|  |
| --- |
| **빅데이터 분석 정의서** |

|  |  |
| --- | --- |
| **과제명 :** | **빅데이터를 이용한 오피니언 마이닝 기반 창업 추천서비스**  **“스타트라이트"** |

**2023.04.28**

**Ⅰ. 개요**

1. 아이디어 주제

: 상권 정보 데이터와 SNS 데이터를 각각 분석하여 창업 업종 등을 추천하는 서비스

2. 개발 목표

: 예비 창업자에게 상권에 맞는 업종을 추천하는 서비스

3. 개발 내용

: SNS 데이터를 이용한 오피니언 마이닝

: 상권 정보 데이터 분석의 결과들을 이용하여 점수로 수치화

**Ⅱ. 기능별 빅데이터 분석 명세서**

|  |  |
| --- | --- |
| **기능명** | 업종 추천\_종합평가 |
| **데이터 준비** | |
| **데이터 정의** | 상권생활인구, 직장인구, 상주인구, 점포, 집객시설, 추정매출 등 상권관련 데이터 |
| **데이터 획득 방법** | 서울 열린데이터 광장(https://data.seoul.go.kr/)에서 상권관련 정보 2017년부터 2022년 분기별 데이터 약 500만개 |
| **2. 전처리** | |
| **전처리 과정** | 1. 데이터의 정합성을 고려해 기준연도, 기준분기, 상권코드를 기준으로 merge 함수로 각 데이터를 합침  2. 상권관련 데이터는 왜곡정규분포 형태로 하한과 상한의 차이가 큰 것을 확인 할 수 있어서 평균을 조정하기 위해 하한에 위치한 데이터 중 일부(약 30%)를 drop함  3. 집객시설의 경우, 상권 데이터에 75%가 존재하지 않는 경우가 있어 해당 데이터(은행 수)에 대해 2개 이상 있는 경우를 모두 2개라고 간주함.  4. 인구관련 데이터의 경우, 데이터의 편차가 심해 하한에 위치한 데이터 중 일부를 drop한 뒤, 사분위수 범위(IQR)의 1.5배만큼 threshold 조건을 부여해 outlier를 제거함,  5. 추정매출의 경우, 회귀모델에서는 데이터의 유효숫자를 만()단위로 고려해 전처리함. 분류모델에서는 데이터의 분포를 확인해 매출을 8구간으로 나눠 각 구간에 대해 데이터 개수가 최대한 균일하게 나눔 |
| **3. 모델 생성 및 학습** | |
| **모델링 목표** | 상권위치(자치구), 집객시설, 인구(직장,상주,생활), 업종(한식, 중식 등)를 독립변수로 고려해 종속변수(추정매출)을 예측하는 모델 |
| **모델링 가능 알고리즘** | lightgbm(회귀), lightgbm(분류) |
| **학습** | GridSearchCV를 이용해 하이퍼파라미터 조정 |
| **4. 검증** | |
| **모델링 검증 방안** | 1. k-fold cross validation으로 테스트와 동시에 검증,  2. 2022년 4분기 실제 강남구 추정매출과 모델을 통한 추정매출에 대한 비교 |
| **모델링 평가 결과** | 1. lightgbm(회귀)의 경우 R2 score : 0.956, rmse : 103.8  2. lightgbm(분류)의 경우 accuracy : 0.45, f1 score : 0.26 ~ 0.72 |

|  |  |
| --- | --- |
| **기능명** | 업종 추천\_SNS 점수 |
| **데이터 준비** | |
| **데이터 정의** | 강남구 소재의 요식업종별 리뷰 데이터 |
| **데이터 획득 방법** | 1.학습용 데이터는 네이버 영화, 쇼핑, 게임 리뷰데이터(약 50만개)  2.업종별 약 10개의 가게에 대해 식당 리뷰글과 별점 크롤링(총 약 3천개) |
| **2. 전처리** | |
| **전처리 과정** | 1. Okt 형태소 분석기로 리뷰를 형태소 단위로 토큰화하고 불용어사전을 통해 조사, 지시대명사, 도메인을 고려한 단어를 제외시킴  2. Counter 패키지로 형태소 단위로 토큰화된 단어의 빈도수를 확인 |
| **3. 모델 생성 및 학습** | |
| **모델링 목표** | 리뷰 감성분석 |
| **모델링 가능 알고리즘** | RNN 모델(BiLSTM) |
| **학습** | Sequential 모델에 임베딩된 단어에 대해 BiLSTM층으로 문장구조 양방향 학습 |
| **4. 검증** | |
| **모델링 검증 방안** | 50만개의 리뷰의 20%를 분할해 감성분석 score 점수 검증 |
| **모델링 평가 결과** | 테스트 정확도 0.79 |