캡스톤 디자인 중간보고서

서울시 창업 위치 추천

32190158 공수빈

32190158 이나영

목 차

1. 프로젝트 목적

'서울시 창업 위치 추천' 프로젝트는 요식업으로 창업 분야를 한정하며 서울시 상권을 분석하여 서울시에 창업을 하고자 하는 사용자들에게 필요한 정보를 제공하는 것을 목적으로 한다.

2. 수행방법

2.1. 초기설정

2.1.1. 흐름

초기 서울시 창업 추천 시스템을 위해 설정한 흐름은 다음과 같다.

1) 사용자가 창업 희망 업종, 행정동 단위의 창업 희망 지역을 입력하고 해당 업종이 존재하는 업종인지를 확인한다. 만약 존재하지 않는다면 분석이 어렵기 때문에 추천이 불가능하다.
2) 네이버 지도에서 사용자가 입력한 업종을 검색하여 '많이 찾는' 순으로 정렬하였을 때(네이버 지도 기능) 상위에 있는 10곳의 위치 정보를 가져온다.
3) 가져온 10곳의 가게 반경 500m 내의 상권을 분석하여 추천 기준을 설정한다.
4) 사용자가 입력한 지역을 일정 단위로 쪼개어 구역으로 나누어 분석을 진행하고, 위에서 설정한 기준을 통과한 구역들 중에서 위험 요소가 있는 곳들을 제외한 뒤 최종적으로 선정된 위치를 사용자에게 추천한다.

2.1.2. 분석방법

흐름 3번에서 진행되는 분석 방법은 다음과 같다.

- **유동인구 분석**

  유동인구를 그 지역의 상권을 이용할 만한 사람의 수라고 정의하여 유동인구가 많을수록 더 많은 소비자가 상권을 이용할 것으로 추정한다. 또한 집객시설은 고객을 유입시키는 흡입력을 가졌기 때문에 집객시설의 수가 많을수록 유동인구가 더 많이 발생할 것으로 추정한다.

1. 대중교통 승/하차 인원

   사용 데이터: 서울시 버스노선별 정류장별 시간대별 승/하차 인원, 서울시 지하철 호선별 역 별 시간대별 승/하차 인원

   위 데이터를 사용해 상위 10개 가게의 반경 500m 내의 대중교통 승/하차 승객 수의 평균을 구한다. 환승역일 경우 단순 환승 인원을 제외하기 위해 승차 승객 수와 하차 승객 수를 더한 수치에서 환승 인원 수치를 빼 계산한다.

2. 집객시설

   사용 데이터: 공공데이터 포털 소상공인 시장 진흥공단 상가(상권)정보 API

   위 데이터를 사용해 상위 10위 가게의 반경 500m 내의 집객시설 별 평균을 구한다. 집객시설이 적은 곳과 많은 곳의 편차를 고려해 가장 큰 수치와 작은 수치를 제외하고 계산한다.

- **교통량 분석**

  교통량을 소비자들이 상권을 쉽게, 편리하게 이용할 수 있는 접근성의 정도로 정의하여 교통량이 많을수록 소비자들이 더 많을 것으로 추정한다.

  사용 데이터: 서울시 노선 별 정류장별 총 버스 운행횟수, 서울시 노선 별 역 별 총 지하철 운행횟수 API

  위 데이터를 사용해 상위 반경 500m 내의 교통량의 평균을 구한다.

상위 10개 가게의 주변 상권 분석 방법은 위와 같으며 위에서 구한 수치들로 추천 기준이 설정된다. 사용자가 입력한 지역을 작은 단위의 구역으로 나누었을 때 이 기준을 통과한 구역들을 선별하여 1차 추천 후보들을 추린다. 이후 흐름 4번에서 언급한 위험 요소가 있는 곳들을 제외한다. 위험 요소는 구역내에 있는 같은 업종의 가게들, 곧 경쟁업체들 중에서 별점이 4.5가 넘는 곳의 비율이 50%가 초과할 경우를 뜻한다. 이 과정을 거쳐 최종 위치를 사용자에게 추천한다.

2.2. 방향성 수정

1. 모든 업종에 동일한 분석 기준을 적용

초기 설정에서 모든 업종에 동일한 분석 기준을 적용하였기 때문에 획일화된 결과를 낼 수 있다는 문제점이 있었다. 처음부터 잘 되는 가게의 기준을 네이버 지도에서 제공하는 '많이 찾는'순 top10으로 정의해 top10의 주변 상권만을 분석했고 소수의 데이터만을 사용해 해당 업종의 전체 특성을 결정한다는 것에서 신뢰성이 떨어지게 된다.

따라서 잘 되는 가게를 정의하지 않고 단위를 상권으로 바꾸어 해당 상권의 패턴이 될 수 있는 데이터들을 분석해 최대한 많은 변수를 반영할 수 있도록 하고 이 데이터들을 클러스터링을 통해 유사한 특성을 가진 데이터끼리 묶어서 해당 업종별로 차별화된 분석 기준을 설정할 수 있다.

2. **높은 수치를 중점으로 두고 분석**

   단순히 수치가 높다고 하여 좋은 지표라고 생각하지 않고 각 업종들의 핵심 특성을 파악하기 위해 변수들 간의 관계를 알아보고 비교를 통해 중요성의 정도를 알아낸다. 이 분석 기준을 적용하여 수치가 높은 것에만 치중하지 않고 업종마다 중요한 변수를 파악하여 점수를 다르게 매길 수 있다.

3. **업종별 어떤 변수가 중요한지 사용자에게 조언**

   업종별 패턴을 분석해 해당 업종에 중요하게 작용하는 변수에 대한 정보를 사용자에게 제공할 수 있다.

2.3. 현재 설정

현재 서울시 창업 추천 시스템을 위해 설정한 흐름은 다음과 같다.

1) 사용자가 창업 희망 업종과(10가지 중에 선택) 창업 희망 지역을(행정동 단위 지역) 입력한다.
2) 시스템에서 사용자가 입력한 지역에 해당하는 모든 상권 데이터를 가져온다.
3) 가져온 상권들이 해당 업종의 어느 *군집에 속하는지 분류한다.
4) 3번의 결과를 해당 상권의 첫번째 가중치로 사용하고, 상권 변화 지표 데이터를 두번째 가중치로 사용하여 점수를 부여한다.
5) 사용자가 입력한 지역의 지도를 가져와 상권 위치를 표시하고 점수를 시각화 하여 보여준다.

* 군집: 해당 업종이 그 상권에서 얼마나 잘되는지를 나눈 데이터.

예를 들어, A상권에서 한식 음식점이 얼마나 잘 될지를 알고 싶은 경우 군집데이터에 A상권의 특성들을 넣으면 80% 잘되는 상권 군집에 속하는지 70% 잘되는 상권 군집에 속하는지 알려준다.

3. 데이터 목록

3.1. 사용 데이터

6번을 제외하고 모두 상권 단위의 데이터이다.

1) 추정 매출 API
   - 요식업 업종별로 분류
   - 매출 건수, 매출 금액, 매출 비율 데이터끼리 분류
   - 점포 수 결측 치 치환
2) 집객시설 API
3) 상권영역 API
   - 상권영역 데이터를 상권코드를 기준으로 매출데이터와 조인해 사용자가 선택한 지역의 상권들을 불러올 때 사용
   - 상권 중심좌표와 면적 유동인구 추정에 사용
4) 생활인구 API
5) 상주인구 API
6) 유동인구
   - 지하철 시간대별 승/하차 인원 정보 API + 지하철역 좌표
   - 버스정류장 시간대별 승/하차 인원 정보 API + 버스정류장 좌표
   - 상권의 중심 좌표를 기준으로 반경 내의 유동인구 추정하기

3.2. 특이사항

행정동 단위 지역에 해당 업종 상권이 하나도 없는 경우가 있는가?

→ 있다. 행정동 코드의 개수가 400개인데 한식 업종의 상권코드와 행정동 코드를 조인한 결과 유니크한 행정동 코드가 399개 존재한다.

행정동 단위 지역에 상권이 하나만 있는 경우가 있는가?

→ 한식, 카페 업종을 제외한 나머지 업종에서 지역에 상권이 딱 하나 존재하는 경우가 있다. 이 경우 다른 상권과의 비교는 불가능하므로 사용자가 선택했을 시 상권이 하나밖에 존재하지 않으므로 비교가 불가능하여 상권의 정보만 제공된다고 알린다.

4. 데이터 패턴 분석

4.1. 지도 학습 vs 비지도 학습

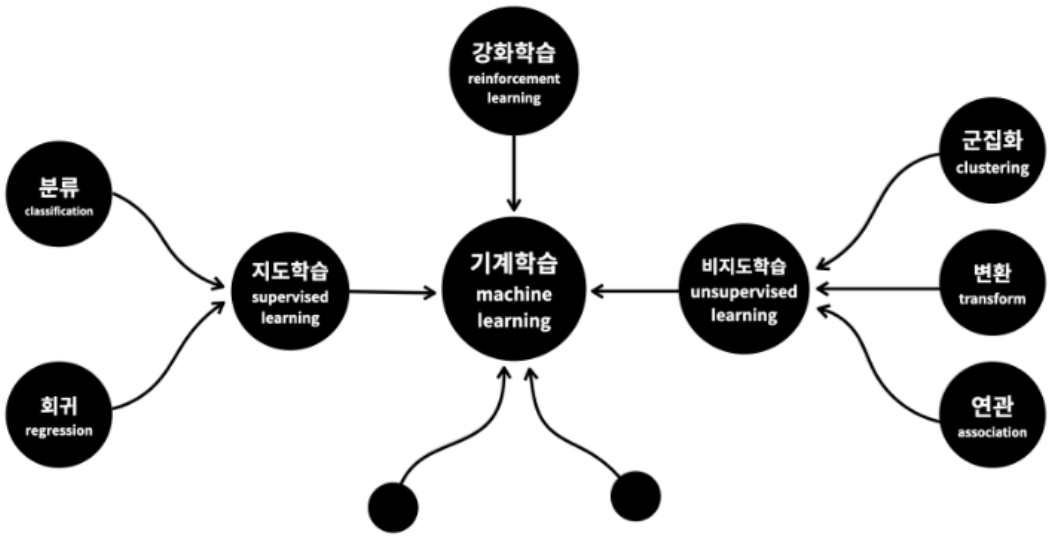

그림 1. 머신러닝의 분류 (출처: 생활코딩)

지도학습(Supervised Learning)

지도학습의 '지도'는 기계를 가르친다는 의미이다. 지도학습의 알고리즘들은 정답이 있는 데이터들을 학습한다. 이런 정답들은 사람들이 직접 데이터에 표기한 것이다. 예를 들어 사진 분류를 지도학습을 통해 해결하고자 한다면 굉장히 많은 개와 고양이들의 사진이 필요하고 각 사진들이 개인지 고양이인지를 표기, 라벨 해주어야 한다. 이렇게 label된 데이터 묶음을 훈련 데이터로 사용하여 기계는 사진이 개인지 고양이인지를 예측한다. 지도학습의 훈련은 이 예측 값과 정답(label)이 같아지도록 지도한다.

비지도학습(Unsupervised Learning)

비지도학습의 알고리즘들은 정답(label) 없이 데이터를 학습한다. 사람이 정답을 정해 놓는 것이 아니라 스스로 데이터를 학습하여 그 속의 패턴(pattern) 또는 각 데이터 간의 유사도를 학습한다. 대표적인 기술로 클러스터링(clustering)과 연관(association)이 있다. Clustering은 서로 가까운 관측치를 찾아 비슷한 특징의 데이터들끼리 군집화 시키는 기술이다. 표로 본다면 비슷한 행을 그룹핑하는 것으로 볼 수 있다. 이 도구를 사용하면 수많은 행이 있는 데이터를 입력하더라도 원하는 만큼의 클러스터를 만들어준다.

연관은 데이터간의 관계(relationship) 혹은 패턴(pattern)을 학습한다. 표로 본다면 연관규칙은 서로 관련이 있는 특성(열)을 찾아주는 기법이다. 연관성을 파악할 수 있다면 추천에도 사용할 수 있다.

4.2. K-means 알고리즘

- **K-Means Clustering**

1) k개의 초기값을 설정한다. 초기값은 사용자가 설정할 수도, 랜덤한 값이 될 수도 있다.
2) 각 데이터들을 가장 가까운 중심점으로 할당한다. 그룹화되는 것이다.
3) 새로 들어온 데이터를 포함해 중심점을 업데이트한다.
4) 더 이상 중심점이 업데이트 되지 않을 때까지 2번~3번을 반복한다.

- **장점**

1) 알고리즘이 매우 간단해 시간적 성능면에서 매우 빠르다. 시간 복잡도가 O(n)이다.
2) 대용량 데이터를 다루기에 적합하다.
3) k값이 많아질 수록 순도 높은 결과를 도출한다.

- **단점**

1) 초기값 설정에 따라 결과가 달라질 있다. 즉, 초기값을 잘 설정하는 것이 매우 중요하다.
2) 군집화 할 클러스터링 개수를 설정하는 것이 어렵다. 최적의 k를 찾기 위해 도움을 줄 수는 있지만 정답이라고 할 수는 없다는 한계가 있다.

5. 향후 추진 계획

- 클러스터링을 통한 업종별, 상권별, 지역별 특성 정리
- 정리한 것을 바탕으로 상권 점수 부여(추천) 알고리즘 구현
- 데이터 시각화, 사용자 UI 구현