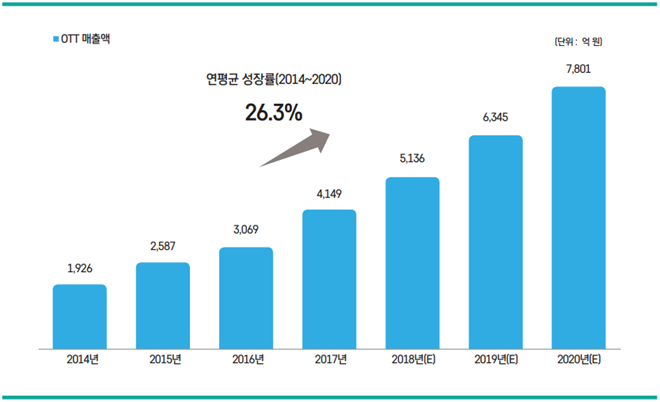
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **K-드라마 성공요인 분석 및 성공여부 예측**     |  |  | | --- | --- | | **지도 교수님** | **김일주 교수님** | | **제출 일자** | **2022-12-23** |  |  |  | | --- | --- | | 3조 | | | 60191664 | 여다은 | | 60191670 | 유은서 | | 60191651 | 류채현 | | 60191650 | 김자윤 | | 60181647 | 안수민 |   **목차**   1. **주제 선정**    1. **주제 선정 이유**    2. **목표** 2. **데이터 분석**    1. **전처리**    2. **데이터 분석**       1. **Data Set 구성**       2. **Data Set 분석** 3. **예측 모델링**    1. **예측 방법론**       1. **Text Mining**       2. **Naïve bayes와 Confusion Matrix**       3. **Logistic Regression과 KNN**       4. **방법론 사용 이유(모델링 목적과 기여도를 설명)**       5. **실제 활용 및 시사점** 4. **결론**    1. **느낀점** 5. **Appendix(부록)** |
|  |

1. **주제 선정**
   1. **주제 선정 이유**

코로나19 이후, 집에 머물러야 하는 상황이 길어지자 집에서 즐길 수 있는 오락거리에 대한 관심이 많아졌다. 이에 따라 집에서도 편하게 즐길 수 있고 다양한 미디어 콘텐츠를 인터넷을 통해 제공하는 OTT(Over the top) 산업이 크게 주목을 받기 시작하면서 K-드라마 또한 국내에서의 흥행을 넘어 전세계적인 인기를 누리고 있다. 이러한 K-드라마의 인기는 드라마 뿐만 아니라 OTT 산업의 발전에 크게 기여했고 본 프로젝트를 통해 드라마의 성공요소를 분석하여 OTT 산업 발전을 위해 유의미한 결과를 도출하고자 한다. 많은 사람들이 OTT 서비스를 이용했다는 것은 다양한 통계 자료가 보여주고 있다. (그림1 참조) 

-드라마의 흥행, 시청률이 OTT 산업 성공에 큰 영향을 미친다는 사실에 기반하여 시청률에 유의한 영향을 미치는 변수들을 도출하고, 도출된 변수들을 활용해 분석적으로 한국 OTT산업의 발전을 도모하고자 한다.

* 1. **목표**

K-드라마 1278편의 데이터를 분석해서 평점을 높게 받은 것들을 중심으로 인기도가 높은지 낮은지로 분류하여 K-드라마의 어떤 요소들이 인기도에 영향을 끼치는지 분석하고자 한다. 평점에 영향을 미치는 요소들 간에 연관관계를 분석하고, 드라마를 방영할 평점을 미리 예측하여 제작하는데 인지도를 미리 파악하여 사람들이 좋아할 만한 드라마를 내놓을 수 있도록 한다. 그렇게 인지도를 높임으로써 한국 드라마의 위상을 높이고 해외 진출을 도모하여 대한민국의 위상을 높이고 해외 진출을 도모하여 대한민국의 OTT플랫폼이 외국에서 큰 성공을 이룰 수 있도록 하고자 한다.

- K-드라마의 데이터를 기반으로 상관분석과 회귀분석을 이용하여 인기있는 드라마와 관련된 변수들과 그 수치(조건)들을 분석하고자 한다.

-‘Genre’와 ‘duration(러닝타임)’, Aired.On(요일), ‘tag’변수를 통해 어떠한 장르와 태그일 때 평점 혹은 시청률이 잘 나올지를 시각화를 통해 알아볼 예정이다.

-마지막으로 연도별 인기있는 장르 추이, 다음 년도 인기 장르 예상, OTT와 지상파 연도별 점유율, 시대를 풍미한 배우 리스트 등을 분석해서 최신 드라마의 트랜드를 알아보고자 한다.

- 미디어의 발전 및 성공은 OTT 서비스의 발전으로 어느 방송사에서 방영하느냐가 중요한 게 아니라 해당 OTT에서만 방영하는 Trendy한 컨텐츠를 방영하는가가 더욱 중요해 졌다. 그러므로 시대의 흐름에 맞게 시청률이 잘 나올 드라마 혹은 영화를 제작하는 것이 중요하다. 최근 드라마 트렌드를 파악하기 위해서 텍스트 마이닝을 통해서 키워드들을 추출하고, 그것으로 어떤 태그들이 주요 트렌드 인지를 파악할 것이다. 분석을 통해 파악한 주요 키워드들을 중심으로 시대의 흐름을 파악할 수 있으며 해당 태그가 포함된 드라마를 제작하도록 하여 OTT 산업의 발전을 도모하고자 한다.

1. **데이터 분석**
   1. **전처리**
2. **결측치 처리 Appendix-(3),(7),(9)**

kdramalist.csv의 컬럼 중 end.airing은 방영이 끝난 날을 가리키는데 중간에 중단된 드라마는 N/A로 아직 방영이 안끝난 드라마는 ?로 나타나 있다. 따라서 우리가 분석하기에는 정확한 데이터 수집이 완료된 상태가 아니기 때문에 제거하기로 결정 했다.

추가로 scored by에 두 레코드에서 N/A가 발견되었는데 이 2가지의 레코드의 다른 변수들에도 N/A가 발견이 되었고 분석하는데 의미가 없다고 판단되어 삭제하기로 결정했다. Genres 컬럼에서 발견된 하나의 N/A 또한 동일한 이유로 삭제하였다.

kdramalist.csv의 컬럼 중 Popularity는 mydramalist.com 홈페이지에 존재하는 전세계 모든 드라마의 시청자수에 따라 순위를 매긴 것이다. 이는 watchers (mydramalist.com에 집계된 시청자 수)컬럼으로 대체가 가능하다. 따라서 curse of dimensionality을 방지하기 위해 사전에 제거하기로 결정했다.

Duration에서도 na값이 발견되었는데 이 결측치에는 중앙값을 계산해서(60분) 넣어주기로 결정을 했다.

1. **차원 축소(컬럼 삭제) Appendix-(4)**

텍스트이(가) 표시된 사진

자동 생성된 설명

“N/A”의 개수(이 csv에서는 결측치를 문자열 N/A로 표현하였다 )를 확인해본 결과, imdb\_description, imdb\_rating, imdb\_user\_count 컬럼에서 너무 많은 “N/A”가 발견되어 분석에 의미없는 컬럼이라는 판단이 들어, 결측치 처리한 후 컬럼 자체를 삭제함.

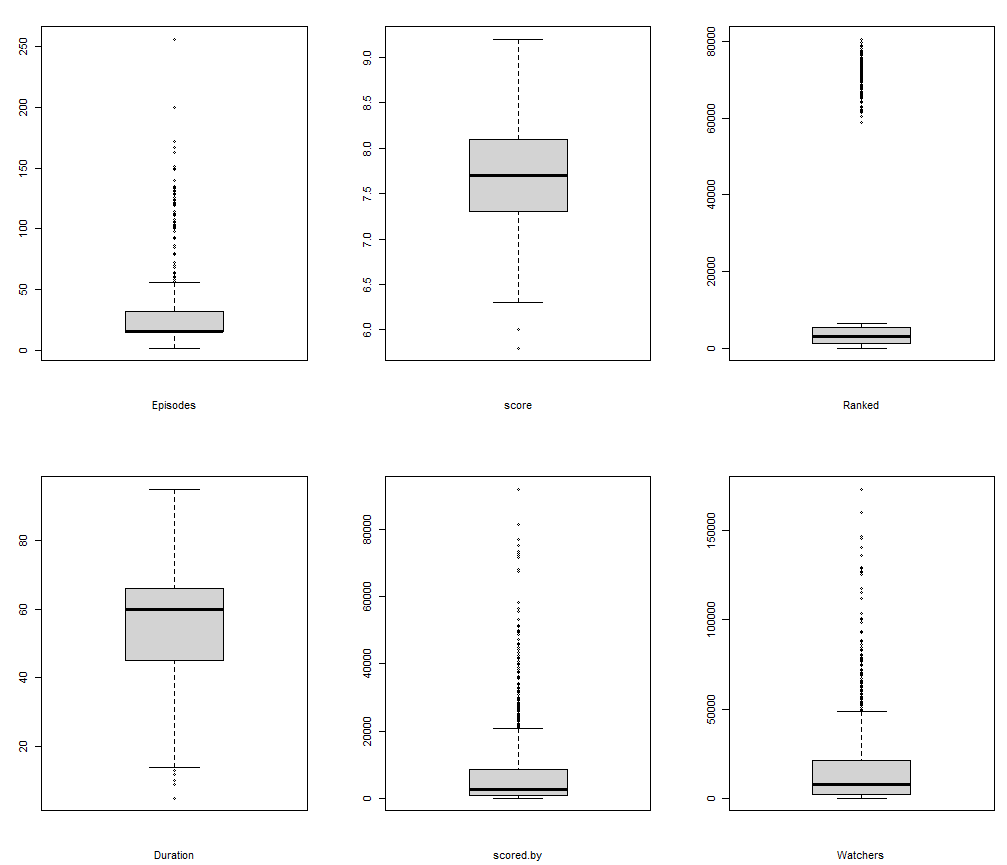
1. **정확성이 낮다고 판단된 레코드의 경우 Appendix-(5)**

Scored(점수)의 경우 50명 이하의 사람이 점수를 매긴 드라마는 scored(점수)에 대한 정확성이 떨어진다고 판단하였다. 그렇기에 이와 연관된 Rank 역시 정확도를 믿을 수 없다고 판단이 되어 50명 이하의 드라마의 레코드는 삭제하기로 결정하였다.

1. **데이터 셋 타입과 형태 변경 Appendix-(6), (8)**

데이터 셋의 타입이 character로 읽어 와지는 문제가 발생해서 각 타입에 맞게 변경을 해주었다. (date나 numeric 등등)

1. **Boxplot 분석 및 이상치 Appendix-(11)**



numeric 변수들의 이상치를 박스플랫으로 측정한 것이다. 이상치가 발견이 되었으나 실제 측정값이기도 하고 해당 레코드를 확인해본 결과 모델링을 하는데 영향을 크게 미칠 수 있는 값들은 아니라고 판단되어 제거하지 않고 사용하기로 결정했다. 이것은 인기가 없는 드라마와 인기가 있는 드라마인지에 따라 발생하는 차이로 판단이 되어 그대로 사용하기로 결정했다.

**2-2 데이터 분석**

**2-2-1 Data Set 구성**

1. **분석할 데이터**

K-drama의 성공요인 분석 및 성공 여부 예측을 위하여 우리가 분석할 Data Set은 ‘kaggle’에서 가져온 데이터는 “kdramalist.csv”라는 공공 데이터이며, 이 안에는 약 1278개의 한국 드라마 목록이 있다.

1. **Data Set 구성**

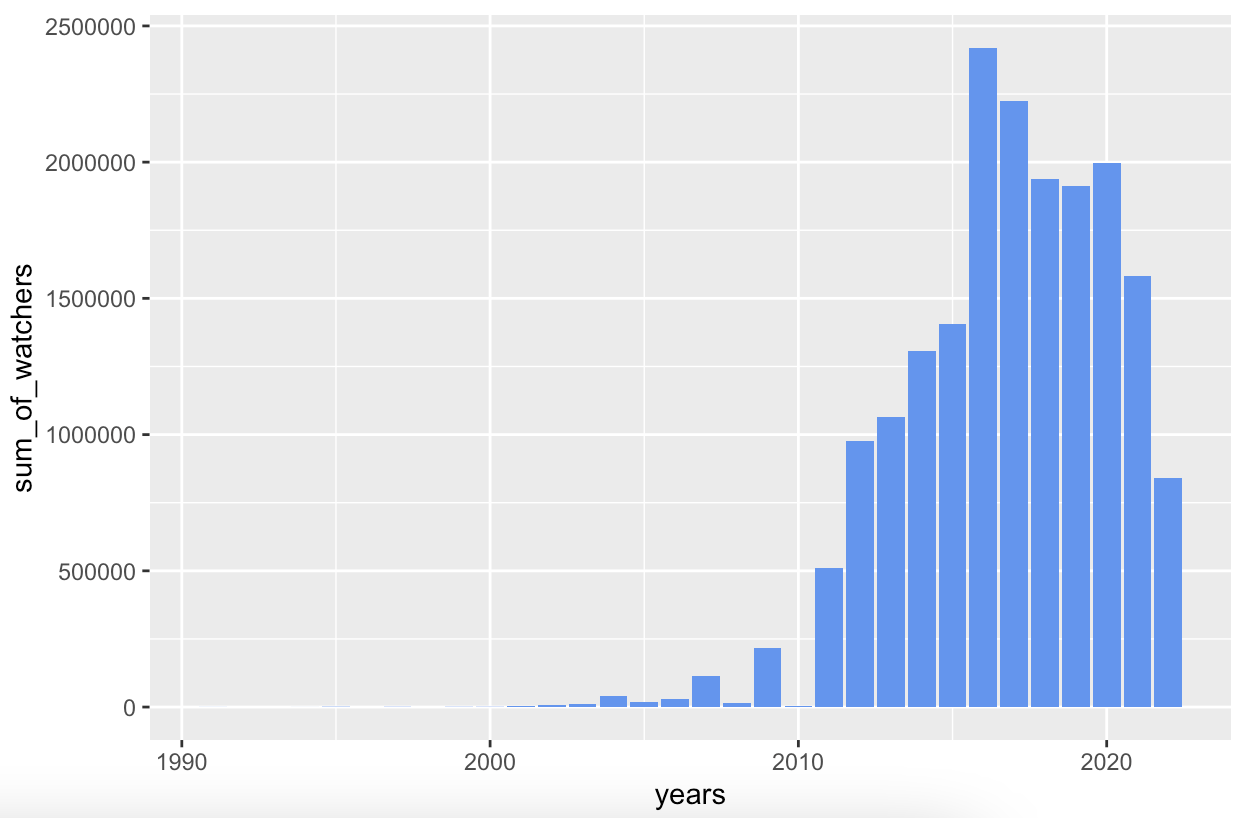
“k-dramalist.csv”안에는 20개의 Column이 존재한다.

<k-dramalist의 Column 목록>

|  |  |  |  |
| --- | --- | --- | --- |
| 번호 | 이름 | 타입 | 설명 |
| 1 | drama\_name | character | 드라마 이름 |
| 2 | Genres | character | 드라마의 장르 |
| 3 | Tags | character | 드라마별 태그(주요 내용을 간략하게 설명할 수 있는 단어들로 구성되어 있음) |
| 4 | Episodes | integer | 드라마의 에피소드 수 |
| 5 | start airing | double | 최초 출시일 |
| 6 | end airing | double | 마지막 에피소프가 처음 공개된 날 |
| 7 | Aired On | character | 드라마 방영 요일 |
| 8 | Original Network | character | TV 방송국 |
| 9 | Duration | double | 상영 길이(러닝 타임) |
| 10 | score | double | 시청자가 평가한 드라마 점수(10점 기준) |
| 11 | Scored by | double | 점수를 매긴 시청사 수 |
| 12 | Ranked | double | Kdramalist를 제외한 다른 모든 드라마와 비교한 kdramalist 순위 |
| 14 | Content Rating | character | 콘텐츠 연령 제한 |
| 15 | Watchers | double | k-dramalist의 시청자 수 |
| 16 | actors | character | 등장 배우 |
| 17 | platforms | character | 방영 플랫폼들 |

**2-2-2 Data Set 분석**

1. **K-drama의 추세 분석 - Appendix-(12)**



K-dramalist의 데이터를 토대로 연도별 watchers 분석해본 결과, 2010년 전후로 watchers수가 급증한 것을 볼 수 있다. 이는 k-drama 의 인기가 급상승했음을 입증해주는 것을 알 수 있다.

1. **K-dramalist.csv의 컬럼 분석**
2. **Appendix-(13),(14)**

텍스트이(가) 표시된 사진

자동 생성된 설명

numeric 변수들의 분석 결과이다.

Numaric Variable(Episodes, Duration, score, scored.by, ranked, watchers)에 대한 summarise를 진행하였다. Episodes의 최솟값은 2, 최댓값은 256, 평균값은 31이다. Duration의 최솟값은 5, 최댓값은 95, 평균값은 56분이다. Score의 최솟값은 5.8, 최댓값은 9.2, 평균값은 7.739이다. Scored.by의 최솟값은 51, 최댓값은 92027, 평균값은 7614이다. Ranked의 최솟값은 4, 최댓값은 80632이다. Watchers의 최솟값은 190, 최댓값은 173027, 평균은 7564이다.

– 이 구성에서 우리가 사용할 Column과 사용하지 않을 Column을 전처리 과정을 통해 나누고 덜어내어 사용하였다.

에피소드 수의 경우 중간 값과 평균값의 차이가 존재하는데 일일드라마의 경우 매일 방영을 하기 때문에 회차가 많고 일주일에 2일만 방영하는 드라마의 경우 회차가 20회 이하로  하기 때문에 차이가 많이 나는 것으로 보인다.

duration과 scored는 평균과 중앙값이 비슷한 경향을 보이고 rank의 경우 dramalist.csv에 있는 드라마 뿐만이 아니라 mydramalist.com에 존재하는 모든 드라마에 대한 순위이기 때문에 순위가 낮은 드라마와 높은 드라마의 순위 차이가 크게 나서 평균과 중앙값의 차이가 보이는 것으로 보인다.

scored.by와 watchers의 경우 인기있는 드라마와 인기가 없는 드라마의 시청자 수 차이가 많이 나서 중앙값과 평균 사이에 데이터 차이가 많이 나는 것으로 보인다.

1. **PCA분석 - Appendix-(10)**

테이블이(가) 표시된 사진

자동 생성된 설명

추가로 PCA 분석도 진행을 했고 그것에 대한 결과이다.

PCA로 어떤 변수가 시청률이 높을 때 영향을 끼치는지 알아보고, 나이브베이즈로 어떤 요소가 들어갔을 때, 각 요소들에 대한 시청률이 높을 지 세부요소들을 계산할 것이다.

1. **Heatmap 분석 - Appendix-(15)**

테이블이(가) 표시된 사진

자동 생성된 설명

다음은 히트 맵 분석으로 각 컬럼 별 상관관계를 분석한 것이다. scored.by와 watchers가 진한 색깔로 나타났으며 이는 높은 상관관계를 보이는 것을 알 수 있다.

1. **모든 Column의 속성** **Appendix-(13),(14)**

텍스트, 영수증, 스크린샷이(가) 표시된 사진

자동 생성된 설명

1. **예측 모델링**

**3-1 예측 방법론**

Classification은 Supervised learning 의 일종으로 예측하고자 하는 대상인 독립변수가 존재하는 상태로 하는 예측기법을 말한다. 우리는 Classification을 사용하여 Score(평점)이 8이 초과인지 아닌지에 따라 아직 평점이 나오지 않은 드라마의 평점이 높을지 낮을지를 예측하는 모델링을 하였다.

**3-0-1 사전 작업- Appendix-(16)**

Classification을 위해 Score(평점)이 8을 초과 했으면 1 아니면 0이라는 레코드를 갖는 컬럼을 추가했다.

**3-1-1 Text Mining- Appendix-(18), (19), (20)**

Tag 변수를 텍스트 마이닝으로 분석하기 위해 tag에 대한 전처리를 추가로 해주었고 텍스트 마이닝을 진행하였다.

그후 tag가 무엇이 있는지 뽑아내어 확인 절차를 먼저 거쳤다. 총 1348개로 가장 많이 중복되는 tag를 15개 가려내었다.

**텍스트이(가) 표시된 사진

자동 생성된 설명**

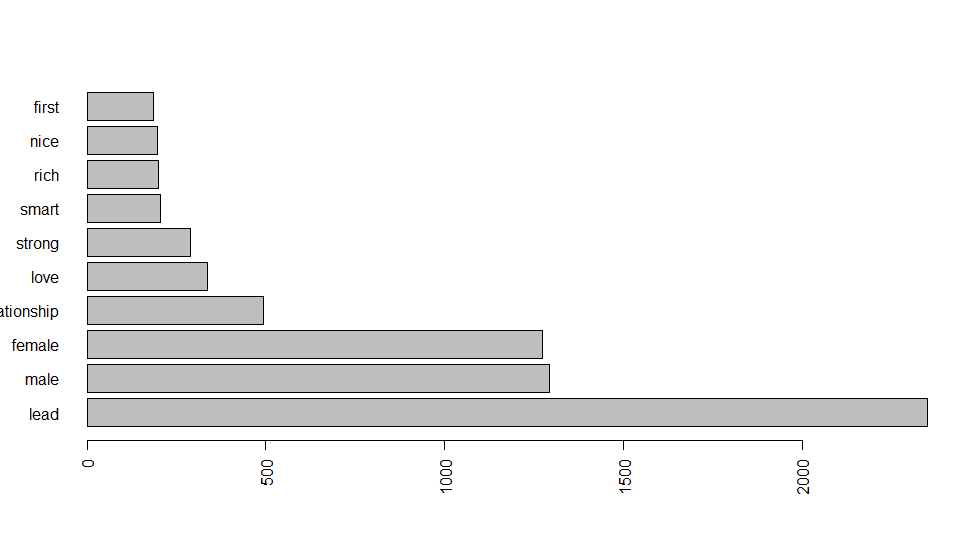
그 후 100번 이상 언급된 단어들만 다시 출력하였다.

**텍스트이(가) 표시된 사진

자동 생성된 설명**

* + 1. **Naïve Bayes와 Confusion Matrix -Appendix- (19), (20), (21)**

드라마의 주요 내용을 간략하게 설명해주는 값인 Tags를 활용하기 위하여, text mining을 통해 얻어낸 가장 많이 나온 상위 10개의 tag를 통해 highRated라는 새로운 변수(score가 8점 초과면 1, 아니면 0)를 예측하는 모델링을 진행하였다. training/validation/testing은 6 대 2대 2 비율로 처리하였다.



위 그래프는 빈도수가 높은 10개의 태그를 그래프로 그려본 것이며, 아래 그림은 wordCloud로 좀 더 쉽게 파악하기 위해 시각화한 것이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

**1) Naïve Bayes vaildation data accuracy**

텍스트이(가) 표시된 사진

자동 생성된 설명

정확도가 0.746인 것을 볼 수 있다.

**2) Naïve Bayes test data accuracy**

텍스트이(가) 표시된 사진

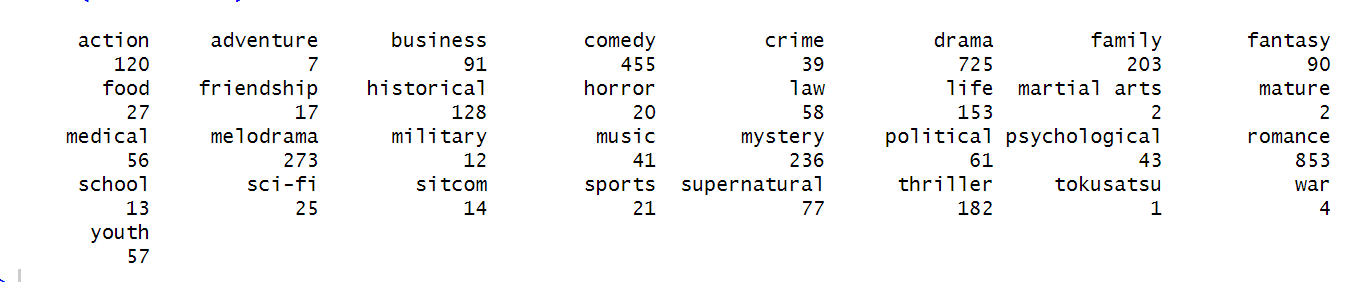
자동 생성된 설명

정확도가 0.7566인 것을 볼 수 있다.

**3-2 Logistic Regression 과 KNN**

**3-2-1 전처리- Appendix-(22)**

numeric에서는 방송을 하기 전에 알 수 있는 정보인 에피소드 수(episode)와 러닝타임(duration)그리고 category data 중에서는 text mining에 사용되어 다른 모델링에 사용되는 Tag를 제외하고 가장 많은 영향을 끼친다고 생각되는 장르(Genres)와  방송 요일(cycle) 2가지까지 추가해서 모델링을 하기로 결정하였다.



장르의 경우 총 33개가 있는데 drama와 romance의 경우 많은 드라마의 장르에 들어가 있어서 의미가 없다고 생각해 제거했고 해당하는 장르가 10개 이하인 경우(war, tokusatsu, mature, adventure, martial arts)도 제외하고 나머지 장르는 비슷한 장르는 하나로 합쳤다.(양이 많아서 밑에 사진 첨부, 각 줄마다 제일 왼쪽에 있는 장르로 통합하고 총 11개의 장르로 줄였다.)

텍스트이(가) 표시된 사진

자동 생성된 설명

요일은 밑에 적혀있는 규칙에 따라 5개의 category로 만든 후 더미 데이터로 변경하여 사용하였다

|  |  |
| --- | --- |
| 월 화 수 목 금 | 일일드라마 daily |
| 금 토 | 금토 드라마 friSat |
| 금 | 금토 드라마 friSat |
| 토 일 | 주말 드라마 weekend |
| 토 | 주말드라마 |
| 일 | 주말드라마 |
| na | 비정기 드라마 irregular |
| 나머지 주간드라마 | 주간 드라마 week |

training/validation/testing은 6 대 2대 2 비율로 처리하였다. cut off는 가장 많이 쓰이는 0.5(50%)로 정했다.

**3-2-2 Logistic validate - Appendix-(23)**

앞에서 처리한 변수들을 가지고 logistic regression으로 모델링을 한 뒤 validation/test 데이터를 이용해서 정확도 검사를 하였다.

**1) Logistic validate accuracy**

텍스트이(가) 표시된 사진

자동 생성된 설명

결과를 보면 정확도가 0.7051인 것을 볼 수 있다.

**2) Logistic test accuracy**

텍스트이(가) 표시된 사진

자동 생성된 설명

결과를 보면 정확도가 0.7202인 것을 볼 수 있다.

결과적으로 약 70초반의 정확도를 보인다는 것을 알 수 있다.

**3-2-2 knn- Appendix-(24)**

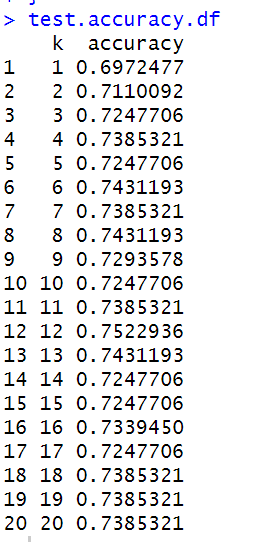
텍스트이(가) 표시된 사진

자동 생성된 설명logistic regression에서 사용한 변수를 가지고 다른 알고리즘인 knn을 가지고 모델링을 한 후 정확 도 검사를 진행하였다.

**1) knn validate accuracy**

K는 1~ 20개까지의 검증 정확도는 Logistic validate accuracy와 비교했을 때보다 낮은 수치를 보여준다

그중 k = 4일 때 약 69의 가장 높은 정확도를 보인다.

**2) knn test accuracy**

Test data를 이용해서 분석을 했을 때, validate를 사용을 했을 때보다 더 높은 정확도를 보여주며 k=12 일 때, 가장 높은 정확도를 보여준다(약 0.75)

* + 1. **방법론 사용 이유(모델링 목적과 기여도를 설명)**

가장 유의미한 모델링은 시청률로 분석을 하는 것이 맞지만 시청률의 경우 방송사나 플랫폼에 따라서 높다의 기준이 다르다(예를 들어 케이블 방송의 시청률은 지상파 보다 낮은 데에도 높다고 하는 경우가 많다) 그렇기 때문에 좀 더 기준이 있는 평점으로 결정을 했고 어느 정도 이상의 평점은 되어야 작품성이 인정을 받을 수 있다고 생각이 되어서 특정 평점을 넘냐를 기준으로 classification을 하게 되었다.

Text Mining을 통해 다양한 태그 속 유의미한 단어를 추출하고 어떤 단어들이 드라마에 많이 사용되었는지 트랜드를 파악한다. 또한 Text Mining을 통해 추출된 단어들이 사용된 드라마 정보를 가지고 실제로 태그가 시청률에 크게 기여하는지 파악하고 분류하기 위해 Naïve Bayes를 사용하였고, 시청률이 높은지 낮은지 예측하는 모델을 만들었다.

드라마에서는 태그가 중요한 요소이긴 하지만 그 외의 장르나 러닝타임 등 다른 요소들도 영향을 많이 끼친다는 생각이 들어 text mining을 통한 모델링 외에 다른 요소들을 통한 분석 및 모델링도 진행을 하였다.

* 1. **실제 활용 및 시사점**

Text Mining을 통해 추출된 단어들로 해당 단어들이 사용된 드라마들이 시청률이 높게 나오는지 낮게 나오는지 파악할 수 있었다. Valid data로 사용했을 때 70%, valid data로 사용했을 때 72% 로 정확도가 높게 나왔다. Tag가 시청률을 결정하는데 어느정도 영향을 미친다는 점은 파악하였으나 정확도가 98%, 90% 처럼 높지 않아 100% 좌우한다고 말할 수 없다.

하지만 OTT산업을 발전시키기 위해서 컨텐츠를 제작하고자 한다면 분석한 결과를 통해 파악된 자주 사용되고, 시청률이 높다고 예측되는 태그를 무시할 수는 없다. 해당 태그를 사용한다면 OTT 산업에서 70%는 보장된 성공을 거둘 수 있을 것이다.

또한 텍스트 마이닝을 통해 지난 5년간, 10년간 자주 사용되었던 드라마 태그들을 파악할 수 있는데 이를 통해 태그들이 외부의 어떤 요소 때문에 흥행에 성공했는지 연관 지어 파악할 수 있을 것이다. 더 나아가 앞으로의 K-drama 트랜드를 쉽게 파악하고 highrated가 ‘높다’라고 예측될 태그들 또한 스스로 예측할 수 있을 것이다.

좀 더 과학적인 분석을 통해 도출된 k-drama의 성공 요인을 가지고 드라마를 제작한다면 적어도 드라마가 성공을 할 지 못할 지 수학적으로 예측할 수 있을 것이다.

로지스틱 회귀분석을 통해 어떠한 예측변수가 종속변수인 higlyRated에 유의미한 영향을 미치는지 알아볼 수 있었다. Summary함수로 모델을 살펴보았을 때, 상영 요일은 거의 영향을 미치지않았고, 장르나 태그가 평점와 연관이 있음을 알고 어느정도 영향을 미친 것을 알 수 있었다.

로디스틱 회귀분석을 이용해 예측 모델링한 결과, valid data를 사용했을 때에는 정확도가 0.7051도 도출되었고, test data를 사용했을 때에는 0.7202의 정확도가 도출됐다. 이를 knn이라는 다른 기법의 정확도와 비교하여 봤을 때, valid data는 knn이 전반적으로 정확도가 높은 것이 확인되었고 test accuracy는 로지스틱 회귀분석 모델링이 더 정확도가 좋았음을 알 수 있었다. 이를 통해 각각에 맞게 어느 기법이 더 성능을 보기에 좋은기 판단할 수 있었다. 또한 이러한 모델링을 통해 어떤 장르와 테그가 평점에 영향을 미치며 그로인해 인기가 좋을 만한 드라마를 제작하는데 기여할 수 있음을 분석해내었다.

가장 정확도가 높은 것은 text mining을 이용한 태그를 통한 나이브 베이즈였으나 다른 요소들로 이루어진 logistic regression도 비슷한 수치를 보였다. 드라마를 제작을 할 때, 태그 분석과 그 외의 요소로 이루어진 것들 2가지를 분석하면 드라마의 평점을 예측을 할 수 있고 이걸 이용해서 더 높은 퀄리티의 드라마를 제작을 할 수 있는데 기여를 할 수 있을 것 같다.

1. **결론**
   1. **느낀점**

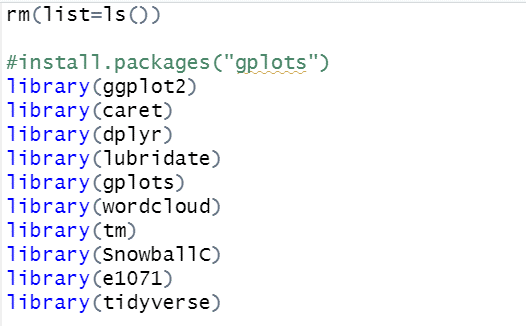
이론상 공부했던 방법론을 패키지를 통해 실제 구현해보고, 결과값을 보며 R에 대한 이해와 이론을 동시에 익힐 수 있었다. Naïve byes로 분류된 데이터를 confusion Matrix로 검증하는 과정을 통해 내가 예측한 데이터가 실제로 맞게 분류되었는지 정확도를 한 번에 파악할 수 있다는 점이 가장 신기하면서도 편리하다는 생각을 했다. TextMining을 통해 tags 데이터에서 빈도수가 높은 태그들을 뽑고, 시각화하는 과정 또한 흥미로웠다. 한 줄에 글처럼 다양한 태그들이 적혀 있는데 그 태그들 중 유의미한 태그를 뽑을 수 있으며 빈도 또한 파악할 수 있었다. 시각화 방법 중 WordCloud를 통해 한눈에 가장 많이 사용된 태그가 무엇인지 파악할 수 있었던 과정이 가장 기억이 남는다. 마지막으로 tag의 값이 실제로 시청률에 미치는 영향도가 엄청 크지 않다는 점이 의외였다.

또한, 로지스틱 회귀분석을 하기 위해 예측 변수인 장르와 드라마 요일을 dummy variables로 만들어보는 과정에서 많은 고민과 어려움을 겪었지만, 변환 후에 이를 회귀분석에 성공적으로 활용할 수 있어 뜻깊었다. 로지스틱 회귀분석에서 활용한 변수들을 이용해 추가적으로 knn 분석을 해보면서, 동일한 예측변수임에도 불구하고 모델링 방법에 따라 서로 다른 accuracy를 나타낸다는 점을 확인할 수 있어 흥미로웠다. 본 연구를 통해 실제 데이터를 활용해 막연하게 예측했던 여러 변수들과 평점의 실제 상관관계를 파악해볼 수 있는 좋은 경험을 한 것 같다.

실제로 분석할 데이터를 탐색하는 것부터 어떤 데이터가 분석할 때 의미가 있을까 혹은 어떤 데이터가 사람들의 흥미를 유발할 수 있는까를 생각해볼 수 있어 좋았다. 이어서 이렇게 우리가 선정한 데이터를 가지고 데이터의 전처리와 데이터분석과 같이 그 Data Set이 어떤것이있고 이 Data Set으로 무슨 유의미한 분석을 만들지 생각해 볼수 있어 좋았다. 마지막으로 Classification의 KNN, 테이스마이닝과같은 기법들을 통해 예측 모델링을 하여 실제로 어떠한 예측이 나오며 우리가 만든 예측 모델링이 어떻게 사용될수있는지를 생각해보며 생활속에서 유의미한 모델링을 만들어 낼 수 있어 좋은 경험을 했다. 이렇게 한 단계씩 차례대로 이해해 나가면서 탐색하고자하는 Data Set이 어떻게 유의미한 분석이 가능할지와 실제로 이러한 분석을 어떻게 적용할지에 대해 직접 생각하고 분석해가는 과정을 겪고 또 다른 데이터도 이런식으로 하나씩 분석해나가면 더 유의미한 분석 모델링이 나올것이라는 기대를 하게 되었다.

**Appendix(부록)**

**1)** **라이브러리 정리** **Appendix-(1)**

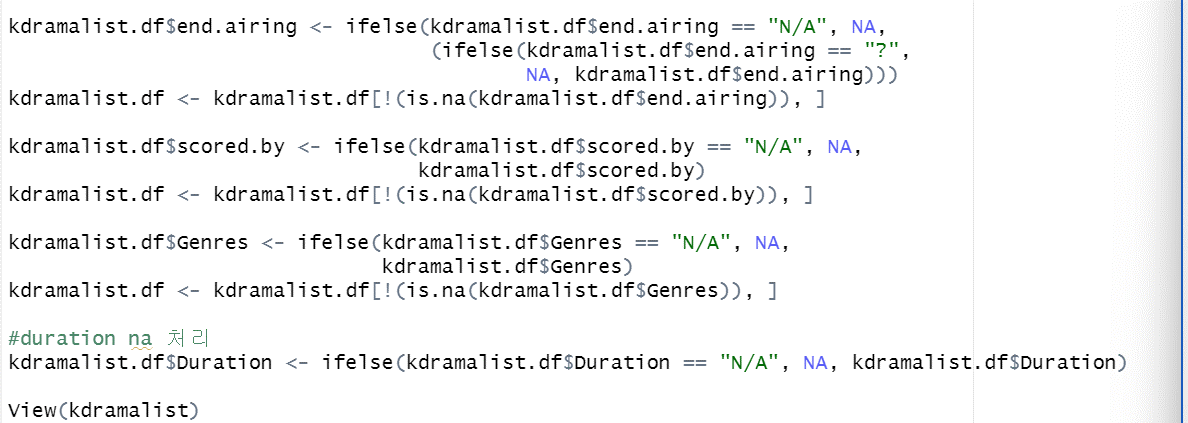
****필요한 라이브러리 올리기.

**2) csv 가져오기 Appendix-(2)**

사용할 kdramalist.df를 만들어준다.

**3) 결측치 처리- csv에 적힌 결측치 표시 “N/A”와 “?”를 r에 맞게 NA로 변경**

**-Appendix-(3)**



**4) 불필요한 컬럼 제거 -Appendix-(4)**

텍스트이(가) 표시된 사진

자동 생성된 설명

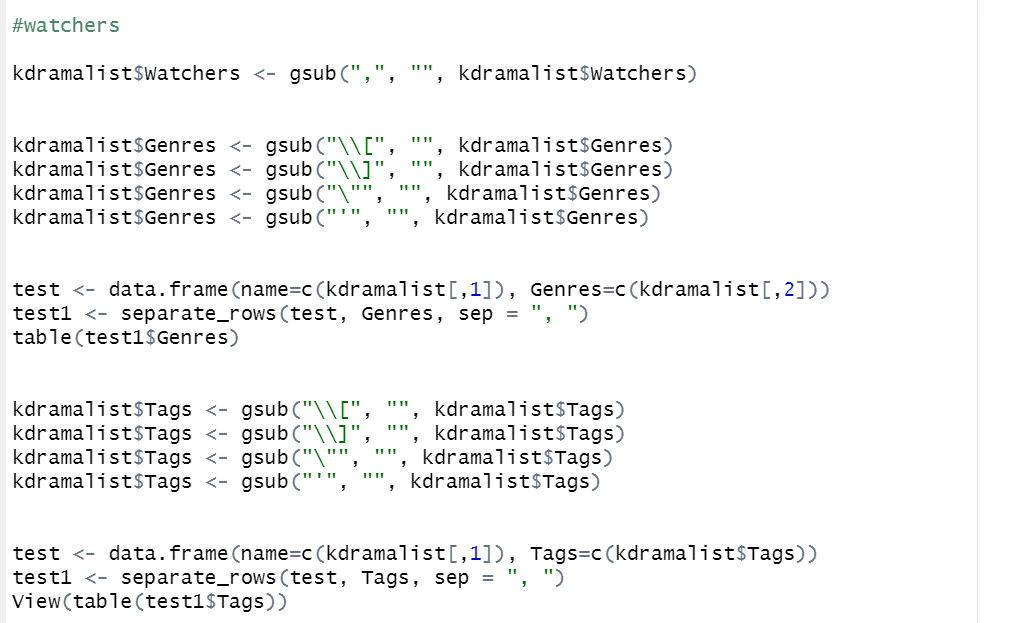
**5) 신뢰도가 떨어진다고 판단되는 레코드 삭제-Appendix-(5)**

텍스트이(가) 표시된 사진

자동 생성된 설명

분석하기 위해 결측치를 제거하거나 사용하지 않은 Column을 제거하는 등 전처리를 한다.

**6) 데이터 셋의 형태를 우리가 사용을 할 수 있는 형태로 수정- Appendix-(6)**



사용할 Column인 태그와 장르를 테이블에 넣어 몇 번 사용되었는지 사용 빈도수를 각 단어별로 체크한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

Tag는 개수가 너무 많아 View의 형태로 필요할 때마다 보도록 했다.

. 테이블이(가) 표시된 사진

자동 생성된 설명

일부를 보여주는 그림이다. 총 1849개의 태그가 몇 번 사용되었는지를 알 수 있다.

**7)N/A 제거가 제대로 되었는지 확인- Appendix-(7)**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**8) csv에서 받아온 형식을 사용목적에 맞게 변경(예를 들어 날짜나 수치형 변수들이 char 타입으로 되어 있어서 각 타입에 맞게 수정)- Appendix-(8)**

**텍스트이(가) 표시된 사진

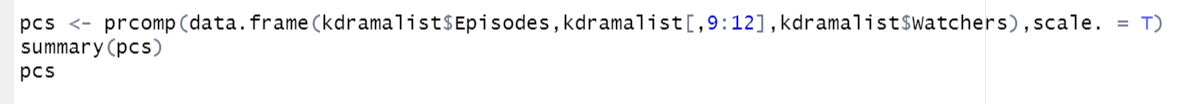
자동 생성된 설명**

+추신: r에서 date의 포멧 중 중인 월(Jan, Feb 등)로 표시하는 %b는 현재 MAC에서만 정상 작동되는 버그가 있습니다.

**9) 결측치 처리(러닝타입(duration)은 중앙값인 60을 넣어주었다)- Appendix-(9)**

****

**10)PCS 분석)- Appendix-(10)**

****

**11)boxplot 분석- Appendix-(11)**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**12) 연도별 드라마 시청자 수 막대그래프- Appendix-(12)**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**13) 각각의 컬럼 타입 확인-Appendix-(13)**

**텍스트이(가) 표시된 사진

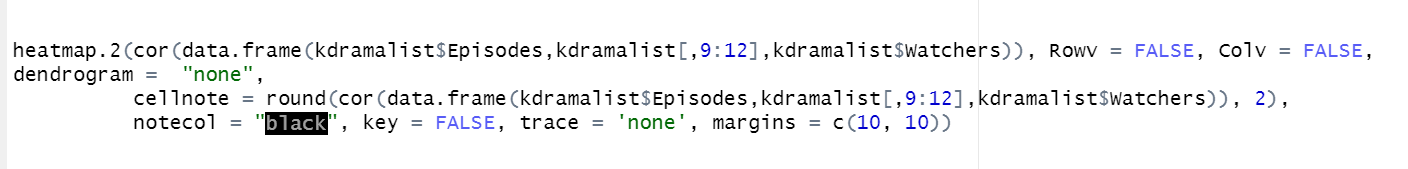
자동 생성된 설명**

**14) numeric 타입 분석-Appendix-(14)**

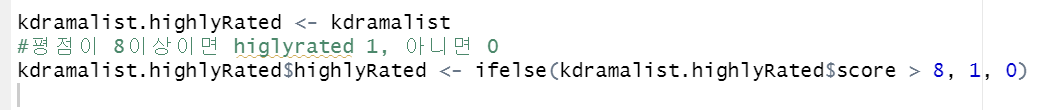
**텍스트이(가) 표시된 사진

자동 생성된 설명**

**15) numeric 타입 히트맵 분석-Appendix-(15)**

****

**16) classification을 위해 score에 따라 2가지로 분류-Appendix-(16)**

****

**17) 장르 변수의 개수를 조건에 따라 줄이거나 통합-Appendix-(17)**

**텍스트, 신문이(가) 표시된 사진

자동 생성된 설명**

**18) 태그 변수를 텍스트 마이닝을 하기 위해 전처리-Appendix-(18)**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**19) 문자열 데이터 정리 및 분석-Appendix-(19)**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**20) 텍스트 마이닝-Appendix-(20)**

텍스트이(가) 표시된 사진

자동 생성된 설명

**21)** **Naive bayes-Appendix-(21)**

텍스트이(가) 표시된 사진

자동 생성된 설명

**22) 장르(genres)와 요일(aired.on) 컬럼을 더미 데이터로 변환-Appendix-(22)**

텍스트이(가) 표시된 사진

자동 생성된 설명

**23) logistic regression과 정확도 검사-Appendix-(23)**

텍스트이(가) 표시된 사진

자동 생성된 설명

**24) KNN과 정확도 검사-Appendix-(24)**

텍스트이(가) 표시된 사진

자동 생성된 설명