**KB국민은행 캡스톤 회의록**

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

|  |  |
| --- | --- |
| **회의안건** | 1. **추정매출, 과밀지수, 블록, 발달상권 ,배후지, 유사업종점포수 등 용어 명확한 정리 필요** 2. **고객들에게 친근한 워딩이 필요함** 3. **상권 내 해당업종의 수익성/활성도/성장성/안정성을 보여주는 것을 최종 Output** 4. **데이터시트 정리하기** |

|  |  |
| --- | --- |
| **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, 매출, 포화지수, 신규창업위험지수, 창업위험도, 활성도, 성장성, 안정성 (창업위험도는 행정동으로했고 나머지는 상권단위로 되어있음.) * 과밀지수가 어떠한 문제점이 있어 행정동을 기준으로 하는 창업위험도를 사용했으며, 어떠한 점이 보완되었는지를 알아야 한다. * 신규창업위험지수가 과밀지수의 하위로 들어감이 옳은 것인가.   경쟁점포가 없으면 과밀이라는 지수가 없는데, 이럴 경우 매출 포화지수를 사용할 수 없기에 신규창업위험지수를 사용한 것 같으나,  신규창업위험지수를 넣음 자체가 혼란스럽다 생각하며 워딩이 문제가 있다.  신규창업위험지수는 모델을 통해 예측을 한 값으로 만든 지수이며,  창업위험도는 다름에 불구하고 워딩이 매우 비슷하게 되어 있다.   * 과밀지수를 만들 때 회귀식을 사용하였는데, 어떠한 변수를 사용하였는지는   알려주고 있지 않기에, 추후 우리가 회귀식을 사용시 어떠한 변수를 사용했는지  앎이 매우 중요할 것 같다.   * 활성도, 성장성, 안정성은 정형데이터를 가지고 통계를 낸 것이므로 깔끔하여   문제없이 그대로 사용할 수 있을 것이다.   * 창업위험도를 빼고, 과밀지수는 수익성을 말하는 것이기에 수익성을 포함하는   워딩을 사용하여 매출을 추정하는 것을 목표로 하여 상권 내 업종이 신규개업  했을 경우, 그 상권 내에서 해당 업종의 수익성/활성도/성장성/안정성을 보여주  고 SNS데이터를 포함하여 사용함이 옳지 않을까.  **소득소비**   * 소득에 대한 부분/ 지출에 대한 부분으로 나뉘어져 있다. * 평균 소득을 골목상권별로 제시한 것이며,   지출의 경우는 여가비/문화비/유흥비/의류 등에 대해 이루어져 있다.   * 반기(6개월)별로 업데이트가 된다고 하였으나, 2년간 데이터가 모두 같다.   **아파트**   * 아파트 면적별, 가격에 따라 가구수 등이 나타나 있다. * 결측치가 있어 전부를 대상으로 하지 못하는데,   이러한 처리를 어떻게 해야 할 지 (모든 데이터가 있는 row만을 하거나  다른 데이터를 import해서 쓸 방법이 있는지)     1. **의견**   좀 더 단순하고 명확해질 필요가 있다. 고객들의 접근성이 떨어짐이 크다.   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 |
|  |  |