**2020 2학기 소프트웨어 응용 텀 프로젝트**

**편의점 위치 추천/분석 서비스 보고서**

2조

컴퓨터과학부 2016920010 김현구

컴퓨터과학부 2016920024 엄현식

컴퓨터과학부 2016920030 이근희

**Ⅰ. 개요**

**1) 선정 배경**

은퇴 이후의 5, 60대 뿐 아니라 2, 30대의 창업에도 편의점 창업의 비중이 크게 증가하여 현재 편의점 시장은 레드 오션으로, 정확한 분석 없이 뛰어든 후 얼마 지나지 않아 폐업하는 사례가 자주 발생한다. 프랜차이즈 편의점에 한하여 자체적으로 분석 서비스를 제공하고 있으나, 창업 유치를 위해 수익성 측면에서 과대 포장을 하는 등 편파적인 모습을 띄고 있어 이 또한 신뢰하기 어려우며, 구체적인 지역을 추천하지 않아 부정확한 판단의 근거로 작용하고 있다. 따라서, 예비 자영업자들의 정착을 위해 공공데이터를 기반으로 정확한 예측 및 창업 위치 추천을 제공하는 서비스의 필요성을 느끼게 되었다.

**2) 프로젝트 목표**

본 프로젝트는 서울특별시에서 제공하는 공공 데이터를 활용하여 지도 학습을 통해 행정동별 편의점 매출과 폐업률을 예측하여 사용자에게 가장 적합한 창업 지역을 추천하는 서비스 구축을 목표로 한다.

**Ⅱ. 구축 과정**

**1) 데이터 정제**

|  |  |
| --- | --- |
| 수집 데이터셋 | 출처 |
| 서울시 지역별 편의점 매출 (2분기) (2018-2020) | 우리마을가게 상권분석 서비스 |
| 서울시 지역별 편의점 개/폐업률 (1,2분기) (2018-2020) | 우리마을가게 상권분석 서비스 |
| 서울시 지역별 평당 임대시세 (1,2분기) (2018-2020) | 우리마을가게 상권분석 서비스 |
| 서울시 지역별 편의점 점포수 (1,2분기) (2018-2020) | 우리마을가게 상권분석 서비스 |
| 서울시 지역별 슈퍼마켓 점포수 (1,2분기) (2018-2020) | 우리마을가게 상권분석 서비스 |
| 서울시 지역 업종별 점포수 (1,2분기) (2018-2020)  (외식업, 주유소, 피시방, 당구장, 노래방, 독서실) | 우리마을가게 상권분석 서비스 |
| 행정동별 서울생활인구 (1-6월) (2018,2020) | 서울 열린데이터광장 |
| 서울시 지역별 주거인구, 직장인구 (1,2분기) (2018-2020) | 우리마을가게 상권분석 서비스 |
| 서울시 행정구역별 주민등록 인구 (남, 여) (1-6월) (2020) | 행정안전부 주민등록 인구통계 |
| 서울시 행정구역별 주민등록 인구 (연령별) (1-6월) (2020) | 행정안전부 주민등록 인구통계 |
| 서울시 세대원수별 세대수 (동별) 통계 (1,2분기) (2020) | 서울 열린데이터 광장 |
| 서울시 행정동 별 지하철역 수 | 네이버지도, 도로명주소 안내시스템 |

**[표 1] 사용한 데이터셋**

데이터 수집에 앞서, 어떤 데이터를 사용할 것인지 선정하는 것이 가장 중요하다고 판단하여, 편의점 매출과 관련된 논문 참고 및 추론 바탕으로 데이터를 수집하였다. 하지만 인구 관련 데이터를 제외하고는 제공되는 데이터가 매우 한정적이었다. 따라서 인구 관련 데이터에 중점을 두어 연령별 통계, 세대별 통계 등을 수집하였고, 인구 외 데이터로는 업종별 점포 수, 임대 시세, 지하철 역 수 등의 데이터를 수집하였다.

또한 지도 학습을 위해 매출과 폐업률에 대한 정보가 본 프로젝트에서 중요도가 가장 높았으나, 그 만큼 시장에서 역시 중요한 데이터였기 때문에 쉽게 구할 수 없어 데이터의 범위가 우리 마을 가게 상권분석서비스에서 제공하는 2018, 2019, 2020년 각 2분기의 서울시 행정동별 매출로 한정되었다. 나머지 데이터셋은 행정동 별 데이터가 대체로 획득가능한 최소 단위였기 때문에 위 데이터셋들을 통해 최적의 행정동을 분석, 추천하는 프로젝트로 진행하도록 결정하였다.

**2) 데이터 정제**

데이터 수집 완료 이후 데이터 파일이 주로 엑셀과 csv 파일 형태로 있어, 해당 파일들의 필요한 부분만 쉽게 제거하기 위해 파이썬 언어와 pandas 프레임워크를 사용하여 데이터를 정제했다. 이를 통해 데이터에서 필요한 부분만을 따로 추출하거나 5살 단위의 연령 구분을 10살 단위의 연령 구분으로 바꾸는 등의 데이터 통합을 진행하고, 2018년부터 2020년까지의 3개의 파일로 나눠 저장했다. 이후 모델 학습에 사용될 데이터 통합을 위해 지역명, 자치구, 행정동 명, 행정동 코드로 이루어진 지역 코드 파일을 미리 제작하고 앞서 따로 정제한 통계 파일들을 행정동 코드나 자치구와 행정동 문자열 값을 통해 병합을 진행했다.

이러한 과정에서 강북구의 경우 2018년에서 2019년으로 넘어가며 기존 번 제1동, 번 제2동, 번 제3동, 수유 제1동, 수유 제2동, 수유 제3동이 동일한 관할구역을 따르지만 각각 이름이 번1동, 번2동, 번3동, 수유1동, 수유2동, 수유3동으로 변경되었고 행정동 코드에서도 변화가 생겼다. 또한 구로구의 경우 2019년에서 2020년으로 오류2동에서 항동이 분리되며 추가적인 행정동이 생겼다. 이 경우 번동와 수유동은 이름과 행정동 코드의 변화를 제외하고는 달라진 점이 없어 이름과 행정동 코드를 통일하여 파일을 통합했고, 구로구 항동의 경우 대부분의 공공 데이터셋에서조차 항동이 분리되지 않고 제공되었기 때문에, 오류2동에 통합하여 데이터를 완성했다.

|  |  |
| --- | --- |
|  |  |

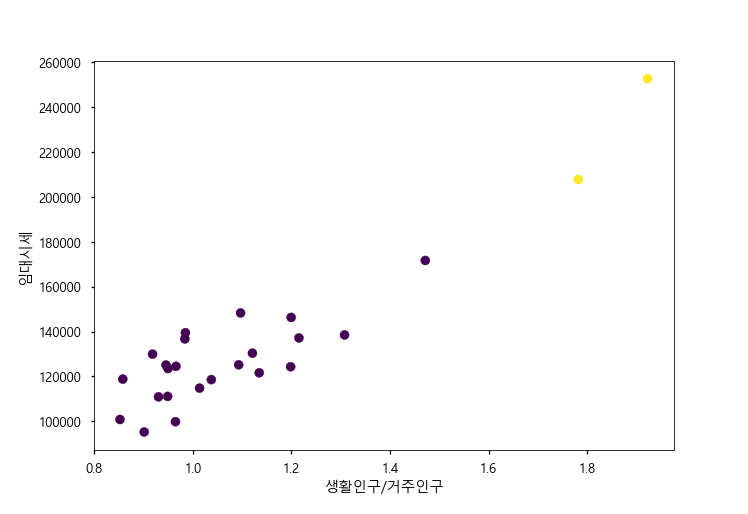
**[사진 1] 동 단위 데이터 병합**

**3) 모델 선정 및 학습**

매출과 폐업률 예측을 위한 회귀 모델은 scikit-learn 라이브러리에서 제공하는 random forest regressor와 PyTorch 라이브러리를 통해 hidden layer 수를 다양하게 하여 구축한 neural network 모델들간의 성능을 mean squared error를 기준으로 평가하여 가장 좋은 성능을 보인 모델을 채택하였고, 모델의 학습은 random forest regressor의 경우 자체적으로 제공하는 fit() 함수를 사용, neural network의 경우 learning rate를 1e-4로 하는 adam optimizer를 통해 학습을 진행했다.

학습 과정에서 많은 시행착오를 거치며 데이터셋과 모델을 계속하여 수정하면서 프로젝트를 진행하였다. 먼저, 많은 기간의 매출 데이터를 얻을 수 있을 것이라고 판단하여 행정동 단위가 아닌 구 단위의 매출, 폐업률 데이터를 통해 모델을 구축하고자 하였으나 확보할 수 있는 데이터가 3개년에 한정되어있어 데이터 수가 너무 적어 사용이 불가능한 수준의 오차가 발생해 행정동 단위의 데이터를 사용하여 모델을 구축하는 것으로 계획을 수정하였다.

다음으로, 주거지역과 상업지역은 매출 결정 요인에 차이를 보일 것으로 판단하여 인구 데이터를 기준으로 k-means 알고리즘을 통해 구 단위 클러스터링을 시도하였으나 다양한 기준을 사용하여도 가장 적합한 클러스터링이 23:2 수준에 그쳐, 역시 모델 학습에 있어 데이터 부족으로 인해 사용이 불가능한 수준의 오차가 발생해 사용하지 않았다.



**[사진 2] 클러스터링 결과**

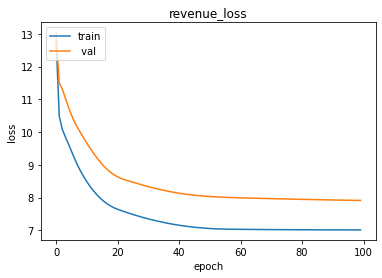
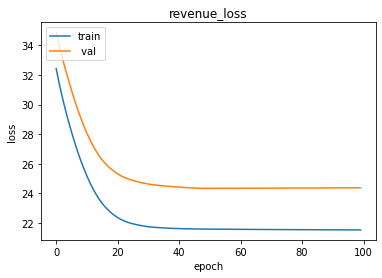
또한, scikit-learn에서 제공하는 feature\_importances를 통하여 전체 데이터가 아닌 중요 데이터만 모델 학습에 사용하고자 했으나 성능에 큰 변화가 없을 뿐 아니라 악화되는 경향을 보여 전체 데이터셋을 사용하였다.

더하여, 프로젝트 계획 단계에서는 2분기의 매출 데이터를 통해 학습한 모델에 3분기의 데이터를 입력하여 매출을 예측하는 서비스를 계획하였으나, 3분기의 데이터가 공개되지 않아 방향성 수정이 필요하여 1분기의 feature들로 2분기의 매출, 폐업률을 예측하는 것 또한 시도하였으나 역시 오차가 크게 나타나 후술할 부정확한 데이터셋 문제 해결에 프로젝트 중점을 두었다.

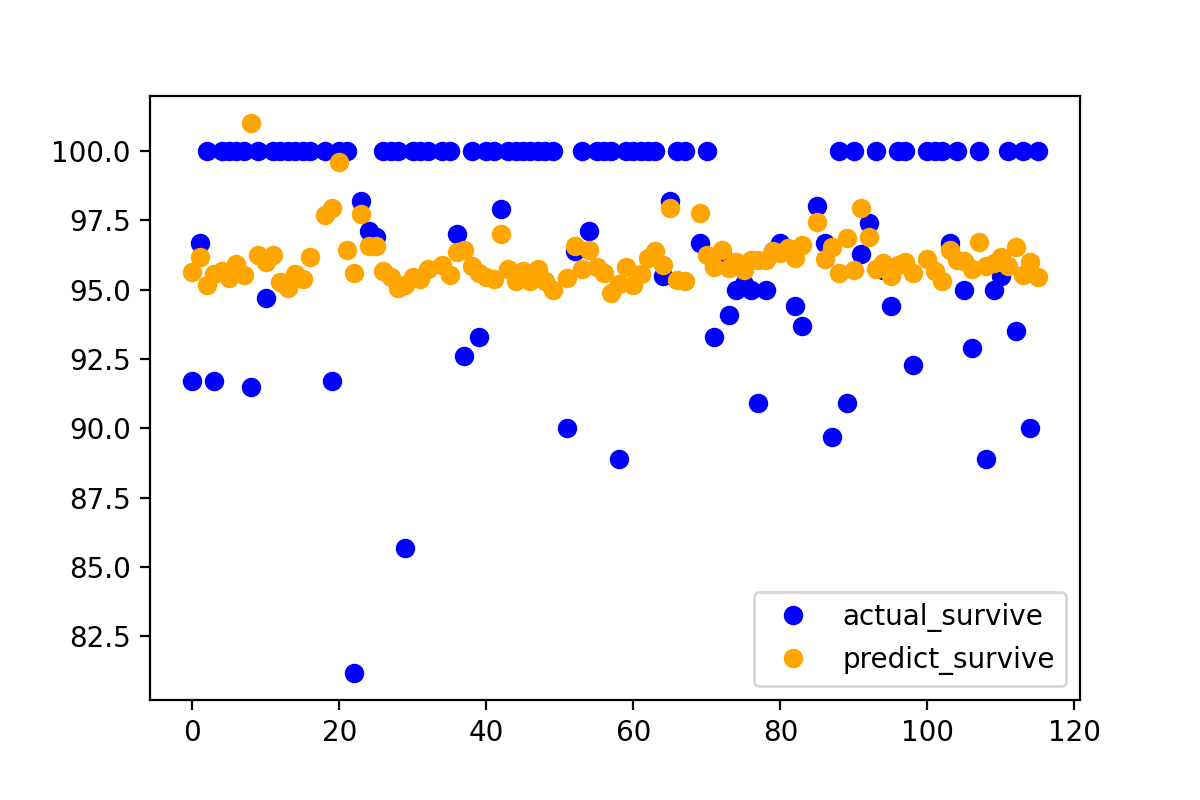
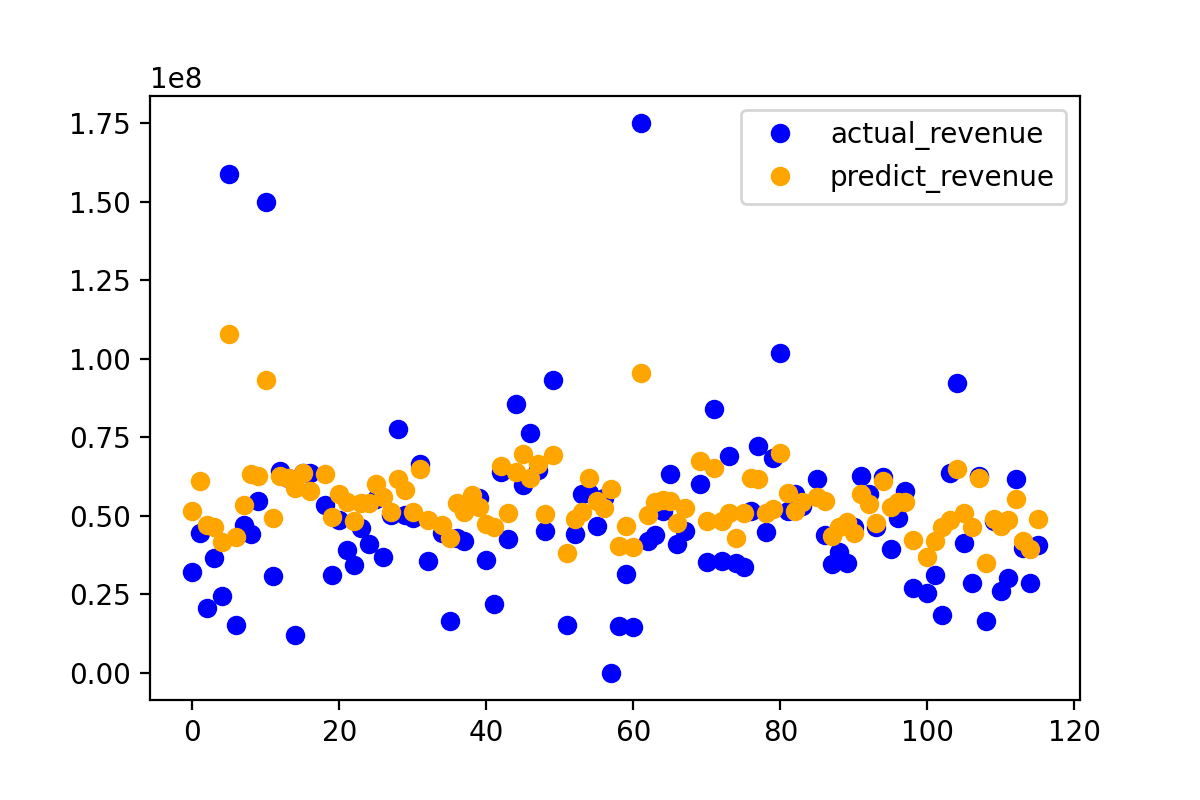
마지막으로, 폐업률 예측 모델 학습에 있어 해당 기간의 실제 값을 사용하였을때 오차가 크게 나타나 이전 분기의 데이터를 추가로 확보하여 데이터 변화량을 feature로 사용하여 약간의 성능 향상을 이루었다.

이러한 과정을 거쳐 최종적으로 26개의 features를 통해 매출과 폐업률을 예측하는 여러 모델을 구축하여 성능 비교 후 최종 모델을 선정하였다. 최종 모델 후보군으로 random forest regressor와 각각 0, 1, 2개의 hidden layer를 갖는 neural network를 비교하였으며, 2개의 hidden layer를 사용할 때 부터 성능이 매우 나쁘고, overfitting 역시 발생하여 그 이상 hidden layer를 사용하여 비교하지는 않았다.

매출을 예측하는 모델은 random forest 모델과 hidden layer가 없는 neural network 모델이 다른 모델에 비해 좋은 성능을 보였으며, train data와 validation data 분할에 따라 두 모델의 성능 우위에 변동이 있어 두 모델의 예측값의 평균을 사용하는 ensemble 방식을 활용하여 새로운 모델을 만들어 사용하였고, 이떄 mean absolute error는 12,227,989 (원)으로 나타났다.

폐업률을 예측하는 모델은 hidden layer가 없는 neural network 모델이 다른 모델에 비해 꾸준히 좋은 성능을 보여 해당 모델을 직접 사용하였으며, 이떄 mean absolute error는 3.77 (%p)로 나타났다.

**[사진 3] 매출(좌), 생존률(우) 에러 그래프**

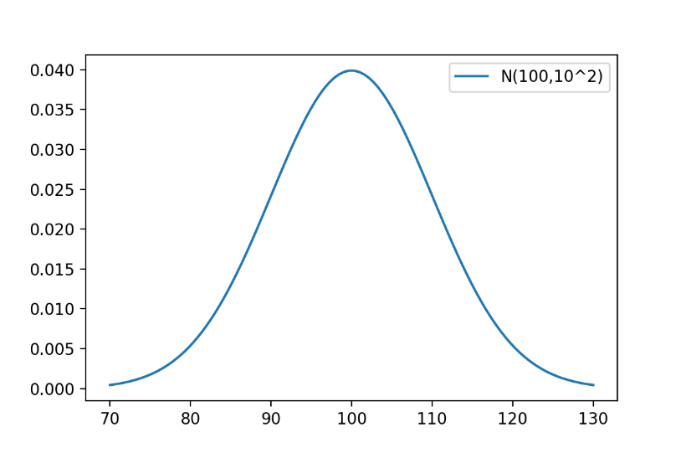


**[사진 4] 실제 매출(좌), 생존률(우)와 예측 비교 그래프**

**4) 점수 산정**

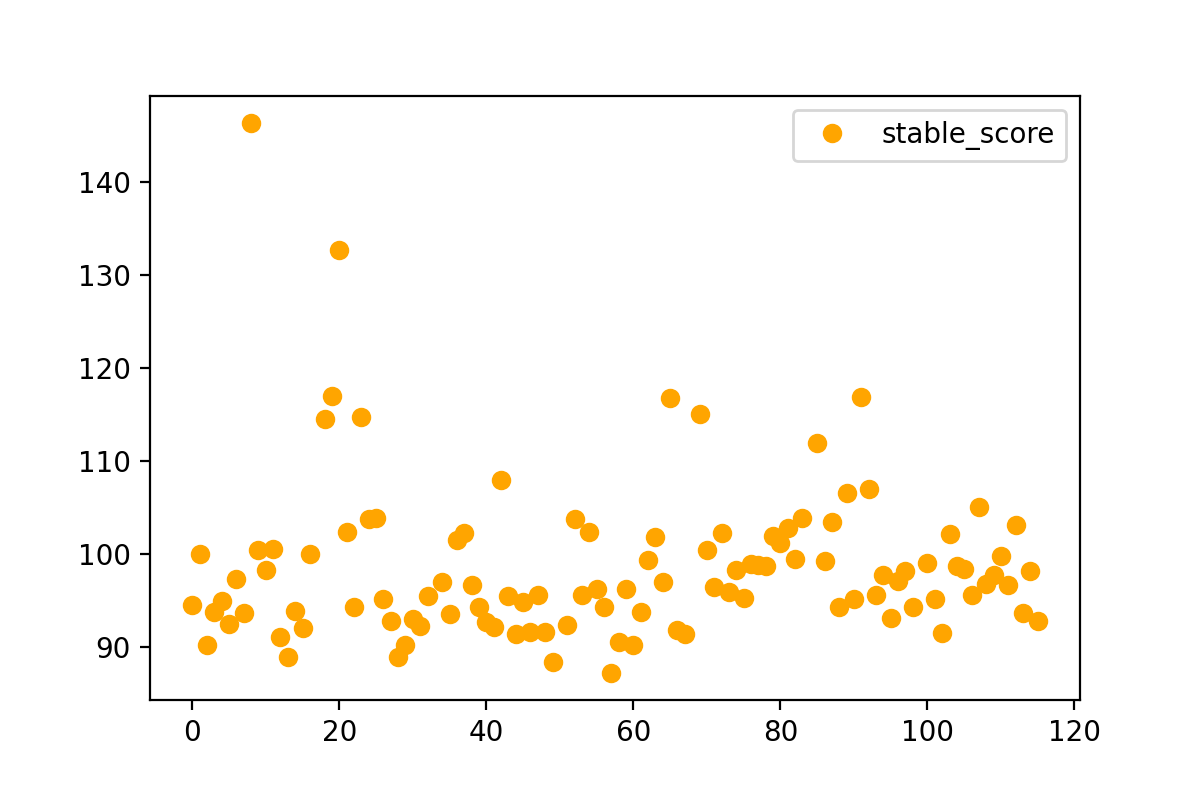
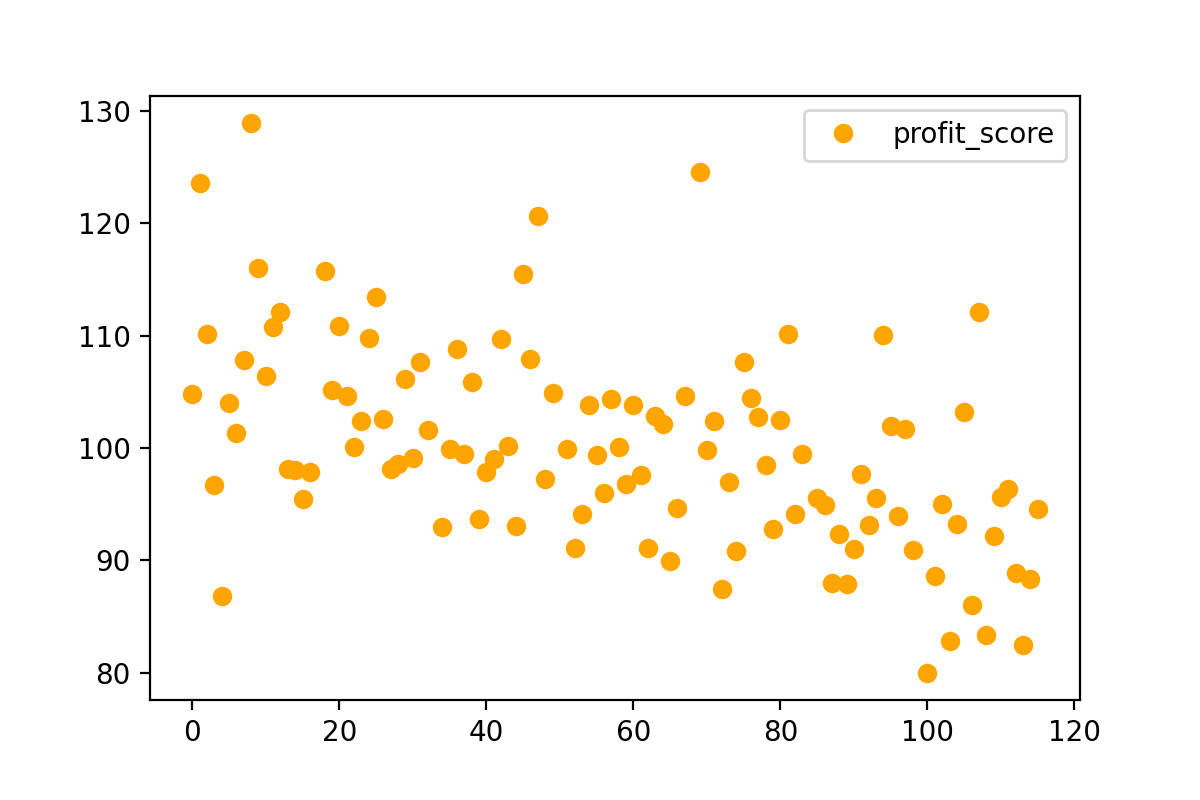
본 프로젝트는 사용자가 수익과 안정성에 대해 판단한 가중치를 입력받아 그에 맞는 행정동 추천을 목표로 하고 있어, 두 데이터의 비교 및 연산이 필수적이다. 그러나 매출은 원 단위, 안전성은 % 단위로 직접 연산이 불가능하여 표준 점수로 널리 사용되는 T점수를 계산하여 T점수에 사용자가 입력한 가중치를 곱한 뒤 두 값을 더하여 최종 점수를 산정하였다.

T-점수란 개별 데이터와 평균 사이의 거리를 통해 계산하는 점수로, 실제로 수능 표준점수 등 척도가 다른 값들을 비교하기 위해 널리 사용되고 있으며 본 프로젝트에서 사용한 계산 식은 다음과 같다.



**[사진 5] T-분포 그래프**

또한, 폐업률에 직접적으로 T점수를 적용할 경우 매출의 경우 점수가 가장 큰 지역, 폐업률의 경우 점수가 가장 작은 지역을 추천해야 하기 때문에 생존률(100 – 폐업률)이라는 지표를 추가로 만들어 점수를 산정하였다.



**[사진 6] 매출(좌), 생존률(우) 점수 분포**

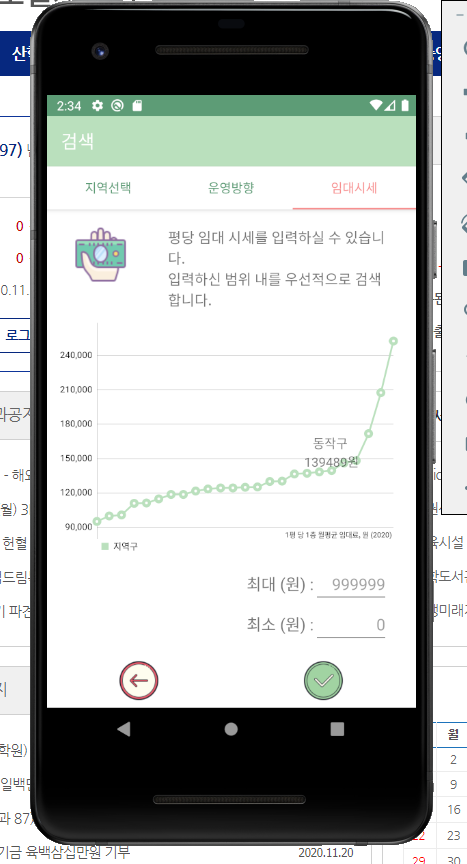
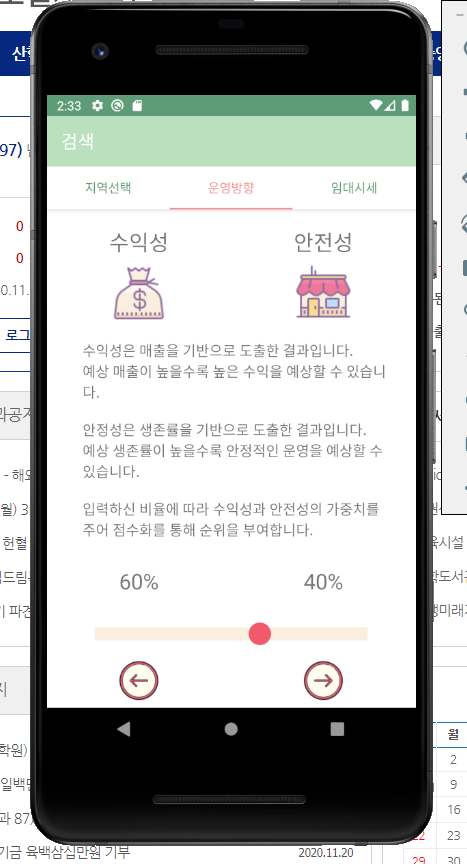
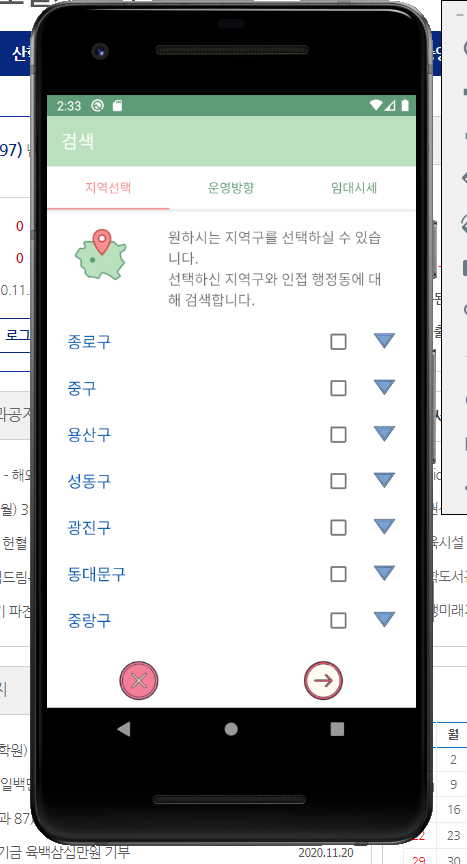
**5) 서비스 구현(구축)**

먼저 서비스를 구축함에 있어 우리가 예측한 데이터를 통한 점수만으로 사용자에게 추천하는 것이 아니라, 사용자의 입력을 추가적으로 받아서 작동하게 하였다. 사용자는 어플리케이션을 실행하면 관심 지역구, 매출과 생존률의 중요도 비율, 임대시세 범위 세 가지를 설정할 수 있다.

관심 지역구는 서울시의 지역구 중 관심있는 구역을 선택하는 것이다. 사용자가 선택한 지역구 내의 행정동과 지역구와 인접해 있는 타 지역구의 행정동 데이터를 그래프형식으로 DB에 저장해두어, 사용자에게 행정동을 추천할 때 해당 범위 내의 행정동만 추천하게 해준다.

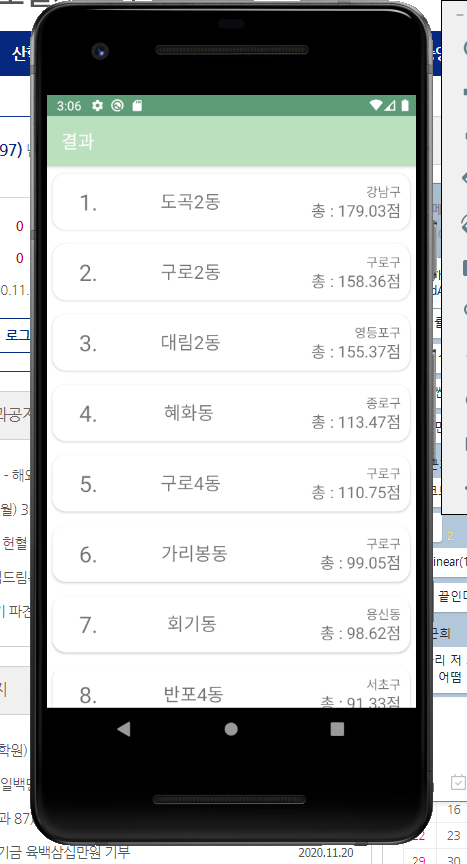
매출과 생존률의 중요도 비율은 수익성과 안정성에 따른 사용자의 선호도 비율을 입력받로록 하였다. 이는 이후 매출에 따른 점수와, 생존률에 따른 점수를 합산할때의 비율로 작용하여 사용자의 성향에 따라 변화를 일으킨다.

마지막으로 임대시세는 다수의 편의점이 1층에 위치하는 관계로, 지역구별 1평당 1층 월평균 임대료를 그래프로 제시하며, 사용자에게 평당 임대료의 상한과 하한을 설정할 수 있게 하였다.



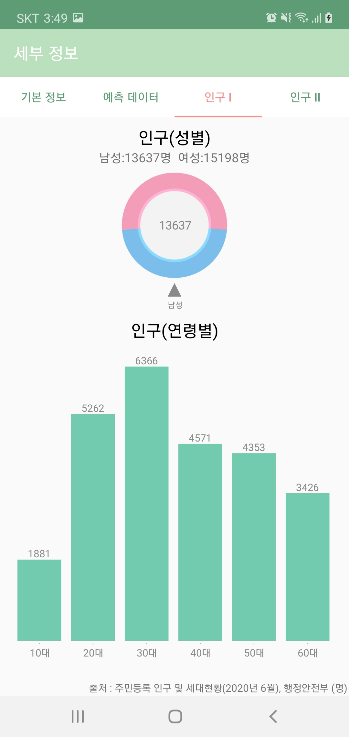
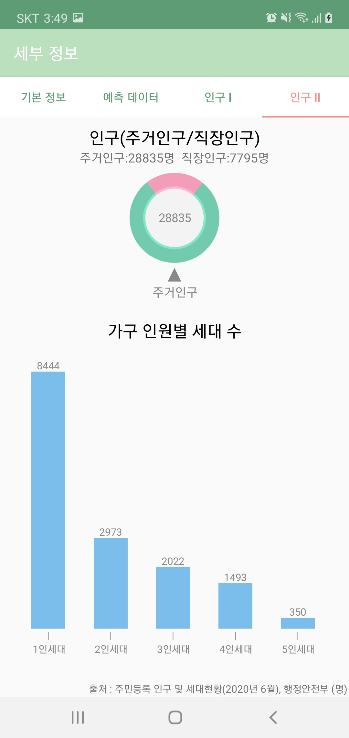
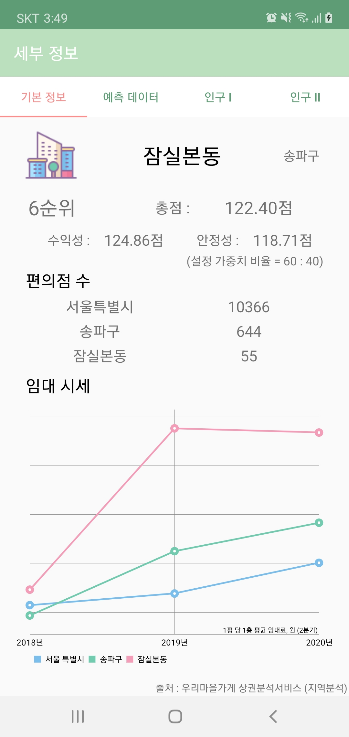
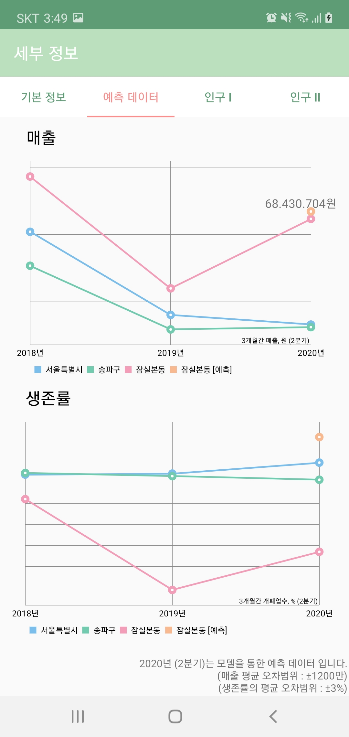
**[사진 7] 어플리케이션의 사용자 입력 화면**

이후 사용자의 입력을 모두 받고 나면, 위와 같이 사용자가 결정한 선택이나 범위를 분석에 적용하여 결과를 도출해낸다. 이 경우 미리 저장되어 있는 데이터베이스의 행정동들을 사용자의 선택을 바탕으로 걸러내거나 점수 총합을 계산한다. 행정동에 관련된 데이터, 전체 랭킹점수는 미리 어플리케이션에 데이터베이스 형식으로 내장되어 있다. 예측 알고리즘을 통하여 실시간으로 서버를 운영하지 못한 팀의 기술적인 한계와, 보유한 데이터의 제공 주기가 최소 1달 이상이기 때문에 실시간 업데이트를 하지 못하는 어려움으로 인하여 부득이하게 미리 데이터베이스를 모바일 어플리케이션에 내장하였다. 이후 나온 결과들을 내림차순으로 정렬하여, 사용자에게 표시하도록 하였다.



**[사진 8] 어플리케이션의 분석 결과 화면**

또한 추가적인 정보 제공을 위해 사용자가 해당 순위 아이템을 클릭할 경우, 해당 행정동의 정보들을 제공한다. 먼저 앞선 순위의 세부 점수와 해당 지역의 편의점 수, 임대 시세 변화 그래프를 보여주고, 추가적으로 매출과 생존률의 이전 데이터와 예측을 담은 그래프, 해당 행정동의 인구 관련 정보들을 제공한다.



**[사진 9] 어플리케이션의 추가 정보 제공 화면**

**Ⅲ. 한계 및 평가**

프로젝트의 주요 기능인 예측 모델의 오차는 1220만, 3.7 %p 수준으로 매우 크게 나타났는데, 이는 프로젝트에 작용한 많은 한계 요인들의 복합적인 작용에 의해 발생하였다.

먼저, 일부 데이터의 부재가 가장 큰 문제로 작용하였다고 생각한다. 본 프로젝트의 핵심 데이터인 매출과 매장의 폐업률 관련 데이터를 제공하는 곳을 찾을 수 없어 ‘우리 동네 상권분석 서비스’에서 일일이 검색을 통해 얻은 2018, 2019, 2020 3개년의 2분기 데이터만을 사용할 수 있었다.

또한, 개별 편의점의 지리적 특성이 반영되지 않은 것도 한계점으로 작용하였다. ‘입지요인이 편의점 성과에 미치는 영향에 대한 연구’(황규성, 2016)의 경남, 울산지역의 편의점에서의 하루평균매출액을 종속변수로 한 회귀분석에서는 중요 유의변수로 유동인구, 경쟁지점, 도로에 접하는 면의 수를 선정하였는데, 이 중 개별점포 주변에서 경쟁하는 경쟁지점이나, 매장의 가시성에 영향을 주는 길과 맞닿은 면의 수와 같은 입지요인, 즉 지리적 특성에 대한 데이터 부재로 본 프로젝트에서는 고려하지 못하였다.

더하여, 상권의 범위 역시 예측의 오차를 크게 만든 요인 중 하나로 평가하였다. ‘편의점의 상권 추정과 매출예측에 관한 연구’(이춘섭 2012)의 편의점 내점객 방문 소요 시간 분포도를 따르면 응답자 중 89%가 매장에 방문하는데 5분 이내로 소요하였다는 것을 알 수 있다. 따라서 해당 연구에서는 1차 상권 비율을 이동속도를 고려하여 150m로 설정하였지만 본 프로젝트에서는 편의점의 매출, 폐업률을 추정하는 단위가 그보다 훨씬 넓은 행정동 단위였기 때문에 실질적으로 내점객들에게 영향을 주는 개별 상권의 특성이 잘 반영되지 않았다고 고려된다. 이는 팀의 기술력과 정보력으로 수집할 수 있는 데이터셋들의 최소단위가 행정동별 단위였기 때문에 팀의 기술, 정보력 역시 한계점으로 평가할 수 있을 것 같다.

마지막으로, 데이터의 대표성 역시 한계점으로 작용했다. 2020년은 코로나 19로 인하여 기존의 사회와는 생활 양상, 경제활동이 많이 달라졌지만, 본 프로젝트에서 사용한 데이터셋으로 표현하기에는 한계가 있었다. 하나의 예시로 전농2동의 경우, 2019년 2분기 매출합이 대략 1억7천만원 정도지만, 2020년 2분기에는 약 1억 1천5백만원으로 감소한다. 이는 기존 매출의 45%인 7천5백만원 정도를 차지하던 10~20대 소비량이 4천만원으로 전년 대비 47% 감소한 것의 영향으로, 코로나로 인한 비대면 수업 등이 원인으로 추정된다. 이러한 특수한 사회적 상황을 단순 인구 데이터로는 대표할 수 없다는 점이 한계로 작용한다. 추가로 폐업률의 경우 기존에 편의점 수가 매우 적은 지역에서 1개의 편의점만 사라지더라도 폐업률이 매우 크게 나타나는 착시 현상이 발생하여 모델에 큰 영향을 준 점도 있다.

결국 이러한 한계점들의 복합적인 작용으로 인하여 프로젝트의 기대목적보다 훨씬 부정확한 예측을 제공하는 알고리즘을 만들었다고 평가한다. 추후 알고리즘의 학습을 위한 행정동별 매출, 폐업률 데이터의 누적과 보다 세부적인 범위의 데이터 제공, 개별 매장 데이터 확보가 이루어진다면 현재 모델보다 훨씬 신뢰성이 높은 추천 결과를 사용자들에게 제공할 수 있을 것이라고 판단한다.

**Ⅳ. 참고문헌**

이철환. (2012). **편의점의 상권 추정과 매출 예측에 관한 연구.** 석사학위논문, 건국대학교 부동산대학원, 서울.

황규성.,&장형진. (2016). 입지요인이 편의점 성과에 미치는 영향에 관한 연구. **부동산학보, 제64권,** 199-211.

**V. 첨부 코드**

* Data\_preprocess: 수집한 데이터 원본 일부가 담겨 있는 DataSet 폴더, 데이터 처리 후 정제, 통합한 파일들 일부가 담겨 있는 ProcessedDataSet 폴더, 데이터를 처리하는데 사용한 코드가 담긴 PythonCode로 구성
* Modeling: 데이터 학습, 예측 및 점수화 알고리즘 코드들이 담겨있는 폴더.

학습을 위해 추가적으로 데이터를 가공한 dataset폴더, 학습된 모델을 저장하는 models폴더, 결과 랭크파일을 저장하는 result 폴더, 코드들이 저장된 src폴더로 구성.

Src 폴더 내부 코드 구성은 모델 선정을 위해 테스트한 코드(model\_test)에 Revenue(매출), Closed(폐업률)로 구분

Revenue 폴더 내에 최종적으로 사용한 앙상블 모델 코드 = OurEnsemble

Closed 폴더 내에 최종적으로 사용한 단일 뉴럴 네트워크 코드 = SingleLayer

model\_training : 선정한 모델을 학습 후 저장하는 폴더

predict: 학습된 모델을 통해 실제 점수가 담긴 랭킹파일을 저장하는 폴더

* WaraConvApp: 모바일 어플리케이션 전체 프로젝트 폴더