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 모았다. 행정안전부에서 제공한 ‘주민등록 인구 및 세대 현황’, ‘지역별 세대원수별 세대 수’, ‘구별 외국인 인구’, 서울시 열린 데이터 광장에서 제공한 ‘서울시 주민등록인구 (동별) 통계’, 우리 마을 가게 상권분석서비스에서 제공한 ‘주거인구, 직장인구’ 데이터를 가져왔다. 인구 외적 요소로는 해당 지역 편의점 수, 대규모 점포 인허가 정보, 서울시 지하철 역 정보, 폐업률 통계, 임대시세를 수집했다.

이때 매출에 대한 정보가 본 프로젝트에서 중요도가 가장 높았으나, 그 만큼 시장에서 역시 중요한 데이터였기 때문에 쉽게 구할 수 없었다. 결국 우리 마을 가게 상권분석서비스에서 2018, 2019, 2020년 각 2분기의 서울시 지역구별 매출밖에 구할 수 없었다. 나머지 데이터셋은 행정동 별 데이터가 대체로 획득가능한 최소 단위였기 때문에 위 데이터셋들을 통해 최적의 행정동을 분석, 추천하는 프로젝트로 진행하도록 결정하였다.

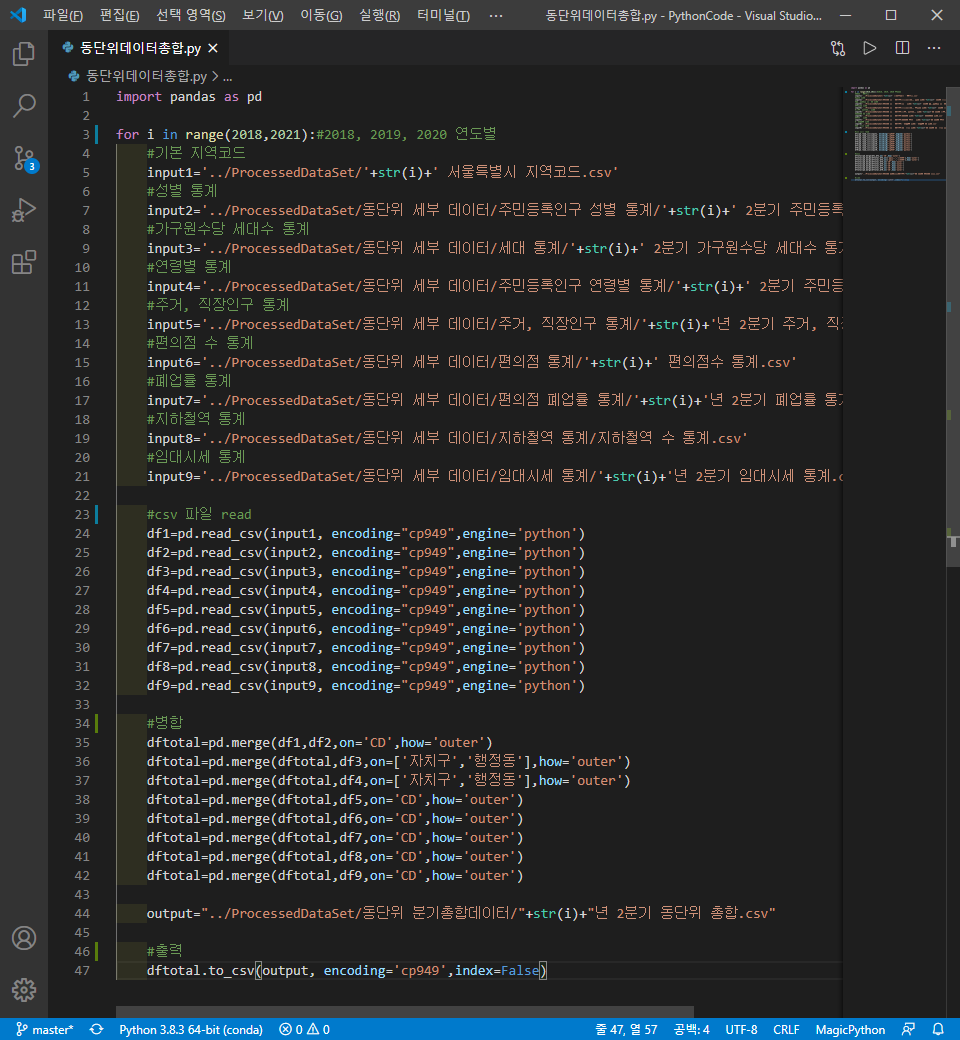


**[사진 1] 데이터 수집**

**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 데이터셋에서 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 떄문에 통일성을 위해 생존률(100-폐업률)을 지표로 점수를 매겼다.

## 추가하면 좋겠지만 어디갈지 모르겠는 내용

(이건 맨 뒷부분에 결과 이후에 피드백에 넣으면될듯)

실시간 데이터가 아닌 미리 구축한 database를 사용하는 이유

* 우리가 사용하는 데이터셋 특징상 최신화 주기가 짧아야 1달임

교수님 피드백 반영 못한거랑 이유

* 못한거
  + 단순 매출이 아닌 순이익?
    - 순이익 계산하려면 예상 매출 – 평균매장넓이\*임대시세 + 원자재가격
      * 평균 매장 넓이 데이터가 없음..

오차가 이렇게 큰 이유?

* 논문 봤는데 매출의 중요 요소에 매장의 가시성 등 구할 수 없었던 데이터도 중요한 요소로 작용한다고함
* 혹시 논문뭔지 있으면 이름이라도 넣는게 좋을듯