2020년 2학기

소프트웨어응용

팀프로젝트 보고서

**편의점 위치 추천/분석 서비스 보고서**

2조

2016920024 엄현식

2016920030 이근희

2016920010 김현구

**Ⅰ. 개요**

**1) 선정 배경**

사회적으로 50, 60대들의 은퇴 이후 편의점 창업이 증가하고 있고, 20, 30대들의 창업에서도 편의점 창업이 증가함에 따라 경쟁이 과열되고 있다. 이러다 보니 상권 자체가 과열되어 폐업을 하는 경우도 적지 않게 발생한다. 물론 특정 편의점 브랜드에 한하여 편의점 창업을 시도하는 사람들에게 자체적으로 분석 서비스를 제공하고는 있으나, 이는 수익성 측면에서 과대 포장을 하는 등 편파적인 모습을 띄고 있다. 그러기에 자체적으로 현재 주어진 공공데이터를 기반으로 편의점 창업에 좀 더 적합한 지역을 추천하고, 다양한 정보를 제공하고자 이번 프로젝트를 계획하였다.

**Ⅱ. 구축 과정**

**1) 데이터 수집**

먼저 데이터 수집 과정에 앞서 예측을 한 부분은, 아무래도 편의점이라는 분야의 매출은 인구와 관계가 높다 생각하여 인구 관련 데이터를 중점적으로 수집하였다. 최종적으로 사용한 데이터셋은 아래와 같다.

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

이때 매출에 대한 정보가 본 프로젝트에서 중요도가 가장 높았으나, 그 만큼 시장에서 역시 중요한 데이터였기 때문에 쉽게 구할 수 없었다. 결국 구할 수 있는 데이터의 범위가 우리 마을 가게 상권분석서비스에서 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.



**[사진 1] 데이터 수집**

**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 2] 동 단위 데이터 병합**

**3) 모델 구축**

매출과 폐업률 예측을 위한 회귀 모델은 Scikit-learn 라이브러리와 PyTorch 라이브러리를 활용해 여러 모델을 구축해본 뒤 가장 좋은 성능을 보이는 모델을 채택하였으며, 성능 평가 지표로는 Mean Absolute Error를 사용했다.

1. Random Forest Regressor



**[사진 3] Random Forest Regressor**

Scikit-learn library를 활용해 구축하였으며, 여러 의사 결정 트리를 생성해 예측하는 앙상블 기법의 대표적인 모델이다. 본 프로젝트의 경우 매출 예측에서 가장 좋은 성능을 보여 매출 예측 모델로 사용하였다.

2. Single-Layer Perceptron



**[사진 4] Single Layer Perceptron**

Pytorch library를 활용해 구축하였으며, 머신 러닝의 가장 기본적인 모델이다. nn.Linear 모듈을 사용하였기 때문에 선형 회귀와 동일한 작동을 한다. 본 프로젝트의 폐업률 예측에서 가장 좋은 성능을 보여 폐업률 예측 모델로 사용하였다.

3. Multi-Layer Perceptron

Pytorch library를 활용해 구축하였으며, nn.Linear 모듈과 Tanh, Dropout을 층으로 쌓아 만들었다. 층수와 dropout의 parameter를 조정해 보았지만 overfitting 문제가 발생하고, 성능 향상 역시 관측되지 않아 본 서비스에서는 사용하지 않았다.

**[사진 5] Multi-Layer Perceptron**

4) 모델 학습

모델 학습 과정에서 여러 시행착오가 발생하여 데이터셋을 수정했다.  
 최초 모델 계획은 2018~2020년 3개년의 구별 매출 평균을 통해 모델을 만들어 일부 동별 매출을 validation set으로 활용하고자 하였으나, train dataset이 너무 적어 오차가 매우 크게 발생했다.

그래서 뭣이냐 동별 매출을 수집해서 라벨로 사용해 매출 예측으로 사용

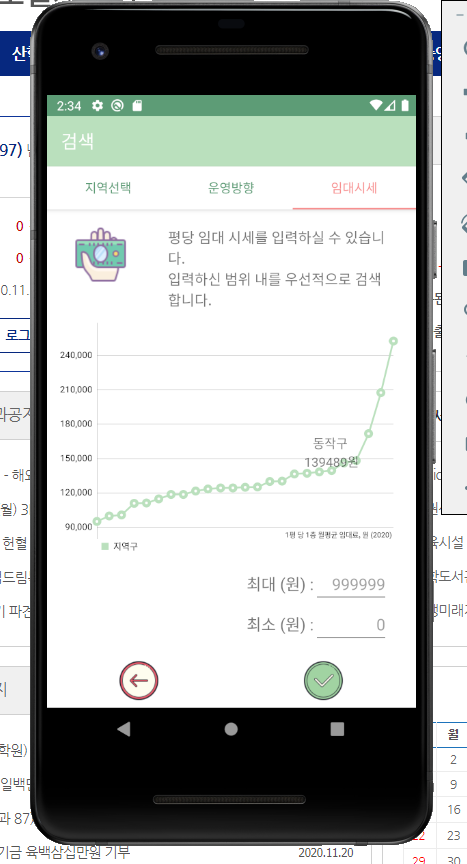
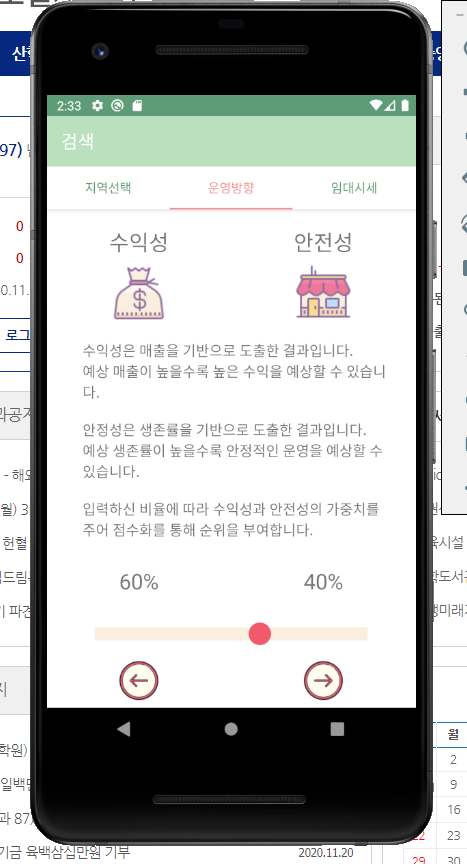
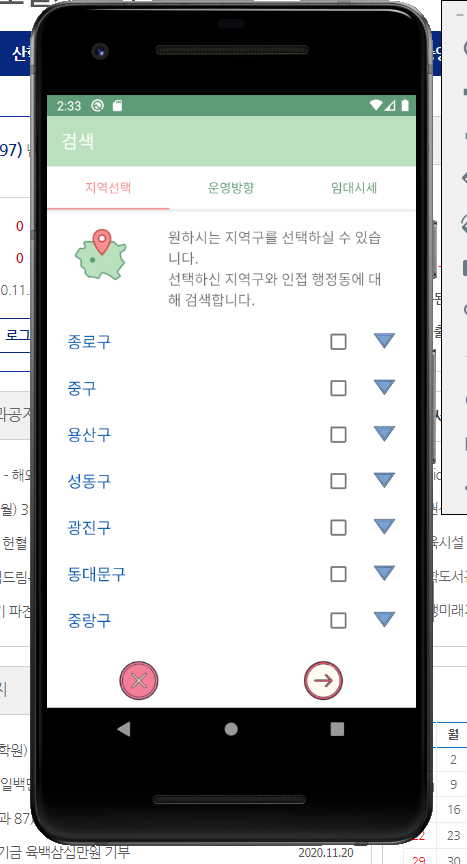
(폐업률은 델타가 중요할거로 예측해서 델타 사용했다 추가 가능 – 근데 결과가 똑같아서 의미가 있나 잘 모르겠음)

5) 점수 산정

매출은 원 단위, 폐업률은 % 단위라 직접 연산이 불가능하기 때문에 Z점수 (Z점수 설명 넣을 예정)를 계산하여 사용자가 입력한 가중치를 곱해 종합 점수를 산정하였다. 또한 폐업률에 직접 Z점수를 계산할 경우 매출은 클수록, 폐업률은 작을수록 유리하기 때문에 통일성을 위해 생존률(100-폐업률)을 지표로 점수를 매겼다.

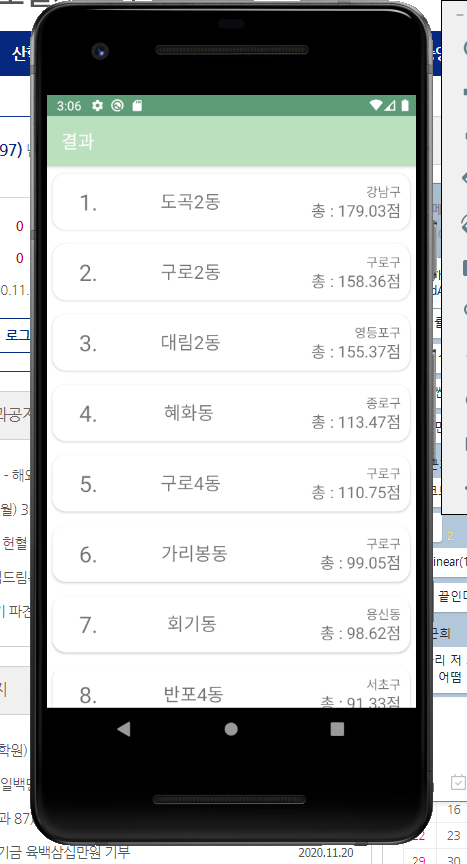
**6) 서비스 구현(구축)**

먼저 서비스를 구축함에 있어 편의점 위치 추천/분석 서비스라는 주제에 맞게 사용자가 편의점을 창업하는 데 있어서 중요한 요소라 생각하는 것을 입력 요소로 정했다. 하지만, 희망하는 편의점의 크기나 브랜드와 같은 세부적인 항목의 경우 분석에 어려움이 있어, 사용자의 입력을 크게 지역선택, 수익성과 안전성을 택하는 운영방향, 평당 임대시세로 정했다. 먼저 지역 선택의 경우 사용자가 선호하는 구를 선택할 수 있게 하여, 하나의 구를 선택할 수도, 또는 여럿의 구를 선택하거나 모든 구를 선택할 수 있게 했다. 이뿐만 아니라 선택된 구들의 분석을 할 때도 단순 해당 지역구의 속한 동들뿐만 아니라 인접 행정동 또한 분석에 포함시켜 결과가 나오도록 했다. 운영 방향의 경우 사용자가 편의점을 운영함에 있어 어느 것을 중시하는 지를 비율로 설정할 수 있도록 했고, 평당 임대시세의 경우 기본적인 지역구들의 1층 평균 임대시세를 그래프로 제공하여 사용자가 희망하는 시세 범위를 더 수월하게 정할 수 있도록 했다.



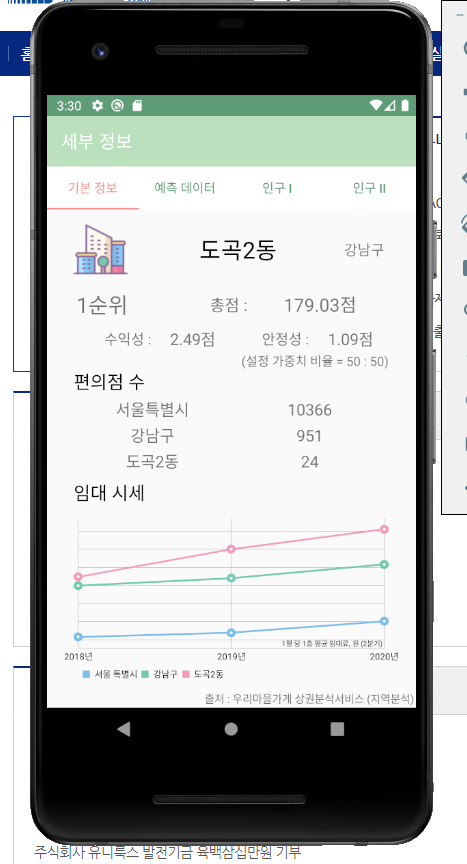
**[사진 6] 어플리케이션의 사용자 입력 화면**

이후 사용자의 입력을 모두 받고 나면, 사용자가 결정한 선택이나 범위를 분석에 적용하여 결과를 도출해낸다. 이 경우 미리 저장되어 있는 데이터베이스의 행정동들을 사용자의 선택을 바탕으로 걸러내거나 점수 총합을 계산한다. 이 때 행정동들에 관한 데이터들을 실시간 데이터가 아닌 미리 구축한 데이터베이스를 사용한 이유는 서비스에서 사용한 데이터셋들 모두 최신화 주기가 짧아야 1달이기 때문에, 데이터 자체를 실시간으로 최신화해야 하는 이유가 없어 사용이 용이하도록 데이터를 미리 저장하는 방법을 사용했다. 이후 나온 결과들을 내림차순으로 정렬하여, 사용자에게 표시하도록 했다.

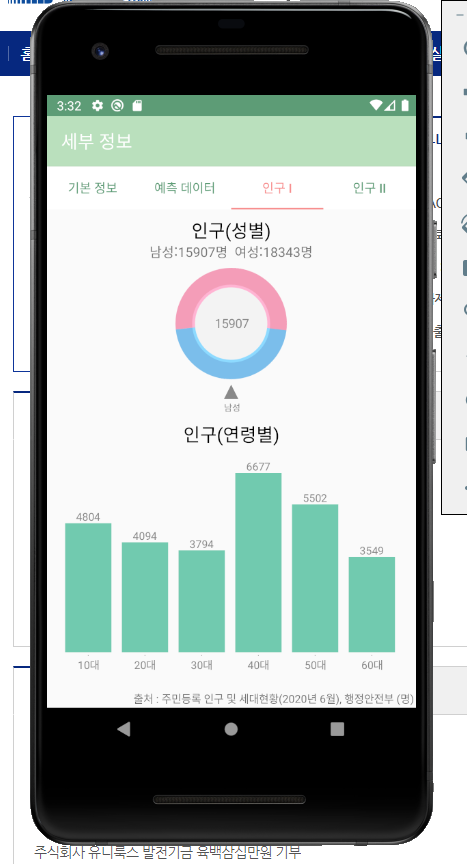


**[사진 7] 어플리케이션의 분석 결과 화면**

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 8] 어플리케이션의 추가 정보 제공 화면1**



**[사진 9] 어플리케이션의 추가 정보 제공 화면2**

**Ⅲ. 한계 및 평가**

프로젝트의 주요 기능인 예측알고리즘의 성능에 대해 프로젝트의 많은 한계요인들이 작용하였다고 생각된다. 프로젝트에서 한계요인으로 평가되는 항목들은 다음과 같다.

우선 데이터의 부재가 가장 큰 문제로 작용하였다고 생각된다. 본 프로젝트에서 가장 중요한 데이터인 매출과 매장의 개,폐업률 관련 데이터를 제공하는 곳을 찾기 힘들었고, 결국 기존의 상권분석서비스에서 일일히 검색을 통하여 얻은 행정동별18,19,20년도 2분기 매출 데이터가 전부였다. 따라서 데이터의 범위가 각 3개년도 2분기로 한정될 수밖에 없었다.

또한 데이터의 신뢰성 역시 한계점으로 평가하였다. 20년도의 경우 코로나19로 인하여 기존의 사회와는 생활양상, 경제활동이 많이 달라졌다. 따라서 우리가 가지고 있는 데이터셋에서는 잘 표출되지 못한 요소들이 존재하였다. 하나의 예시로 전농2동 같은 경우, 19년2분기 매출합이 대략 1억7천만원 정도지만, 20년 2분기에는 1억 1천5백만정도로 감소한다. 이는 기존에 매출의 45%인 7천오백만원 정도를 차지하던10~20대 소비량이 4천만원으로 전년대비 47% 감소하였는데, 코로나로 인한 비대면 수업 등의 영향이 적용된다고 추정된다. 이러한 특수한 사회적 현상으로 인해 데이터의 신뢰성이 다소 떨어진다고 평가했다. 추가적으로 데이터 수가 적은 부분에서는 훨씬 비율이 크게 계산되는 데이터 착시현상도 관측할 수 있었다.

개별 편의점의 지리적 특성이 반영되지 않은 점도 한계점으로 평가하였다. ‘입지요인이 편의점 성과에 미치는 영향에 대한 연구’(황규성, 2016)의 경남, 울산지역의 편의점에서의 일평균매출엑을 종속변수로 한 회귀분석에서는 중요 유의변수로 유동인구, 경쟁지점, 접면수를 뽑았다. 이중 개별점포 주변에서 경쟁하는 경쟁지점이나, 매장의 가시성에 영향을 주는 접면수(길에 노출된 면수)와 같은 입지요인, 즉 지리적 특성에 대해서 본 프로젝트에서는 고려하지 못하였다.

상권의 범위 역시 예측의 오차를 크게 만든 요인 중 하나로 평가하였다. ‘편의점의 상권 추정과 매출예측에 관한 연구’(이춘섭 2012)의 편의점 내점객 방문 소요시간 분포도를 따르면 응답자중 89%가 매장에 방문하는데 5분이내로 소요하였다는 것을 알 수 있다. 따라서 해당 연구에서는 1차 상권비율을 이동속도를 고려하여 150m로 설정하였다. 하지만 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 데이터 셋들의 최소단위가 행정동별 단위였기 때문에 팀의 기술, 정보력 역시 한계점으로 평가할 수 있을 것 같다.

결국 이러한 예외상황으로 인한 데이터의 낮은 신뢰성, 지리적 특성 미반영, 넓은 상권 범위와 학습데이터의 부재가 함께 작용하면서 적은 데이터와 적은 특성들로 매출과 폐업률을 예측하게 되었고, 결국 프로젝트의 기대목적보다 훨씬 성능이 낮은 예측알고리즘을 만들어내는 한계점으로 작용했다고 평가한다.

추후 예측알고리즘의 학습을 위한 행정동별 매출, 폐업률 데이터가 많이 제공되어 학습을 더욱 점진적으로 진행하고, 분석 반경을 행정동보다 좁히고 구체적 입지요인을 데이터셋에 반영한다면, 훨씬 신뢰성 높은 추천 결과를 사용자들에게 제공할 수 있을 것이라고 생각한다.

또한 창업희망자들에게 임대시세와 인건비에 따른 필요수익량 계산, 편의점 가맹점별 창업비용, 방식등의 차이와 점포매물 정보등의 부가적인 서비스를 추가함으로써 편의점 창업 희망자들이 더욱 안전하고 신중하게 창업을 할 수 있도록