Analysis – 퓨처스리그 영화 관객수 예측

팀 명	팀 원	학 교
우리들의 일그러진 스님	강수현, 여현웅, 한선웅, 김형규	동국대학교

팀 소개

팀 소개

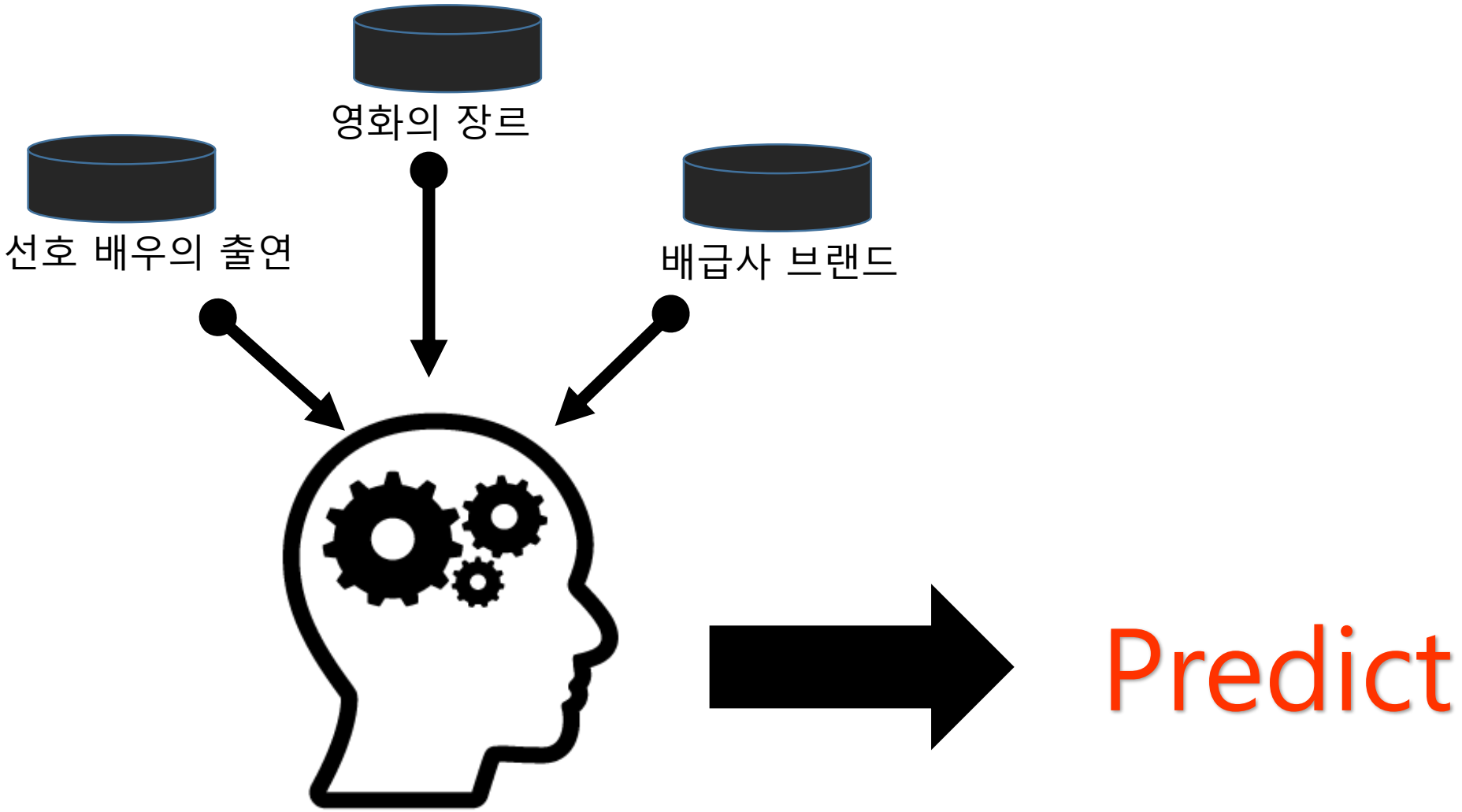
- **분야 및 주제 명칭**
  - ✓ **Analysis –** 퓨처스리그(영화 관객수 예측)

- **팀 명 및 구성원**
  - ✓ 팀 명 : 우리들의 일그러진 스님
  - ✓ 구성원 : 팀장 - 강수현 / 팀원 - 여현웅, 한선웅, 김형규

- **구성원 별 역할**
  - ✓ 강수현 : 시계열 데이터 **set** 전처리
  - ✓ 여현웅 : 모델 개발
  - ✓ 한선웅 : 비시계열 데이터 **set** 전처리
  - ✓ 김형규 : **DB** 및 **ppt** 제작

프로젝트 기획

프로젝트 기획



프로젝트 기획

실제 관객이 영화를 선택하는 기준?

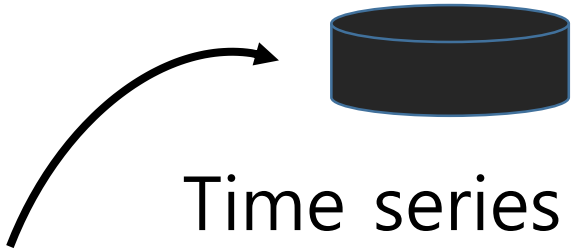
그 시기 가장 인기 있는 영화 위주의 선택

그 시기 개봉된 영화 중 끌리는 것

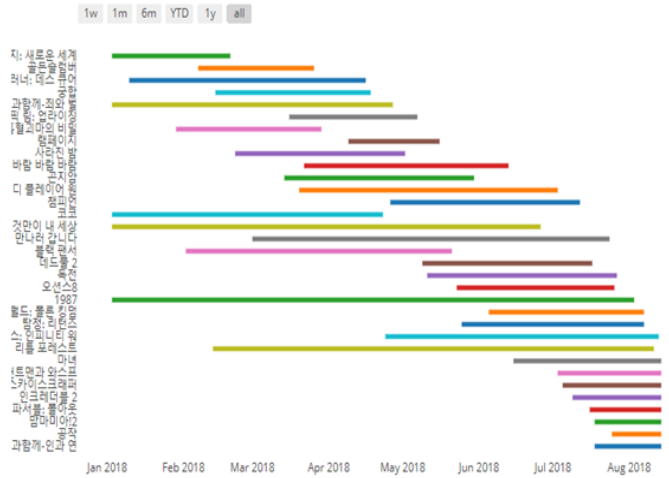
프로젝트 기획

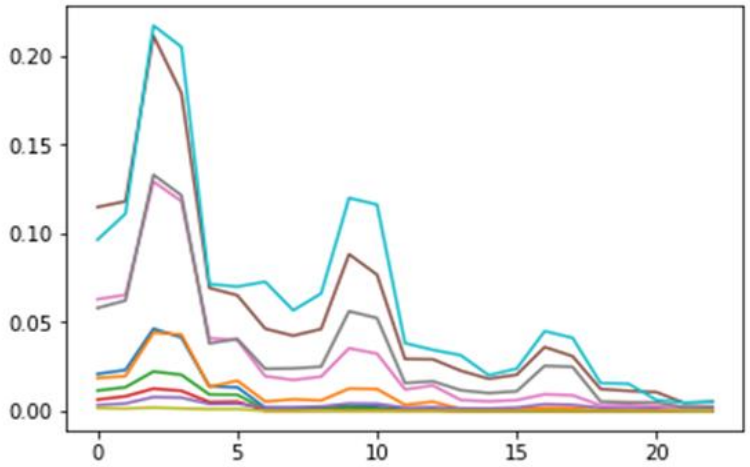
!!!

관람객 변화 추이가 비슷하네?

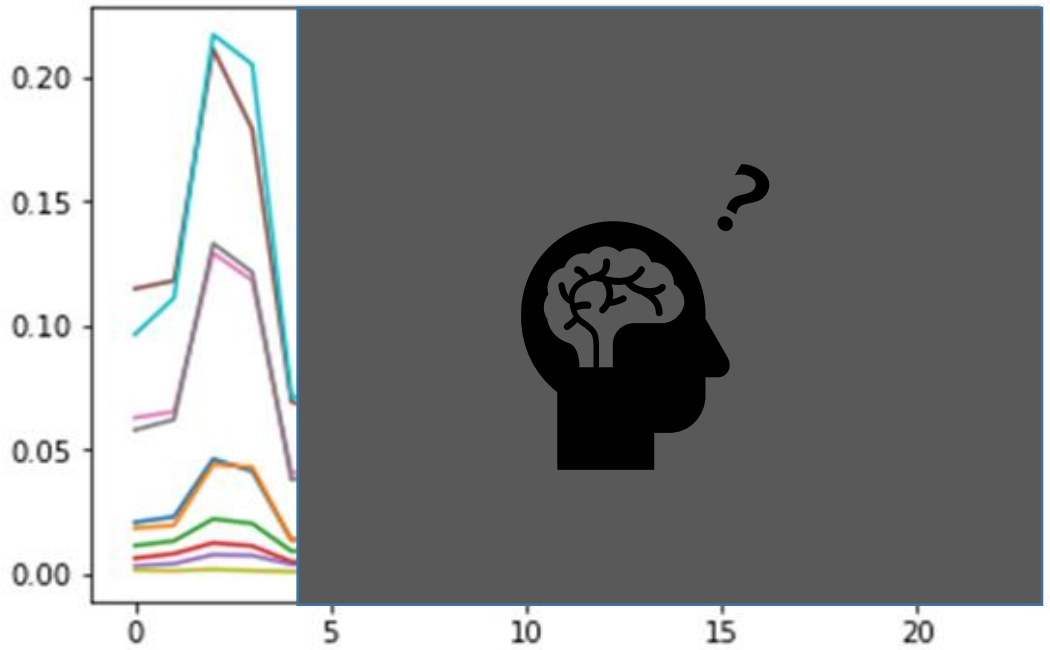


Time series Data





프로젝트 기획



특정 기간에 데이터를 통해 뒷부분의 그래프를 예측 할 수 있다면
총 관객수도 알 수 있지 않을까?

Time Series + **Other Data etc** + **Deep learning**

데이터 전처리

Time Series Data

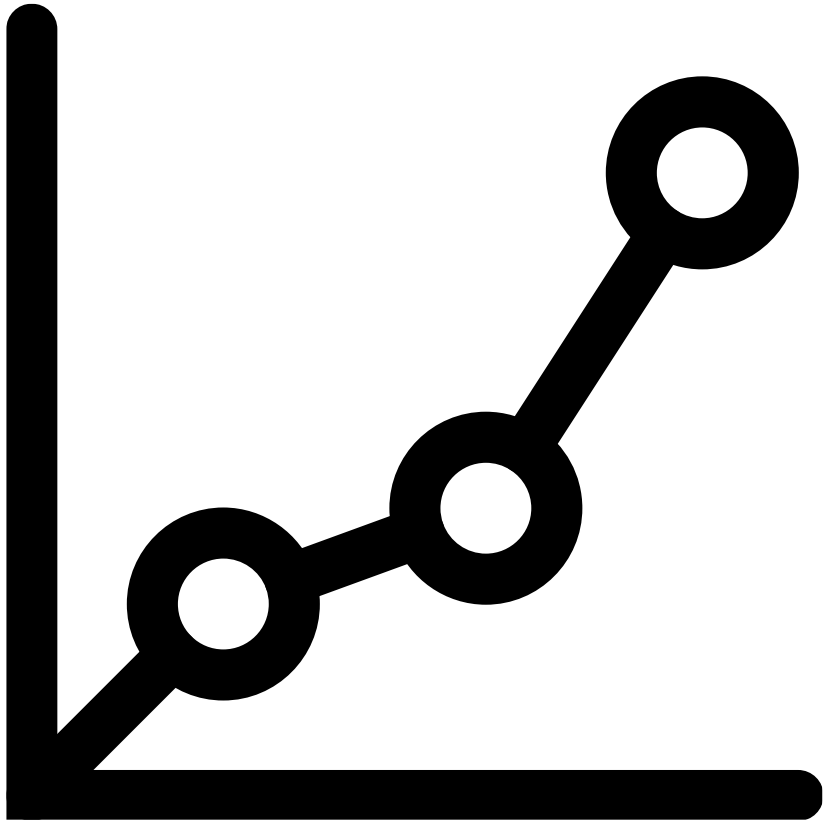
- 영화 별 일일 관객수
- 영화 별 일일 상영횟수
- D-Day

Nontime Series Data

- 개봉일
- 제작 국가
- 관람 등급
- 장르
- 개봉 전 평점
- 감독
- 배우

Time Series Data

Feature

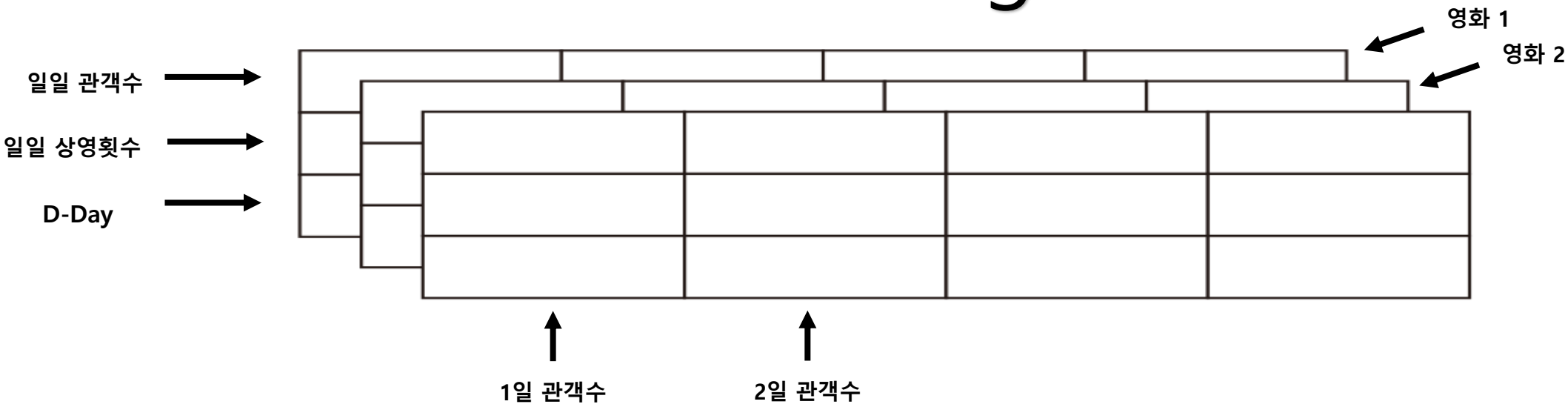


총 1047 편

Feature	대상 기간	수집 범위
영화별 일일 관객수	2015년 ~ 2018년	• 관객 수 10,000명 이상 • 각 영화 별 개봉일 ~ 종영일
영화별 일일 상영횟수		
D-day		

Time Series Data

Data Handling



Feature	범위 조정	결측치 처리
영화별 일일 관객수	0에서 1의 값을 가지도록 스케일링	개봉전 22일간 또는 개봉후 40일간 데이터가 없는 경우 0으로 대체. (Zero padding)
영화별 일일 상영횟수		
D-day	-	

Time Series Data

Data Handling

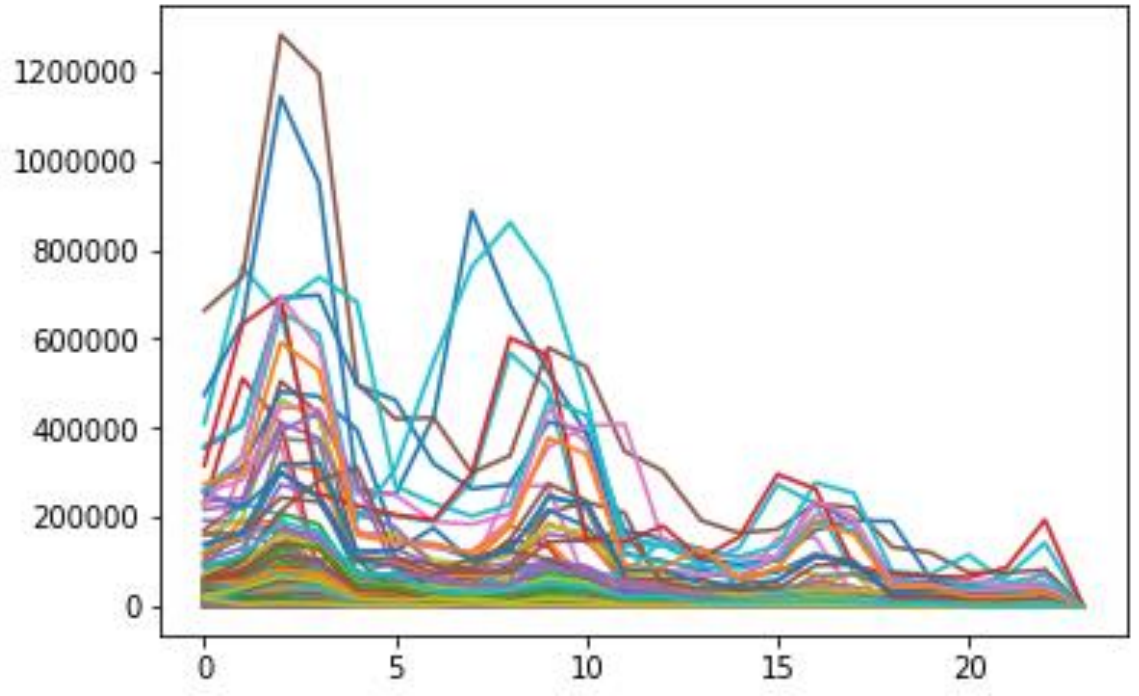
랜덤하게 200개의 영화 Data 추출

↓

Data imbalanced 확인

↓

Oversampling 수행 결정



Time Series Data

Oversampling

DeepAR : Probabilistic Forecasting with Autoregressive Recurrent Networks

(Valentin Flunkert, David Salinas, Jan Gasthaus)

v_i = i번째 영화에 대한 관객수 평균.

$z_{i,t}$ = i번째 영화 t일 차의 관객수.

p_i = 모든 영화 데이터셋중 i번째 영화가 선택될 확률

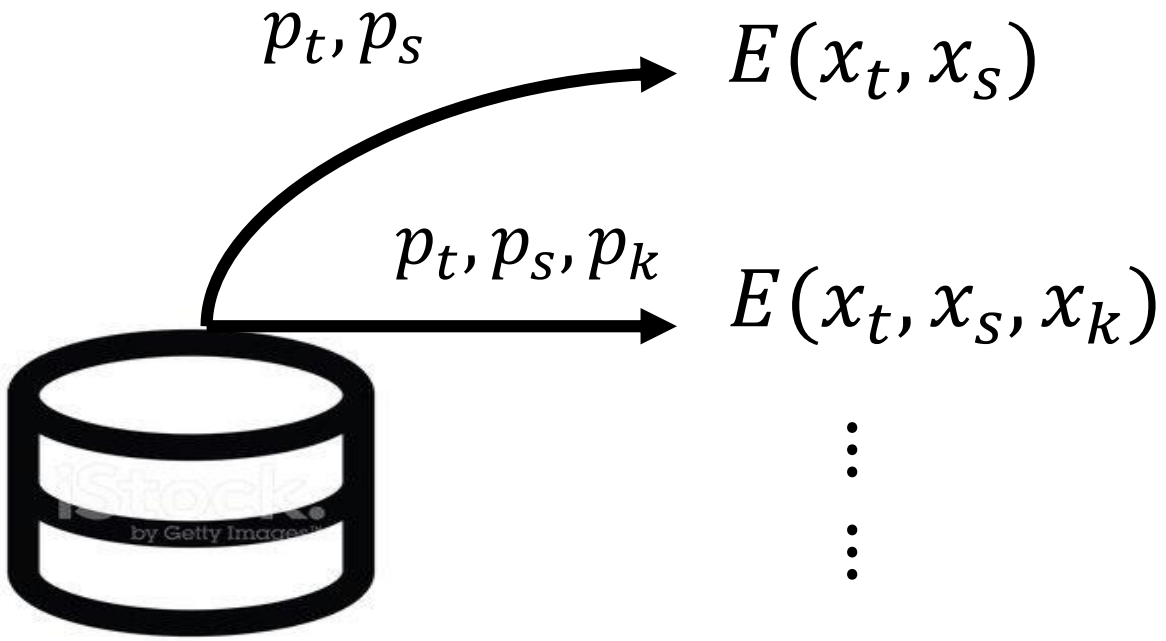
$$v_i = 1 + \frac{1}{t_o}\sum_{t=1}^{t_o} z_{i,t}$$

$$p_i = \frac{v_i}{\sum v_i}$$

→ Probabilistic을 기반으로 Sampling 수행

Time Series Data

Oversampling



- 관객수가 많은 영화
  - ✓ 선택될 확률 ↑
- 관객수가 낮은 영화
  - ✓ 선택될 확률 ↓
- 임의의 데이터들을 선택
  - → 값들의 평균 계산
- 그 평균을 새롭게 생성한 데이터로 사용.

Nontime Series Data

Feature

변수	대상 기간	설명	수집한 이유
Open	2015년부터 2018년에 개봉한 영화	개봉일	개봉한 요일의 영향이 있을 것이다.
Nation		제작 국가	언어의 차이가 관객수의 차이를 만들어 낼 것이다.
Grade		관람 등급	영화의 주고객층이 달라지기 때문에 관객수에 영향을 미칠 것이다.
Genre		장르	인기있는 장르가 따로 있을 것이다.
Score		개봉전 평점	개봉전 영화에 대한 관심
Director		감독	흥행한 영화를 만들어본 감독은 추후작도 흥행할 가능성이 높을 것이다.
Actors		배우들	인기있는 배우들이 출연한 영화도 인기있을 것이다.

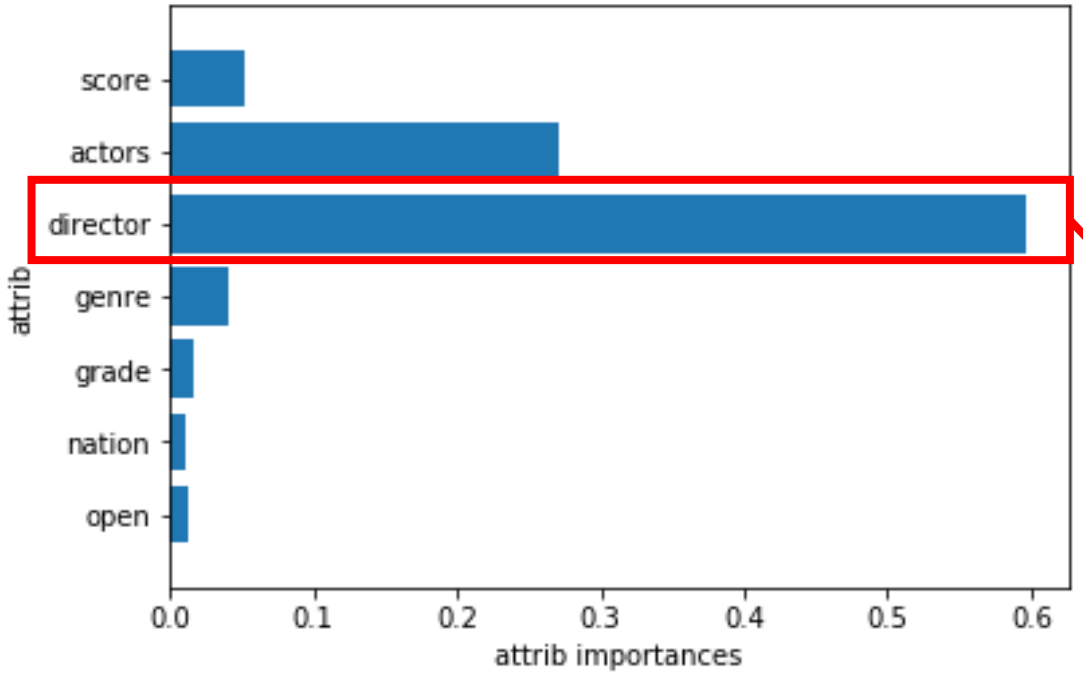
Nontime Series Data

Data Handling

Director	최근 3작의 관객 동원 수 평균
Nation	(한국, 영미권, 일본, 중국, 기타) Class로 나누고 Label Encoding
Grade	(전체, 12세, 15세, 18세, 청불) Class로 나누고 Label Encoding
Genre	두개 이상의 장르를 가진 경우 앞의 하나만 남긴 후 Label Encoding
Score	Real Data 사용
Actors	주연 배우의 최근 3작의 관객 동원 수의 합 → 7개의 Class로 나누고 Label Encoding

Nontime Series Data

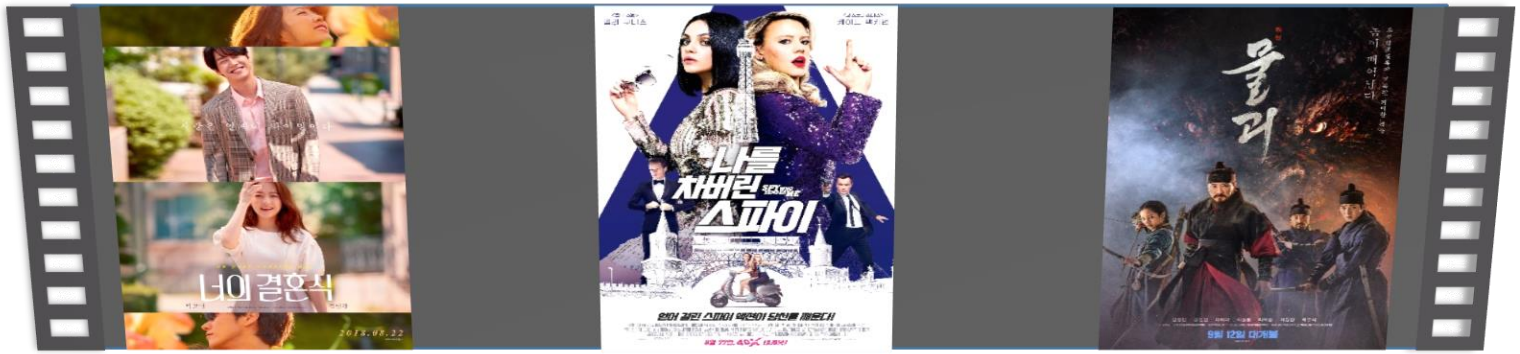
Data Handling



- **Random Forest의 Feature importance를 확인한 결과 감독의 관객 동원수가 관객수에 중요한 영향을 끼치는 것으로 확인되어 분석에 사용하기로 결정**

학습 알고리즘

학습 알고리즘



RNN

RNN

RNN
Encoder-Decoder

일별 데이터

	date	audience	play	director
3	2018-06-28	1,834	288	235499
4	2018-06-29	1,093	296	235499
5	2018-06-30	2,007	274	235499
6	2018-07-01	1,576	252	235499
7	2018-07-02	689	200	235499
8	2018-07-03	676	198	235499
9	2018-07-04	114	35	235499
10	2018-07-05	506	23	235499

누적 관객 수

accum_audience
6088876

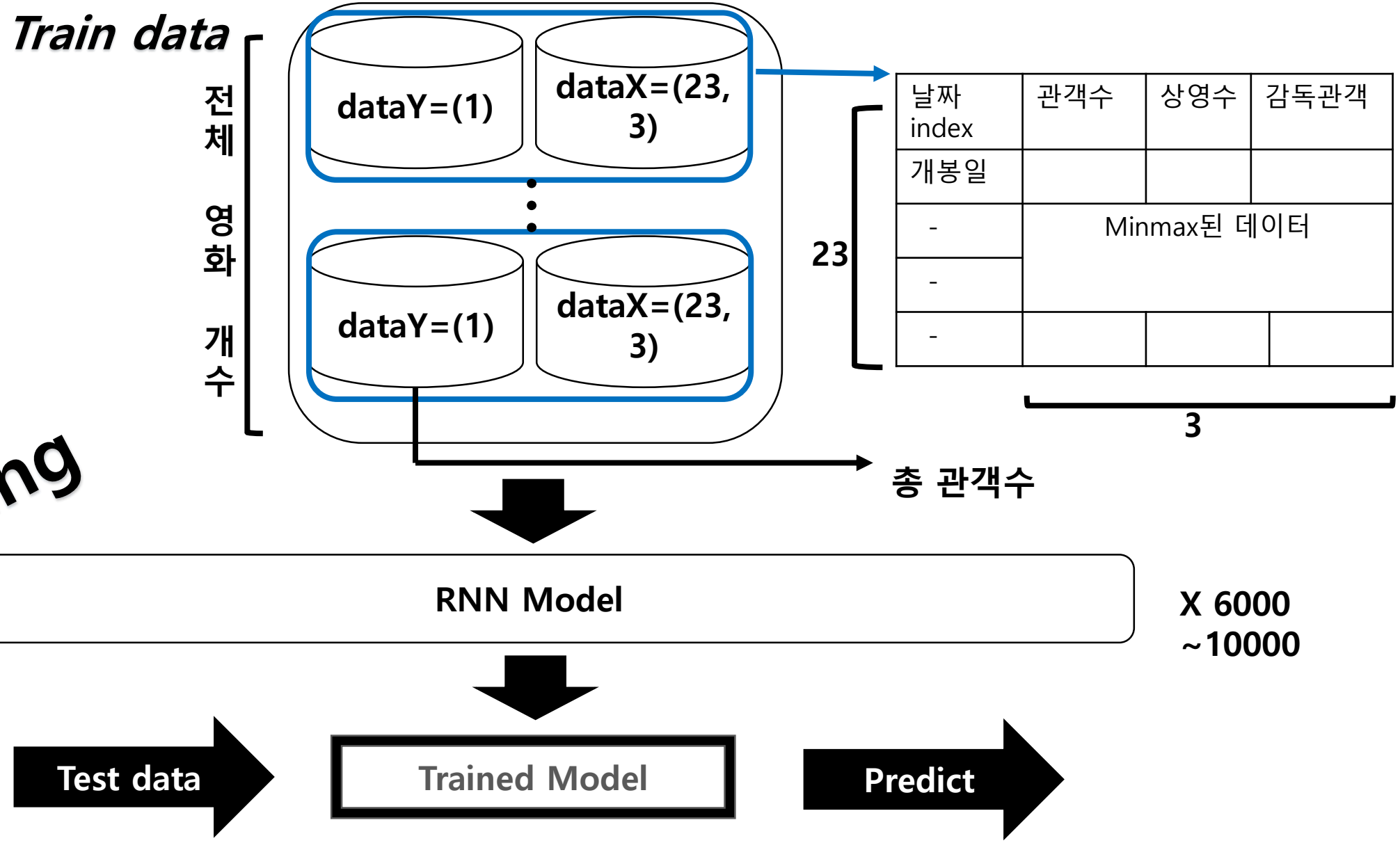
2일치 데이터

	date	audience	play	director
3	2018-06-28	1,834	288	235499
4	2018-06-29	1,093	296	235499

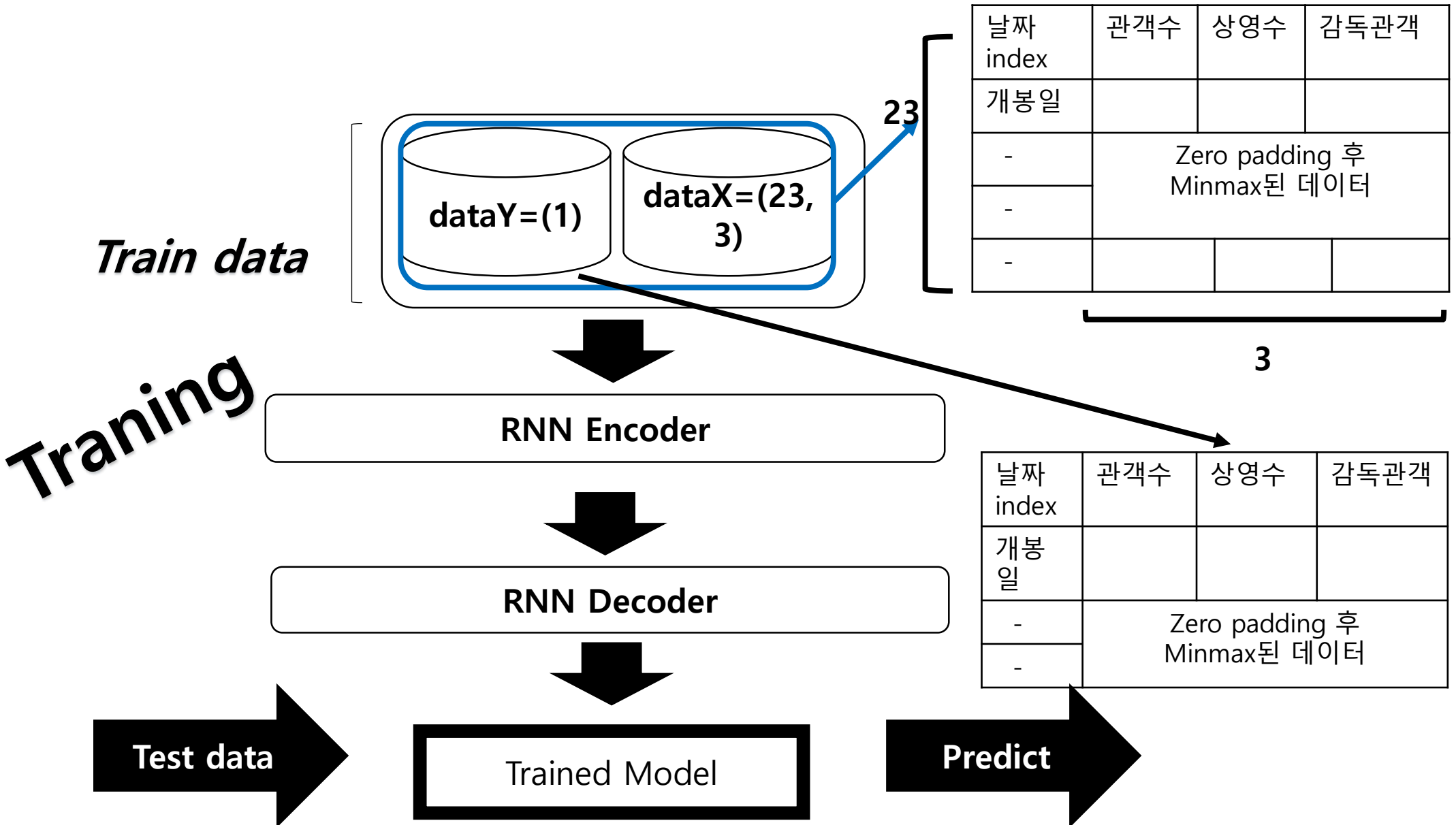
누적 관객 수

accum_audience
6088876

학습 알고리즘


Train data
전체 영화 개수
dataY=(1)
dataX=(23, 3)
dataY=(1)
dataX=(23, 3)
23
날짜 index
관객수
상영수
감독관객
개봉일
-
-
-
Minmax된 데이터
3
총 관객수
Traning
RNN Model
X 6000 ~10000
Test data
Trained Model
Predict

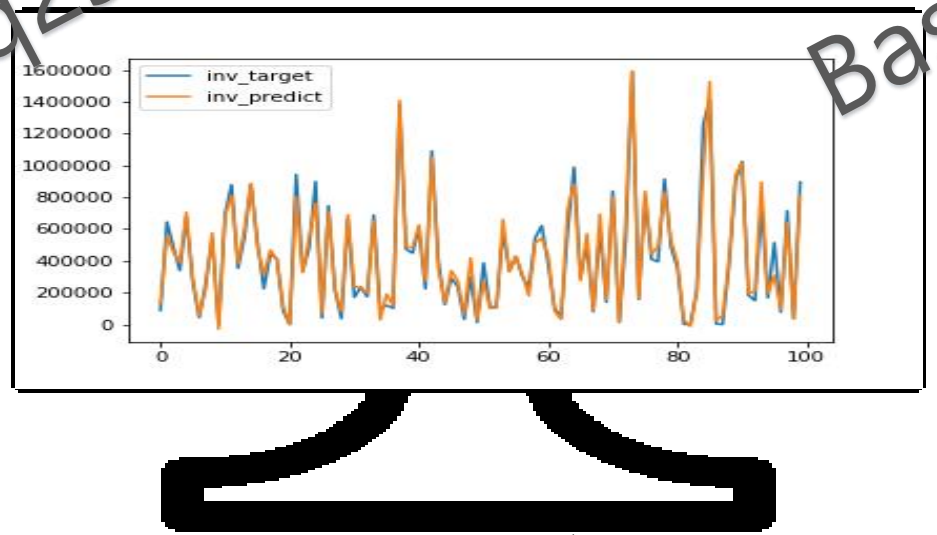
학습 알고리즘


Train data
dataY=(1)
dataX=(23, 3)
23
3
날짜 index
관객수
상영수
감독관객
개봉일
-
-
-
Zero padding 후 Minmax된 데이터
Traning
RNN Encoder
RNN Decoder
Test data
Trained Model
Predict
날짜 index
관객수
상영수
감독관객
개봉일
-
-
Zero padding 후 Minmax된 데이터

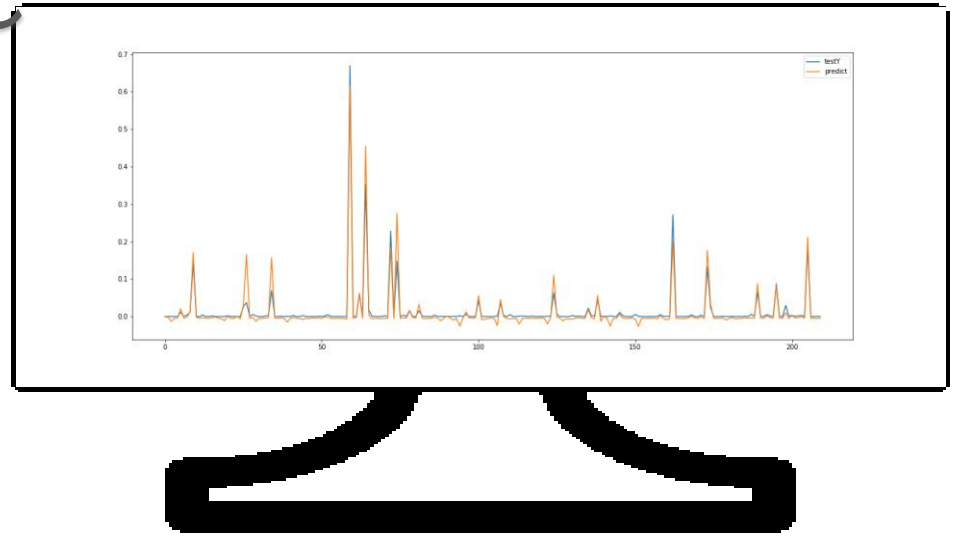
결과 및 결론

결과

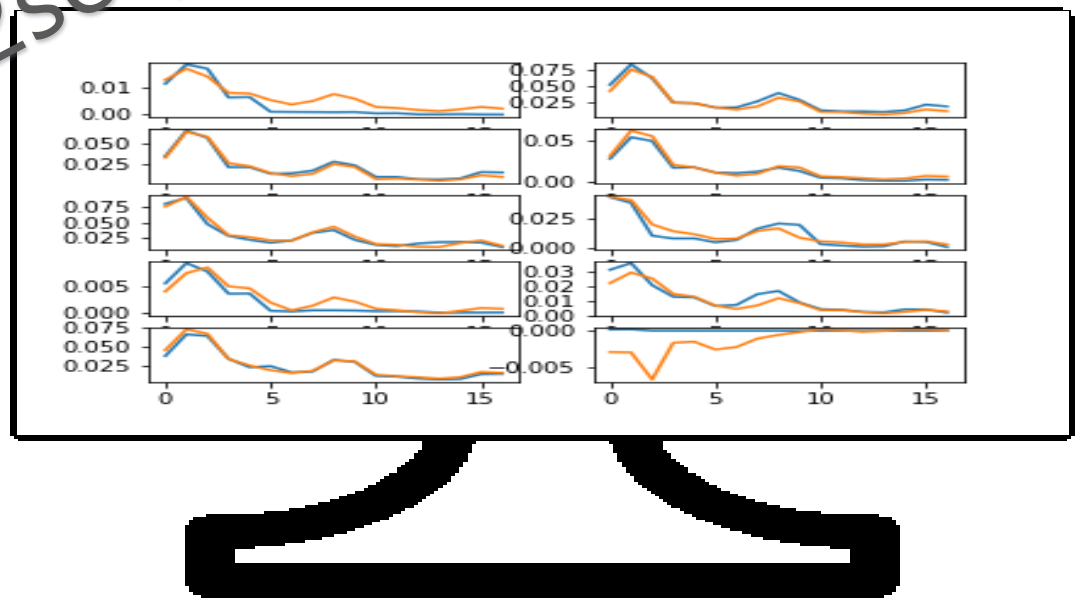
Seq2seq

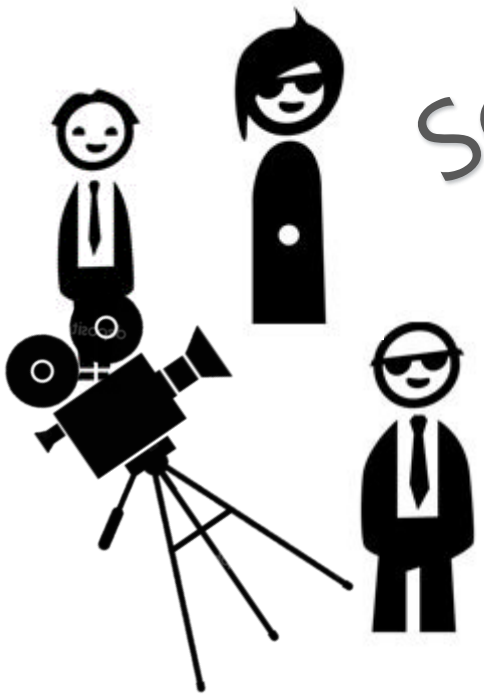


Basic-rnn



Seq2seq





결론

	lstm	Seq2seq
	예측 관객수	예측 관객수
너의 결혼식	2,767,164	X
나를 차버린 스파이	271,311	X
물괴	X	1,234,491

참고 문헌

참고문헌 및 자료
(1) http://www.kobis.or.kr/kobis/business/mast/mvie/searchMovieList.do
(2) https://movie.naver.com/
(3) Valentin Flunkert, David Salinas, Jan Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. arXiv:1704.04110, 2017

-END-