|  |
| --- |
| **제11회「2023 빅콘테스트」데이터 분석 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 ☑ 표시 | | |
| **참가분야** | □ 생성형AI 분야 □ 데이터신기술 분야  □ 정형데이터 분석 분야 □ 비정형데이터 분석 분야  ☑ 빅데이터플랫폼 활용 분야 | | |
| **세부리그**  \*해당시 체크 | □ 어드밴스드 리그 □ 스타터 리그  \*정형데이터 분석분야에 한함(선택) | | |
| ☑ 지정주제 리그 □ 자유주제 리그  \*빅데이터플랫폼 활용분야에 한함(선택) | | |
| **개인/팀여부** | □ 개인 ☑ 팀(총 4명) | **개인/팀명** | 신사임당 |
| **지도교사명** | \*스타터 리그에 한함(선택) | | |
| **대표ID** | vicky4150@korea.ac.kr | | |

※ **5장 내외로** 목차는 준수하여 자유롭게 작성

※ 제출자료는 평가에 반영 예정

1. **분석 주제명**

지역경제 활성화를 위한 필지, 업종별 소상공인 매출등급 예측모형 개발

1. **분석 배경 및 목적**

통계청(2021)에 따르면 국내 사업체 중 87.9%가 소상공인으로 분류되며, 국내 종사자 중 37.7%가 이 업종에서 일하고 있어 소상공인이 국민 경제의 중요한 주춧돌 역할을 한다. 특히 서울시의 소상공인은 전체 소상공인 사업체수 대비 21.6%를 차지하고 있어 지역 경제에 큰 영향을 미친다. 현재 급변하는 경제 환경 하에서 대기업의 강점인 규모의 경제와 같은 ‘정태적 우위’보다는 끊임없는 혁신을 통한 ‘동태적 우위’가 더 중요해진 시대이며 여기에서 소상공인이 중요한 역할을 하고 있다.

또한, 한국은 OECD 회원국 중에서 자영업자 비중이 평균 15.9%보다 약 2배 높은 28.8%로 경쟁이 치열한 환경임을 나타낸다. 특히 요식업(음료 및 음식점업)과 같이 진입장벽이 낮아 창업과 폐업 비율이 높은 업종에서는 정확한 매출 진단이 필요한 실정이다.

이 분석의 목적은 먼저 소상공인 중에서 특히 요식업의 매출등급을 예측하는 모델을 개발하여 업종의 건전성을 평가하고 매출을 안정화하는 방법을 찾는 것이다. 그 후, 모델의 예측 결과를 활용하여 폐업 위험에 노출된 소상공인을 식별하고 조기에 지원하는 방안을 모색하여 폐업을 예방하고 지역 상업용 부동산의 공실률을 낮추는데 기여한다. 더 나아가, 매출 예측 결과를 활용하여 상권 활성화를 위한 전략을 개발하고 어떻게 요식업 업종을 유치하고 지원할 것인지에 대한 아이디어를 제시하여 상업용 부동산 시장을 활성화시키고 국내 경제에 기여하고자 한다.

1. **분석 내용 요약**

서울시 소상공인의 매출등급 예측을 위해 필요한 변수들을 추가하고, 변수 간의 연관성 파악 및 결측치 처리와 같은 데이터 전처리를 수행한다. 주로 트리 기반 모델을 활용하여 모델을 구축하고 파라미터 튜닝을 통해 예측 성능을 최적화한다. 이렇게 구축된 모델을 통해 어떤 변수가 매출등급 예측에 중요한 역할을 하는지 확인하고, 이 정보를 활용하여 소상공인들에게 전략적인 조언을 제공할 수 있다.

1. **분석방법 및 계획**
   1. **데이터 전처리 방법 및 활용 내용** 
      * 1. 데이터 가공

* 필지번호(PNU코드) 기준의 데이터와 행정동코드 기준의 데이터를 병합하기 위해 두 코드 사이의 관계를 파악해야 한다. 행정동 코드는 행정기관 주소이지만, PNU 코드(19자리)는 지적도 관련코드로 정해진다. 추가한 데이터들이 대부분 행정동을 기준으로 하고 있기 때문에, 필지와 행정동 공간 데이터를 활용하여 QGIS로 필지번호 당 행정동명을 매칭하였다.
* 이때 원본 데이터 필지번호와 매칭되지 않는 104개의 필지번호에 대해, 필지번호 앞 10자리(법정동)를 슬라이싱하여 이를 기준으로 매칭 후, groupby 하여 평균값으로 대체하였으며 10자리로 매칭되지 않는 데이터들에 대해서는 필지번호 앞 5자리를 기준으로 매칭하였다.
  + - 1. 임대료 데이터 결측값 처리
* 결측값을 살펴보면, ‘전체 임대료’, ‘1층 임대료’, ‘1층 외 임대료’ 중 ‘1층 임대료’ 또는 ‘1층 외 임대료’에서 결측치가 있거나, 2022년 모든 분기에 데이터가 결측치이고 2023년 1분기에만 값이 있는 총 2가지의 경우가 있었다. 첫번째 경우인 ‘1층 임대료’ 또는 ‘1층 외 임대료’에서 결측치가 있는 경우에는 결측값 해당 분기 기준, 전월 대비 전체 임대료 변화율을 반영하여 (전분기\*전체 임대료 변화율)을 통해 결측값을 대체했다. 또한 두번째 경우인 2022년 모든 분기에 데이터가 결측치이고 2023년 1분기에만 값이 있을 때는 생활물가지수(CPI)를 활용하여 2022년 1분기부터 2023년 1분기까지의 분기별 CPI 평균값의 변화량을 계산하여 2023년 1분기부터 역산하여 결측값을 대체했다.
  + - 1. 같은 행정동명 구별
* 서울시 상권분석 데이터에 관악구, 강남구 둘 다 신사동이 있어 구분이 필요하였고 이를 위해 관악구 신사동을 신사동1로, 강남구 신사동을 신사동2로 정의하기로 하였다.
* 필지번호와 행정동을 매칭하는 파일 역시 관악구 신사동에 포함되는 필지의 행정동명은 신사동1, 강남구 신사동에 포함되는 필지의 행정동명은 신사동2로 처리하였다.
  + - 1. 변수 간 상관계수 분석 및 변수선택
* 점포수 데이터 내에서 ‘일반점포수’와 ‘전체점포수’의 상관계수가 0.9, ‘프랜차이즈점포수’와 ‘전체점포수’의 상관계수가 0.74이다. 전체점포수가 ‘일반점포수’+’프랜차이즈점포수’여서 전체점포수 변수를 빼고 ‘일반점포수’, ‘프랜차이즈점포수’ 변수만 사용하기로 결정하였다.
* 65세 이상 인구수에서 성별을 나누는 것은 중요하지 않다고 판단하여 ‘65세이상\_남’ 변수와 ‘65세이상\_여’ 변수를 합친 변수인 ‘65세이상’ 변수만 사용하기로 하였다.
* 임대료 데이터에서 ‘전체 임대료’ 변수가 ‘1층 임대료’, ‘1층 외 임대료’ 변수들을 포함하는 변수여서 ‘전체 임대료’는 빼고 ‘1층 임대료’, ‘1층 외 임대료’ 변수만 사용하기로 결정하였다.
  + - 1. 데이터 표준화

범주형 변수인 ‘대학교여부’, ‘업종코드’ 변수를 제외하고 모두 Standard Scaler를 적용하여 데이터를 표준화시킨다.

* 1. **분석에 활용되는 추가데이터 (출처 기재)**

|  |  |  |  |
| --- | --- | --- | --- |
| **원본 데이터** | **사용할 변수** | **출처** | **상세 설명** |
| 점포수 | 전체점포수,  프랜차이즈점포수,  일반점포수 | 서울시 상권분석 서비스 - 지역·상권별 현황 |  |
| 개폐업수(률) | 개업수, 폐업수, 개업률, 폐업률 | 서울시 상권분석 서비스 - 지역·상권별 현황 |  |
| 인구수 | 길단위유동인구, 주거인구, 직장인구 | 서울시 상권분석 서비스 - 지역·상권별 현황 |  |
| 임대시세 | 전체 임대료,  1층 임대료,  1층 외 임대료 | 서울시 상권분석 서비스 - 지역·상권별 현황 |  |
| 부동산거래량(금액) 대비 유동인구 | 부동산거래대비 유동인구 | 부동산 빅데이터 플랫폼 | 유동인구 점수 |
| 젠트리피케이션 위험지수 | 젠트리피케이션 | 부동산 빅데이터 플랫폼 | 젠트리피케이션 점수 |
| 공실률 대비 매매가  임대료 | 공실률 대비 매매가 임대료 | 부동산 빅데이터 플랫폼 | 매매가 점수 |
| 서울시 생활물가지수 통계 | 생활물가지수 | 서울 열린데이터광장 |  |
| 행정안전부\_지역별(행정동) 성별 연령별  주민등록 인구수 | 65세이상, 65세이상\_남,  65세이상\_여 | 행정안전부 | 지역별(행정동별), 성별, 연령별로 제공되는 주민등록 인구수 중에서 만 65세 이상 인구수만 집계하여 사용 |
| 수도권 지하철  공간데이터 | 지하철 개수 | 서울특별시  빅데이터 캠퍼스 | 지하철 좌표 기준으로 500m 크기의 버퍼(buffer)를 생성한 후, 필지번호별로 포함하고 있는 지하철 버퍼의 개수를 집계하여 사용 |
| 서울시 대학교  공간데이터 | 대학교여부 | 서울특별시  빅데이터 캠퍼스 | 대학교 좌표 기준으로 500m 크기의 버퍼를 생성한 후, 필지번호별로 대학교 버퍼를 포함하고 있는지(포함하면 1, 포함하지 않으면 0) 여부를 계산하여 사용 |
| 스타벅스 매장 위치 | 스타벅스 개수 | 스타벅스 홈페이지 크롤링 | 필지번호별로 포함하고 있는 스타벅스 매장의 개수를 집계하여 사용 |
| 서울교통공사 연도별 일별 시간대별 역별  승하차 인원 | • 출근 시간대(7am~9am) 역별 승차 인원  • 출근 시간대(8am~10am) 역별 하차 인원  • 주말(10am~20pm) 역별 하차 인원 | 서울 열린 데이터 광장 | 1~9호선 |

* 1. **분석에 적용·활용할 통계·분석 기법, 방법론**
     + - 분석방법

1. Pearson 상관계수를 이용하여 높은 상관관계를 보이는 변수들을 1차적으로 제거한다.

2. 매출등급 예측을 위한 모델로는 feature importance를 뽑아낼 수 있는 트리 기반 모델(XGBoost, Random Forest, LightGBM 등)을 활용한다.

3. 결정된 모델의 feature importance를 계산하여 낮은 중요도를 갖는 변수들을 2차적으로 제거하여 유의미한 변수들을 추려낸다.

4. 파라미터 튜닝을 통해 모델의 예측정확도를 향상시킨다.

* 1. **분석 결과에 대한 시각화 방법 등**

Tableau를 통한

* 공간별(필지번호, 행정동, 시군구) 매출등급 시각화
* 중요도가 높은 변수들을 기준으로 공간별(필지번호, 행정동, 시군구) 시각화

1. **분석결과 활용 및 시사점**
2. 적용대상, 활용방안 및 기대효과
   1. 정부

* 예측 매출등급이 낮은 필지에 대한 특징 및 원인을 분석하여 더 효율적인 소상공인 지원방안을 낼 수 있고 이를 통해 지역경제 활성화를 이뤄낼 수 있다.
  1. 부동산 업계
* 부동산 회사 및 중개 업체는 소상공인의 매출등급 예측을 통해 해당 필지의 임대 및 인수를 검토하고 소상공인에게 조언을 제공한다.
* 필지 별 예상 매출등급으로 해당 필지에 부동산 투자 수익성을 평가하고, 어느 지역에 부동산 투자를 할지 의사결정에 도움을 주어 투자의 효율성을 높인다.
  1. 소상공인
* 필지 별 업종별 매출등급 예측모형을 활용하여, 사업체를 낼 때 장소, 업종 등의 선택 시 합리적인 선택을 할 수 있다.
* 금융사 또한 모델 기반의 예상 매출등급을 활용하여 금융 상품을 낼 수 있으므로 소상공인이 금융자금에 더 쉽게 접근 가능하며 높은 신용평가와 더 나은 조건의 대출을 받을 가능성이 높아진다.
* 분석결과를 토대로 효율적인 경영 전략 수립 및 자금 관리가 가능하며 특정 지역에 맞춤형 비즈니스 전략을 개발할 수도 있다. 이는 상권 활성화를 촉진시킬 수 있으며 결과적으로 지역경제 발전에 기여할 수 있다.

1. 시사점

앞선 활용방안과 같이 정부, 부동산 업계, 소상공인이 예측 모형을 활용하여 유기적으로 긍정적인 영향을 주고받을 것으로 보인다. 이러한 시너지 효과를 통해 지역 경제 활성화, 더 나아가 국가 경제 활성화에 이바지할 수 있을 것으로 기대된다.