제2회 KUIAI 해커톤

도로명주소별 입지 점수 예측 모델 개발 및 입지 점수 등급 시각화

분석 결과보고서

팀 명 : 아이티에스

참여자 : 안정수

구병모

김가영

목차

1. 개발배경

가. 개발/분석배경

본 분석의 대상은 판매시설, 제1종 근린시설, 제2종 근린시설로 모두 매출이 발생하는 가게 혹은 점포들을 포함한 건물들입니다. 모든 판매시설들은 매출 극대화를 목표로 하고 이를 위해서는 최적 입지에 대한 의사결정이 필수적입니다. 점포 입지는 한번 정해지고 나면 쉽게 바꿀 수 없는 장기적인 성격을 가지는 고정 투자이므로 무엇보다 신중하게 결정을 내려야 합니다. 본 분석에 들어가기에 앞서 상권 및 입지에 대한 정확한 정의를 내려 분석 범위를 한정하겠습니다.

먼저 상권(Trading Area)의 개념은 여러 문헌들에서 다양하게 정의되고 있습니다. Huff(1964)는 지정기업이나 다수기업에서 판매하는 상품이나 서비스를 판매할 확률이 '0' 이상인 잠재적 고객을 포함하고 있는 지리상으로 묘사된 지역을 상권으로 정의하고 있습니다. 상권을 매출액이 발생하는 구역으로 정의하면서, 상권이 공간적으로 고정되어 있는 것이 아니라 동적인 특성을 가지고 있는데, 이러한 동적인 특성 중에는 마케팅 전략이나 가격, 점포규모, 경쟁, 교통 접근성 등의 변화에 따라 움직이고 있는 가변적 특성을 가지고 있습니다.

다음으로 입지는 경제활동을 잘하기 위해 적당한 사업장의 장소를 찾아서 선택하는 것으로, 사업장(점포)이 소재(所在)하는 위치조건을 의미합니다. 산업, 생산, 창업 중에 어떤 목적을 갖느냐에 따라서 그 의미와 특성은 조금씩 차이가 생길 수 있습니다. 본 분석에서는 매출이 발생하는 사업장의 입지 분석에 집중하기에 창업을 목적으로 합니다. 일반적으로 입지조건이 좋으면 상권은 좋게 되고, 반대로 상권이 좋으려면 입지조건이 좋아야 합니다. 따라서 입지는 상권과 혼용되어 사용되기도 하는 불가분(不可分)의 관계입니다.

(사진출처: 매경비즈)

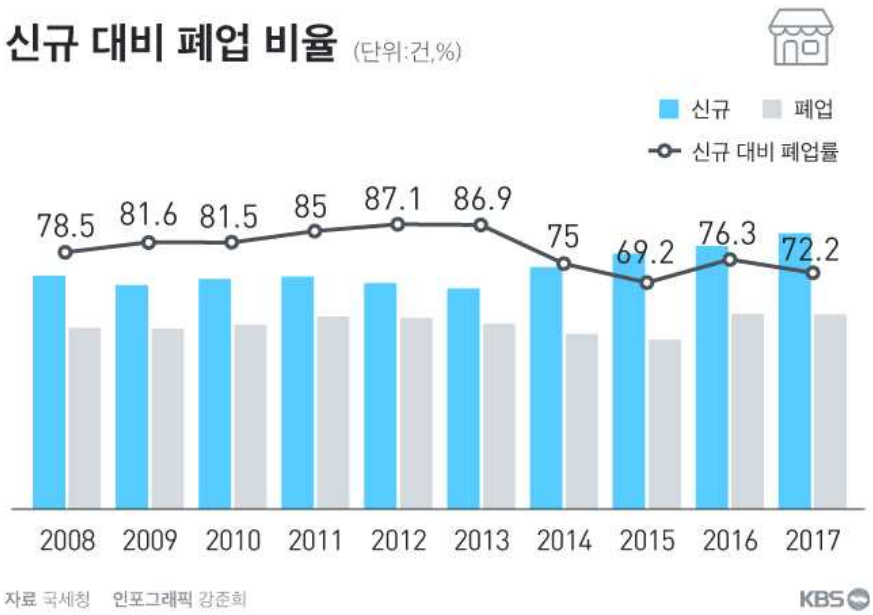
본 분석은 매출이 발생하는 시설을 포함한 건축물을 대상으로 상권 분석을 통해 입지 점수 및 등급을 부여합니다. 따라서, 타 목적을 위한 입지를 선정할 경우 본 분석에서 고려한 변수 이외의 변수들을 선정해서 입지를 선정해야 합니다.

나. 개발/분석 필요성

정부와 카드사가 분석한 결과 자영업자 소멸률은 13.3%에 달하고, 창업 후 5년 생존율이 26.9%로 신생업체의 3/4 가량이 5년내에 폐업하는 것으로 나타났습니다. 이러한 자영업의 위기는 국가 사회의 기본단위인 개인과 가정 경제를 위협하고, 금융기관의 대출부실과 실업 증가, 소득감소로 이어지는 악순환을 유발할 수 있습니다. 정부와 지자체에서는 동종 업종 거리 제한, 대형마트 의무 휴업, 자영업자 금융 지원 등 다양한 정책을 시행하고 있으나 실질적인 효과는 크지 않은 것으로 평가됩니다.

폐업의 원인은 다양합니다. 높은 창업비용과 과도한 타인자본 조달로 인한 고정비 부담, 판매율 저조 등을 꼽을 수 있습니다. 본 분석에서 주목한 점은 창업 준비 기간을 짧게 갖고 급하게 창업을 하여 해당 업종의 사업이 적절하지 못한 입지에서 이뤄지는 것입니다. 입지는 사업의 성공과 직결되어 있습니다. 구매 의지가 있는 소비자가 자주 왕래하고 다른 사업장과 시너지 효과가 날 수 있는 입지를 찾는 것이 중요합니다. 예비 창업자는 교통의 편리성, 창업을 하고자 하는 사업장의 주변 상권의 영향 등을 고려할 필요가 있습니다.

본 분석은 경제적 악순환을 야기하는 창업의 낮은 생존율 문제를 해결하기 위해 상권, 지하철역, 학교 등과 같은 다양한 변수들을 고려하여 매출액을 예측한 입지를 선정하고자 합니다.



(사진출처: KBS)

2. 분석개요

가. 분석목표

본 연구는 서울시 행정구역별 상권 데이터를 토대로 건축물의 도로명 주소 검색 시 입지 score를 도출하는 '건축물 입지 점수 예측 모델' 구축을 목표로 합니다. 본 조의 목표는 해당 도로명 주소를 검색 시 지도에 500m까지의 입지 점수를 반영하고 이를 시각화하여 추가적인 시각 자료로 제공합니다. 본 모델은 입지 점수를 등급화시켜 500m 반경 이내에 범위에 해당하는 건축물들을 분류하여 시각화하는 작업을 추가적인 목표로 합니다.

대한민국 수도인 서울은 수도권 집중도가 전 세계에서 가장 높은 나라입니다. 수도권 집중도가 높은 일본과 프랑스가 30%, 영국은 23%, 이탈리아는 11%인 반면, 우리나라는 인구의 50%가 서울에 거주하고 있습니다. 이는 국토의 10분의 1인 크기에 비해 여러 인구 및 자원과 시설 등이 집중 포집 되어있다고 해석될 수 있습니다. 따라서, 본 연구가 갖는 의의는 대한민국에서 발생하는 경제활동 중 큰 비중을 포함한 서울 지역 판매 시설들의 입지 점수를 산정하여 예비 창업자 혹은 기존의 사업자들에게 상권분석과 다양한 변수들을 고려한 입지 분석의 자료로서 활용되는 것입니다. 추가적으로 서울 이외의 다른 지역에서도 본 분석을 바탕으로 활발한 연구가 이루어지기를 기대합니다.

나. 분석내용

본 연구는 서울시 건축시설의 상권 데이터를 이용하여 분석을 진행합니다. 본 분석은 2022년 1월 10일 - 2022년 1월 13일 기간 동안 진행되었습니다.

본 연구에서 기본적으로 사용한 데이터셋은 총 9개의 데이터셋으로 판매시설, 제1종 및 2종 근린생활시설, 상권 코드별 좌표, 대규모 점포, 생활인구, 지하철역 그리고 학교 정보 등 다양한 인구, 경제 그리고 시설 데이터를 포함하였고, 해당 데이터 셋에서 유의미한 데이터만을 추출하여 새로운 데이터셋을 형성하여 분석을 진행하였습니다.

본 분석에서는 건축물 생애 이력 데이터에서 판매시설과 근린생활시설의 주소를 토대로 500m 이내 상권의 평균 매출액을 반응변수로서 표현합니다. 이외에 추가적인 시설물과의 거리 및 인구 데이터 등을 추가적인 설명 변수로 제안합니다.

본 조는 도로명 주소를 좌표화하여 위도 경도로서 변환 후, 각종 시설물과의 하버사인 거리를 구하여 구체화하고 최종적인 입지선정을 위한 분석을 진행합니다.

본 조는 기존 데이터셋에서 데이터들간의 관계성, 중요성, 필요성 등을 복합적으로 평가하여 필요한 변수들만을 추출하여 최종적인 입지 점수를 도출하는 근거로서 활용합니다.

다. 활용방안

본 연구는 서울시 상권 개발에 직·간접적으로 활용될 수 있는 지표들로서 충분히 제공할 수 있습니다. 본 조가 제안하는 입지 점수는 상권의 핵심이 되는 지수인 '상권평가지수'에서 부족한 유동성을 보완할 수 있습니다. 상권평가지수란 상권 내 음식, 서비스, 소매 등의 전반적인 업종의 경기에 더불어 상권의 인구 수, 집객 시설, 교통 시설 등 다양한 시설 서비스를 종합하여 산출한 등급을 의미합니다. 이는 점수가 절대적인 등급이 낮을수록 상권이 활성화되어 있음을 뜻합니다. 위 등급은 일반적으로 카드거래건수의 추이를 이용하여 상권 내 가맹점의 평균 구매빈도를 증감률을 지수화한 정보로 정의합니다. 이는 거래횟수가 많아질수록 상권이 활성화되었다고 판단하기 때문에, 본 연구에서 이용하는 상권 내 평균 생활 인구수와 매출액

을 토대로 유동성 지수를 제안합니다. 또한, 지하철역과의 거리가 가까울수록 자기 차량을 이용하는 빈도가 줄어들기 때문에, 수동적인 유동성 지수가 증가할 것이며, 이에 따른 자체적인 워킹 포인트를 지정할 수 있습니다. 본 연구에서는 워킹 포인트를 지정한다면, 특정 워킹 포인트를 기점으로 상권을 세운다면 상대적인 매출 증대를 기대할 수 있다고 생각합니다.

본 조는 해당 연구를 음식점 및 기타 레스토랑 상권의 입지선정에 활용할 수 있다고 판단하였습니다. 해당 입점 성공 요인으로는 (1) 시계성(가시성)의 확보, (2) 건물의 1층을 확보, (3) 랜드 마크 건물 내에 입지선정 및 주 동선상에 위치, (4) 도로에서 매장으로의 즉각적인 유입 확보를 고려합니다. 본 조에서 목표로 하고있는 HEATMAP을 이용한 지도 시각화와 건물의 층별 개수를 변수로 한 모델은 해당 건물의 층별 분석과 더불어 해당 도로 분석 그리고 연면적 분석을 통해 중점적인 건물의 분석을 기대할 수 있다고 생각합니다. 일반적으로 적용되는 허프의 확률 모델에서는 특정 도시의 상점 간의 중력모형을 이용하여 해당 건물의 면적과 건물 사이의 거리만을 표현하고 있습니다. 본 조는 입지선정에 있어 이러한 요인들을 넘어 언덕이 존재하는 상권, 배후가 단절된 상권, 연속성이 없는 상권 등에 대한 부분을 시각화 및 데이터 중복 완화를 이용해 충분히 활용 가능하다는 것을 추가적으로 제안하고 싶습니다. 본 조는 앞서 기술한 매장 면적과 거리만을 이용하는 허프모델을 넘어 하버사인 거리로 구축되는 상권 주변의 학교 정보 및 지하철역까지의 기타 시설까지 고려한 허프모델로 보다 넓게 확장할 수 있습니다. 이는 물론, 해당 지역의 특성에 맞게 중요도를 일종의 수치로 표현할 수 있다고 생각하지만, 서울시에서는 반영하기 최적화된 모델이라고 판단됩니다. 뿐만 아니라, 행동반경, 대중교통 그리고 추가적인 기타 시설을 중시하는 지역에서는 활용 가능한 모델임을 제안합니다.

3. 상세분석구조

가. 구축 모델 및 분석 구조

본 조의 최종 구축 모델은 크게 세가지 흐름에 따라 개발되었습니다.

(1) 데이터셋 구축

(2) 데이터 학습 및 입지점수 예측모델 개발

(3) 최종 구현 함수 개발

(1) 데이터셋 구축

모델 개발에 필요한 데이터셋 구축 과정을 순서에 따라 기술하겠습니다.

(1-1) 생애이력 데이터 결합

건물의 용도별로 구별로 나누어져있는 엑셀 파일들을 결합하는 과정을 거쳤습니다.

(1-2) 좌표 정보 추출

건물의 생애이력 데이터의 도로명주소를 기준으로 좌표를 도출했습니다. 해당 과정에서 도로명 주소의 수가 약 12만건으로 카카오 api를 통해 도출해낼 수 있는 좌표의 개수가 한계가 있기 때문에 구글 api를 사용했습니다.

지하철역, 학교, 대규모 점포, 상권코드별 좌표는 카카오 api를 통해 좌표를 도출했습니다.

(1-3) 거리 계산, 평균 생활인구, 평균 매출액(y) 추출

먼저 건물과 가장 가까운 지하철역, 학교, 대규모 점포와의 최단 거리를 계산하기 위해서 택한 방법은 다음과 같습니다. 각각의 건물별로 일정 크기의 버퍼(ex. 5km, 10km)를 그려서 해당 버퍼에 속한 지하철역, 학교, 대규모 점포들과 건물 간의 거리 중 최소값을 선택했습니다. 건물별로 모든 지하철역, 학교, 대규모 점포들과의 거리를 비교한 결과 소요 시간이 컸기에 버퍼를 그리는 방식을 택했습니다. 또한 거리는 유클리디언이 아닌 하버사인(haversine) 거리를 사용했습니다. 단순히 두 좌표 사이의 직선거리를 구하는 것은 지구 곡률을 무시한 결과로서 왜곡된 결과를 산출할 수 있기에 지구 곡률을 고려하여 두 위경도 좌표간의 거리를 구하는 공식인 하버사인 공식을 사용했습니다.

	도로명주소	위도	경도	xy_tuple	geometry	haver_dis_subway
0	서울특별시 강동구 양재대로 1540	37.542920	127.142450	(37.5429196, 127.1424496)	POINT (968414.138 1949348.008)	0.316026
1	서울특별시 강동구 양재대로 1449	37.535248	127.138685	(37.5352477, 127.13868470000001)	POINT (968078.277 1948498.129)	0.307749
2	서울특별시 강동구 양재대로128길	37.544508	127.144746	(37.5445084, 127.14474569999999)	POINT (968617.642 1949523.507)	0.207319
3	서울특별시 강동구 천호대로 1156	37.533434	127.139603	(37.533434299999996, 127.13960279999999)	POINT (968158.619 1948296.632)	0.487983

위와 같이 도로명주소별 가장 가까운 지하철역과의 하버사인 거리가 구해졌고, 이는 학교와 대규모 점포에서 동일하게 이루어졌습니다.

평균 생활인구와 평균 매출액(y)을 구하기 위해서 상권코드별로 2020년 총 매출액과 총 생활인구를 합산한 결과를 활용했습니다. 다음으로 도로명주소별 500m 반경 내에 있는 상권코드를 구하여 해당 상권들의 총 매출액과 총 생활인구의 평균을 내서 평균 생활인구, 평균 매출액(y)으로 사용하였습니다.

(1-4) 건물별 생애 이력 데이터 결합

앞서 만든 데이터셋과 건물별 생애 이력 데이터를 결합하여 분석에 활용할 최종 데이터셋을 구축하였습니다. 건물별 생애 이력 데이터 중 건물 용도, 연면적, 층수, 주차장 개수, 승강기 개수만을 추출하여 분석에 활용했습니다. 위 과정들을 통해 최종 완성된 데이터셋을 다음과 같습니다.

	도로명주소	위도	경도	haver_dis_subway	haver_dis_school	haver_dis_mart	mean_sales	mean_pop	용도	연면적(㎡)	자주식주차장	기계식주차장	승강기개수	층수
0	서울특별시 강동구 양재대로 1540	37.542920	127.142450	0.316026	0.459123	0.110586	2.551361e+10	18225491.8	판매시설	1543.47	21	0	0	2
1	서울특별시 강동구 양재대로 1449	37.535248	127.138685	0.307749	0.416656	0.139283	4.715353e+10	20419730.5	판매시설	2772.93	13	0	1	3
2	서울특별시 강동구 양재대로 128길	37.544508	127.144746	0.207319	0.498036	0.232098	2.896793e+10	22305878.2	판매시설	3640.85	0	0	0	4
3	서울특별시 강동구 천호대로 1156	37.533434	127.139603	0.487983	0.432536	0.346491	3.770563e+10	15761681.4	판매시설	1625.67	16	0	1	3

(2) 데이터 전처리 및 예측모델 개발

(2-1) 데이터 결측치 처리

최종 데이터셋에서 평균 매출액과 평균 생활인구 컬럼 정보에 결측치가 있었습니다. 이는 도로명주소의 500m 반경에 상권이 하나도 없을 경우 계산이 되지 않았기 때문입니다. 해당 경우 상권이 하나도 없어서 입지 점수를 산출할 수 없기에 빼는 것이 적합하지만 input으로 해당 도로명주소가 들어오면 오류가 날 수 있기에 0으로 처리해주었습니다.

(2-2) 중복 도로명 주소 처리

하나의 도로명주소에 여러개의 건물이 할당될 수 있습니다. 해당 경우 본 조는 건물보다 도로에 집중해서 값을 할당해야한다고 파악했습니다. 따라서 도로명주소를 그룹으로 하여 중복된 도로명주소의 건물들의 컬럼 정보를 평균을 내서 사용했습니다. 중복된 도로명주소 중 하나의 건물을 선택하는 것이 아니라 평균을 낸다면 해당 도로의 모든 건물을 고려할 수 있을 것으로 판단했기 때문입니다.

(2-3) 예측모델 개발

해당 모델은 도로명 주소별로 평균 매출액(y)을 예측하는 모델입니다. 따라서 회귀모델을 개발해야합니다. 본 조는 단순선형회귀, 라쏘회귀, 릿지회귀, XGBoost, LightGBM, K-nearest Neighbors 회귀모델을 고려하여 비교했습니다.

성능	단순선형회귀	라쏘	릿지	XGBoost	LightGBM	Knn
r2-score	0.2654	0.2654	0.2654	0.79550	0.6878	0.87594
mae	36520919362	36520919362	36520745612	21500392828	25816916270	6887593272
mse	5.753138795414207e+21	5.753138795442346e+21	5.753178453315655e+21	1.601665333 9250584e+21	2.444480046 997519e+21	9.716139353 286528e+20
rmse	75849448221	75849448221	75849709645	40020811260	49441683294	31170722406

그 결과 가장 높은 성능을 보인 K-nearest Neighbors 회귀모델을 선택해서 학습을 진행하고 예측값을 산출해냈습니다.

(3) 최종 함수 구현

위 과정에서 예측모델을 통해 도로명주소별 입지 점수(예측 평균 매출액)를 산정했습니다. 해당 점수를 순위화한 다음 4개 구간으로 나누어 '최상, 상, 하, 최하' 등급을 부여했습니다. 마지막으로 도로명주소를 input으로 받으면 예측한 입지 점수를 반환하고 입지 등급에 따른 500m 반경 시각화와 입지 점수를 기준으로한 히트맵 시각화 지도를 반환하는 함수를 개발했습니다.

```
def analyze_score():
    location = input('도로명 주소 입력::')

    score = data_score[data_score['도로명주소'] == location]['y_score']
    score = float(score)

    col = data_score.loc[data_score[data_score['도로명주소'] == location]['color'].index[0], 'color']

    latitude, longitude = data_score[data_score['도로명주소'] == location]['위도'], data_score[data_score['도로명주소'] == location]['경도']

    global seoul

    seoul =  folium.Map(location=[latitude, longitude], zoom_start=16) #확대정도 지정

    folium.Marker([latitude, longitude],
              popup=location,
              tooltip=location,
              icon=folium.Icon('red', icon='star'),
              ).add_to(seoul)

    seoul.add_child(plugins.HeatMap(zip(data_score['위도'],
                                        data_score['경도'],
                                        data_score['y_score']), radius=20))

    folium.Circle([latitude, longitude],
                        color=col,
                        fill_color=col,
                        radius = 500,
                        tooltip=location).add_to(seoul)


    print('입지 점수는 {} 입니다'.format(score))

    return seoul
```

나. 사용기술 및 알고리즘

본 조의 최종 구축 모델을 개발하는 과정에 따라 사용된 라이브러리, 알고리즘을 순서에 따라 기술하겠습니다.

(1) 데이터셋 구축

데이터셋을 구축하는 과정에서는 먼저 데이터를 결합할 때는 Pandas 패키지를 활용했습니다. 다음으로 좌표를 도출하기 위해서 googlemaps 패키지를 활용하여 구글 api를 통해 좌표를 도출하고, 카카오 api를 통해 좌표 정보를 받아오는 함수를 구현해서 도출했습니다.

건물과 가장 가까운 지하철역, 학교, 대규모 점포와의 최단 거리를 계산하기 위해서 GeoPandas, shapely, haversine 패키지를 사용했습니다. GeoPandas 형식에 맞춰 좌표계를 수정하고 shapely 패키지 중 버퍼를 그리는 함수를 사용하여 도로명주소별로 버퍼를 그리고 버퍼 안에 해당하는 지하철역, 학교, 대규모 점포를 구하여 haversine 함수로 최소값을 도출했습니다. 평균 매출액과 평균 유동인구를 구하기 위해 500m 반경 내 상권코드를 찾는 방식도 위와 동일합니다.

(2) 데이터 전처리 및 예측모델 개발

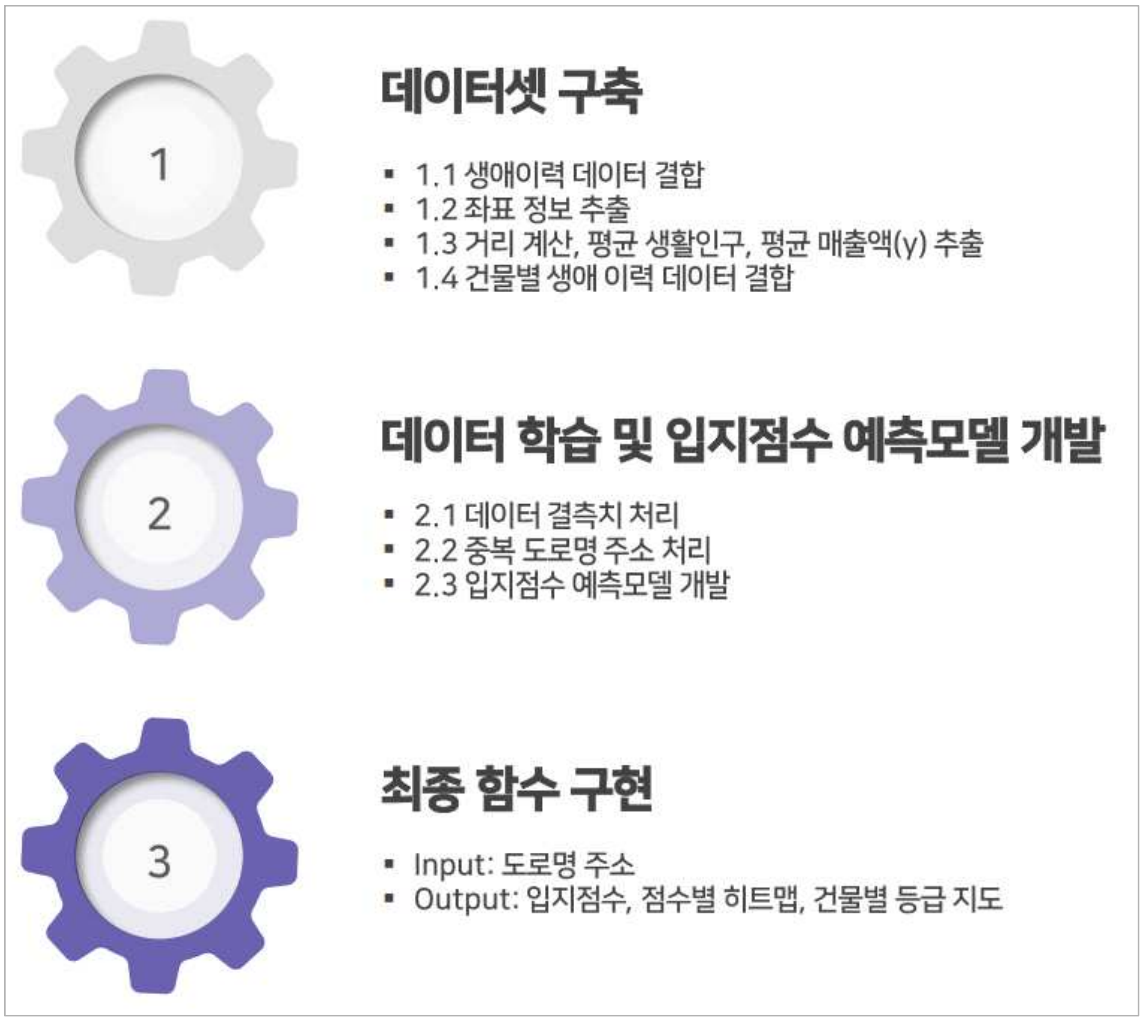
전처리 과정은 데이터셋이 Pandas의 DataFrame 객체로 되어있기에 해당 문법에 맞게 처리하였습니다. 예측모델을 개발하기 위해서는 sklearn, xgboost, lightgbm, numpy 패키지를 활용했습니다.

(3) 최종 구현 함수 개발

최종 구현 함수에서 지도 시각화를 하는 과정이 필수이기에 folium 패키지를 활용했습니다.

4. 개발결과

가. 전체 모델 구현 흐름도



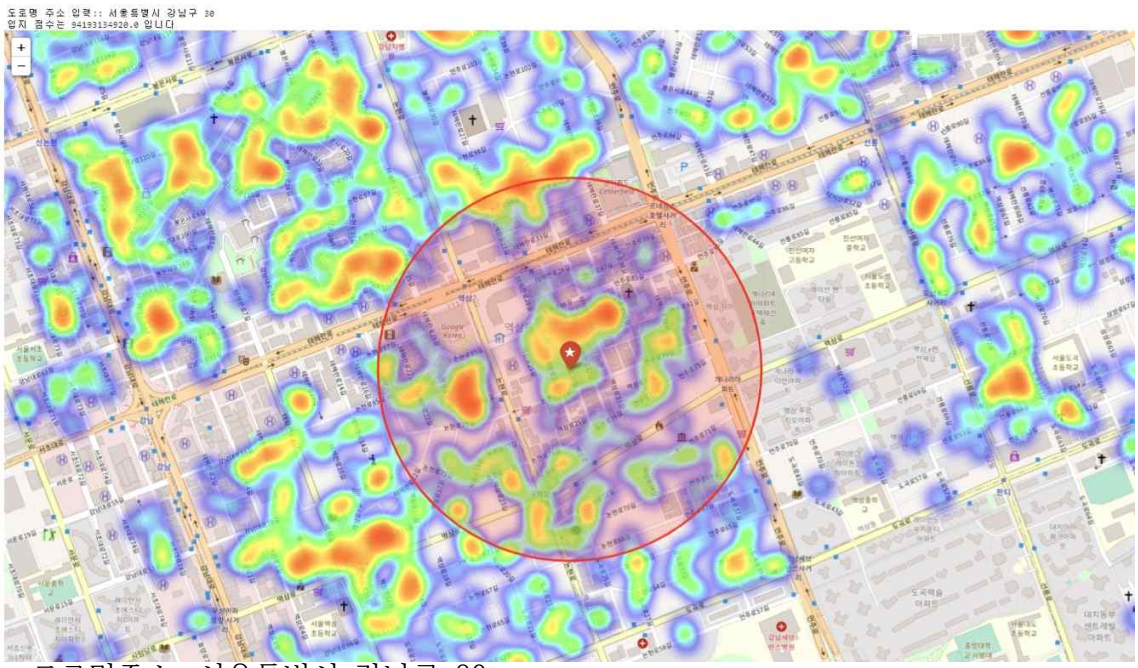
전체 모델 구현 흐름도는 위와 같습니다. 크게 데이터셋 구축, 데이터 학습 및 입지점수 예측모델 개발, 최종 구현 함수 개발의 흐름입니다.

첫 번째, 데이터셋 구축 과정은 4단계로 진행했습니다. 판매용도별, 지역구별로 나누어져있던 생애이력 데이터를 결합했습니다. 다음으로 생애이력 데이터에 있는 도로명주소를 기준으로 좌표를 추출하고, 지하철역, 학교, 대규모 점포, 상권들의 좌표를 추출했습니다. 추출된 좌표들을 하버사인 공식을 통해 건물과 지하철역, 학교, 대규모 점포간의 최단 거리를 산출하고, 500m 반경 내의 상권들을 찾아내서 평균 매출액과 평균 생활인구를 계산했습니다. 마지막으로 건물별 생애 이력 데이터에서 사용할 컬럼들을 선택해서 결합했습니다.

두 번째, 데이터 학습 및 입지점수 예측모델 개발은 3단계로 진행했습니다. 먼저 건물의 500m 반경 내에 상권이 전혀 없다면 평균 생활인구와 평균 매출액이 결측치가 되기에 0을 넣어줬습니다. 다음으로 중복된 도로명주소가 있다면 해당 도로에 있는 건물들의 정보를 평균내서 사용했습니다. 마지막으로 후보 예측모델 중 K-nearest Neighbor 회귀모델을 선택해서 예측 입지 점수를 도출해냈습니다.

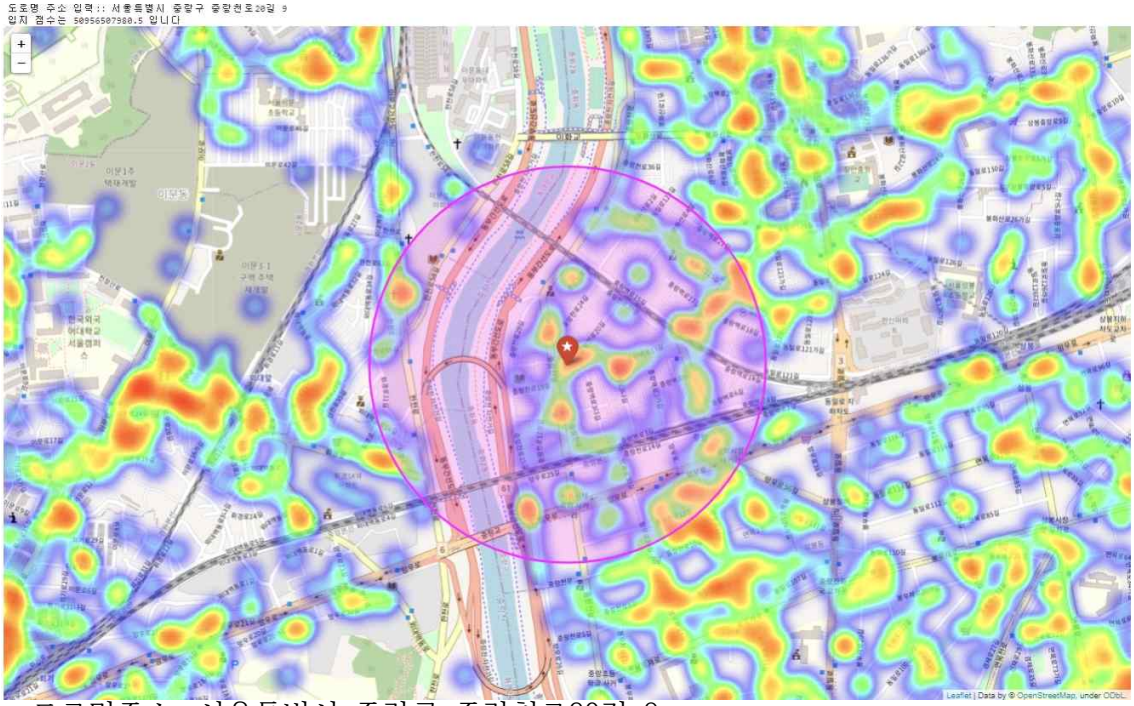
세 번째, 위 예측모델을 기반으로 계산된 입지점수를 활용하여 최종 함수를 구현했습니다. input으로 도로명주소를 넣으면 해당 도로명주소의 입지점수, 점수별 히트맵, 건물별 등급 지도를 반환합니다.

나. 결과



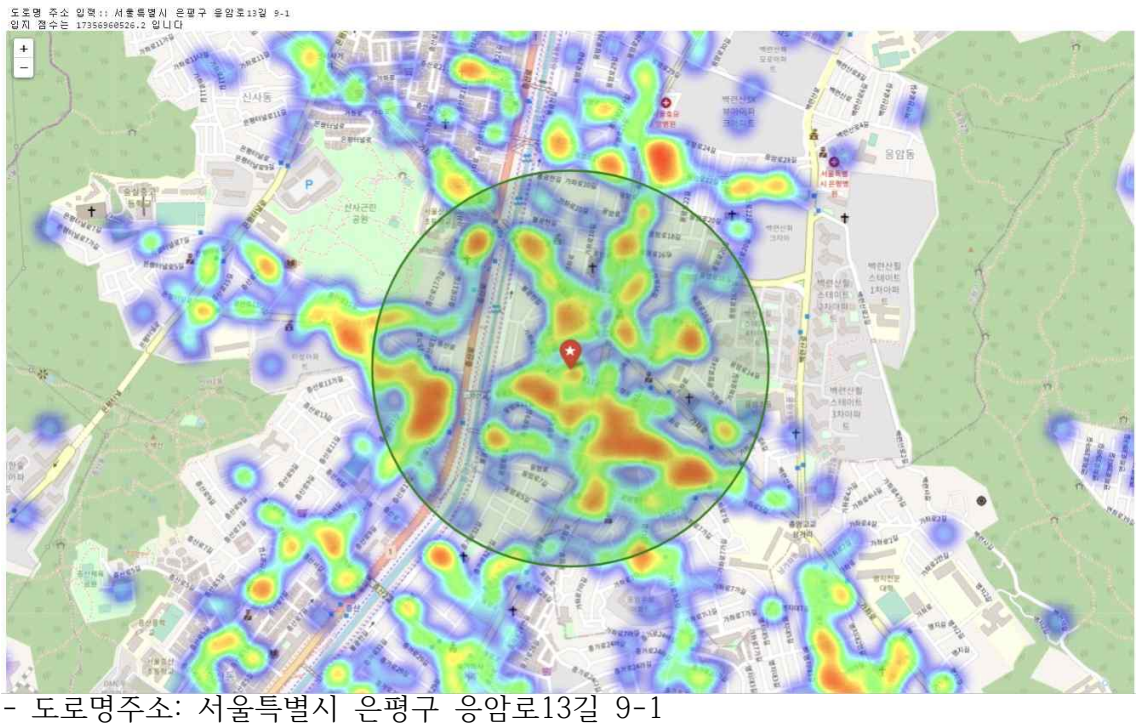
- 도로명주소: 서울특별시 강남구 30

- 등급: 최상



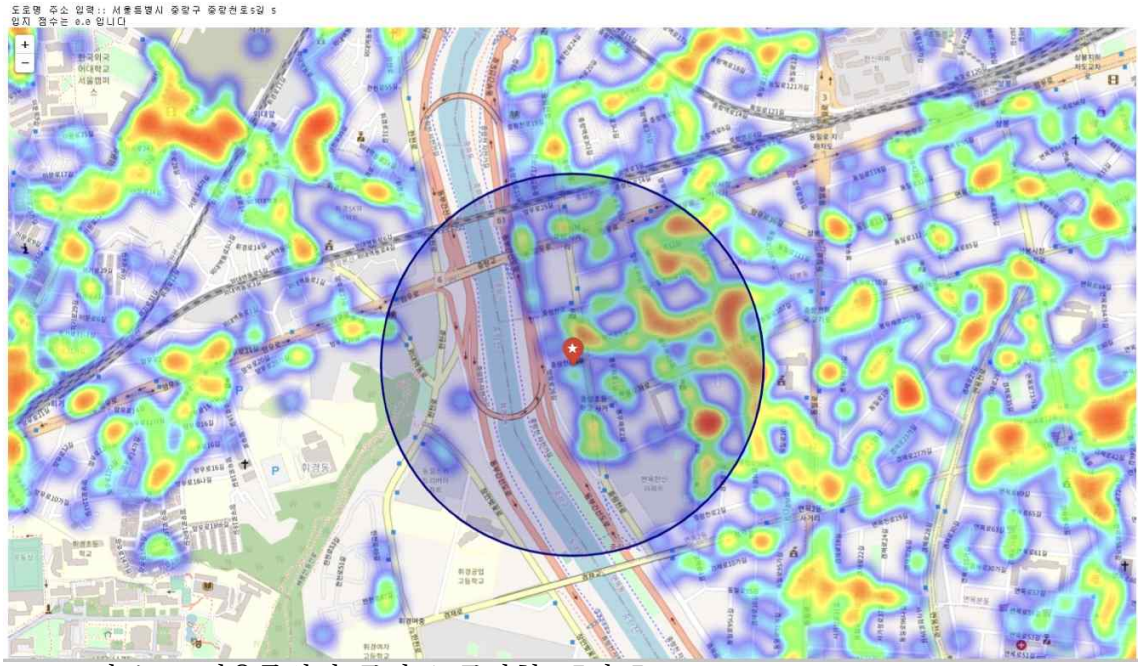
- 도로명주소: 서울특별시 중랑구 중랑천로20길 9

- 등급: 상



- 도로명주소: 서울특별시 은평구 응암로13길 9-1

- 등급: 하



- 도로명주소: 서울특별시 중랑구 중랑천로5길 5

- 등급: 최하

위 결과물은 input으로 도로명주소를 입력하면 나오는 output들입니다. 등급에 따라 순차적으로 빨간색, 분홍색, 녹색, 남색을 부여했습니다. 해당 지점의 500m 반경을 표시하고 주변 상권들을 입지 점수 기준으로 히트맵화했습니다.

본 조에서 구현한 모델의 한계점들을 기술하겠습니다.

첫 번째로, 데이터 결측치 처리 방식입니다. 해당 모델은 생애이력 데이터에 있는 어떠한 도로명주소를 넣어도 output이 나오도록 설계했습니다. 따라서 도로명주소 500m 반경 내에 상권이 없어서 평균 매출액과 평균 생활인구가 기록되지 않은 경우 0으로 넣고 모델을 학습시켰습니다. 하지만 이는 기록되지 않았을 뿐이지 매출액이 0원이었다고 해석하기는 힘듭니다.

두 번째로, 다양한 변수들을 고려하지 못했다는 점입니다. 시간 제한이 있는 해커톤의 특성상 모델 구현을 최우선의 목표로 잡고 일정을 진행했기에 사용할 변수를 빠르게 결정해야했습니다. 이에 따라 평균 매출액의 경우 분기별, 업종별, 성별, 시간대별로 세분되어있었으나 2020년 한 해 동안의 상권코드별 총 매출액으로 학습을 진행했습니다. 이는 추가적인 분석을 통해 개선될 수 있습니다.

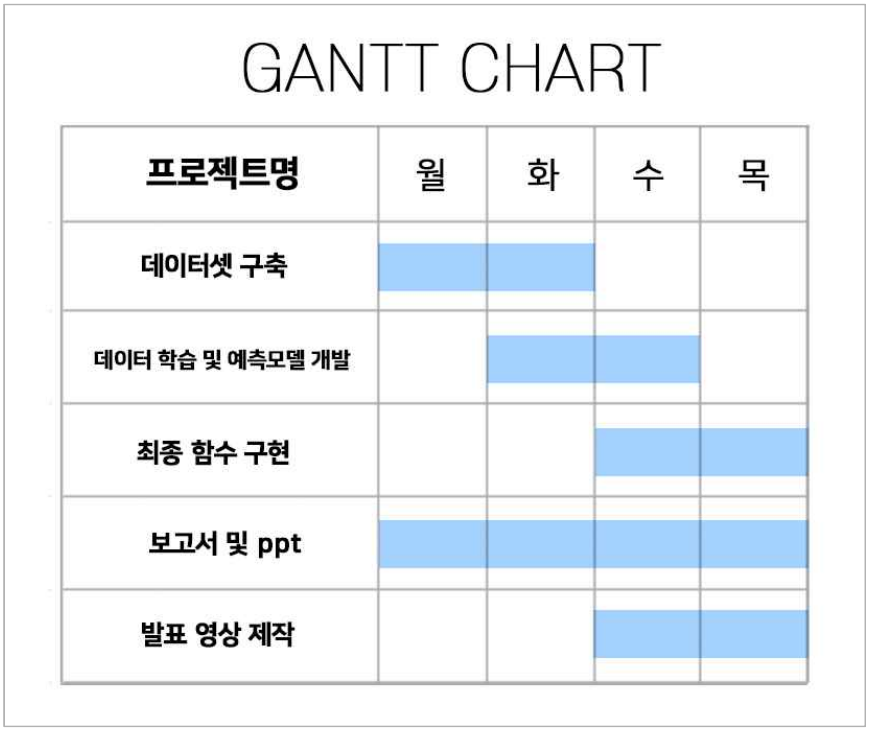
세 번째로, 중복 데이터 처리 방식입니다. 본 분석에서 중복된 도로명주소가 있다면 해당 거리에 여러 건물들의 정보를 평균을 내서 사용했습니다. 구현한 함수가 input으로 해당 도로명주소를 받아서 처리하기 위와 같은 결정을 내린 것입니다. 그러나 더 정확하고 엄밀한 모델을 구현하기 위해서는 중복된 도로명주소를 특정 건물의 도로명주소로 바꾸어 좌표를 계산해서 모델에 입력하는 방식으로 바뀌어야할 것입니다.

5. 구성원별 역할 및 개발일정

- 구성원별 역할

구성원	역할
안정수	좌표 추출, 데이터셋 결합 및 전처리, 보고서 제작
구병모	좌표 추출, 계산 컬럼 도출, 모델링 및 함수 구현, 보고서 제작
김가영	좌표 추출, 데이터셋 결합 및 전처리, PPT 제작

- 개발일정



6. 참고문헌

논문

E Maria, E Budiman, Haviluddin, M Taruk. Measure distance locating nearest public facilities using Haversine and Euclidean Methods. Journal of Physics: Conference Series. 2020;1450(1):1. Accessed January 13, 2022.

김황배, & 김시곤. (2006). 접근성이론과 GIS 공간분석기법을 활용한 행정기관의 입지선정. 대한토목학회논문집 D, 26(3D), 385-391.

심재헌, & 이성호. (2008). 대형할인점의 입지선정을 위한 의사결정에 관한 연구. 대한토목학회논문집 D, 28(5D), 705-712.

Huff, D. L.(1963). "A probability analysis of shopping centre trade areas," Land

Economics, Feb : 81-90.

홍준혁.(2019). 빅 데이터를 활용한 서울시 요식업 분석 : 데이터 마이닝을 활용한 서울시 내 요식업 창업 전략수립을 위한 연구.

김일광.(2018). 우리나라 자영업 업체 현황과 재무특성에 관한 연구. 지역산업연구 제 41권 제 3호, 343-364.

김진철 & 양현철. 빅데이터를 활용한 소상공인 영업지원 점포 분석 사례 연구, 한국정보처리학회 2015년도 추계학술발표대회 2015 Oct. 28, 2015년 pp.1244 - 1247

유지은, & 양혜원. (2011). 패스트푸드점 입지요인이 방문의도에 미치는 영향. *한국외식산업학회지*, *7*(1), 23-41.

태경섭, & 임병준. (2010). 상권경쟁을 고려한 신규점포의 입지선정에 관한 연구: 서울시 대형마트를 대상으로. *대한지리학회지*, *45*(5), 609-627.

라이브러리

- 파이썬 라이브러리 Folium

https://python-visualization.github.io/folium/

- 파이썬 라이브러리 Shapely

https://pypi.org/project/Shapely/

- 파이썬 라이브러리 GeoPandas

https://geopandas.org/en/stable/