한, 정밀도와 재현율을 구하는 함수에서는 옵션을 'weighted'로 설정한다. 리셀 시장 자체가 리셀이 가능한 상품이 주를 이루기 때문에 데이터 비율 역시 리셀 가능 제품이 70% 이상으로 불균형하다는 특징을 가지고 있어 이러한 특징을 반영하였다.

2.3.5 DNN 모델의 실험 결과

정확도와 F1 score를 기준으로 평가했을 때 가장 성능이 뛰어난 모델의 설정값은 batch_size = 10, epochs = 200, validation_split = 0.2, earlyStopping 옵션의 patience = 30으로 [그림2-]의 코드를 통해 구현하였다. 또한, 입력변수로 적용된 특성은 로지스틱 회귀모델을 통해 도출된 coefficient 기준으로 상위 8개에 해당하는 컬럼(라인, color5, converse, 콜라보(유명도), jordan, 콜라보 유무, color3, col_num(사용 색상수))과 10번째로 큰 값을 가진 'nike' 컬럼이다. 해당 설정값과 입력 변수를 모두 동일하게 적용하여 발매일부터 발매 89일 후에 해당하는 90개의 모델을 구현하고 각 모델별 성능을 도출하였다.

```
hist = model.fit(
    train_x, train_y,
    batch_size = 10,
    epochs = 200,
    validation_split = 0.2,
    shuffle = False,
    callbacks=[tf.keras.callbacks.EarlyStopping(monitor='val_loss', patience = 30)],
    verbose=1)

# 테스트 데이터로 성능평가
score = model.evaluate(test_x, test_y)
print('test_loss: ', score[0])
print('test_acc: ', score[1])
```

[그림 2-6] DNN 모델 실행 코드

day	accuracy	precision	recall	f1-score
0	84.7%	82.5%	84.7%	83.6%
1	79.2%	76.3%	79.2%	77.7%
2	75.0%	75.0%	75.0%	75.0%
3	75.0%	73.1%	75.0%	74.0%
4	72.2%	71.7%	72.2%	71.9%
5	72.2%	67.3%	72.2%	69.7%
6	72.2%	69.6%	72.2%	70.9%
7	72.9%	71.5%	72.9%	72.2%
8	77.8%	79.0%	77.8%	78.4%
9	75.0%	73.9%	75.0%	74.4%
10	71.5%	71.3%	71.5%	71.4%
11	72.2%	70.9%	72.2%	71.5%
12	71.5%	68.8%	71.5%	70.2%
13	70.1%	72.3%	70.1%	71.2%
14	70.1%	66.2%	68.8%	67.5%
15	69.4%	67.2%	69.4%	68.3%
16	68.8%	63.9%	68.8%	66.2%
17	68.8%	67.9%	67.4%	67.6%
18	67.4%	72.9%	68.1%	70.4%
19	68.1%	70.0%	70.1%	70.1%
20	70.1%	72.9%	66.7%	69.6%
21	66.7%	60.7%	61.8%	61.2%
22	69.4%	69.4%	69.4%	69.4%
23	70.1%	69.2%	70.1%	69.6%
24	66.0%	67.7%	66.0%	66.8%

25	70.8%	70.6%	70.8%	70.7%
26	69.4%	68.3%	69.4%	68.8%
27	68.1%	63.6%	68.1%	65.7%
28	69.4%	67.9%	69.4%	68.7%
29	67.4%	62.4%	67.4%	64.8%
30	70.8%	74.7%	70.8%	72.7%
31	68.8%	71.2%	68.8%	70.0%
32	68.1%	72.3%	68.1%	70.1%
33	72.2%	71.0%	72.2%	71.6%
34	67.4%	67.0%	67.4%	67.2%
35	70.8%	69.5%	70.8%	70.2%
36	70.1%	73.7%	70.1%	71.9%
37	70.8%	69.3%	70.8%	70.1%
38	70.1%	75.7%	70.1%	72.8%
39	69.4%	70.3%	69.4%	69.9%
40	70.1%	69.4%	70.1%	69.8%
41	68.8%	66.4%	68.8%	67.5%
42	68.1%	67.2%	68.1%	67.6%
43	65.3%	65.7%	65.3%	65.5%
44	69.4%	69.3%	69.4%	69.4%
45	73.6%	72.9%	73.6%	73.3%
46	67.4%	66.9%	67.4%	67.1%
47	67.4%	66.4%	67.4%	66.9%
48	71.5%	69.2%	71.5%	70.3%
49	72.2%	76.5%	72.2%	74.3%
50	68.1%	68.0%	68.1%	68.0%
51	67.4%	65.9%	67.4%	66.6%
52	70.1%	70.1%	70.1%	70.1%
53	69.4%	67.7%	69.4%	68.6%
54	72.9%	76.0%	72.9%	74.4%
55	72.2%	72.4%	72.2%	72.3%
56	75.7%	73.7%	75.7%	74.7%
57	69.4%	69.1%	69.4%	69.3%
58	72.9%	72.6%	72.9%	72.8%
59	69.9%	69.0%	69.9%	69.4%
60	66.4%	66.1%	66.4%	66.3%
61	70.6%	68.9%	70.6%	69.8%

62	66.9%	62.3%	66.9%	64.5%
63	69.7%	72.9%	69.7%	71.3%
64	68.3%	66.6%	68.3%	67.4%
65	67.6%	64.8%	67.6%	66.2%
66	69.7%	69.0%	69.7%	69.4%
67	68.3%	67.5%	68.3%	67.9%
68	67.6%	73.3%	67.6%	70.3%
69	67.6%	67.2%	67.6%	67.4%
70	68.3%	66.6%	68.3%	67.4%
71	69.0%	67.4%	69.0%	68.2%
72	64.8%	65.4%	64.8%	65.1%
73	64.8%	62.7%	64.8%	63.7%
74	65.5%	66.7%	65.5%	66.1%
75	69.7%	68.8%	69.7%	69.2%
76	66.9%	63.3%	66.9%	65.0%
77	69.0%	66.8%	69.0%	67.9%
78	69.7%	68.3%	69.7%	69.0%
79	70.4%	74.2%	70.4%	72.3%
80	69.7%	69.0%	69.7%	69.3%
81	69.0%	64.8%	69.0%	66.8%
82	67.6%	63.6%	67.6%	65.6%
83	66.9%	65.0%	66.9%	65.9%
84	64.1%	61.0%	64.1%	62.5%
85	68.3%	67.6%	68.3%	67.9%
86	64.1%	62.1%	64.1%	63.1%
87	68.3%	68.4%	68.3%	68.4%
88	66.9%	72.8%	66.9%	69.7%
89	65.9%	66.6%	65.9%	66.2%

[표 2-] DNN 모델 성능 평가 결과

정확도와 F1 score를 모두 고려하였을 때 가장 높은 성능을 발휘하는 모델은 발매 당일의 리셀 가능 여부 예측 모델로 정확도 84.7%, F1 socre 83.6%이다. 전체적인 모델의 성능은 발매일로부터의 경과일수가 증가할수록 낮아지는 경향을 보였으며 가장 낮은 성능의 모델은 발매일로부터 84일 뒤 예측을 수행한 모델이다. 평균적으로 약 70%의 정확도와 F1 score로 리셀 가능 여부를 예측하였다.

제 4 절 LSTM, GRU, AdaBoost-GRU 앙상블 모델을 활용한 가격 예측 모델

2.4.1 단일변량 모델링 진행

2.4.1.1 LSTM

수집한 가격 데이터만을 Input 데이터로 하여 LSTM 시계열 가격 예측을 진행하였다. LSTM 모델의 파라미터를 test_size window_size를 각각 (전체일수*0.2), 20으로 설정하였다. 여기서 test size = (전체일수*0.2)는 학습이 과거부터 (전체일수*0.2)일 이전의 데이터를 학습하게 되도록 설정해주는 것이고 window size란 예측에 반영할 과거 데이터의 기간을 의미하는 것으로 window size = 20으로 지정하면 과거 20일 데이터를 기반으로 내일 가격을 예측한다. 파라미터 설정 후 train, test set 분리를 수행하였다.

```
from sklearn.model_selection import train_test_split

feature_cols = ['가격','종가','검색량','평균기온','소비자물가데이터']
label_cols = ['가격']

train_feature = train[feature_cols]
train_label = train[label_cols]

train_feature, train_label = make_dataset(train_feature, train_label, 20)

x_train, x_valid, y_train, y_valid = train_test_split(train_feature, train_label, test_size=0.2)
x_train.shape, x_valid.shape
```

```
((442, 20, 5), (111, 20, 5))
```

[그림 2-7] LSTM 모델 설계 구현 코드

train, test set 분리 완료 후 modeling 작업을 수행하였다. 모델 구현에 있어 설정해야 하는 설정 변수로 batch size, epoch를 고려하였으며, 여러 번의 모델링 진행 후 MSE, MAE, RMSE와 같은 성능지표를 사용해 최적의 파라미터 값을 도출해서 모델링을 진행하였다.

```
from keras.layers import Flatten
model = Sequential()
model.add(LSTM(64,
               input_shape=(train_feature.shape[1], train_feature.shape[2]),
               activation=None,
               return_sequences=True)
          )
```

[그림 2-8] LSTM 모델링 구현 코드

LSTM 모델 실행 코드는 [그림2-]과 같다. loss함수로 'mean_squared_error'를 사용하였으며, optimizer로는 'adam'을 적용하였다. 해당 모델을 활용해 리셀 가격 데이터만을 사용하여 제품별로 실행 결과를 도출하였다.

```
import os

model.compile(loss='mean_squared_error', optimizer='adam')
early_stop = EarlyStopping(monitor='val_loss', patience=5)

model_path = 'model'
filename = os.path.join(model_path, 'tmp_checkpoint.h5')
checkpoint = ModelCheckpoint(filename, monitor='val_loss', verbose=1, save_best_only=True, mode='auto')

history = model.fit(x_train, y_train,
                                    epochs=200,
                                    batch_size=16,
                                    validation_data=(x_valid, y_valid),
                                    callbacks=[early_stop, checkpoint])
```

[그림 2-9] LSTM 모델 실행 코드

2.4.1.2 GRU

두 번째로 GRU 모델을 도입하여 예측 모델을 구현하였다. 다음 [그림] 코드는 해당 GRU모델에서 가장 최적의 성능을 보인 파라미터를 도출하여 구축한 결과이다. 또한, 타모델과 성능 비교를 위해 train, test set은 동일한 데이터셋으로 통일하여 모델링을 진행하였다.

```
model = Sequential()

model.add(GRU(256,activation='tanh',input_shape=x_train[0].shape))

model.add(Dense(1,activation='linear'))

model.summary()

model.compile(loss='mse',optimizer='adam',metrics=['mse'])
early_stop = EarlyStopping(monitor='val_loss',patience=5)

model_path = 'model'
filename2 = os.path.join(model_path, 'tmp_checkpoint_gru.h5')
checkpoint = ModelCheckpoint(filename2, monitor='val_loss', verbose=1, save_best_only=True, mode='auto')

model.fit(x_train,y_train,validation_data=(x_valid,y_valid),epochs=100,batch_size=16,callbacks=[early_stop])
```

[그림2-]

2.4.1.3 AdaBoost-GRU 앙상블 모델

타 모델들과 동일하게 Test, train set, window size를 설정하도록 하여 성능을 비교하는데 동일한 조건을 가지도록 설정해주었다.

```
feature_cols = ['가격']
label_cols = ['가격']
train_feature_1 = train_1[feature_cols]
train_label_1 = train_1[label_cols]

train_feature_1, train_label_1 = make_dataset(train_feature_1, train_label_1, 20)
x_train_1, x_valid_1, y_train_1, y_valid_1 = train_test_split(train_feature_1, train_label_1, test_size=0.2)

test_feature_1 = test_1[feature_cols]
test_label_1 = test_1[label_cols]

test_feature_1, test_label_1 = make_dataset(test_feature_1, test_label_1, 20)
```

[그림 2-] AdaBoost-GRU 앙상블 train,test set & window size 설정

그 후 AdaBoost-GRU 앙상블 모델의 세부 파라미터 조정을 통해서 단일변량일 경우의 최적 파라미터 값을 구했고 그 결과 gru unit = 512, epoch = 100, batch size = 16, n_estimator = 30, activation function = tanh으로 도출되었다.

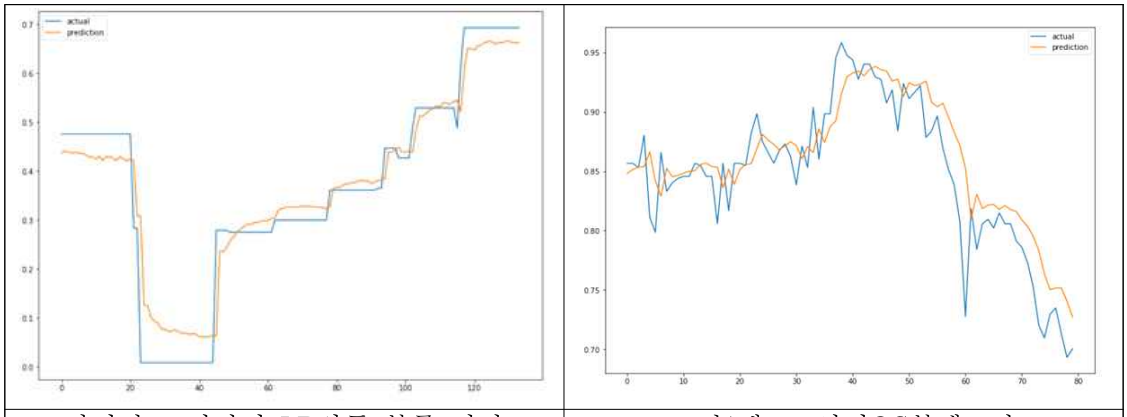
```
model = Sequential()
model.add(InputLayer(input_shape = x_train_1[0].shape))
model.add(GRU(units=512))
model.add(Dropout(0.5))
model.add(Dense(units=1))
model.compile(loss='mean_squared_error', optimizer='adam')

GRU_Predictors = KerasRegressor(build_fn=lambda:model, epochs=100, batch_size=16)
final_model_1 = AdaBoostRegressor(GRU_Predictors, n_estimators=30, random_state=42)

final_model_1.fit(train_feature_1, train_label_1)

preds_1= final_model_1.predict(test_feature_1)
```

[그림2-] AdaBoost-GRU 앙상블 모델 구축 코드

2.4.2 단일변량 모델링 결과 및 비교

2.4.2.1 LSTM

LSTM 모델을 통해 예측을 진행한 결과는 아래와 같다. 두 가지 제품 데이터를 사용하여 결과를 도출하였다.

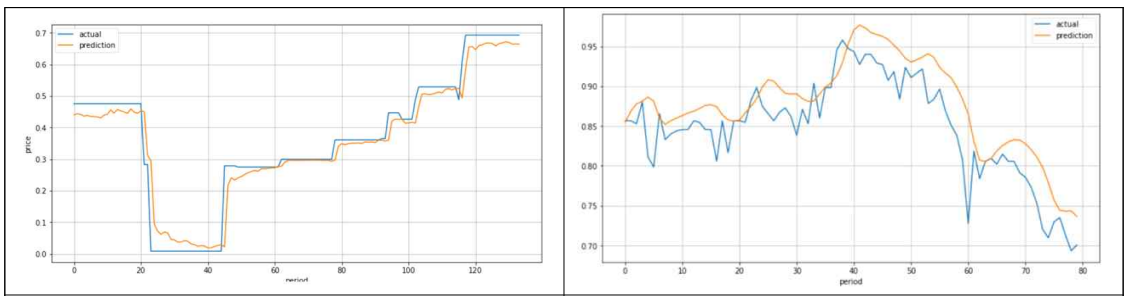


나이키 x 사카이 LD와플 블루 멀티	조던1레트로하이OG블랙모카

[표2-] LSTM 모델 구현 결과 그래프

2.4.2.2 GRU

GRU 모델을 통해 예측을 진행한 결과는 아래와 같다. 두 가지 제품 데이터를 사용하여 결과를 도출하였다.

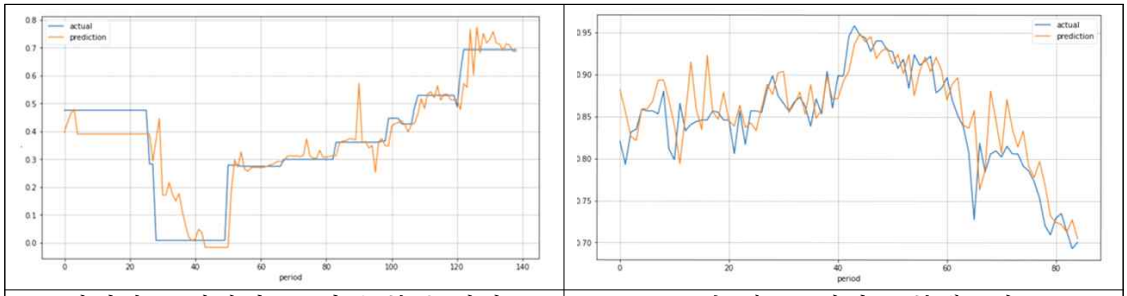


나이키 x 사카이 LD와플 블루 멀티	조던1레트로하이OG블랙모카

[표2-] LSTM 모델 구현 결과 그래프

2.4.2.3 AdaBoost-GRU 앙상블 모델

AdaBoost-GRU 앙상블 모델을 통해 예측을 진행한 결과는 아래와 같다. 두 가지 제품 데이터를 사용하여 결과를 도출하였다.



나이키 x 사카이 LD와플 블루 멀티	조던1레트로하이OG블랙모카

[표2-] AdaBoost-GRU 앙상블 모델 구현 결과 그래프

2.4.2.4 모델 구현 결과의 정략적 비교

제품의 예측 결과를 정량적인 평가 지표를 활용하여 비교하기 위해 MSE, MAE, RMSE, Adjusted R2 수치를 도출하고 실제값과 예측값 사이의 오차값과 모델의 설명력을 표현하였다.

	LSTM	GRU	AdaBoost-GRU
MSE	0.0067	0.0023	0.0069
MAE	0.0064	0.0023	0.0065
RMSE	0.0588	0.0485	0.0335
Adjusted R2	0.9164	0.9471	0.6764

[표2-] 모델별 나이키 x 사카이 LD와플 블루 성능평가지표

	LSTM	GRU	AdaBoost-GRU
MSE	0.0046	0.0015	0.0741
MAE	0.0092	0.0015	0.2148
RMSE	0.0304	0.0392	0.0836
Adjusted R2	0.7642	0.5745	0.8189

[표2-] 모델별 조던1 레트로 하이OG 블랙 모카 성능평가지표

위의 결과는 2가지 제품을 3가지 모델에 각각 적용하여 예측을 수행한 결과이다. 모델마다 동일하게 들어가는 파라미터들에 대해서는 같은 값으로 설정하여 동일한 조건에서 모델의 성능을 가할 수 있도록 하였다. 그 외에 추가적인 각 모델의 파라

미터는 여러 번의 시도를 통해 최상의 성능을 내는 파라미터 값으로 설정하였다. 정량적인 결과를 정확하게 비교하기 위해서 시계열 데이터 분석 시 사용하는 평가 지표인 mse, mae, rmse, adjusted R2 값들을 각각 도출하였다.

성능 지표를 종합적으로 비교해 보았을 때 mse, mae 값이나 adjusted R2 값이 가장 좋은 모델은 GRU 모델 혹은 LSTM 모델이다. 하지만 GRU나 LSTM으로 도출된 결과를 자세히 확인해보면 예측값이 실제값을 뒤로 조금 밀린 것과 같은 값으로 예측하는 경향성을 발견할 수 있다. 즉, 가격이라는 단일 변수로 GRU나 LSTM 모델을 실행할 경우 성능은 좋지만, 그 이유가 바로 전날의 가격을 따라서 예측하기 때문이라는 문제점이 존재한다. 즉, 추세선을 잘 따라가기는 하지만 바로 전날의 가격을 따라서 예측하는 것이므로 예측 수행의 의미가 떨어지고 실제 환경에서의 예측에 사용하기는 그 활용성이 떨어진다고 볼 수 있다. 따라서 본 연구에서는 LSTM, GRU 모델보다 정량적인 평가 지표 비교에선 우위에 있지 않지만, 실제값들이 밀린 듯한 예측이 아니라 실제로 모델 학습을 통한 예측값이 도출되는 결과를 보이는 AdaBoost-GRU 앙상블 모델을 활용하는 것이 더 적합하다고 판단하였다. 따라서 다음 단계에서는 AdaBoost-GRU 앙상블 모델을 활용한 다변량 예측을 수행하여 해당 모델의 성능을 개선하고 예측력이 향상시키는 것을 목표로 한다.

2.4.3 AdaBoost-GRU 앙상블 다변량 예측 모델 구현

2.4.3.1 다변량 모델과의 비교 필요성

해당 연구에서 가격이라는 변수와 더불어 가격에 유의미한 영향을 줄 수 있는 다른 시계열 데이터들을 추가하여 입력 변수로 사용하는 다변량 시계열 예측을 진행하고자 한다. 비슷한 예시로 주식 종가 예측을 진행할 때 과거의 주식 종가 데이터뿐만 아니라 거래량, 시가, 고가, 저가, 대비와 같은 주식 종가와 연관성이 있는 추가적인 시계열 데이터를 입력값으로 넣어서 예측을 수행하는 것을 들 수 있다. 본 연구 역시 리셀 가격이라는 변수에만 의존한 예측이 아닌 리셀 가격에 유의미한 영향을 줄 수 있는 변수를 탐색하여 예측에 함께 반영함으로써 정확도 및 예측력 향상을 꾀한다. 그 후 최종적으로 다변량 예측 결과와 앞서 수행한 단일변량 모델 예측 결과의 비교 및 분석을 수행하여 다변량 예측의 필요성에 대한 검증을 진행한다.

2.4.3.2 AdaBoost-GRU 앙상블 다변량 모델 구현

다변량 모델에서도 입력값인 feature_cols에 가격 외에 추가적인 데이터인 가격변동량, 종가, 검색량, 소비자물가데이터를 추가한 후 단일변량 모델링과 동일하게 train, test size, window size 같은 파라미터 값을 설정한다. 다음 [그림2-]는 다변량 예측 모델 설계 코드이다.

```
from sklearn.model_selection import train_test_split

feature_cols = ['가격','가격변동량','종가','검색량','소비자물가데이터']
label_cols = ['가격']

train_feature = train[feature_cols]
train_label = train[label_cols]

train_feature, train_label = make_dataset(train_feature, train_label, 20)

x_train, x_valid, y_train, y_valid = train_test_split(train_feature, train_label, test_size=0.2)
x_train.shape, x_valid.shape
```

[그림2-]다변량 입력값 모델 설계 코드

2.4.4 AdaBoost-GRU 앙상블 다변량 예측 모델 최적 변수 조합 결정

최종적인 다변량 모델중 최적의 성능을 보이는 변수 조합을 찾기 위해서 다양한 조합의 입력값들을 적용한 모델별 성능을 도출하였다. 변수 조합의 변화에 대한 성능을 비교하는 과정이므로 입력 변수 조합을 제외하고는 동일한 파라미터 값을 사용해서 분석을 진행하였다.

Case 1. 가격, 가격변동량, 포털 검색량, 주식 종가를 사용
Case 2. 가격, 가격변동량, 포털 검색량, 소비자 물가데이터 사용
Case 3. 가격, 가격변동량, 주식 종가, 소비자 물가데이터 사용
Case 4. 가격, 가격변동량, 포털 검색량 사용
Case 5. 가격, 가격변동량, 소비자 물가데이터 사용
Case 6. 가격 가격변동량, 주식 종가 사용
Case 7. 가격, 가격변동량, 주식 종가, 소비자 물가데이터, 포털 검색량 사용

위와 같은 7가지 변수 조합을 AdaBoost-GRU 모델에 입력 변수로 적용하여 도출한 성능 지표는 다음 [표2-]와 같다.

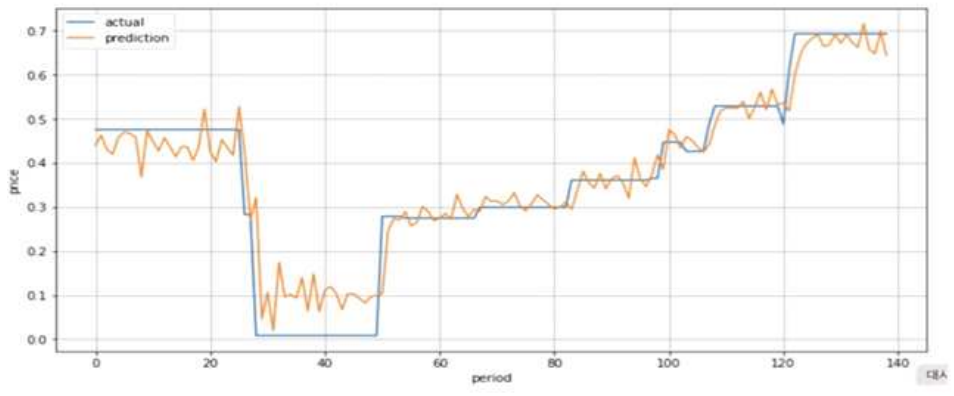
	Case1	Case2	Case3	Case4	Case5	Case6	Case7
MSE	0.0069	0.0735	0.0073	0.0721	0.0698	0.0726	0.0666
MAE	0.2097	0.2156	0.2139	0.2129	0.2085	0.2124	0.2053
RMSE	0.0576	0.0625	0.0544	0.0708	0.0663	0.065	0.0652
Adjusted R2	0.9138	0.8988	0.9231	0.8701	0.886	0.8904	0.8898

[표2-] 나이키 x 사카이 LD와플 블루 멀티 변수 조합별 평가 성능지표

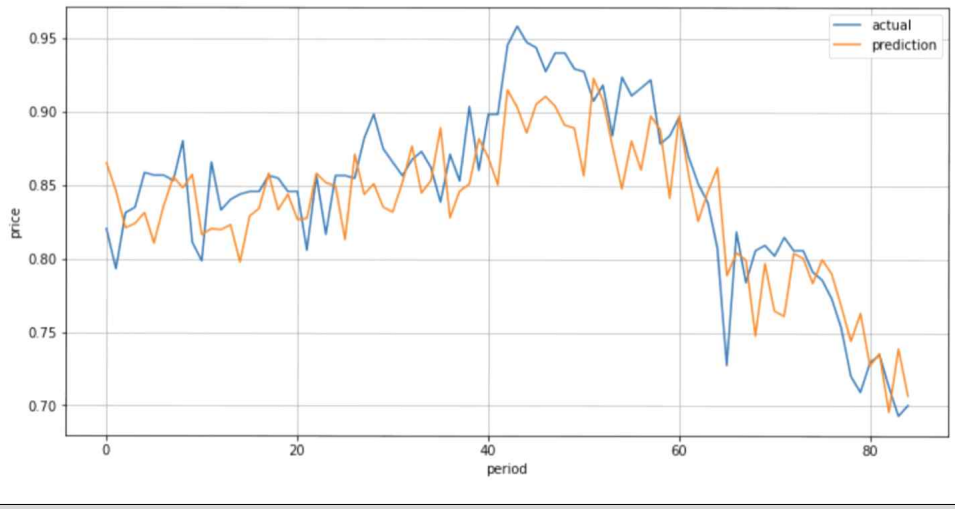
	Case1	Case2	Case3	Case4	Case5	Case6	Case7
MSE	0.0062	0.0057	0.0076	0.0065	0.0052	0.0066	0.0084
MAE	0.0062	0.0595	0.0684	0.0636	0.0567	0.064	0.0711
RMSE	0.0334	0.0383	0.0539	0.0318	0.0355	0.0365	0.0457
Adjusted R2	0.6784	0.5768	0.1647	0.7079	0.6360	0.6151	0.3991

[표2-] 조던1 레트로 하이OG 블랙모카 변수 조합별 평가 성능지표

2가지 제품을 모델에 적용한 성능 지표들을 종합적으로 고려해보았을 때 가장 최적의 성능을 내는 변수 조합은 Case1에 해당하는 [가격, 가격변동량, 검색량, 종가]임을 확인할 수 있다. 해당 조합을 적용했을 때의 두 가지 제품 데이터에 대한 가격 예측 그래프는 다음과 같다.



나이키 x 사카이 LD와플 블루 멀티



조던1레트로하이OG블랙모카

[표2-] Cas1 변수 조합을 적용한 예측 결과 그래프

2.4.4 AdaBoost-GRU 앙상블 단일변량, 다변량 모델 성능 비교

앞서 2.4.2에서 구현한 단일변량 모델과 전 단계에서 구현한 다변량 모델 중 가장 성능이 좋은 모델의 성능을 비교한 결과이다.

	단일변량	다변량
MSE	0.0069	0.0062
MAE	0.0654	0.0062
RMSE	0.0335	0.0334
Adjusted R2	0.6764	0.6784

[표2-] 나이키 x 사카이 LD와플 블루 멀티 단일변량, 다변량 결과 성능지표

	단일변량	다변량
MSE	0.0741	0.069
MAE	0.2148	0.2092
RMSE	0.0836	0.0576
Adjusted R2	0.8189	0.9138

[표2-] 나이키 x 사카이 LD와플 블루 멀티 단일변량, 다변량 결과 성능지표

다변량 입력값을 사용해서 AdaBoost-GRU 앙상블 모델 구현을 진행한 결과 단일변량 모델링을 진행했을 경우보다 mse, mae, rmse, adjusted R2 값이 모두 조금씩 개선되었다. 즉, 실제값과 예측값의 차이를 나타내는 정확도 측면이나 해당 예측의 설명력을 보여주는 성능 지표 모두 단일변량보다 다변량을 활용했을 때 더욱 향상된 결과를 보였다. 결론적으로 단일변량인 리셀 가격 데이터에 의존해서 예측을 수행하는 것보다 다변량 시계열 데이터를 사용하여 예측을 모델을 수립하고 예측을 수행했을 때 모델의 성능 및 예측력이 개선됨을 확인할 수 있었다.

제 3 장 결론

제 1 절 리셀 가능 여부 예측 모델 결과분석

본 연구에서는 한정판 스니커즈의 특성 및 거래 가격 데이터를 수집하여 발매일부터 발매 89일 뒤까지 총 90일에 대한 리셀 가능 여부를 예측하는 딥러닝 모델을 구현하였다. 해당 모델은 딥러닝 기법 중 입력층과 출력층, 그 사이에 여러 개의 은닉층을 사용한 DNN(심층 신경망)을 활용하였다. 또한, 데이터의 모든 특성(feature)을 분석에 반영하는 것보다 예측이 미치는 영향력이 큰 변수를 선별하여 반영했을 때 모델의 성능이 향상될 것이라는 가정하에 로지스틱 회귀모델을 활용해 특성별 coefficient(계수)를 도출하였다. 다양한 파라미터와 모델 구성을 조합한 실험 결과 가장 좋은 성능을 보인 모델의 구성 및 파라미터와 성능은 다음과 같다.

입력 변수	라인, col_num(사용 색상수), color3, color5, converse, nike, jordan 콜라보(유명도), 콜라보 유무
모델 구조	7개 층으로 구성 입력층(노드 수: 18) 은닉층(노드수: 18, 32, 16, 32, 16) 출력층(노드수: 2)
모델 파라미터	batch_size = 10, epochs = 200, validation_split = 0.2, earlyStopping 옵션의 patience = 30
모델 성능	최고 정확도 84.7% / 최고 F1 score 83.6% 최저 정확도 64.1% / 최저 F1 score 62.5% 평균 정확도 및 F1 score 70%

[표3-]

실험 결과, 전체 16개의 특성을 모두 활용한 모델의 최고 정확도는 77% 수준이었지만, coefficinet 기반으로 다양한 조합을 실험한 결과 16개 중 9개의 특성을 선별하여 모델 입력 변수로 활용한 모델은 최고 정확도 84% 수준이었다. 이를 통해 모든 변수를 활용한 모델보다 영향력이 큰 변수를 선별하는 과정을 거쳐 수립, 구현된 모델이 더 좋은 성능을 발휘한다는 가정이 타당함을 증명하였다.

또한, 기존 연구에서 10가지 머신러닝을 사용한 예측 모델의 결과와 비교를 진행하였다. 기존 논문에서 모든 일자에 해당하는 모델 성능이 아닌 일부 모델의 결과만을 공개하였기 때문에 전체 결과가 아닌 발매 당일과 발매 후 30일 차 예측 모델에 대한 성능을 비교하였다.

사용기법	정확도	순위	F1 score	순위
Bagging	73.5%	3	89.6%	1
Ridge	74.4%	4	88.3%	2
Logistic Regression	73.6%	2	88.3%	2
AdaBoost	76.2%	7	88.1%	4
Random Forest	74.4%	1	85.3%	5
ExtraTrees	73.1%	8	83.3%	7
K-Nearest Neighbors	72.3%	11	82.9%	8
Decision Tree	73.0%	9	82.4%	9
Support Vector Machine	69.9%	5	81.6%	10
Multi-Layer Perceptron	45.9%	10	81.6%	10
Deep Neural Network	84.7%	6	83.6%	6

[표3- 발매당일(day0)의 성능 비교]

사용기법	정확도	순위	F1 score	순위
Bagging	73.5%	5	77.6%	1
Ridge	74.4%	2	77.1%	2
Logistic Regression	73.6%	4	76.7%	3
AdaBoost	76.2%	1	76.0%	4
Random Forest	74.4%	3	76.0%	4
ExtraTrees	73.1%	6	75.7%	6
K-Nearest Neighbors	72.3%	8	70.5%	8
Decision Tree	73.0%	7	68.7%	9
Support Vector Machine	69.9%	10	66.6%	10

Multi-Layer Perceptron	45.9%	11	nan	-
Deep Neural Network	70.8%	9	72.7%	7

[표3- 발매 후 30일 차(day30)의 성능 비교]

해당 연구에서 구현한 모델 중 발매 당일에 해당하는 모델은 정확도 기준으로 전체 11가지 모델 중 6번째, F1 score 기준으로도 역시 6번째 성능을 기록하였다. 발매 후 30일 차에 해당하는 모델의 경우 정확도 기준 9번째, F1 score 기준 7번째로 높은 성능을 기록하였다.

기존 연구 결과와 비교했을 때 해당 연구의 DNN 모델이 가장 뛰어난 모델이라고 할 수는 없지만, 그동안의 리셀 시장 연구에서 한 번도 활용되지 않았던 딥러닝 기법을 활용한 연구를 수행했다는 점, 최근에 창립된 플랫폼의 데이터를 활용하여 데이터셋 규모 자체에 작다는 한계에도 불구하고 머신러닝 모델과 거의 동등한 수준의 성능을 발휘하는 모델을 구현하였다는 점, 그리고 1개월 예측만이 가능한 기존 연구에 비해 3개월이라는 긴 기간을 평균 약 70%의 정확도로 예측할 수 있다는 점에서 본 연구를 긍정적으로 평가할 수 있고, 유의미한 연구 결과를 얻었다고 분석된다.

제 2 절 리셀 가격 예측 모델 결과분석 및 최종 모델 확정

해당 연구에서는 3가지 딥러닝 모델을 활용하여 가격 예측을 진행한 결과를 토대로 가장 성능이 뛰어난 딥러닝 기법을 선정하고 해당 기법을 활용하여 단일변량, 다변량 활용 예측 모델의 성능을 비교, 분석하는 연구를 제안하였다.

먼저 단일변량 시계열 예측으로 LSTM, GRU, AdaBoost-GRU 모델을 본 결과 성능의 평가 지표인 mse, mae, adjusted R2 값은 LSTM, GRU 모델이 우수했지만, 그래프를 확인했을 때 불안정하고 변동성이 큰 시계열 데이터의 성질에 의해 직전 날의 가격과 비슷하게 예측하는 Shifting 현상이 발견되었다. 이런 경우 성능 지표에서 높은 수치를 보이더라도 좋은 성과를 보여도 해당 수치가 실질적인 예측력을 대변한다고 보기 어렵고 실제로 의미 있는 예측을 하고있다고 볼 수 없다. 따라서 이런 Shifting 현상 없이 학습을 토대로 예측치를 도출하는 AdaBoost-GRU 모델을 선정하여 다음 분석을 수행하였다.

다양한 변수 조합의 적용한 모델 실험 결과 중 가장 성능이 높았던 상위 2가지 조합은 가격, 가격 변동량, 포털 검색량, 주식 종가를 활용한 경우와 가격, 가격 변동량, 주식 종가, 소비자 물가 데이터를 활용한 경우이다. 자세한 성능 지표 값은 다음과 같다.

입력변수	MSE	MAE	RMSE	adjusted R2
가격, 가격 변동량, 포털 검색량, 주식 종가	0.006	0.006	0.03	0.68
가격, 가격 변동량, 주식 종가, 소비자 물가 데이터	0.008	0.07	0.05	0.16

[표3-]조던1 레트로하이 OG블랙모카

입력변수	MSE	MAE	RMSE	adjusted R2
가격, 가격 변동량, 포털 검색량, 주식 종가	0.006	0.2	0.05	0.91
가격, 가격 변동량, 주식 종가, 소비자 물가 데이터	0.007	0.21	0.05	0.92

[표3-]나이키 x 사카이 LD와플 블루 멀티

가격, 가격변동량, 검색량, 주식 종가 데이터를 입력값으로 사용한 모델링 결과를 보면 조던1 레트로하이 OG블랙모카 제품에서 mse, adjusted R2 값이 0.006, 0.68 정도의 상대적으로 좋은 성능이 도출되었으며, 나이키 x 사카이 LD와플 블루 멀티 제품에서도 mse, adjusted R2 값이 각각 0.06, 0.91 수준으로 정확도나 설명력 측면에서 좋은 성능을 보인다고 할 수 있다. 또한, 가격, 가격변동량, 주식종가, 소비자 물가 데이터를 활용한 모델링에서는 나이키 x 사카이 LD와플 블루 멀티의 예측 결과에서 adjusted R2 값이 가장 높게 나왔지만 mse와 mae 부분에서 가격, 가격변동량, 검색량, 주식 종가의 입력값을 가지는 모델링 결과보다 정량적인 정확도가 떨어지므로 최적의 조합으로 선정하기에는 어려움이 있다고 판단하였다. 그 외의 나머지 변수 조합을 통한 모델 실험 결과들에서도 가격, 가격변동량, 검색량, 주식 종가를 입력값으로 사용한 모델보다 성능이 떨어졌다. 따라서 결론적으로 다변량 시계열 분석에서 가장 최적의 결과를 도출할 수 있는 입력 데이터 조합은 리셀 거래 가격, 가격변동량, 포털 검색량, 주식 종가 데이터이다.

결론적으로 3가지 딥러닝 모델 중에는 AdaBoost-GRU 모델에 리셀 거래 가격, 가격변동량, 포털 검색량, 주식 종가 데이터 조합을 활용한 경우의 모델이 가장 최상의 예측 성능을 기록하였고, 단일변량보다 다변량을 활용할 경우 성능이 개선될 것이라는 가정 역시 증명해낼 수 있었다.

제 3 절 본 연구의 최종 결론 및 의의

본 연구에서는 한정판 스니커즈의 특성 데이터를 활용한 리셀 가능 여부 예측과 리셀 거래 가격 데이터를 비롯한 다양한 사회·경제적 지표 데이터를 복합적으로 활

용한 리셀 가격을 예측 모델을 구현해보는 연구를 제안하였다. 이를 위해 먼저 국내 1위 한정판 스니커즈 플랫폼 'KREAM'으로부터 약 480개의 제품 특성 데이터와 4만 5천 건의 거래 데이터, 2가지 제품의 전체기간 거래 데이터를 수집하였다.

리셀 가능 여부 모델의 경우 수집된 데이터를 전처리한 결과 총 16개의 feature로 정리하였다. 해당 결과를 바탕으로 로지스틱 회귀 예측 모델을 활용하여 예측 결과에 각 변수가 미치는 영향력을 확인함으로써 주요 변수들을 선정하였으며, 딥러닝 기법인 DNN 모델을 활용해 리셀 가능 여부 예측을 위한 학습 및 분석, 개선을 통해 최종 모델을 수립하고 결과를 도출하였다. 최종적으로 test set을 90개의 모델에 적용하여 성능을 확인할 결과 최대 약 84%의 정확도, 83%의 F1 score로 리셀 가능 여부를 예측할 수 있었다.

리셀 가격 예측 모델의 경우 LSTM을 비롯해 GRU, AdaBoost-GRU 앙상블 모델을 활용하여 시계열 가격 예측을 수행하였다. 다양한 딥러닝 기법들을 적용한 결과 단일 변량인 리셀 가격만을 가지고 예측을 진행했을 때보다 다양한 입력 변수를 추가한 다변량 모델이 더욱 좋은 성능을 발휘함을 확인할 수 있다. 또한, 각 모델들에 다양한 입력 변수 조합을 시도해보며 성능을 비교해본 결과, 가격, 가격변동량, 포털 검색량, 주식 종가 데이터를 입력 변수로 사용한 AdaBoost-GRU 앙상블 모델이 가장 좋은 성능의 모델이라는 연구 결과를 얻었다.

본 연구의 의의는 다음과 같다. 첫째, 딥러닝 기법을 리셀 시장 분석에 처음 도입했다는 점이다. 기존에 이루어진 스니커즈 리셀 시장 연구에서는 데이터 부족, 변수 활용 방식 등의 이유로 머신러닝 기법만을 활용하여 연구를 수행하였다는 한계점이 있었다. 본 연구에서는 이전까지 딥러닝이 적용되지 않았던 분야에 DNN, LSTM을 비롯한 다양한 딥러닝 기법들을 적용해 리셀 가능 여부와 리셀 가격 예측에 대한 분석을 시행했다는 점에서 의의가 있다.

둘째로 기존 연구들과는 다르게 요인 분석을 통한 주요 요인을 선별해 모델을 설계하였다는 것이다. 변별 없이 모든 특성을 모델의 입력 변수로 활용할 경우 고차원 입력 데이터, 유의미하지 않은 변수의 개입 등으로 인한 모델의 성능 저하 문제가 발생할 수 있다. 본 연구에서는 이러한 문제를 해결하고 예측의 정확도를 높이기 위해 리셀 가능 여부와 리셀 가격에 대한 예측 모델 설계에서 로지스틱 회귀 모델을 활용하거나 변수들에 대한 다양한 조합을 사용한 실험을 통한 성능 개선을 수행하였다. 결론적으로 모든 특성을 반영했을 때보다 유의미한 변수를 선별하여 활용함으로써 정확도를 더욱 향상시킬 수 있었다.

셋째로 리셀 가능 여부에 대한 예측 모델의 경우 3개월 상당히 긴 예측 기간을 예측한다는 것이다. 발매 시기가 각 운동화들마다 모두 달라 기존에 이루어졌던 연구들에서는 짧은 기간의 예측만 진행했다는 한계를 지니고 있었지만, 본 연구에서는 데이터 수집 및 전처리 과정에서 발매 일자로부터 3개월 이상의 거래 데이터가 존

재하는 제품들만 선별하여 데이터셋을 구축하여 예측 기간에 있어 본 연구만의 차별성을 지닐 수 있었다.

넷째로 일반적으로 단일변량인 가격만을 가지고 시계열 예측을 진행하는 기존의 분석 방식들과는 다르게 다변량 입력값을 통해 분석을 수행했다는 것에 의의가 있다. 가격 변수만을 사용해서 시계열 예측을 진행하면 과거 바로 전날의 값을 따라서 예측하는 경향이 생기므로 mse, mae, rmse 등의 평가 지표 값에서 성능이 좋은 것처럼 도출된다. 하지만 대다수의 예측 결과 그래프를 확인해보면 실제값과 비교해서 옆으로 조금씩 밀린 듯한 형태의 결과가 존재했고, 이러한 결과는 의미 있는 결과라고 볼 수 없다. 따라서 본 연구에서는 다변량 입력값을 통해 단순히 과거 가격 변화에만 의존하는 것이 아닌 다양한 수치를 기반의 의미 있는 예측 결과를 도출하였으며, 다양한 입력변수 조합의 비교를 통해 정확도를 더욱 향상시켰다.

제 4 절 한계점 및 향후 연구방향

본 연구의 한계점은 다음과 같다. 첫째로 절대적인 데이터 양의 부족이다. 이것은 국내에서 리셀이 주목받고, 스니커즈 리셀 시장이 확대되기 시작한 지 얼마 되지 않았고 데이터셋을 수집하기 위해 본 연구에서 활용했던 'KREAM'의 경우 2020년 창립된 플랫폼이기 때문에 그 이후에 출시 및 거래된 제품들로 데이터셋을 구축할 수밖에 없었다. 가격예측에 활용한 거래 데이터 역시 플랫폼이 생겨난 이후 거래 내역만 확보할 수 있어 긴 시간의 시계열 데이터를 활용하지 못하였다. 하지만, 시장 내 거래가 활발하게 이루어지고 있는 만큼 향후 더 많은 제품 및 거래 데이터가 축적된다면 모델의 성능을 더욱 향상시킬 수 있을 것이라 기대된다. 둘째로 스니커즈라는 한 품목에 대해서만 분석을 진행했다는 점이다. 리셀 시장에서는 스니커즈 외에도 의류를 비롯한 악세사리, 명품 등 다양한 품목이 거래되고 있다. 하지만 본 연구에서는 현재 국내 시장에서 가장 크고 활발한 리셀 분야인 스니커즈 시장을 분석 대상으로 선택하였다. 각 품목에 따라 영향을 미치는 요인이나 가격 추이가 본 연구의 대상과는 많이 다를 수 있기 때문에 추후 리셀 시장 전체로 범위를 넓혀 연구 진행이 필요하며, 이를 위해서는 앞서 언급한 것과 같이 충분한 데이터 확보가 중요할 것이다.

참 고 문 헌

[1] 이윤화, “[리셀의 세계]①한정판 되판다고 다 돈이 되는 건 아닙니다”, 이데일리, 2020.05.22.
[2] 신은빈, “‘한정판’딱지 붙으면 몇 배씩 뛴다...MZ세대의 짭짤한 제테크 ‘리셀’”, 매일경제, 2021.12.17.
[3] “MZ세대 ‘스니커테크’ 열풍”, 동아일보, 2021.12.11.
[4] “소유보단 경험” 유통가 리셀 시장 급부상, 이유는?, 아주 경제, 2022.01.16.
[5] 이재영, 「스니커즈 리셀(Resale) 현상 분석(제1보)」, 『패션과 니트』, pp. 67-80(14), 한국니트디자인학회, 2021.
[6] 김누리, 「머신 러닝 기법을 활용한 한정판 운동화 리셀 여부 예측 및 수익성 평가」, 서울대학교 석사학위논문, 2020.
[7] 윤현섭, 강주영, 「XGBoost 모형을 활용한 가격 상승 요인 탐색 및 예측을 통한 리셀 시장 진입 장벽 해소에 관한 연구」, 『한국 전자거래학회지』, pp. 155-174(20), 한국전자거래학회, 2021.
[8] 이도연, 장병희, 「딥러닝을 이용한 음악흥행 예측모델 개발 연구」, 『한국콘텐츠학회논문지』, pp. 10-18(9), 한국콘텐츠학회, 2020.