DATAMINING

4조
32152827 정치외교학과 유승훈
32154499 경영학과 조현진
32160175 응용통계학과 고준성

index

01. 주제 선정 및 분석 목적

주제 : 영화개봉 전 “관객 수” 예측

주제 선정 배경

- 영화산업의 특성상 초기 투자비용이 크고 손익분기점을 넘길지 불확실함.

- 영화의 정보들로 개봉 전 관객수를 예측할 수 있다면
  투자사와 배급사, 제작사가 유용하게 활용할 수 있을 것.

분석 목적

- 영화 데이터로 데이터마이닝 기법을 활용하여 관객수를 예측

- 제작된 모델들을 통해 어떤 변수가 영화관객수에 주요한 영향을 끼치는지 확인

02. DATA 소개

데이터 구성

원래 가지고 있던 2015.01.01 ~ 2017.07.05 상영 영화 데이터
2017.07.06 ~ 2018.04.20 상영 영화 데이터를 새로 추가

총 obs : 10,746
이 중 60%인 6431개를 train, 20%인 2181, 2134개를 각각 test, valid set으로 활용

활용 패키지

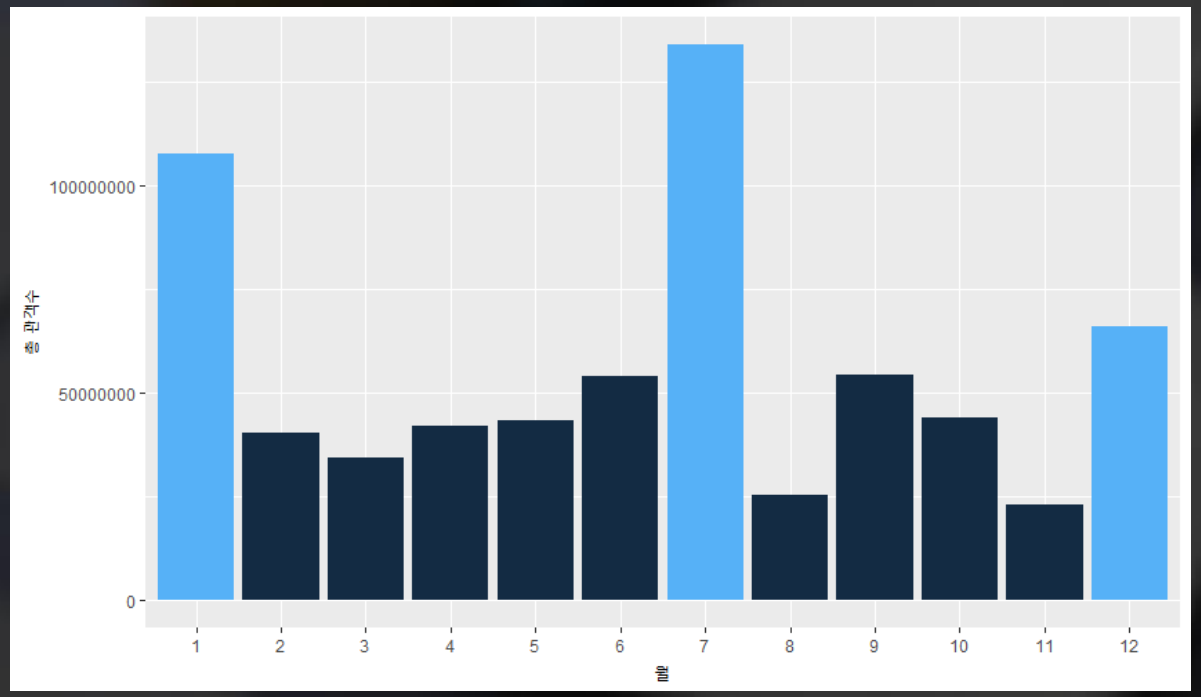
용도	패키지 명
데이터 불러오기	data.table, readxl
데이터 전처리	dplyr, tidyr, stringr, lubridate
시각화	ggplot2, ellipse, lattice
모델링	FNN, tree, neuralnet, randomForest
기타	progress, pbapply, rlist

02. DATA 소개

대상 영화 : 2015.01.01 ~ 2018.04.30 개봉

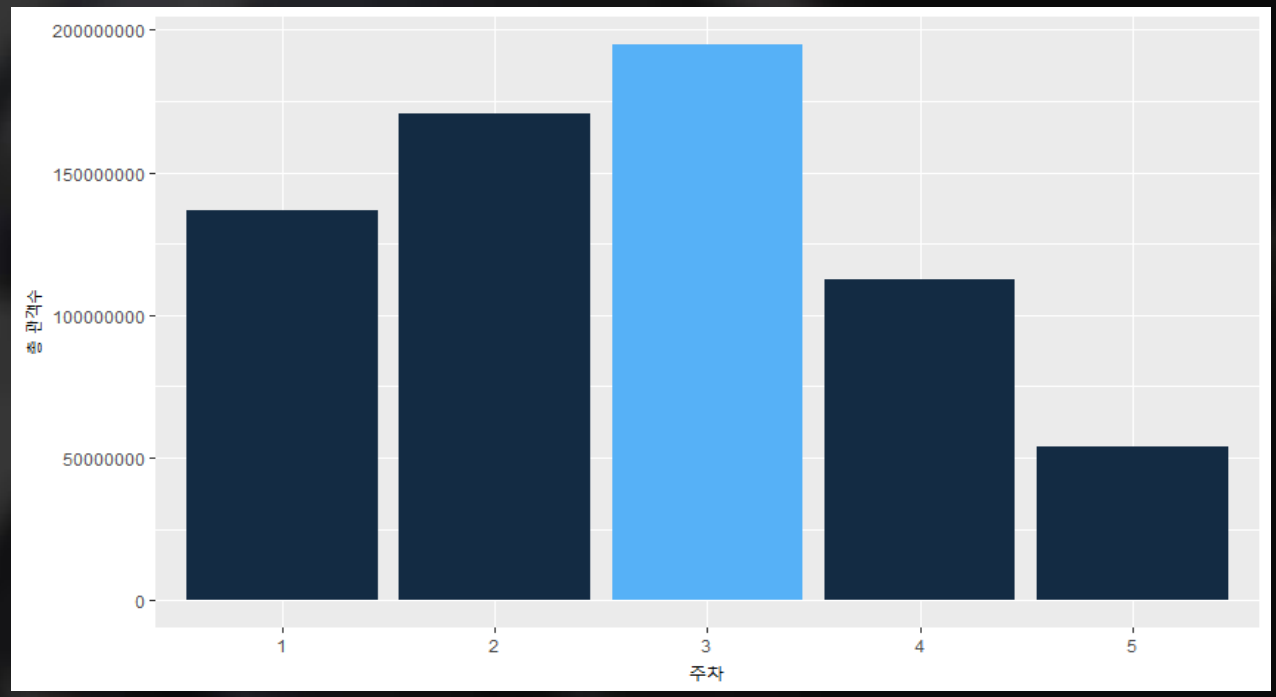
변수명	설명	비고
people	총 영화 관객수	종속변수
play_t	상영횟수	영화가 총 상영된 횟수
play_wk	상영 주	영화가 몇주간 상영되었는가
re_open	재개봉 여부	첫개봉 = 0, 재개봉 = 1
top_country	대표국적	이외 모든 국가 = 0, 한국, 미국, 일본 = 1
record_month	개봉 계절	12,1,2 = 1 / 3,4,5 = 2 / 6,7,8 = 3 / 9,10,11 = 4
month_week	개봉주차	
top_sup	상위 배급사 해당 여부	이외 모든 배급사 = 0, 상위 5개 배급사 = 1

03. EDA



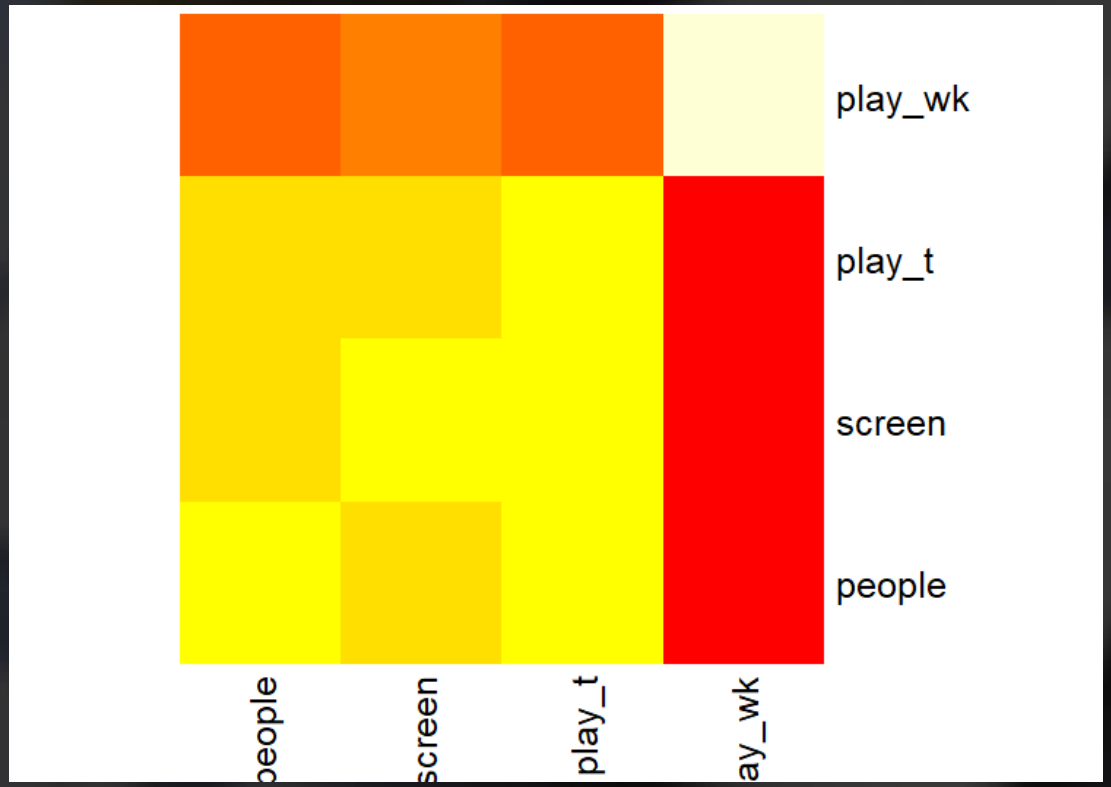
사람들은 1,7,12월에 개봉한 영화를 가장 많이 봤다

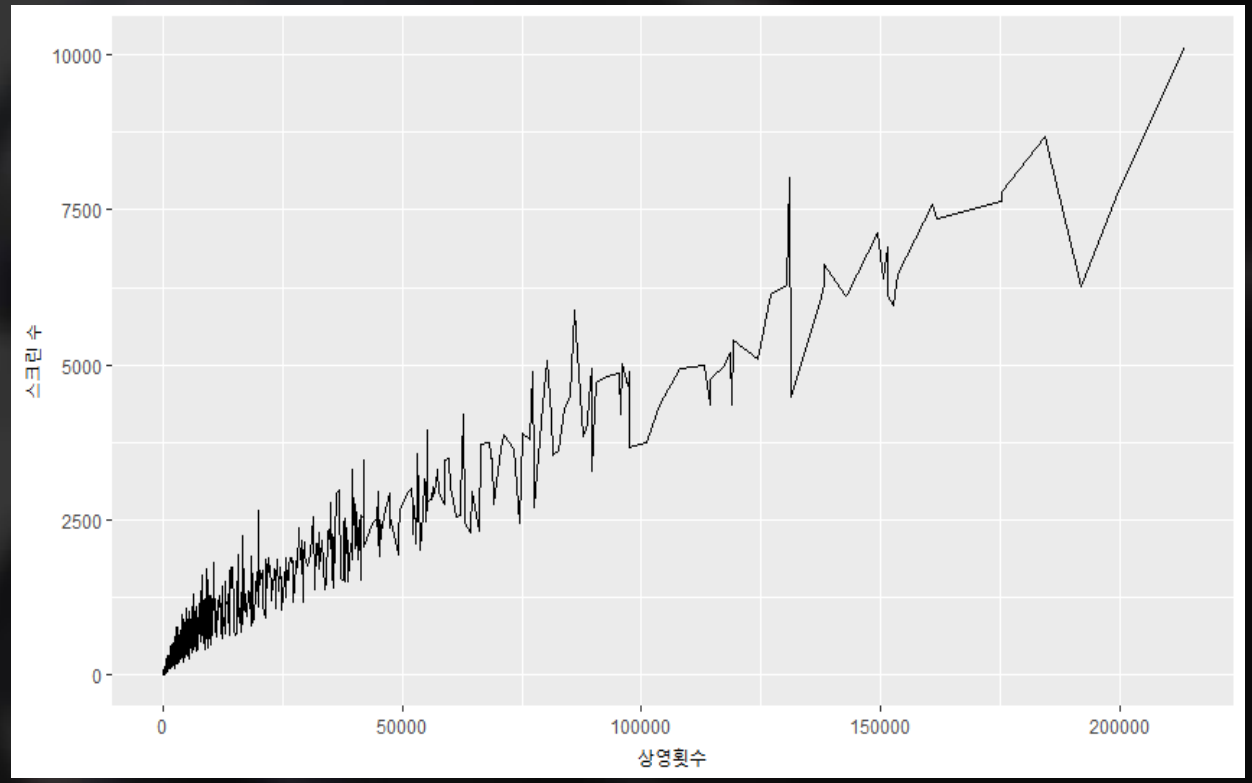
1월 : 검사 외전, 국제시장, 공조
7월 : 베테랑, 암살, 택시운전사, 군함도
12월 : 신과 함께, 히말라야, 1987



사람들은 3주차에 개봉한 영화를 가장 많이 봤다

03. EDA

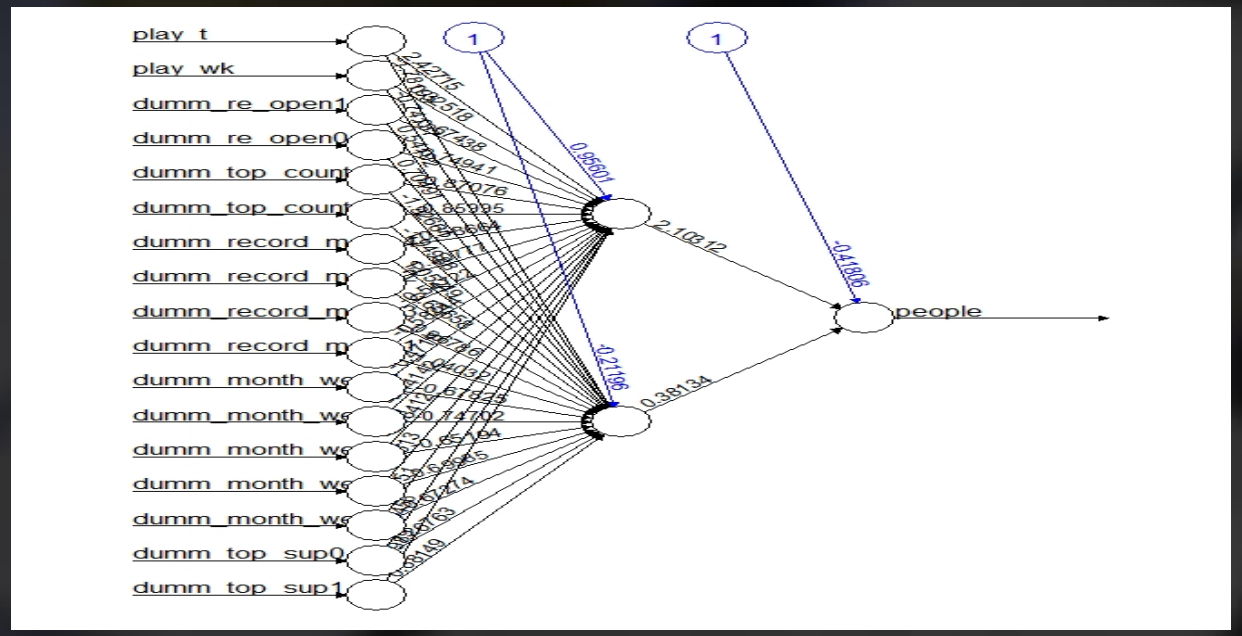




상영횟수 – 스크린 수
상관계수(Corr) : 0.968

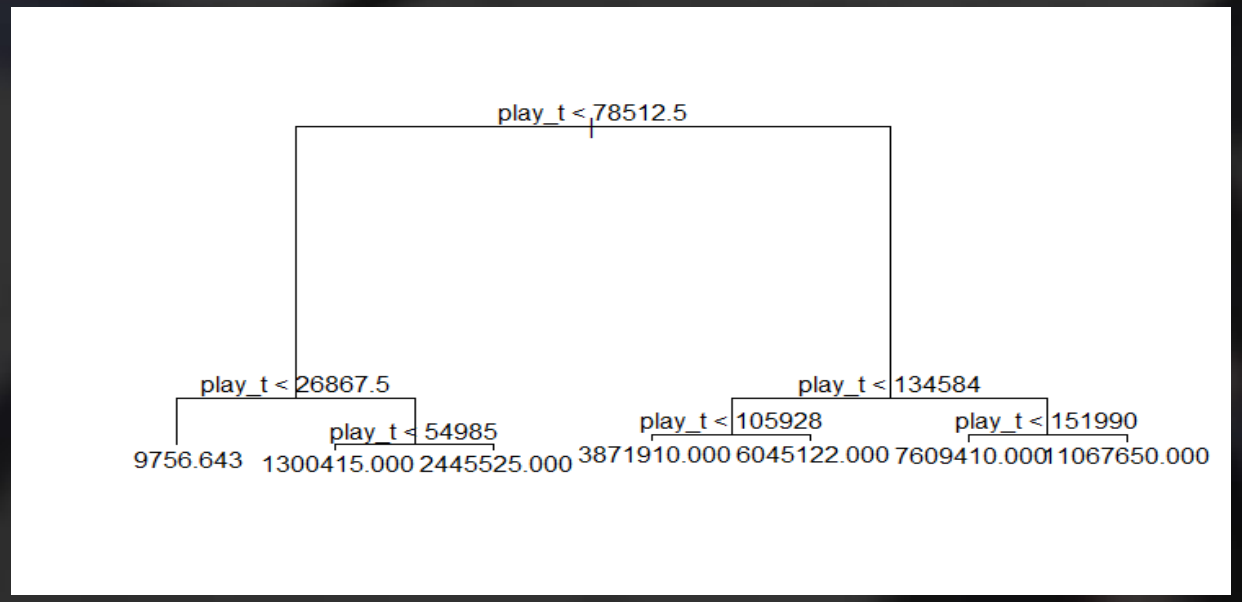
변수 간의 상관관계가 매우 높음
차원의 저주를 피하기 위해 상영횟수만 분석에 사용

04. 결과해석 및 활용



인공신경망

valid data에서 hidden node가 2개일때
가장 좋은 예측 성능을 보임



의사결정나무

모든 node가 play_time으로 분리됨

04. 결과해석 및 활용

최종 성능

모델	test_rmse	test_rsquare	test_adj_rsquare
Neuralnet	95719.587	0.9571	0.9568
Knn.regression	116660.676	0.9363	0.9361
Tree	146460.754	0.8995	0.8992
RandomForest	165144.393	0.8723	0.8718

knn regression, tree, randomforest는 모두 독립변수가 8개였으나,
Neuralnet은 가변수처리를 해 주어 17개의 독립변수를 활용하였음
따라서 독립변수 개수의 영향을 제거한 adj_rsquare도 계산하여 비교함

04. 결과해석 및 활용

분석결과

- Knn 회귀, 인공신경망, Tree, RandomForest 중 인공신경망이 가장 낮은 RMSE와 가장 높은 예측력을 가짐

- 다른 모델들에 비해 인공신경망을 사용했을 경우 평균 4만 5천명 가량 더 정확한 예측결과를 보이며 모델의 완성도 측면에서도 평균 5% 더 나은 성능을 보일 것으로 기대됨

- 인공신경망모델이 사용한 모델 중 가장 최선의 모델인 것

- 변수 중 가장 큰 영향을 끼치는 변수는 상영시간으로 상영 비용 대비 이익을 계산함으로써 이익을 최대화할 수 있으리라 생각됨

04. 결과해석 및 활용

활용방안

- 합리적인 의사결정

예측관객 수를 이용해 최적의 스크린 수 혹은 상영 횟수를 배정하여 건설적인 의사결정이 가능함

- 마케팅

영화의 흥행 요인 중 하나는 사람들의 입소문을 통한 영화의 인지도이다. 주차 별 관객 수를 고려해 이에 맞춰 지속적으로 SNS 등으로 영화를 알린다면 인지도를 높여 관객을 추가적으로 확보할 수 있음

- 불확실성 해소

영화산업 특성 상 초기 투자비용이 크고 불확실하기 때문에 과거 데이터로부터 예측된 관객 수를 고려하여 투자 계획안을 세운다면 보다 합리적인 의사결정이 가능함

05. 후기 및 한계점

본 관객 수 예측 분석의 한계점

- RandomForest 의 Parameter 를 잘 조정하여야 했으나, 전문지식의 미비로 성능 부족의 결과

- 데이터의 분포 분야가 한정되어 있어, 미디어 언급, SNS 등 소셜 데이터가 활용되지 못함

- 상영시간에 따른 비용 증가에 관한 데이터가 없어 분석 목적에 완벽히 부합하지는 못함