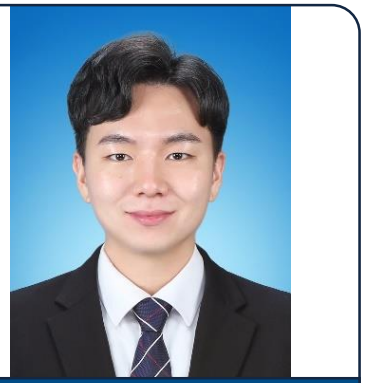
팀원소개

< D.N.A_Kyonggi >



심우열(팀장)

경기대학교 응용통계학과 4학년

dnrkd852@naver.com

분석 총괄 / 시계열예측 / 모델링

박현용

경기대학교 응용통계학과 4학년

etotmetotm@naver.com

크롤링 / 모델링

고수영

경기대학교 응용통계학과 3학년

rhtn2711@naver.com

자료조사 / 데이터구축 /

결과보고서 작성

박세영

경기대학교 응용통계학과 3학년

ptpdud1010@naver.com

자료조사 / 데이터구축 /

결과보고서 작성

제 8회 데이터 분석 경진대회

퓨쳐스 리그

KGU 경기대학교

스포츠투아이(주)

2020 빅콘테스트

목차

개요

분석배경 및 주제

분석 프로세스

데이터 탐색

기존연구조사

데이터 탐색

데이터 전처리 & 병합

파생변수

Train Set

최종변수

Test Set

시계열 분석 및 모델 개발

결과

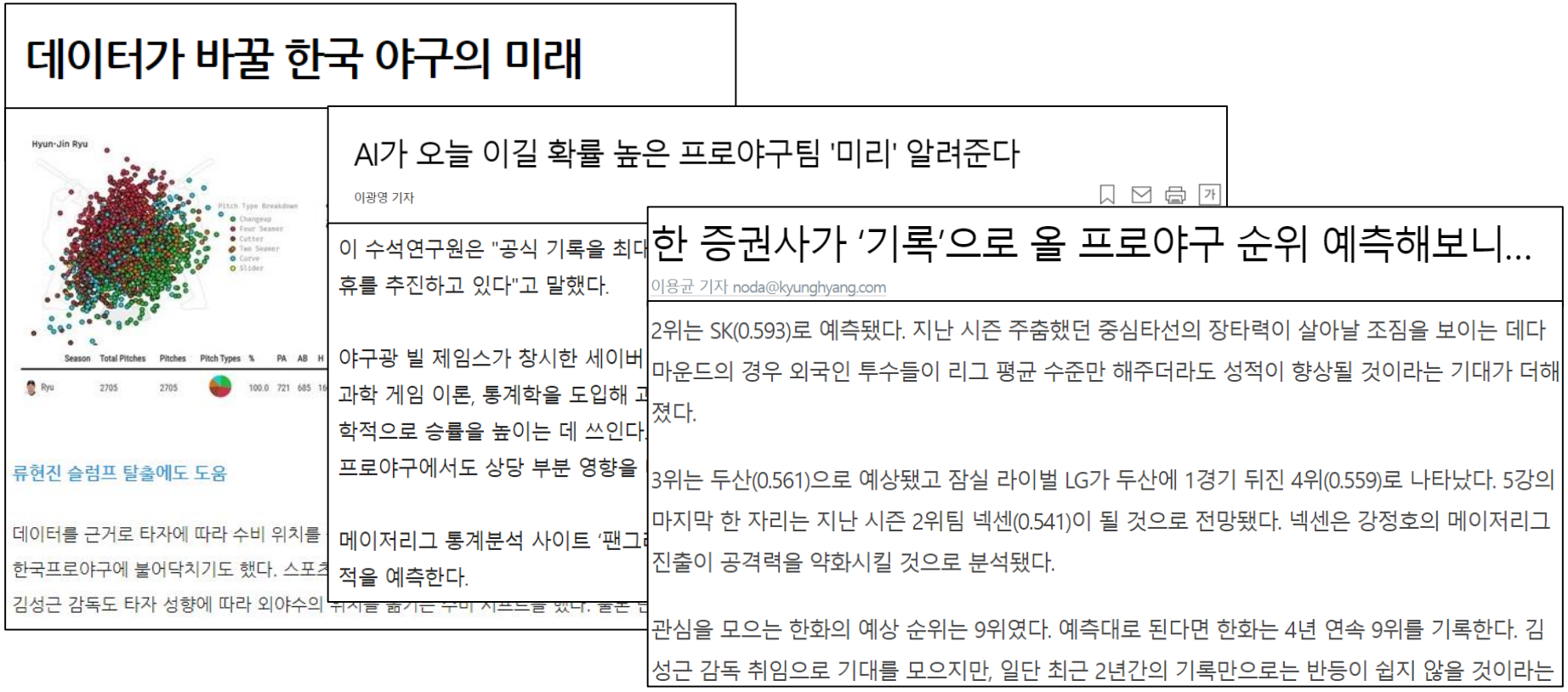
예측 결과

개요

- 분석배경 및 주제
- 분석 프로세스

분석배경 및 주제

▶ 야구 데이터 분석을 통해 팀 전략 계획을 수립 및 선수들 관리에 활용



데이터가 바꿀 한국 야구의 미래

Hyun-Jin Ryu

Pitch Type Breakdown
Changeup
Four Seamer
Cutter
Two Seamer
Curve
Slider

Season	Total Pitches	Pitches	Pitch Types	%	PA	AB	H
Ryu	2705	2705		100.0	721	685	16

류현진 슬럼프 탈출에도 도움

데이터를 근거로 타자에 따라 수비 위치를
한국프로야구에 불어닥치기도 했다. 스포츠
김성근 감독도 타자 성향에 따라 외야수의

AI가 오늘 이길 확률 높은 프로야구팀 '미리' 알려준다

이광영 기자

이 수석연구원은 "공식 기록을 최대
휴를 추진하고 있다"고 말했다.

야구광 빌 제임스가 창시한 세이버
과학 게임 이론, 통계학을 도입해 고
학적으로 승률을 높이는 데 쓰인다.
프로야구에서도 상당 부분 영향을

메이저리그 통계분석 사이트 '팬그
적을 예측한다.

한 증권사가 '기록'으로 올 프로야구 순위 예측해보니...

이용균 기자 noda@kyunghyang.com

2위는 SK(0.593)로 예측됐다. 지난 시즌 주춤했던 중심타선의 장타력이 살아날 조짐을 보이는 데다 마운드의 경우 외국인 투수들이 리그 평균 수준만 해주더라도 성적이 향상될 것이라는 기대가 더해졌다.

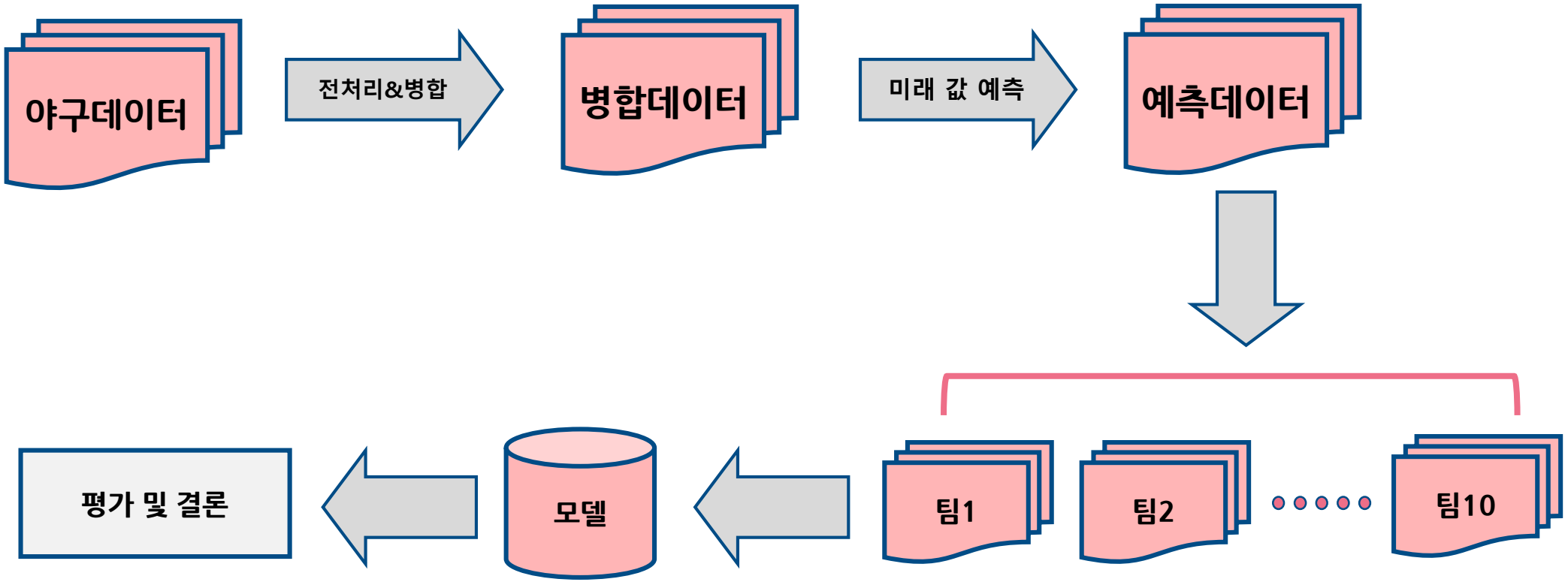
3위는 두산(0.561)으로 예상됐고 잠실 라이벌 LG가 두산에 1경기 뒤진 4위(0.559)로 나타났다. 5강의 마지막 한 자리는 지난 시즌 2위팀 넥센(0.541)이 될 것으로 전망됐다. 넥센은 강정호의 메이저리그 진출이 공격력을 약화시킬 것으로 분석됐다.

관심을 모으는 한화의 예상 순위는 9위였다. 예측대로 된다면 한화는 4년 연속 9위를 기록한다. 김성근 감독 취임으로 기대를 모으지만, 일단 최근 2년간의 기록만으로는 반등이 쉽지 않을 것이라는

야구데이터를 활용 하여, 정규 시즌 잔여 경기에 대한 각 팀별 승률, 타율 및 방어율 예측

분석 프로세스

▶ 전체적인 분석 방향 및 절차 설명



데이터탐색

- 기존 연구 조사
- 데이터탐색

기존 연구 조사

▶ 야구데이터 분석을 통한 승부예측 사례

논문	논문 저자	내용
인공신경망을 이용한 KBO 프로야구 승부예측 연구	노언석	프로야구경기의 승패 예측을 위해 선수들이 기록한 날짜별 데이터를 기반으로 인공신경망을 이용하여 경기를 예측하는 모델을 제시 (선발투수의 세부 기록과 나머지 투수들의 기록을 분리하여 적용)
데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구	오윤학 김 한 윤재섭 이종석	2013년도 시즌 국내 프로야구 팀과 선수들의 누적데이터를 통해서 다양한 데이터 마이닝 기법으로 다음 경기의 승패를 예측
기계학습 및 시뮬레이션을 이용한 프로야구 경기 예측 및 시뮬레이션 게임 시스템 개발	정태충	기계 학습 기법을 이용하여 프로야구 경기와 관련된 여러 가지 정보를 예측하여 볼 수 있는 시스템을 생성하고 야구 관련 정보들을 전산화함

✓ 개인 선수의 데이터 정제를 통한 변수 생성 결정
✓ 개인 선수의 세부 기록인 세이버메트릭스 활용 결정
✓ 개인선수 누적 데이터를 통한 평가지표 활용 결정

데이터 탐색

▶ 기존 제공 데이터 구조 확인

팀
- 2016~2020
- 팀코드 설명 데이터
- 제공 변수

: 팀명, 팀코드

경기
- 2016.04.01~2020.07.19
- 날짜 별 경기정보
- 제공변수

: 게임키, 일자, 원정팀코드, 홈팀코드, 더블헤더코드, 요일, 구장

선수
- 2016.04.01~2020.07.19
- 시즌 별 선수 기본 데이터
- 제공변수

: 시즌, 선수 코드, 선수명, 팀코드, 포지션, 나이, 연봉

등록선수
- 2016.04.01~2020.07.19
- 날짜 별 선수 기본 데이터
- 제공변수

: 일자, 팀코드, 선수코드, 등록말소

개인타자
- 2016.04.01~2020.07.19
- 경기 별 개인타자 기록 총합
- 제공변수

: 게임키, 일자, 팀코드, 상대팀코드, 더블헤더코드, 초말, 선수코드, 선발, 타순, 타자, 타수, 타점, 득점 안타, 등

팀타자
- 2016.04.01~2020.07.19
- 경기 별 팀타자 기록 총합
- 제공변수

: 게임키, 일자, 팀코드, 상대팀 코드, 더블헤더 코드, 초말, 타자, 타수, 득점, 안타, 2루타, 3루타 등

개인투수
- 2016.04.01~2020.07.19
- 경기 별 개인투수 기록 총합
- 제공변수

: 게임키, 일자, 팀코드, 상대팀코드, 더블헤더코드, 초말, 선수코드, 선발, 구원, 완투, 종료, 결과, 홀드 등

팀투수
- 2016.04.01~2020.07.19
- 경기 별 팀투수 기록 총합
- 제공변수

: 게임키, 일자, 팀코드, 상대팀 코드, 더블헤더 코드, 초말, 완투, 결과, 홀드, 이닝 *3, 투구수, 타자, 타수, 안타 등

데이터 탐색

▶ 데이터 분석 방향성

01 팀데이터가 각 팀의 경기력 대표성 충분하지 않다고 판단
▶ **개인선수 데이터는 승패예측, 팀데이터는 팀타율 및 방어율 예측 주요데이터로 활용**

02 세이버메트릭스가 산발적인 개인지표를 종합한다고 파악
▶ **세이버메트릭스 변수 추가**

03 팀 경기력의 흐름 파악 필요성 인지
▶ **개별 경기지표 외에 누적지표 추가생성**

누적지표? 경기지표?

누적지표 : 날이 거듭함에 따른 개인 기록 지표의 누적합을 구하여 세이버메트릭스 생성

경기지표 : 해당 날짜의 경기 기록만을 이용하여 세이버메트릭스 생성

04 7월 19일 이후의 기록들에 대해 예측할 필요성
▶ **통계적 기반의 시계열 분석 모형 ARIMA추정을 통한 미래 시점 예측**

05 한 경기의 승/패 예측을 통해 최종 승률을 예측
▶ **전 시점의 영향력과 여러 변수들을 고려할 수 있는 딥러닝 단방향 모델 LSTM 사용**

데이터 전처리 & 병합

- 파생변수
- Train Set
- 최종변수
- Test Set

파생변수

▶ 세이버메트릭스 : 야구 통계분석 방법

변수명	설명	수식
SECA	타율측정에서 제외되는 볼넷, 사구, 도루의 가치를 고려하는 지수	$\frac{\text{2루타}+2\times\text{3루타}+3\times\text{홈런}+\text{볼넷}+\text{도루}+\text{도루실패}}{\text{타수}}$
BABIP	타자가 친 공이 페어지역 안에 떨어진 경우만 나타내는 지수	$\frac{\text{안타}-\text{홈런}}{\text{타수}-\text{삼진}-\text{홈런}+\text{희생플라이}}$
ISO	순수하게 장타력만을 평가하기는 어려운 부분 존재, 순장타율은 SLG의 계산식에서 분자의 계수를 하나씩 빼 준 것과 동일한 효과	$\frac{\text{2루타}+2\times\text{3루타}+3\times\text{홈런}}{\text{타수}}$
GPA	OPS의 단점을 보완하여 출루율에 1.8 가중치 부여, GPA는 OPS보다 실제 득점과의 상관관계가 더 높고, 계산하기 쉬움	$\frac{1.8\times\text{OPS}}{4}$ * $\text{OPS}=\frac{\text{안타수}+\text{볼넷}+\text{사구}}{\text{타석}}+\frac{\text{1루타}+2\times\text{2루타}+3\times\text{3루타}+4\times\text{홈런}}{\text{타수}}$

파생변수

▶ 세이버메트릭스 : 야구 통계분석 방법

변수명	설명	수식
XR	팀 득점의 공헌도 지표	$1루타\times0.5+2루타\times0.72+3루타\times1.04+홈런\times1.44 +(볼넷+사구+고의볼넷)\times0.34+고의볼넷\times0.25 +도루\times0.18-도루실패\times0.32 -(타수-안타수-삼진)\times0.09-삼진\times0.098-병살타\times0.37 +희비\times0.37+희타\times0.04$
wOBA	득점과 가장 상관관계가 높은 지표이며 타자의 타석당 득점 기대치, 득점 공헌도, 출루가 득점이 될 확률	$\frac{0.72(볼넷-고의볼넷)+0.75\times사구+0.92\times실책 +0.90\times1루타+1.24\times2루타+1.56\times3루타+1.95\times홈런}{타석-고의볼넷}$
EOBP	안타로 인한 출루의 효과를 제거하고 사사구로 인한 출루의 비율만 계산	$\frac{안타수+볼넷+사구}{타수+볼넷+사구+희비}$

Train Set

▶ 개인선수 데이터 분류 기준 설명

개인타자 DATA

교체선수에 대한 고찰

- 같은 경기에 중복되는 타순번호 선수 기록이 존재
- 교체선수는 대부분 타수 및 타석이 0 인 선수 데이터
  - ▶ 세이버메트릭스 지표에 문제점을 발생 (결측, 무한)
- 타순번호가 같으면 같은 선수라고 인식하여 처리

클린업트리오와 그 외

- 주력선수와 그 외 선수들을 분리하여 팀의 경기력을 파악
- 각 그룹의 기록을 평균값으로 데이터 정제

* 클린업트리오
: 팀에서 주력 타자라고 할 수 있는 타순 번호 3,4,5번의 타자들을 일컫는다.

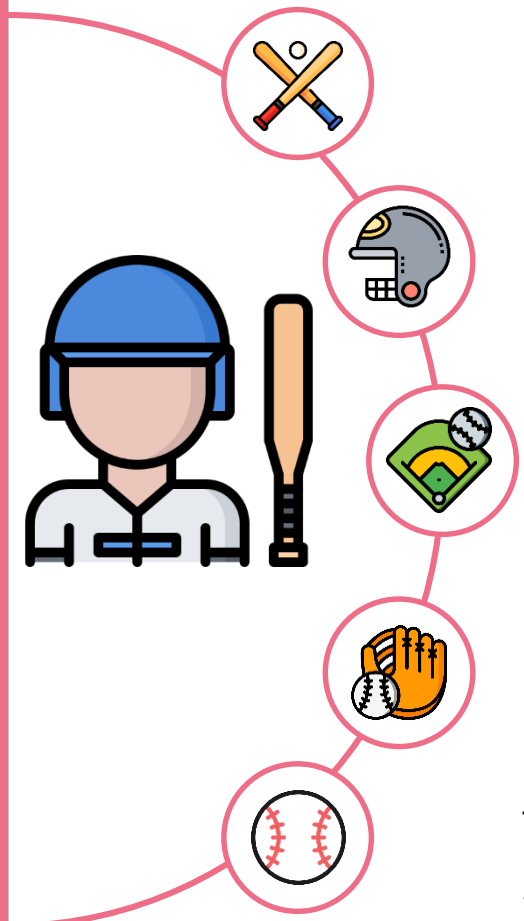
개인투수 DATA

선발투수

- 선발투수의 수비력이 팀의 승패 결과의 중요요인
- 선발투수는 기록 그대로를 활용
- 구원투수는 평균값으로 데이터 정제

Train Set

▶ Train Set 구축 기준 및 세부설명



데이터 행 구성기준 변경

기존의 데이터 : 홈팀기준으로 홈팀 지표만 기록

변경 데이터 : 팀을 기준으로 상대팀의 지표까지 포함하여 기록 (홈팀 여부 변수 추가)

세이버메트릭스

개별 기록들에 대한 정보를 내포한 세이버메트릭스 지표만 예측에 사용

선수들의 컨디션 파악

요일변수 : 해당팀의 전날 휴무에 따른 경기력 파악

더블헤더 : 연속 경기에 따른 경기력 파악

누적 지표 변수 생성

날을 거듭함에 따른 누적기록을 구하여 팀 전체의 누적지표 변수 생성

Train Set 팀별 분리

승패 예측 모델의 효율성을 위하여 분리

Train Set

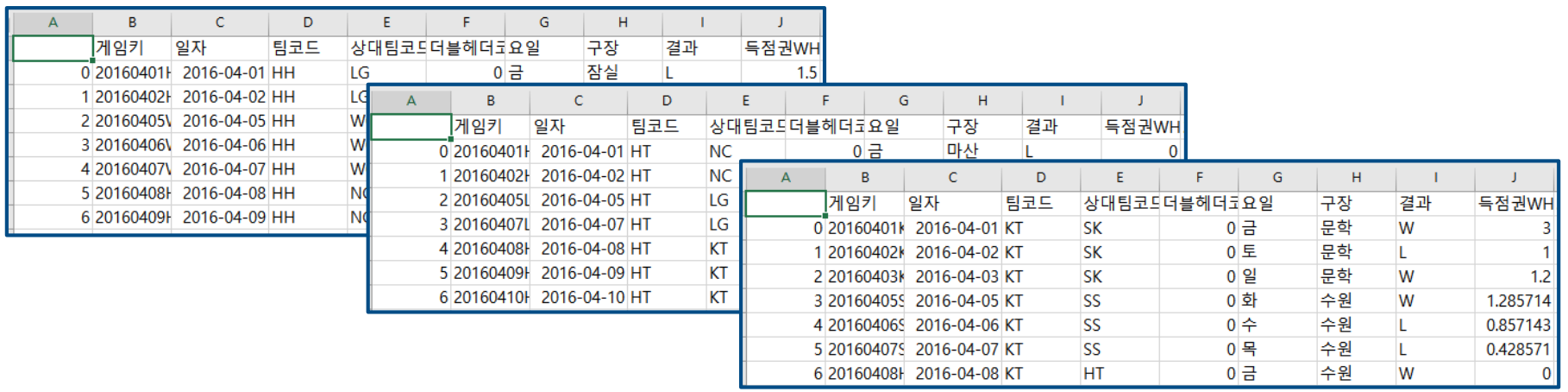
▶ TrainSet 실데이터 개형

게임키	일자	팀코드	상대팀코드	더블헤더코드	요일	구장	결과
20160401HHLG0	2016-04-01	LG	HH	0	금	잠실	W
20160401HTNC0	2016-04-01	NC	HT	0	금	마산	W
20160401KTSK0	2016-04-01	SK	KT	0	금	문학	L
20160401LTWO0	2016-04-01	WO	LT	0	금	고척	L
20160401OBSS0	2016-04-01	SS	OB	0	금	대구	L
20160401HTNC0	2016-04-01	HT	NC	0	금	마산	L
20160401KTSK0	2016-04-01	KT	SK	0	금	문학	W
20160401LTWO0	2016-04-01	LT	WO	0	금	고척	W

…

RCa_etc	XRa_etc	wOBAa_etc	ISOa_etc	EOBPa_etc	home
-0.646830065	0.998	0.25	0	0.077586207	1
-1.215589744	2.14	0.247307692	0.153846154	0.064102564	1
-0.156880658	4.92	0.371111111	0.222222222	0.025641026	1
0.023982906	3.638	0.358461538	0.076923077	0.051282051	1
-0.410044444	1.214	0.2028	0	0.05952381	1
-0.031944444	3.554	0.40375	0.166666667	0.128146453	0
-1.433666667	5.018	0.47625	0.291666667	0	0
-0.076111111	2.498	0.33	0.041666667	0.083333333	0

Rows : 6400
Columns : 89



…

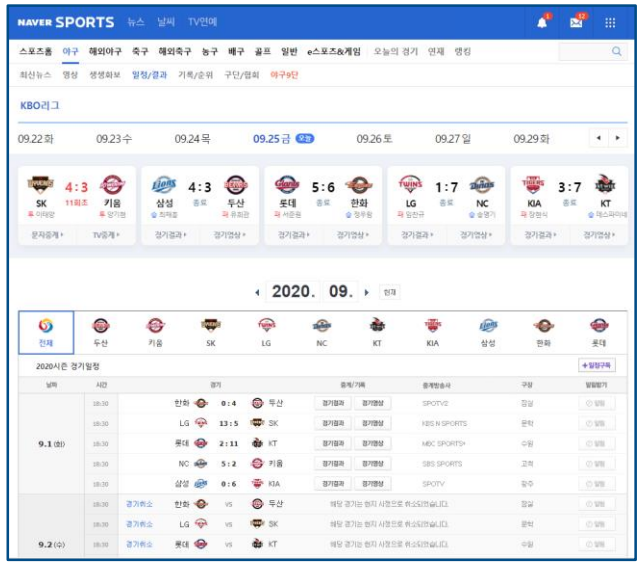
10개 팀으로 분리하여
Train Set 구축

최종변수

▶ 한 행에 하나의 경기로 홈팀과 원정팀 데이터 두 군데에 똑같은 경기가 기록되어 있음

Test Set

▶ 크롤링 및 추가 데이터를 통한 Test Set 구축



2020 KBO 정규시즌 잔여경기

일자	요일	AWAY	HOME	구장
9/29	화	KIA	키움	고척
9/29	화	KT	삼성	대구
9/29	화	롯데	LG	잠실
9/29	화	두산	한화	대전
9/29	화	SK	NC	창원
9/30	수	KIA	키움	고척
9/30	수	KT	삼성	대구
9/30	수	롯데	LG	잠실
9/30	수	두산	한화	대전
9/30	수	SK	NC	창원
10/1	목	KIA	키움	고척
10/1	목	KT	삼성	대구
10/1	목	롯데	LG	잠실
10/1	목	두산	한화	대전

362	20200929	HT	WO	화	고척	0
363	20200929	KT	SS	화	대구	0
364	20200929	LT	LG	화	잠실	0
365	20200929	OB	HH	화	대전	0
366	20200929	SK	NC	화	창원	0
367	20200930	HT	WO	수	고척	0
368	20200930	KT	SS	수	대구	0
369	20200930	LT	LG	수	잠실	0
370	20200930	OB	HH	수	대전	0
371	20200930	SK	NC	수	창원	0
372	20201001	HT	WO	목	고척	0
373	20201001	KT	SS	목	대구	0
374	20201001	LT	LG	목	잠실	0
375	20201001	OB	HH	목	대전	0
376	20201001	SK	NC	목	창원	0
377	20201002	HH	LT	금	사직	0
378	20201002	HT	OB	금	잠실	0
379	20201002	LG	KT	금	수원	0
380	20201002	SS	NC	금	창원	0
381	20201002	WO	SK	금	문학	0
382	20201003	HH	LT	토	사직	0

NAVER SPORTS 야구 페이지 크롤링 / 빅콘 추가 데이터를 통해 미제공 날짜 경기데이터 추출

▶ Test Set으로 활용하여 예측 정확도 파악

시계열 분석 및 모델 개발

- ARIMA
- LSMT

ARIMA

7월 19일 이후 기록되지 않은 세이버메트릭스/팀타율/팀방어율 지표를 시계열 분석을 통해 예측

- ARIMA모델은 시계열 분석 기법의 한 종류로, 과거의 관측값과 오차를 사용해서 현재의 시계열 값을 설명
- 팀별 경기에 대해 Timestep=1 인 일별 경기로 변환
- ARIMA(p,q,d)의 p,q,d의 최적의 모수(최소 AIC) 추정 실시
- 7월 19일까지의 데이터를 Test(최근 31경기), Train(나머지 경기)으로 나누어 검증
- 년도마다 팀의 성적이 크게 차이나는 경우도 있기 때문에 팀방어율과 팀타율은 모든 년도를 고려한 데이터와 최근 동향인 19/20년도만을 사용한 데이터를 RMSE를 통해 비교

ARIMA

▶ RMSE비교를 통해 최종 예측을 진행

✓ 팀타율

(RMSE는 소수 넷째자리에서 반올림)

팀명	NC	LG	키움	삼성	KIA	KT	롯데	두산	한화	SK
16-20 RMSE	0.003	0.006	0.009	0.008	0.0003	0.008	0.003	0.009	0.004	0.013
19-20 RMSE	0.010	0.008	0.012	0.016	0.018	0.018	0.011	0.011	0.017	0.003

✓ 팀방어율

(RMSE는 소수 넷째자리에서 반올림)

팀명	NC	LG	키움	삼성	KIA	KT	롯데	두산	한화	SK
16-20 RMSE	0.574	0.481	0.222	0.406	0.332	0.187	0.315	0.296	0.230	0.225
19-20 RMSE	0.772	0.290	0.310	0.267	0.644	0.273	0.272	0.305	0.826	0.580

SK는 과거 기록에 비해 올 시즌 타격 성적이 매우 안좋음

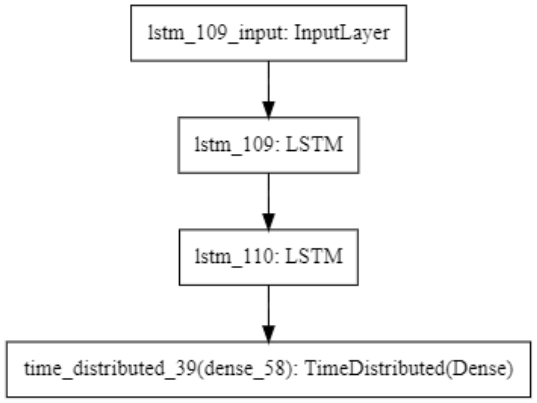
▶ SK 팀타율은 19-20년도 데이터만 사용

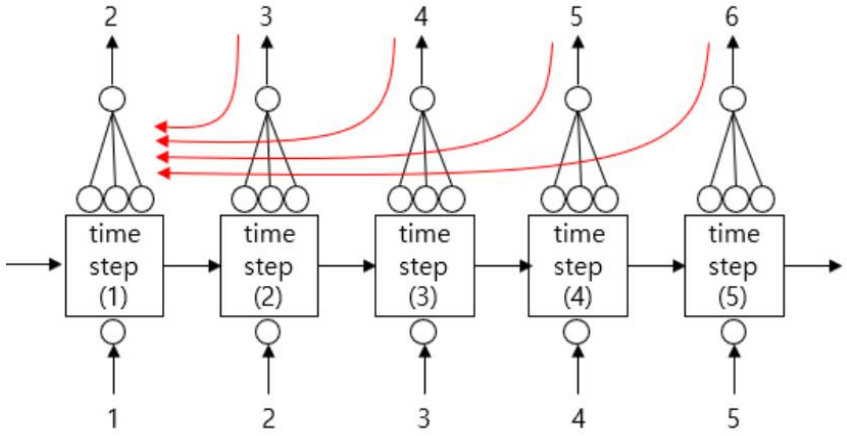
LSTM

LSTM을 사용하여 경기의 결과를 예측

LSTM

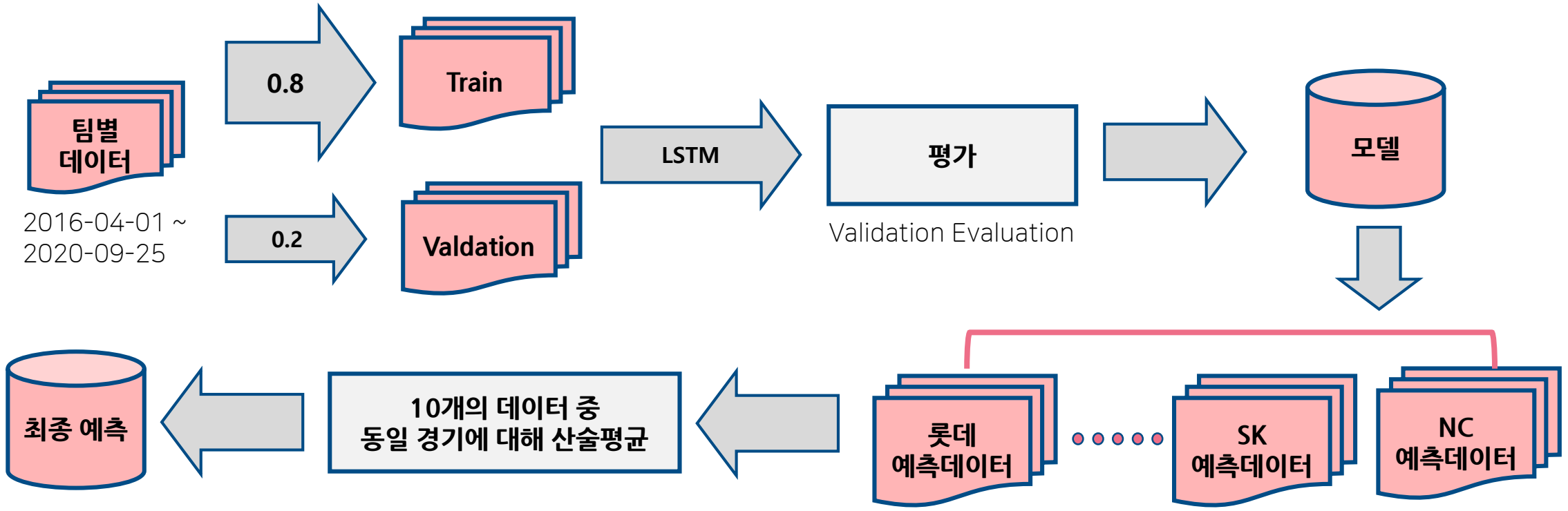
- RNN의 long term 면에서 경사를 소실하는 문제를 개선하기 위한 대표적인 RNN 방법
- 직전 정보만 참고하는 것이 아닌, 그 전 정보를 고려할 수 있어 흐름이 중요한 야구에 적합하다고 판다
- ARIMA모델을 통해 예측된 세이버메트릭스 정보들과 기본 경기정보들을 사용하여 long term의 데이터 Dependency하는 방식으로 진행
- 단층-단방향 &Many-to-Many 방식의 LSTM 모델을 적용하여 각 스텝에서 cost를 계산하고 하위스텝으로 오류를 전파





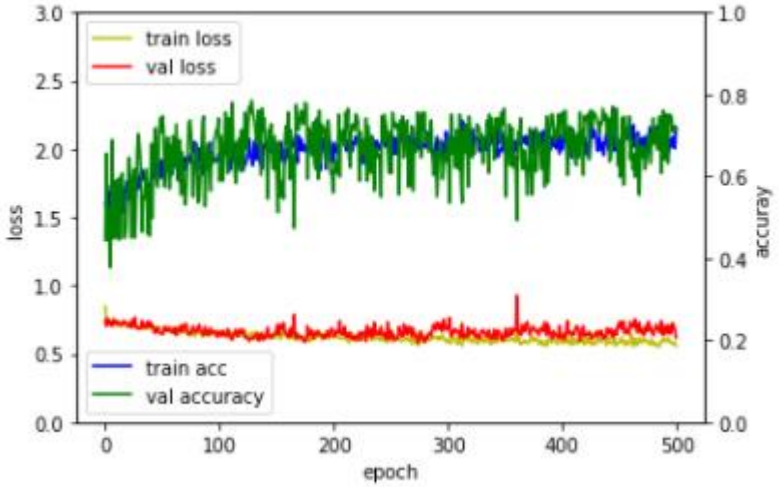
LSTM

▶ LSTM을 통한 승패 예측 과정

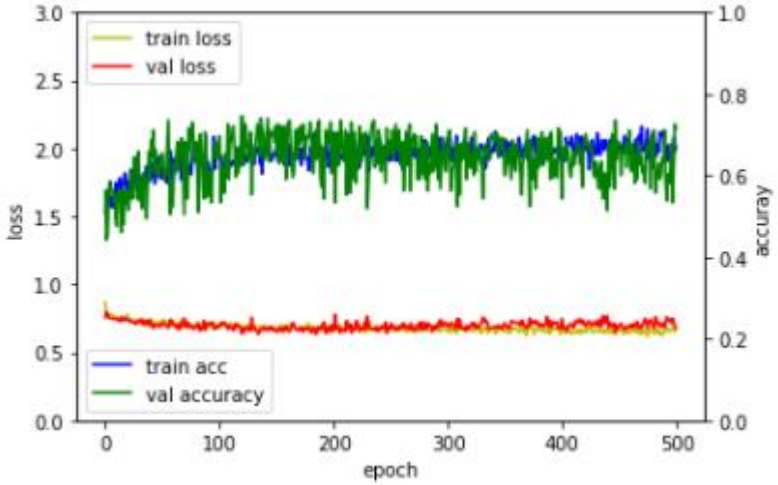


LSTM

▶ 팀별 데이터 LSTM Accuracy



SK (epoch = 500, batch_size=32)



삼성 (epoch = 500, batch_size=32)

✓ Validation evaluation

팀명	NC	LG	키움	삼성	기아	KT	롯데	두산	한화	SK
Accuracy	66.67 %	62.50 %	61.42 %	71.63 %	64.25 %	54.32 %	59.62 %	67.67 %	69.71 %	71.77 %

최종 예측

▶ 최종 승패 확률

일자	홈팀코드	원정팀코드	home_W	away_W	draw_AM
20200926	KT	LG	0.360340	0.628749	0.010911
20200926	HT	LT	0.569746	0.422822	0.007433
20200926	HH	NC	0.317002	0.669385	0.013613
20200926	SS	SK	0.426467	0.562085	0.011448
20200926	OB	WO	0.616978	0.374708	0.008315
20200927	KT	LG	0.382438	0.605407	0.012155
20200927	HT	LT	0.574088	0.418535	0.007377
20200927	HH	NC	0.352424	0.633987	0.013589
20200927	SS	SK	0.455816	0.532374	0.011810
20200927	OB	WO	0.545956	0.444790	0.009254
20200927	OB	WO	0.532965	0.457831	0.009204

결과
LG 승
HT 승
NC 승
SK 승
OB 승
LG 승
HT 승
NC 승
SK 승
OB 승
OB 승

산술평균으로 나온 최종 확률로 각 경기의 결과를 예측하고 각 팀의 승률을 구함

결과

- 예측결과

예측결과

팀명	승률	타율	방어율
NC	59.71223021582733 %	0.279056436497983	4.3924443404908065
LG	65.18518518518519 %	0.28658636928041104	4.175828373219966
키움	60.99290780141844 %	0.2808069366941738	4.7360896342994545
삼성	45.25547445255474 %	0.2709702710278604	5.029784779948445
기아	50.74626865671642 %	0.27754615324803267	4.78632405579775
KT	50.73529411764706 %	0.2762898694092557	5.150387975703577
롯데	47.01492537313433 %	0.2770736249082532	5.146300739461721
두산	56.61764705882353 %	0.287739818346788	4.364229372771637
한화	27.536231884057973 %	0.23484778358090252	5.070280844605222
SK	29.71014492753623 %	0.24625401705138966	5.106417099227543