**KB국민은행 캡스톤 회의록**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **회의일시** | **2018.09.28** | **부서** |  | **작성자** | 정상대, 오동건 |
| **참석자** | 송예은, 오동건, 김하준, 정상대 | | | | |

|  |  |
| --- | --- |
| **회의안건** | 1. **서울시 우리마을 가게 데이터 설명.** 2. **종합 의견** 3. **향후 과제** 4. **데이터 질의** |

|  |  |
| --- | --- |
| **회의내용** | **내용** |
| 1. **데이터 설명**   **예은**  **상주인구**   * 변수 : 년월, 상권코드, 상권코드명, 상주인구수, 성별에 따른/ 남성&나이/ 여성&나이, 가구 수, 비 아파트 가구 수, 아파트 가구 수, 인구 수 * 데이터 제공주기 : 반기(6개월마다) * 상권 1742개, 배후지상권은 1742개보다 적다.   집객시설   * 변수 : 년월, 상권코드, 상권코드명, 관공서, 은행, 병원, 학교, 백화점 등등..   변수별 개수가 카운팅되어 있으며 NA가 많다   * 데이터 제공주기 : 1년 * 배후지는 공항 등 특수한 건 없지만 나머지는 다 들어있다.   **상대**  **점포**   * 변수 : 년월코드, 상권코드, 상권코드명, 서비스업종코드, 서비스업종코드명,   점포수, 유사업종점포수, 개업율, 개업점포수, 폐업률, 폐업점포수   * 상권배후지의 경우 점포수~폐업점포수 변수가 모두 0값임에 불구하고   우리마을가게 상권분석서비스 측에서는 옳은 것이라 함.   * 상권데이터 내에서는 0값만 있는 건 아니나, 폐업률이 100%인 row가 1677개,   개업률이 100%인 row가 1986  **직장인구**   * 변수 : 년월코드, 상권코드, 상권코드명, 총 직장 인구수, 남성/여성 직장 인구수.   연령대별 직장인구수, 연령&성별 직장인구수   * 0값이 많으므로 log를 취해 데이터를 보았으나, 상권과 상권배후지의 차이로 인한 특징은 존재하지 않았음   **하준**  **유동인구**   * 성별 /연령별/ 시간대별/ 요일별/ 성별&연령대별 유동인구 * 유동인구 추정에 대한 것까지 대리님께 드리도록 하자. * EDA를 조금 더 해봤으면 좋겠다. Gather를 통해 분석을 좀 더 편하게 하고,   Spread를 통해 그래프를 볼 수 있었다.  **매출**   * 성별 /연령별/ 시간대별/ 요일별/ 성별&연령대별/ 주중/ 주말/ 매출금액/ 매출비율/ 매출비율건수/ 당월매출금액/ 당월매출건수/ 평균영업개월수/ 생존율 * 생존율 - NA값이 잡혀 있음.   (1년미만/ 1-2년/ 2-3년/ 3-5년/ 5년이상에 따른 생존율, 0~100%)   * 데이터가 너무 커 merge시 문제가 있기에 merge를 위해 분산처리를 고려해야할 필요도 있다.   **동건**  **지수지표**   * 과밀지수(신규창업위험지수) 창업위험도, 활성도, 성장성, 안정성   (창업위험도 행정동 단위. 그 외 골목상권 단위.)   * 과밀지수를 보안하기 위해 행정동을 기준으로 하는 창업위험도를 새로 만들어서 사용. 과밀지구의 문제점은 나와있지 않고 어떤 점을 보완했는지도 확인 불가. * 과밀지수 안에 신규창업지수가 포함되는 데, 부적절하다고 판단. 하위로 들어가는 것이 부적절하다. 신규창업위험지수는 골목상권 내에서 신규창업하는 점포와 같은 업종이 없을 때 사용한다. 그러나 경쟁점포가 없으면 과밀을 말하기 부적절. 신규창업위험지수 네이밍에 문제가 있다. * 신규창업위험지수와 다르게 창업위험도도 있는데, 두 지표는 서로 상이다. 신규창업지수는 회귀식을 이용해 매출을 추정해서 모델링. 그러나 창업위험도는 개폐업률, 생존률 등을 통해 모델링. 따라서 또 다른 지표인 안정성과 더욱 유사. * 활성도, 성장성, 안정성은 정형데이터를 가지고 통계를 낸 것이므로 깔끔하여   문제없이 그대로 사용할 수 있을 것이다.   * 결국, 과밀지수는 매출을 예측하기 때문에 수익성을 포함하는 네이밍이 적절   했을 경우, 그 상권 내에서 해당 업종의 수익성/활성도/성장성/안정성을 보여주  고 SNS데이터를 포함하여 사용함이 옳지 않을까.  **소득소비**   * 기준년월/상권코드/상권코드명/월평균소득금액/소득구간/지출총금액/식료품,의류,생활용품,의료비,교통,여가 등등 항목에 따른 지출 금액 * 제공단위는 골목상권의 평균 소득과 지출. * 반기(6개월)별로 업데이트가 된다고 하였으나, 2년간 데이터가 모두 같다. * 상권코드 결측치(300개 가량)   **아파트**   * 기준년월/상권코드/상권코드명/아파트단지수/아파트면적별 세대수(5범주)/아파트 가격(7개범주)/아파트평균면적/아파트평균시가 * 상권코드 결측치(300개 가량)  1. **의견**  * 지수지표들끼리 중복되는 경우가 많고 네이밍이 부적절한 지표도 있다. 또한 내용이 어렵기 때문에 보다 직관적인 네이밍이 필요. * 상권 내 해당업종의 수익성/활성도/성장성/안정성을 보여주는 것을 최종 Output * 활성도, 성장성, 안정성은 정형데이터를 통해 인덱스 가공. 인덱스 모델링은 기존의 상권분석 사이트, 논문을 활용해서 합리적인 기준 마련. * 수익성은 창업 시 예상 매출을 의미. 통계 모형 또는 머신러닝을 통한 학습결과를 바탕으로 예측.  1. **다음 할 일**  * 매출, 유동인구, 지수지표는 중요하다고 판단되기에 데이터를 모두 각자 보도록 하자. * 데이터를 통해 만들어 낸 지수지표의 문제, 데이터를 통해서 말 할 수 있는 것 파악 * 분산처리  1. **데이터 질의**  * 1. 추정매출에 관련되어 이것이 카드사데이터를 통한 관측인지 추정인지   답: 관측입니다. 추정이라는 표현을 쓴 이유는 모든 카드사를 대상으로 한 것이 아니기 때문에 카드사의 점유율로 가중치를 두어서 매출을 추정한 것 입니다[[1]](#endnote-1).   * 2. 과밀지수가 어떠한 문제점이 있어 행정동/시군구을 기준으로 하는 창업위험도를 만든것인지   .   * 3. 점포데이터 내 유사업종점포수에서 유사업종이라는 기준이 어떻게 되어 있는 것인가 * 4. 블록이라는 것이 어떤것인지 * 5. 신규창업위험지수를 회귀식으로 추정할 때, 어떠한 변수를 사용했는지 (지수지표 알고리즘을 보았으나, 정확하게 나와있지는 않다.) |
|  |  |

1. 4.지수지표 알고리즘 개요 및 도움말 용어 정의10\_공개 [↑](#endnote-ref-1)