EDA Report

|  |  |
| --- | --- |
| **Reference** |  |
| [UCI archive] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29  [Kaggle] https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018 | |

|  |  |
| --- | --- |
| **Dataset** | |
|  | |
| The number of training data **:** 161297 | The number of test data **:** 53766 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset info** | | Feature types | |
| Number of features  Total missing (%) | 7  0.5579% | Numeric  Categorical  Date  Dropped | 2  4  1  0 |

|  |
| --- |
| **Data analysis** |
| 전체 데이터 구성은 uniqueID를 가진 환자가 가지고 있는 증상에 필요한 약을 구입한 뒤에 특정 날짜에 review와 rating을 남김. 그리고 다른 사람이 해당 리뷰를 보고 도움이 되었는지에 대해 usefulCount feature에 점수(1점 추가)를 줌. |

|  |  |  |
| --- | --- | --- |
| **Feature Info** | | |
| Feature | Type | Description |
| uniqueID | Numeric | Identify individual data |
| drugName | Categorical | Name of drug |
| condition | Categorical | Name of condition |
| review |  | Patient review |
| rating | Categorical | 10star patient rating |
| date | Date | Date of review entry |
| usefulCount | Numeric | Number of users who found review useful |

|  |
| --- |
| **Condition, drugName** |
| The number of unique condition **:** 937  The number of unique drugName **:** 3671  Below 4 charts show the pattern between drugName and condition |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |  |
| --- | --- |
| **rating** | |
|  | |
|  |  |
|  | |

|  |
| --- |
| **date** |
| ->31일이 있는 달은 적기 때문에, 31일에 쓰여진 리뷰는 다른 날(day)보다 리뷰의 개수가 적음 |

|  |
| --- |
| **usefulCount** |
| 텍스트이(가) 표시된 사진  자동 생성된 설명  The usefulCount shows extreme between min and max. Also, standard deviation is high.  Below chart shows the relationship between usefulCount and date |

|  |
| --- |
| Relationship between usefulCount and rating      ->이거 한페이지에 따로 각 feature별로 분포도 다 같이 나타내면 될 듯 |

|  |
| --- |
| review |
| There is some html tag, error strings, uppercase word, and emotional parenthesis in reviews |
| "If not for this antibiotic my husband would have been dead several times over. Incidentally **I&#039;m** a Registered nurse. I see where a number have people have attributed wild **&quot;**side effects**&#039;** to this drug. Many are confusing symptoms of their illness with side effects of the antibiotic. Just because you experience something unusual for you when your sick doesn**&#039;**t mean its the fault of the antibiotic. Being ill is an abnormal state for ones body yet some of these people seem to think that any effect of the illness they weren**&#039;**t expecting **MUST BE** attributable to something else and often they blame the treatment rather than the illness. If you**&#039;**ve been prescribed Levaquin I urge you to take many of these **&#039;**reviews**&#039;** with more than a grain of salt! An **EXCELLENT** drug!" |
| Below 3 charts show N-gram      1, and 2- gram shows same distribution both positive(rating>5) and negative(rating<=5) at top 5 |

|  |
| --- |
|  |

|  |
| --- |
| Correlation heatmap of the features engineered |
|  |

|  |
| --- |
| 데이터셋의 한계점및 생각한 대안 |
| 저희 5명이서 각자 나뉘어 kaggle에 1위 부터 5위 까지 랭크된 포스트를 분석 하였습니다. 그러나 모든 모델이 비슷하게 리뷰를 이용해서 감성(positive, negative)분석을 진행하였고 최종 예측으로 rating을 맞추는 모델들이었습니다.  이미 진행된 모델들이 모두 리뷰를 사용하여 평점을 예측하는데 목적이 있어서, 리뷰데이터가 없다면 평점을 예측하기 어렵다는 단점이 있습니다. 이는 챗봇으로 사람들에게 증상을 입력받아 약을 추천해주는 서비스에는 적합하지 않은것 같습니다.  대안   1. 리뷰데이터로 단어를 학습 ex) (input) 배가 아프다 -> 학습된 데이터에서 유사도가 높은 약을 추천. (Word2Vec 활용) 2. 다른 추가적인 데이터셋 찾기 3. 챗봇에만 집중. rating+머신러닝을 활용한 점수 개발로 순위 매기기   -> 미리 증상별로 준비된 순위에 대해 단순 쿼리로 결과 보여주기  4. 웹사이트에서 블로그, 카페에서 의약품에 대한 후기 크롤링  -> RNN 모델에 적용하여 리뷰만으로 평점 예측하여 데이터 셋 구축 |