



BIG data specialist

오윤후



I N D E X



1. 자기 소개

2. 프로젝트 요약



About




오 윤 후

dhdbsgn111@naver.com

<https://github.com/chalchichi>

-인하대 통계학과 졸업예정



A b i l i t y

Software

R

Rstudio 개발환경에서
ggplot2, randomForest, dplyr, Xgboost 등 라이브러리 활용

PYTHON

JUPYTER NOTEBOOKS 에서 Tensorflow 등 딥러닝 라이브러리 활용

C

R의 .c를 활용하여 반복문 수행

OS

LINUX

우분투 리눅스 사용 중

NO.1

포르투갈 은행의 정기에금 가입 여부 예측 분석



기간: 18.03~18.06

분석 소프트웨어 : R

역할: 데이터 가공 및 분석

설명 : Portuguese banking institution의 마케팅 캠페인 중 정기에금 가입여부를 예측하는 모델을 구성 했습니다.

- 고객의 직업, 나이 ,대출금액, 접근 횟수, 학력 등 20가지 변수를 사용하여 예측 모델 구성했습니다.
- 상대적으로 많은 결측값을 가진 변수들을 제거하고 경제지표들을 추가하여 새로운 데이터 구성하였습니다
- 트리 기반 배깅 , 랜덤 포레스트를 활용한 분석 담당하였고 다른 팀원의 딥 러닝 ,군집분석, 판별분석, 로지스틱 회귀분석등을 사용한 모형과 RMSE를 비교 하여 랜덤포레스트를 사용한 모형과 로지스틱 회귀분석을 사용한 모형 채택

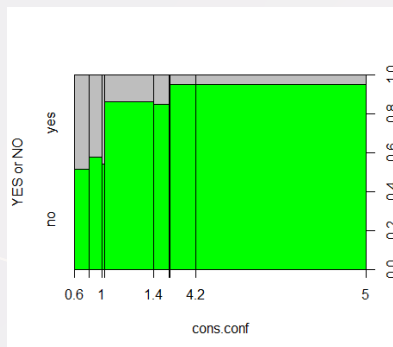
NO.1

프로젝트 과정, 결과



데이터 분석 과정 및 결과

	no	yes
-3.4	617	454
-3	84	88
-2.9	1069	594
-1.8	7723	1461
-1.7	370	403
-1.1	334	301
-0.2	9	1
-0.1	3451	232
1.1	7523	240
1.4	15368	866



Confusion Matrix and Statistics

Prediction \ Reference	no	yes
no	3587	413
yes	38	80

Accuracy : 0.8905
 95% CI : (0.8805, 0.8999)
 No Information Rate : 0.8803
 P-Value [Acc > NIR] : 0.02216

 Kappa : 0.2261
 Mcnemar's Test P-Value : < 2e-16

 Sensitivity : 0.9895
 Specificity : 0.1623
 Pos Pred Value : 0.8968
 Neg Pred Value : 0.6780
 Prevalence : 0.8803
 Detection Rate : 0.8711
 Detection Prevalence : 0.9713
 Balanced Accuracy : 0.5759

 'Positive' Class : no

90%대의 예측력을 보였습니다.

<깃허브 링크>

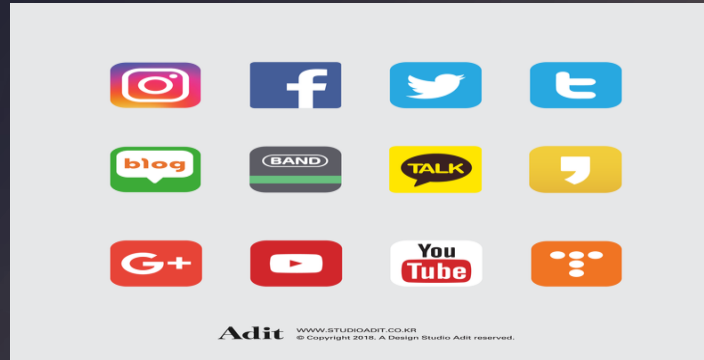
데이터 분석 과정 : <https://github.com/chalchichi/R.MNIST/blob/master/serch.hwp>

데이터 분석 코드 : <https://github.com/chalchichi/R.MNIST/blob/master/porutugal.r>

분석된 데이터+결과 : <https://github.com/chalchichi/R.MNIST/blob/master/project.zip>

NO.2

SNS 사용 시간에 따른 온라인 쇼핑몰 이용에 관한 연구



기간: 18.05~18.07

분석 소프트웨어 : R

역할: 자료 탐색 및 연구 계획

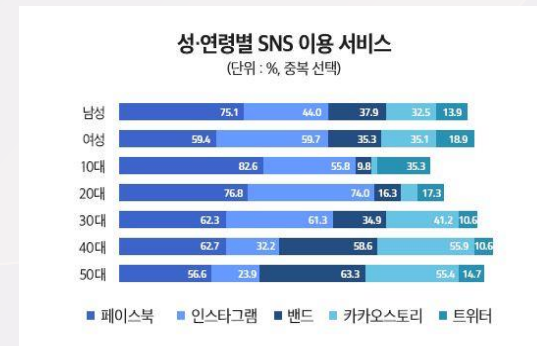
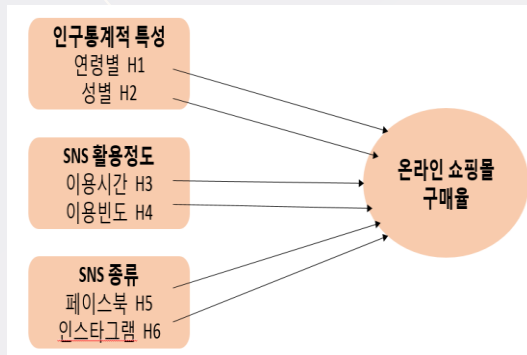
설명 : 선행되었던 연구들이 SNS내의 광고 형태와 매체에 노출되는 광고효과 , 즉 광고 제작의 주체인 기업의 입장에 치중되어 이와 대비되어 **광고에 노출되는 '개인에 대한 연구'** 가 필요하다고 판단 연구 계획과 기초자료조사를 맡았었습니다.

- 인구통계적 특성, sns 활용 정도, sns 종류 등을 변수로 선정하여 조사하였습니다.
- 사회조사데이터로써 하나의 가설이 아닌 다양한 가설들을 사용하여 분석하였습니다.
- Google Analytics를 통해 실제 쇼핑몰의 sns광고 효과를 분석하였습니다.

NO.1

프로젝트 과정, 결과

연구 분석 과정 및 결과



SNS와 다양한 변수들의 결합에 의해 구매력이 변화하는 정도를 측정하였습니다.

구매력의 영향을 미치는 다양한 요인들을 파악하게 되었습니다.

<깃허브 링크 >

연구 계획서: <https://github.com/chalchichi/R.MNIST/blob/master/Research%20plan.pptx>

INDEX



1. 영화 관객수 예측 모형
2. 병원 진료 환자수 예측
3. 프로그래머의 특성 분석

개인 프로젝트

NO.1

영 화 관 객 수 예 측 모 형



기간: 18.08~18.09

분석 소프트웨어 : PYTHON , R

역할: 데이터 수집, 가공, 분석

설명 : 9월 13일에 위의 3개의 영화의 9월 30일 까지의 누적 관객수를 예측하는 모형을 만들었습니다.

- 영화진흥 포털의 API키를 받아 자료를 수집하였습니다.
- 나를 차버린 스파이, 너의 결혼식의 경우 개봉하고 2주 후 영화관객수의 흐름과 주말, 공휴일 등 추가적인 변수를 활용할 수 있었지만 물괴의 경우 영화관객수라는 변수가 없어 다른 방식으로 구성하였습니다.

분석과정

<너의 결혼식, 나를 차버린 스파이>



-LSTM모형 사용

반응 변수가 시계열 적인 성질을 가지고 있다고 판단하였습니다. 반응 변수 이외의 공휴일, 좌석 수 등 다른 설명 변수도 활용 가능한 상태이므로 고전적인 시계열 분석을 하지 않고 딥 러닝 모형중 시계열데이터에 적합한 LSTM모형을 적용하였습니다.

-예측값 사용

좌석수, 상영관 수 등 13일 이후에는 미리 알 수 없는 설명변수들이 존재 하였으나 설명변수간의 회귀분석을 통해 예측된 값으로 대체하였습니다.

<물괴>



-2016년 9월의 상영 정보를 활용하여 회귀 분석

개봉한 당일 날 관객수 이외의 해당 영화에 직접적으로 관련된 다른 데이터는 활용 할 수 없었습니다.

따라서 2016년 9월의 상영 영화들의 데이터들을 활용하여 회귀분석을 진행하였습니다.

-추석연휴가 있어서 두가지 경우로 분석을 진행하였습니다.

NO.1

프로젝트 과정, 결과

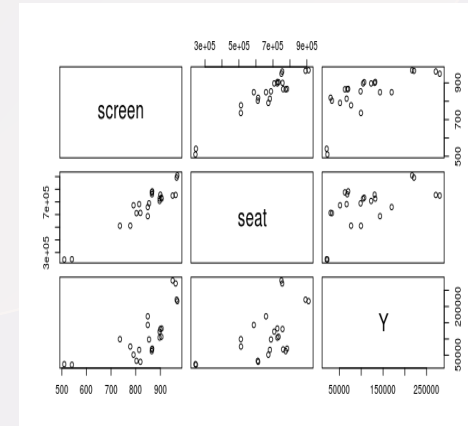
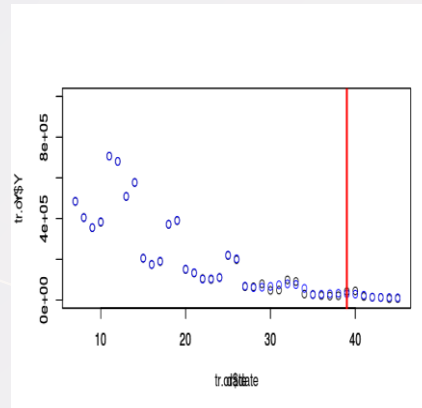


데이터 분석 과정 및 결과

Assessment

1. 결과

-너의 결혼식 : 2845488명
 -나를 차버린 스파이 : 254871명
 -물괴: 995021명



<깃허브 링크 >

PYTHON 코드 :

<https://github.com/chalchichi/R.MNIST/blob/master/LSTM%20in%20tensorflow.ipynb>

R코드 : <https://github.com/chalchichi/R.MNIST/blob/master/movie%20code.R>

API활용 데이터 추출 코드 : <https://github.com/chalchichi/R.MNIST/blob/master/movieAPI.R>

분석된 데이터 : <https://github.com/chalchichi/R.MNIST/blob/master/movie.zip>

개인 프로젝트

NO.2

감기 진료 환자 수 예측 모형



기간: 18.03~18.07

분석 소프트웨어 : R

역할: 데이터 가공, 분석

설명 : 2013년 01월 01일부터 2013년 11월 30일까지의 일별 자료를 이용하여, 일별 감기진료건수를 반응변수로 하고 나머지 변수들을 설명변수로 사용하는 예측 모형을 학습하였습니다

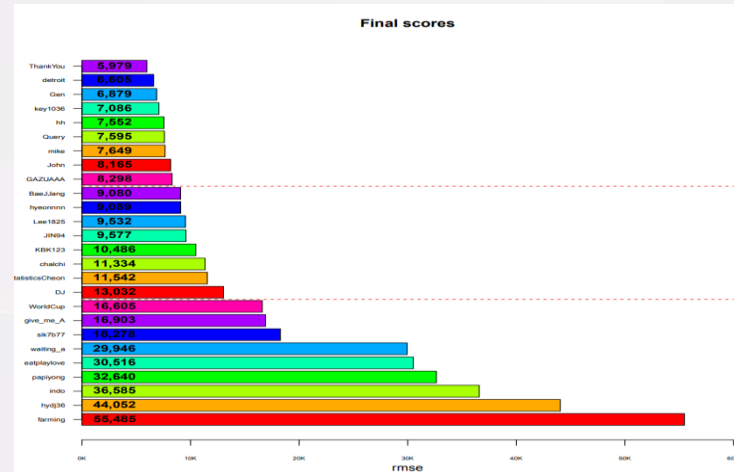
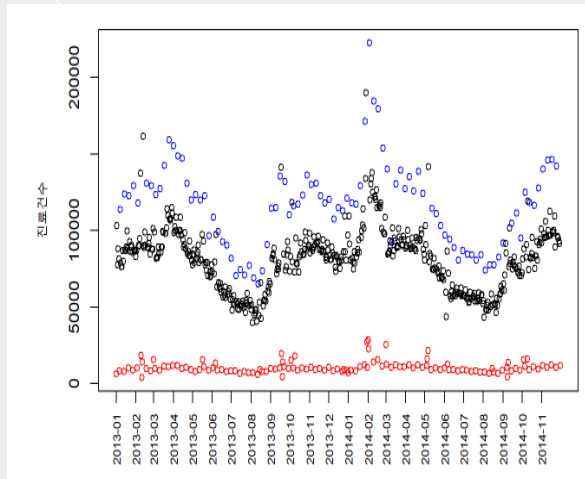
만들어진 예측모형에 2014년 12월 01일부터 2014년 12월 31일까지의 일별 기상자료 및 SNS자료를 입력하여 해당 기간의 일별 감기 진료 건수의 예측 값을 얻는 모형을 만들었습니다.

- 요일, 공휴일 등 추가적인 변수를 사용하여 예측력을 향상시켰습니다.
- 2014년 12월의 1일부터 30일까지 하루 단위의 환자수를 예측하여야 하였고 많은 수의 범주형 자료가 섞여 있어서 앙상블 모형을 사용하였습니다.

NO.1

프로젝트 과정, 결과

데이터 분석 과정 및 결과



RMSE 9000~11000정도의 예측 성능을 만들었습니다.

<깃허브 링크>

R코드 : <https://github.com/chalchichi/R.MNIST/blob/master/ANSENBLE.R>

분석된 데이터 : [https://github.com/chalchichi/R.MNIST/blob/master/dat_pred%20\(1\).csv](https://github.com/chalchichi/R.MNIST/blob/master/dat_pred%20(1).csv)

프로 게이머의 전적 분석



기간: 17.06~17.07

분석 소프트웨어 : R

역할: 데이터 수집, 가공, 분석

설명 : 과거 전전데이터를 비공식 API로 수집하여 데이터가 존재 하는 모든 선수들을 분석하여 성적의 영향을 주는 요인을 분석하고

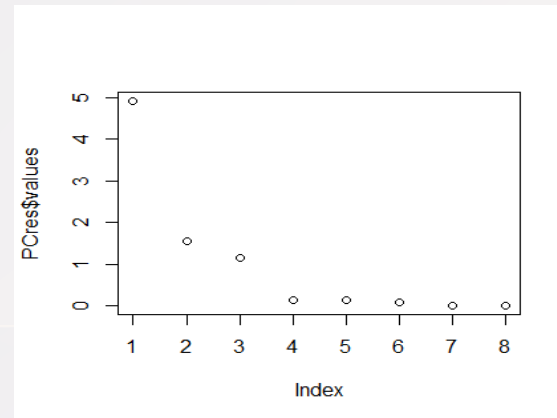
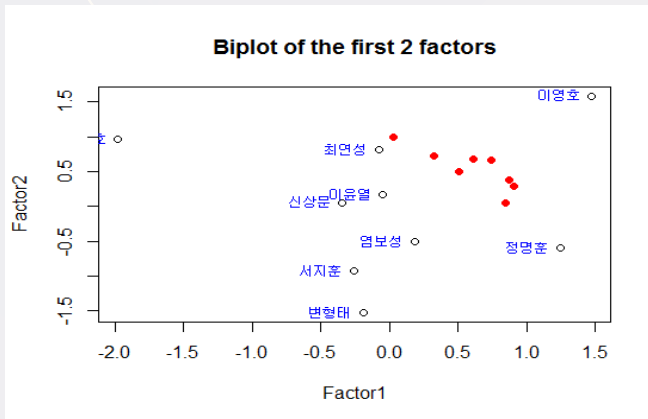
- www.ygosu.co.kr에서 자료수집
- PCA를 활용하여 각 분야별 좋은 성적을 낸 선수들의 특징을 찾아 내는 분석을 실행하였습니다.

NO.3

프로젝트 과정, 결과



데이터 분석 과정 및 결과



게임 전문 사이트에서 비공식 API를 통해 새로운 데이터를 추출하여 분석을 실시하였고 수학적인 모델을 통해 기존의 통념을 증명하고 다양한 분석결과를 만들 수 있었습니다.

<깃허브 링크 >

R코드 : <https://github.com/chalchichi/R.MNIST/blob/master/starcraft.R>

API를 통한 데이터 수집 코드: <https://github.com/chalchichi/R.MNIST/blob/master/rvest.R>

감 사 합 니 다

