「2020 빅콘테스트」 데이터 분석 계획서

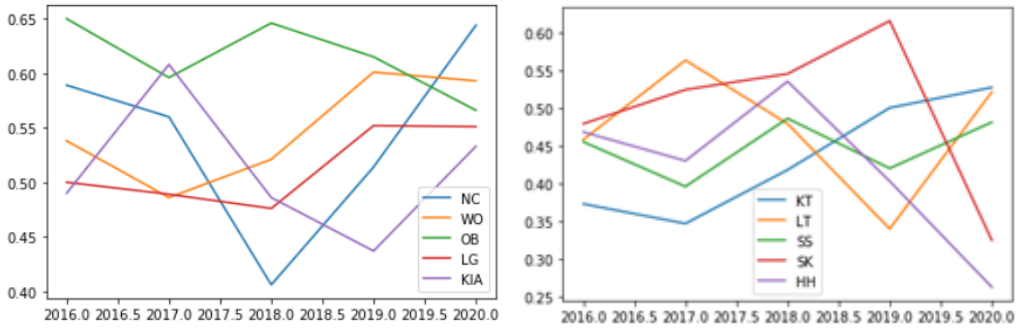
참가분야	☐ 혁신아이디어분야 ☐ 데이터분석분야		
세부분야	☐ 퓨처스리그 ☐ 챔피언리그		
개인/팀 여부	☐ 개인 ☐ 팀(구성원 4명)	개인/팀 명	SVT
대표ID	harin_1029@naver.com		

※5장 내외로 목차는 준수하여 자유롭게 작성

분석 주제 명	KBO 정규시즌 잔여기간 팀 별 **승률, 타율 및 방어율(평균자책점)** 예측 "불꽃 튀는 순위 경쟁을 펼치는 **마지막 한 달**, 내가 좋아하는 KBO 팀은 어떤 성적을 낼까?"
분석 배경	**※ 선택한 주제(분야)의 데이터 분석 제안에 대한 배경을 기술해주십시오** 프로 야구 경기 결과에 영향을 주는 요인으로는 팀과 선수의 경기 능력을 비롯해 부상, 날씨, 컨디션 등 여러 변수들이 있다. 과거에는 주로 스포츠(야구) 업계에서 오랫동안 일을 한 사람들이 자신들의 경험을 바탕으로 선수들을 영입하고 경기 전략을 수립하였다면, 최근에는 **빅데이터, 머신러닝, 인공지능 기법**들을 통해 과거 데이터를 통계적으로 분석하여 모델을 만들어 최적의 팀을 꾸리고 있다. 또한, 관람자가 단순히 경기를 관람하는 수동적인 형태에서 탈피하여 경기결과를 예측하고 배팅을 하는 등의 참여하는 형태로 진화하고 있다. 일방적으로 관람자에게 경기 내용에 관련된 정보만 제공해 주던 TV 중계의 기능 역시 혁신적인 IT 기술의 등장에 따라 경기 내용을 통계적으로 분석하고 결과를 예측하는 수준으로 발전하고 있다. 이에 따라 국내외에서는 다양한 기계학습 기법을 이용하여 야구 경기 결과를 예측하기 위한 연구가 활발하게 진행되고 있다. 이러한 흐름에 따라 본 팀은 KBO 리그의 과거 야구 데이터를 토대로 데이터 분석을 진행하여 **승률, 방어율, 타율에 영향을 주는 요인**을 탐색하고 정규시즌 잔여 경기에 대한 각 팀 별 승률, 타율 및 방어율(평균자책점)을 보다 정확하게 **예측할 수 있는 모형**을 구축하고자 한다.
분석 내용 요약	**※ 분석 내용을 200자 내외로 간략하게 요약 기술해 주십시오.** 2016-2019 시즌 각 720경기와 2020 시즌 320경기의 팀 통산 전적 데이터, 연도별 전적 데이터, 팀 및 선수 기록 데이터 등을 활용하여 **2020 시즌 '마지막 20일(약 100경기)'의 팀 별 승률, 타율 및 방어율(평균자책점)** 예측하고자 한다. "승률 = 승수 / (승수+패수), 타율 = 안타 수 / 타수, 방어율 = (자책점*9) / 이닝 수"이기 때문에 본 팀은 시즌 잔여 경기 별 득점, 실점, 안타 수, 타수, 자책점, 이닝 수를 예측하는 모형을 구축하고자 한다.

분석 방법 및 계획

※ 대회기간 동안 선택한 주제(분야)에 대한 분석 방향에 대하여 기술해주십시오.
- 분석에 활용되는 추가데이터(출처 기재)
- 분석에 적용·활용할 통계·분석 기법, 방법론
- 분석 결과에 대한 시각화 방법 등



"2018년 8월 11일 기준 KBO 정규시즌 1~10위 팀의 2016-2020 정규시즌 승률" 그래프는 각 팀의 경기 능력이 시즌 별로 매우 다름을 보여준다. 선수의 부상, 감독 및 구단의 변화, 선수 트레이드, 스프링 캠프, FA 등 아주 다양한 이유가 있다.

➔ **시즌(년도) 별로 관여하지 않으면서 모든 시즌에서 타겟변수와 공통적인 특징을 가지는 입력변수를 활용**할 것이다.

스포츠 경기는 최근 경기 추세의 영향을 가장 크게 받을 것이라고 생각한다. 2020 시즌의 주어진 데이터와 예측해야 할 경기 사이에는 약 300 경기의 공백이 존재한다.

➔ **제공받은 데이터 외의 일정 경기(20200720-20200928)를 추가로 수집**할 것이다.

분석 주제는 '불꽃 튀는 순위 경쟁을 펼치는 마지막 한 달'이라는 제한된 기간의 승률, 방어율, 타율을 예측하는 것이다. 승률은 승수 / (승수+패수)이다. 결국 승수와 패수를 알기 위해서는 각 경기에서의 승/패 결과를 알아야 한다. 경기의 승/패는 득점>실점이면 승, 득점<실점이면 패, 득점=실점이면 무승부라는 변하지 않는 규칙이 존재한다.

➔ 좀 더 정교한 예측을 위해, 각 경기 별 **득점과 실점을 예측하는 모형을 구축**할 것이다. 이와 마찬가지 이유로 방어율은 **자책점과 이닝 수**를, 타율은 **안타 수와 타수를 예측하는 모형을 구축**할 것이다.

팀 데이터 셋의 사용은 **"1. 최근 경기 추세 반영 2. 상대 팀과의 전적 반영"** 이 두 가지 이유로, 시즌 별 경기를 아래와 같이 7묶음으로 나눠 한 묶음(100경기) 예측을 위한 입력변수를 직전 200경기 혹은 100경기 데이터를 이용하여 입력변수를 만들 것이다. 구체적으로, 최근 경기와 상대 팀과의 전적 이 두 가지의 가중치를 둔 가중평균을 사용할 것이다. 개인 데이터 셋의 사용은 **"1. 개인 선수(선발)의 변화 및 중요성 2. 부상 및 등록/말소"** 이 두 가지 이유로 앞 묶음에서 해당 선발 선수의 경기의 데이터를 이용하여 입력변수를 만들고, 등록/말소에서 "말소 후 10일 간 경기 출전 불가"와 선발 로테이션을 이용하여 선발을 예측할 것이다.

➔ 시즌 별 500경기씩(2020년도는 400경기) 총 2400경기, 즉 4800개의 데이터를 이용하여 분석 데이터 셋을 형성하여 모델 학습과 테스트를 거쳐 최종적으로 2020 시즌 마지막 100경기를 예측할 것이다.

2016	120	100	100	100	100	100	100
2017	120	100	100	100	100	100	100
2018	120	100	100	100	100	100	100

2019	120	100	100	100	100	100	100
2020	120	100	100	100	100	100	**100**

제공받은 변수(득점, 안타 수, 타율 등)들뿐만 아니라 야구를 통계학적/수학적으로 분석하는 방법론인 **'세이버메트릭스'의 지표를 적극적으로 활용**할 것이다. 또한, 외부변수로는 선수들의 컨디션에 영향을 줄 수 있는 **날씨와 이동 거리**를 활용할 것이다. 이동 거리는 '네이버 지도 길 찾기 추천 거리'을 이용하여 구한 후 범주형 변수로 변환하여 사용할 것이다. 날씨는 기상청을 통해 얻은 실제 날씨 값을 넣어 시각화 해본 후, 데이터 분석에는 중/장기 예보 값을 사용할 것이다.

분석 기법

본 팀은 머신러닝에서 트리 기반 XGBoost, LightGBM과 딥러닝 기반 LSTM, GRU를 사용하고 Random Search와 Bayesian Optimization을 이용하여 하이퍼 파라미터를 최적화할 것이다.

① **XGBoost**

: 트리 기반의 앙상블 학습에서 가장 각광받고 있는 알고리즘 중 하나로, 일반적으로 뛰어난 예측 성능을 발휘하고 병렬 수행 및 다양한 기능으로 GBM에 비해 빠른 수행 성능을 보장한다. 또한, 자체에 과적합 규제 기능이 있고 tree pruning으로 더 이상 긍정 이득이 없는 분할을 가지치기 해서 분할 수를 더 줄이는 추가적인 장점을 가지고 있다. 반복 수행 시마다 내부적으로 학습 데이터 세트와 평가 데이터 세트에 대한 교차 검증을 수행해 최적화된 반복 수행 횟수를 가질 수 있다. 지정된 반복 횟수가 아니라 교차 검증을 통해 평가 데이터 세트의 평가 값이 최적화 되면 반복을 중간에 멈출 수 있는 조기 중단이 있고 결손값을 자체 처리할 수 있는 기능을 가지고 있다.

② **LightGBM**

: XGBoost와 함께 부스팅 계열 알고리즘에서 가장 각광받고 있는 알고리즘이다. XGBoost의 단점을 보완한 알고리즘으로, XGBoost에 비해 학습 및 예측 수행 시간이 더 빠르고 메모리 사용량이 더 적다. 또한, 카테고리형 피처의 자동 변환과 최적 분할이 가능하다.

③ **LSTM**

: RNN에서 발생하는 vanishing/exploding gradient problem을 해결하기 위해 제안된 알고리즘으로, 기본적인 RNN의 구조에 memory cell이 은닉층 뉴런에 추가된 것을 볼 수 있다. Memory cell은 추가된 forget gate, input gate, output gate를 의미한다. Forget gate는 과거의 정보를 어느 정도 기억할지 결정하고, 과거의 정보와 현재 데이터를 입력받아 sigmoid를 취한 뒤에 그 값을 과거의 정보에 곱한다. 따라서, sigmoid의 출력이 0일 경우에는 과거의 정보를 완전히 잊고, 1일 경우에는 과거의 정보를 온전히 보존한다. Input gate는 현재의 정보를 기억하기 위해 만들어졌다. 과거의 정보와 현재 데이터를 입력 받아 sigmoid와 tanh 함수를 기반으로 현재 정보에 대한 보존량을 결정한다. Output gate는 과거의 정보와 현재 데이터를 이용하여 뉴런의 출력을 결정한다.

④ **GRU**

: LSTM의 장기 의존성 문제에 대한 해결책을 유지하면서, 은닉 상태를 업데이트하는 계산을 줄인 기법이다. 다시 말해서, 성능은 LSTM과 유사하면서 복잡했던 LSTM의 구조를 간단화시킨 것이다. LSTM의 forget과 input gate 역할을 하는 update gate와 별도의 reset gate로 구성되어 있다. 데이터 양이 적을 때는, 매개 변수의 양이 적은 GRU가 조금 더 낫고, 데이터 양이 더 많으면 LSTM이 더 낫다고 알려져 있다.

Hyper parameter 최적화 방법

① **Random Search**

: Grid Search에 비해 불필요한 반복 수행 횟수를 대폭 줄이면서, 동시에 정해진 간격

(grid) 사이에 위치한 값들에 대해서도 확률적으로 탐색이 가능하므로, 최적 hyper parameter 값을 더 빨리 찾을 수 있는 것으로 알려져 있다.

② **Bayesian Optimization**

: 불필요한 hyper parameter 반복 탐색을 줄여 보다 빠르게 최적 hyper parameter를 찾을 수 있는 최적화 방법이다. 알려지지 않은 목적 함수를 최대로 하는 최적해를 찾는 기법이다.

과적합 방지 및 성능 향상

① **Drop out**

: 인공 신경망의 뉴런을 확률적으로 사용하지 않음으로써 과적합을 방지하는 기법이다.

② **Ensemble**

: 다양한 종류의 여러 estimator를 결합하여 더 좋은 estimator를 만드는 것이다. 종류는 estimator들을 어떻게 결합할 것인지에 의해 결정된다. 대표적으로 배깅, 부스팅, 보팅, 스태킹이 있다.

분석 결과에 대한 시각화 방법

✓ Rader Chart를 이용한 팀 경기력 비교

✓ Word Cloud를 이용한 선발 로테이션

예시) HT와 HH의 선발 투수

✓ Line Chart를 이용한 팀 경기력 추세 확인

분석결과 활용 및 시사점	**※ 분석 결과에 대한 활용 방안, 적용대상, 결과 적용 시 기대효과 및 시사점 등에 대하여 기술해 주십시오.** '정규시즌 잔여기간 팀 별 승률, 타율, 방어율 예측'이라는 부분은 많은 팬들이 단순히 경기를 관람하는 것뿐 아니라 경기 결과를 예측하고 배팅을 하는 등의 참여하는 형태로 이루어질 수 있고, 일방적으로 관람자에게 경기 내용에 관련된 정보만 제공해주던 TV 중계 역시 경기 내용을 통계적으로 분석하고 결과를 예측하며 많은 팬들이 새롭고 다양한 관점으로 더욱 더 즐길 수 있다. 또한 이를 통한 콘텐츠가 더 많아질 수 있고, 궁극적으로는 이러한 것들이 산업 전반의 활성화를 이룰 수 있을 것이다. ➔ 본 팀은 Contents, Data, Analysis 이 세 가지를 모두 활용하여 **프로야구의 산업 활성화를 이끌 수 있는 아이디어**를 제안한다. **'YAnalysis SNS 대학생 서포터즈' 운영** KBO 대학생 객원마케터가 16기까지 진행되어 왔고, 올 해에는 KBO 퓨처스리그 제 1기 대학생 기자단이 생기는 등 야구 관련 대외활동이 KBO, 각 구단 등에서 조금씩 생겨나고 있다. 현재 프로야구는 2017년 최다 관중 수 840만 명을 기록한 후 매 년 줄어들고 있는 추세이다. 특히나 올 해인 2020 시즌은 코로나 상황으로 인해 직관 자체가 불가하게 되면서 오프라인 응원 문화가 자리 잡혀있는 프로야구에 굉장한 타격이 있을 것이라고 생각한다. 본 팀은 이러한 상황에 맞게 야구와 분석에 관심이 많고, SNS 및 방송 활동이 활발한 대학생들을 대상으로 서포터즈를 모집하여 프로야구의 산업 활성화를 이끌 수 있는 여러 활동들을 추천한다. ① 직접 예측한 경기 직관하기 ② '세이버메트릭스 지표' 파헤치기 ③ 프로야구의 이슈 알리기 ④ NC 팬과 OB 팬이 함께 하는, NC vs OB 본격 편파 중계 **팀 전력 분석 및 전략 수립**

※ 제출자료는 최종 출품작 평가 시 활용될 수 있음