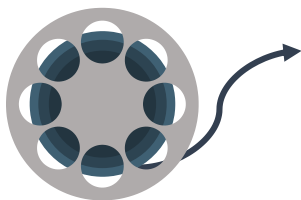


Analysis – 퓨처스리그

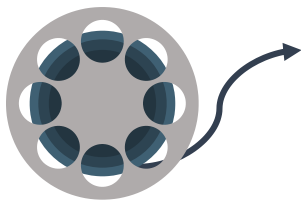
영화 관객수 예측

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



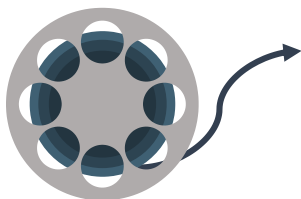
CONTENTS 1

팀 소개 및 역할



CONTENTS 2

프로젝트 기획

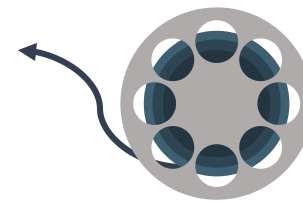


CONTENTS 3

데이터 전 처리

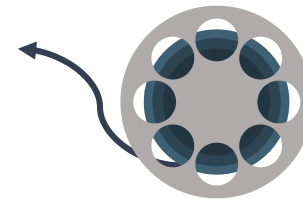
CONTENTS 4

학습 알고리즘



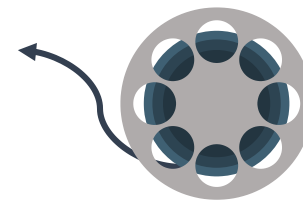
CONTENTS 5

결과 및 결론



CONTENTS 6

개선점



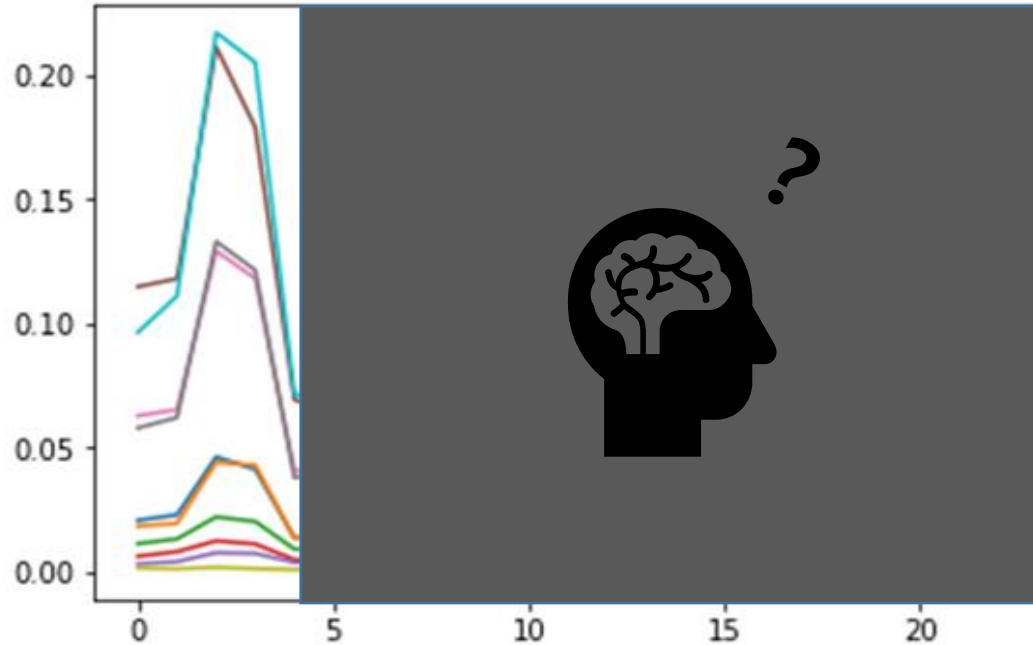
팀 소개 및 역할

팀 소개

- 분야 및 주제 명칭
 - ✓ **Analysis** – 퓨처스리그(영화 관객수 예측)
- 구성원
 - ✓ 구성원 : **김형규**, 강수현, 여현웅, 한선웅 (from. Dongguk AI Association-DAIA)
- 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

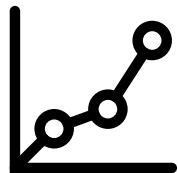


너의 결혼식, 나를 차버린 스파이: 23일
물과: 2일

너의 결혼식, 나를 차버린 스파이: 40일째
물과: 19일째

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

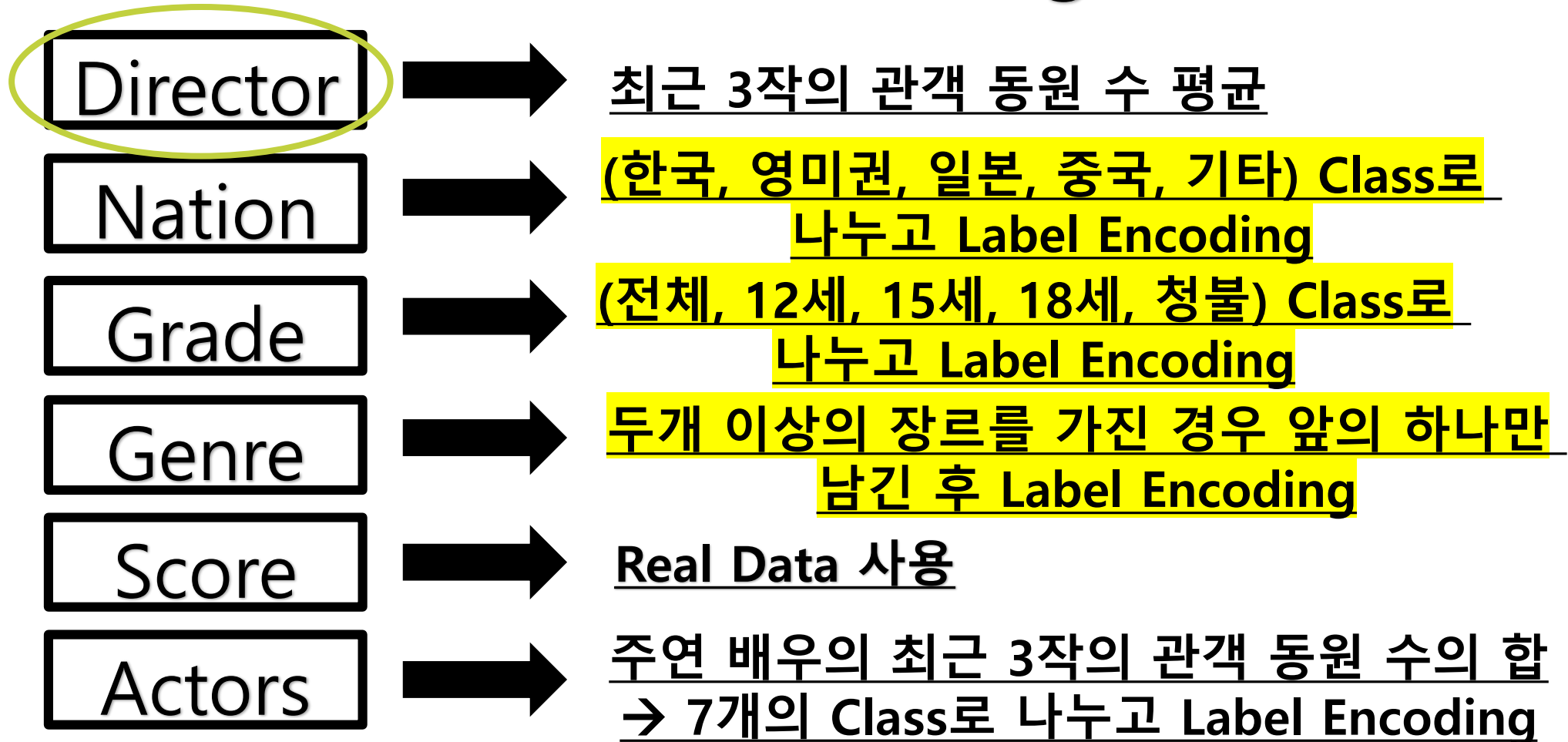
Feature

총 1047 편

Feature	대상 기간	수집 범위
영화별 일일 관객수	2015년 ~ 2018년	<ul style="list-style-type: none">• 관객 수 10,000명 이상• 개봉일 ~ 종영일
영화별 일일 상영횟수		
D-day		

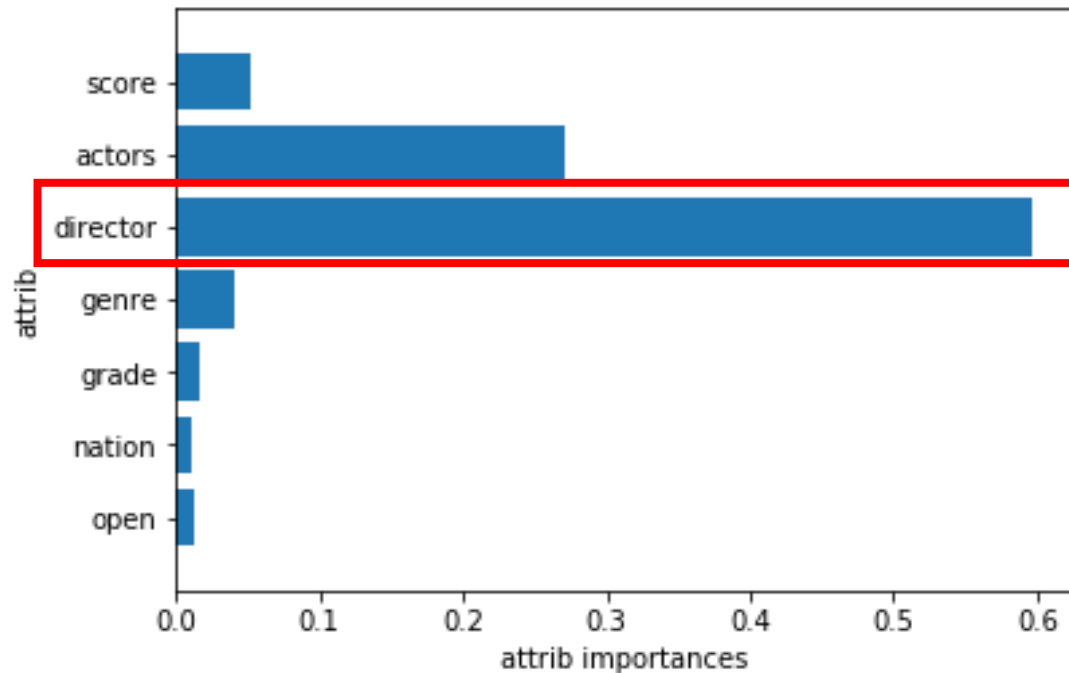
Nontime Series Data

Data Handling



Nontime Series Data

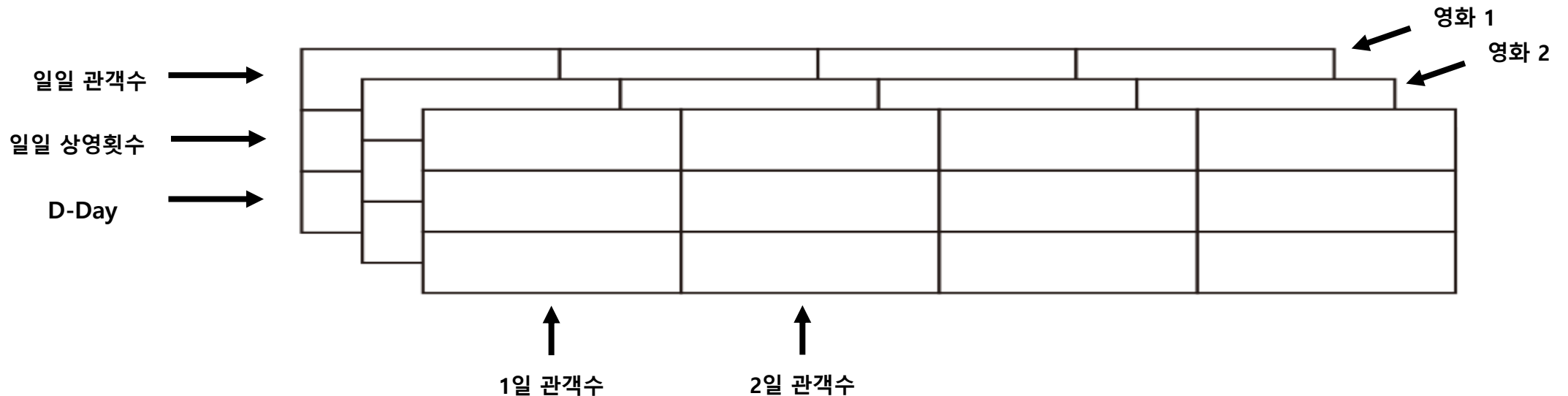
Data Handling



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

Time Series Data

Data Base

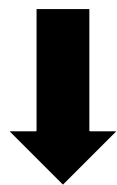


Feature	범위 조정	결측치 처리
영화별 일일 관객수	0에서 1의 값을 가지도록 스케일링	개봉전 22일간 또는 개봉후 40일간 데이터가 없는 경우 0으로 대체. (Zero padding)
영화별 일일 상영횟수		
D-day	-	
director		

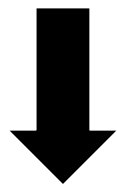
Time Series Data

Data Sampling

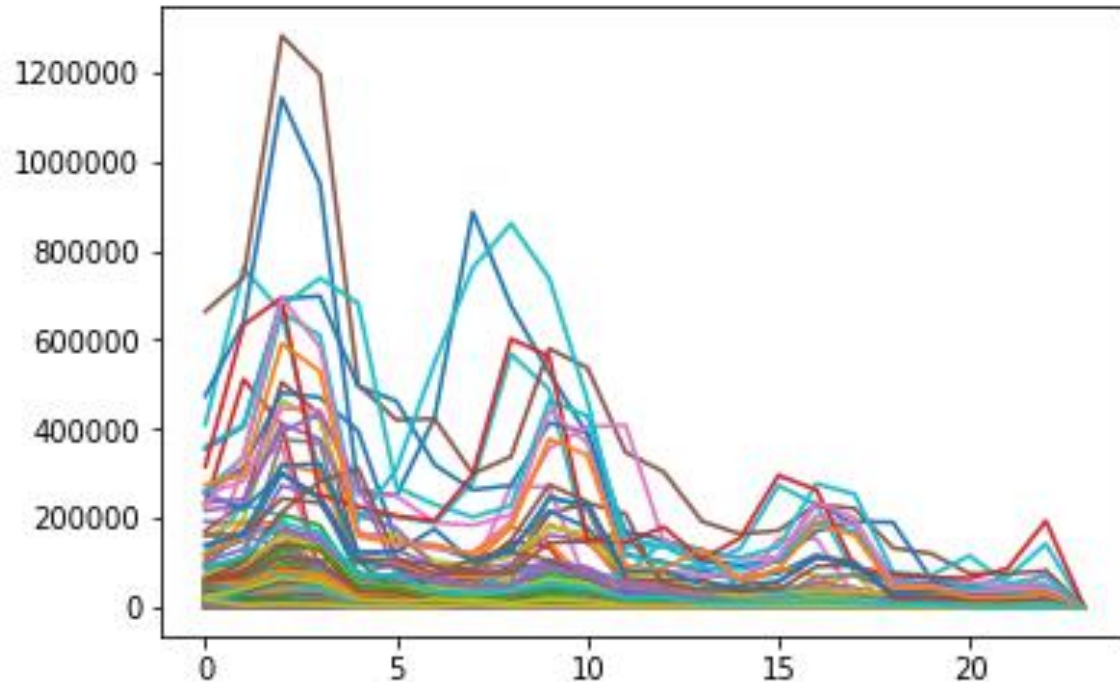
랜덤하게 200개의 영화 Data 추출



Data imbalanced 확인



Oversampling 수행 결정



Time Series Data

Oversampling

DeepAR : Probabilistic Forecasting with Autoregressive

Recurrent Networks

(Valentin Flunkert, David Salinas, Jan Gasthaus)

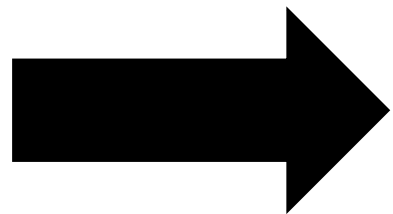
v_i = i번째 영화에 대한 관객수 평균.

$z_{i,t}$ = i번째 영화 t일 차의 관객수.

p_i = 모든 영화 데이터셋중 i번째 영화가 선택될 확률

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

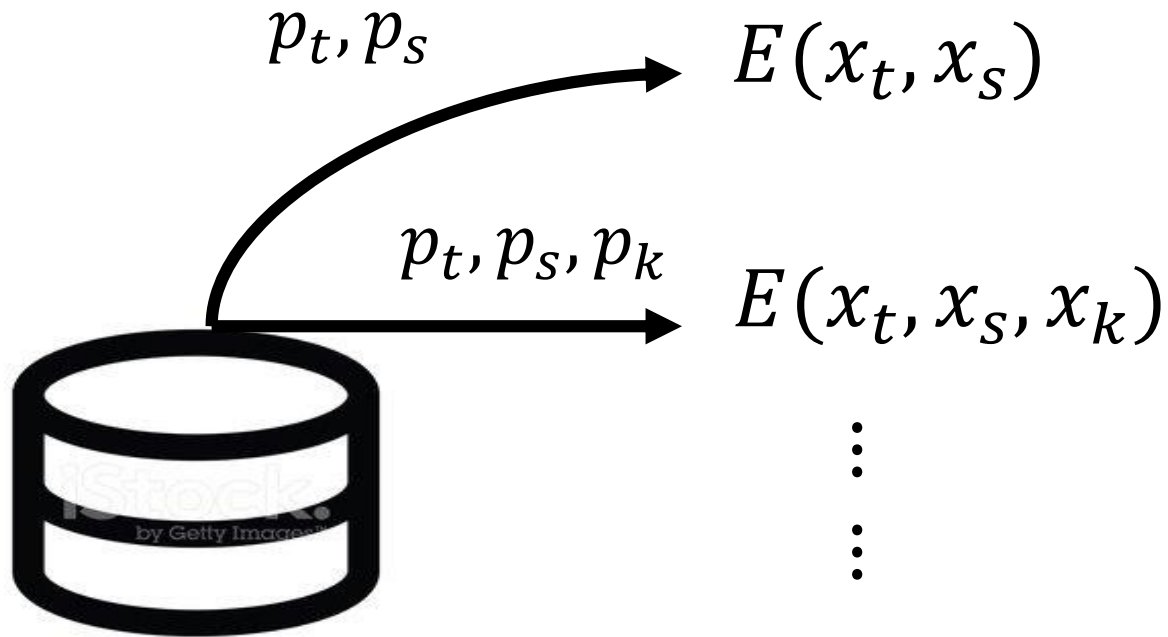
$$p_i = \frac{v_i}{\sum v_i}$$



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.

학습 알고리즘

학습 알고리즘



RNN



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

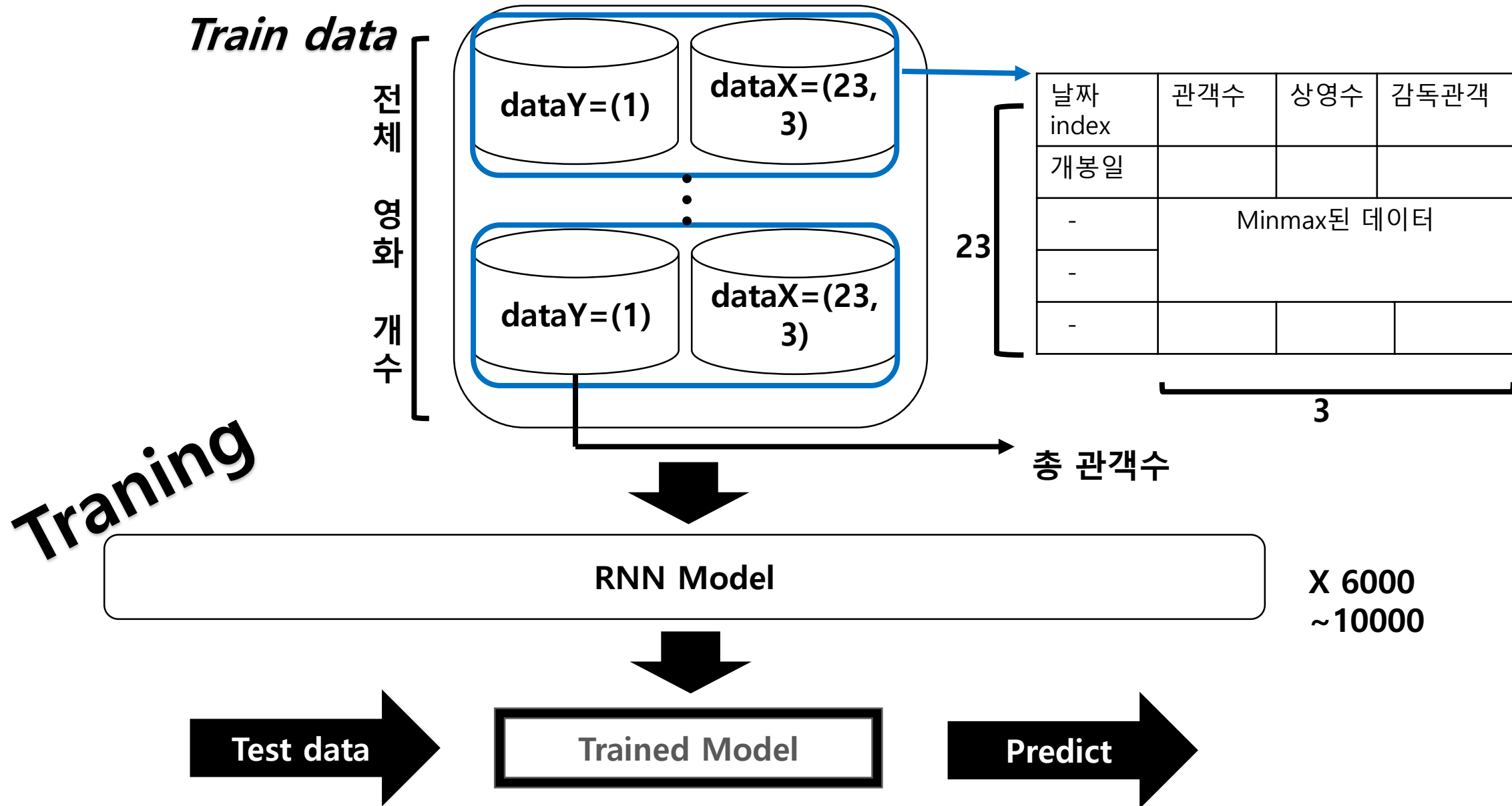
2일치 데이터

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

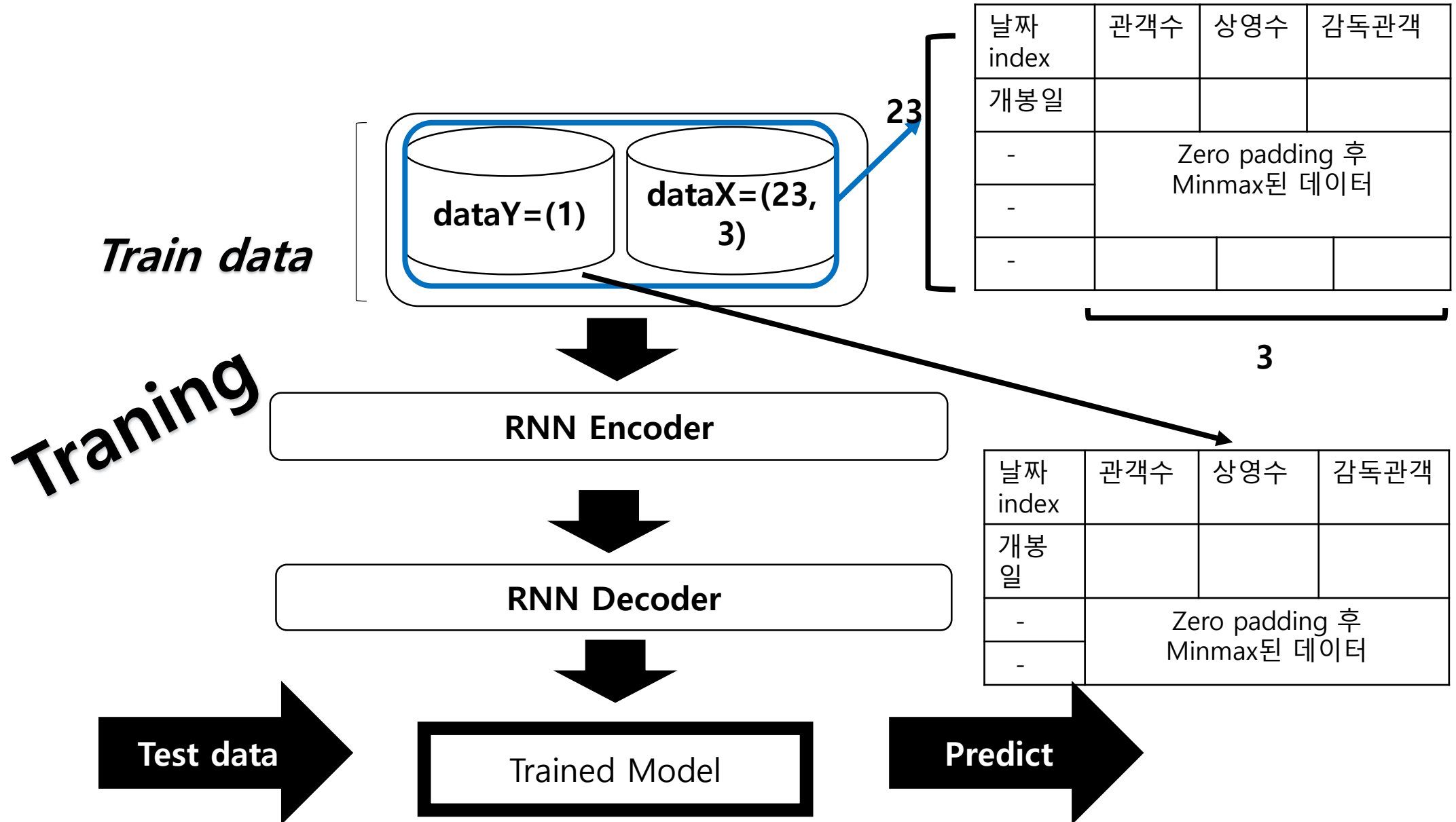
누적 관객 수

accum_audience
6088876

학습 알고리즘



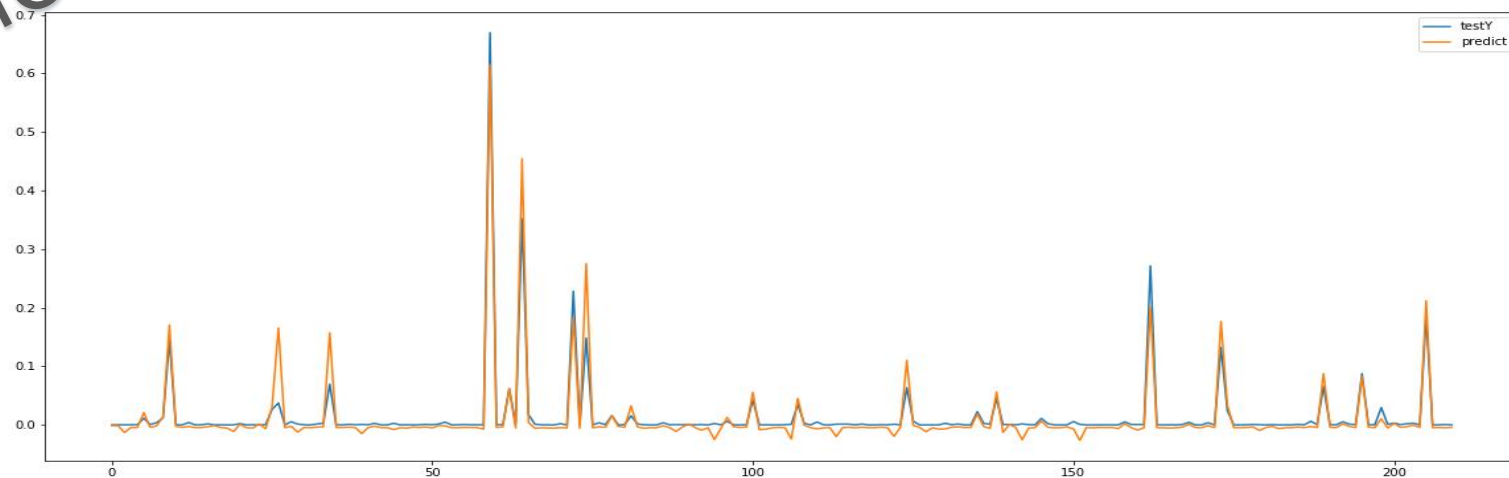
학습 알고리즘



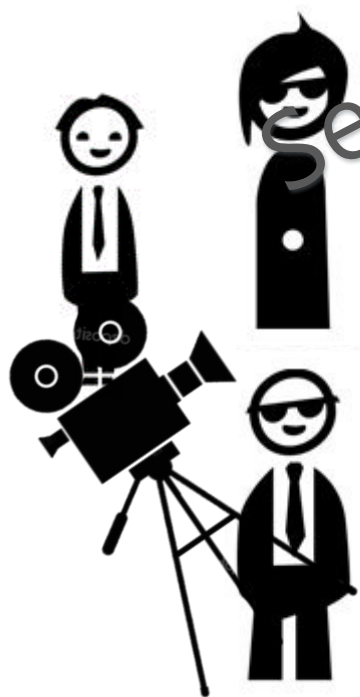
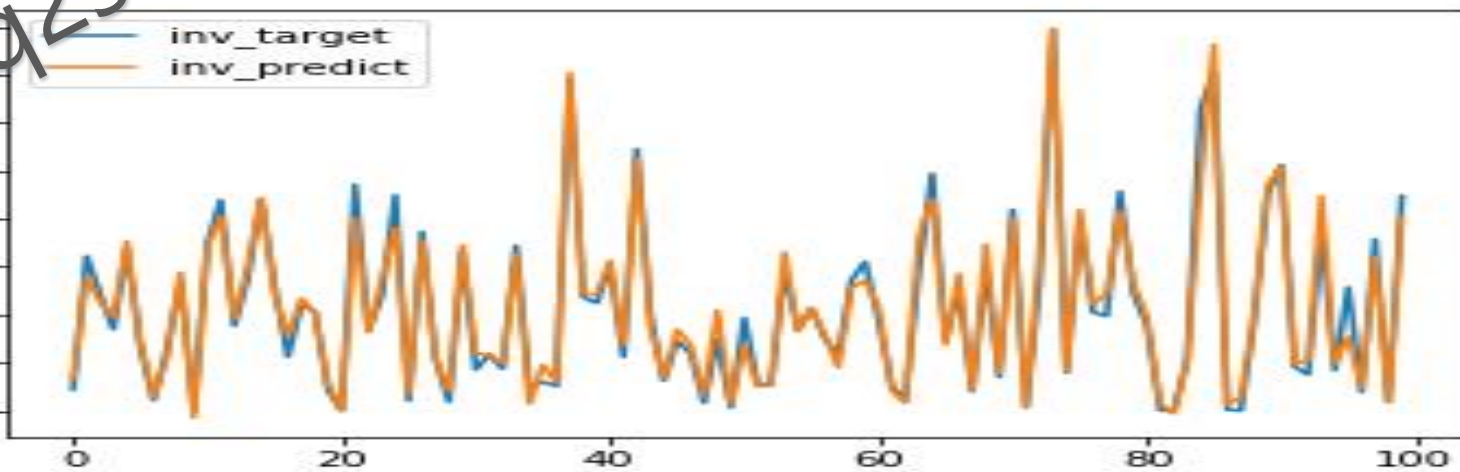
결과 및 결론

Basic-rnn

결과



Seq2seq



결론

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



개선방향

너의 결혼식	• 2015년 이전 데이터 조금 더 확보
나를 차버린 스파이	
물괴	• 머신러닝 기법 적용 • Non time series를 더 많이 적용



상세역할

- 알고리즘 선별
 - ✓ RNN
 - ✓ LSTM
 - ✓ Random Forest (feature importance)
- 데이터 수집
 - ✓ 영화진흥위원회 – 일별 관객수, 일별 상영횟수
- 데이터 전처리
 - ✓ Zero padding + windowing, database
- Feature importance check
 - ✓ Time Series(D-day) + Non Time Series (director)
- PPT

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-

