0918 회의

2020년 9월 18일 금요일 오후 10:32

1. 모델링 코드 소개 + 피드백

1) LGBM

: 월 단위가 아닌 일 단위로 예측하는 것이 좋을 듯

: 추후, 클래스 생성 시 파라미터로 설정할 수 있도록 하기

: regression 사용 불가) 정규성 만족 X

2) LSTM

: 승률만 돌릴 수 있도록 구성

: softmax 사용 필요 ? <-> 팀별 승률 비교를 위하여 softmax 사용 x

2. EDA

1) 개인타자

: 배팅 순서에 따른 타율 확인, 선수 별 평균적 타순 배치 확인, 타순 대비 평균 득점권 타율 (이상치 존재), 30 경기 이상 출전 선수 평균 타율 분포, 선발 여부에 따른 타율, 상관관계 분석

: feature 추가 제안(경기별 상위 10% 선수들의 평균 타율) -> 기준을 조금 낮추어(ex. 상위30%) 활용하면 될 듯!

2) 개인투수

: 30 경기 이상 출전 투수 EDA, 완투 여부와 ERA 관계, 마지막에 던진 것과 ERA 관계, 득점권과의 상관관계 분석

-> 피드백

: 타자와 투수의 경기 수가 너무 차이남. 30경기보다는 10경기로 하는 것이 좋을 것 같음.

: 개인 투수는 ERA가 극명히 달라지는 문제점 존재. 사용 방법을 고민하는 것이 중요함.

: 상관관계가 너무 낮은 것 아닌지?

3) 팀타자

: 년도별 분석 진행

: 전체타율 분포. 평균 타율, 실책, 상관관계 분석, 정규성 확인 등

4) 팀투수

: 완투 여부 개수 확인(완투한 선수들의 ERA 평균 매우 낮음), 이상치 확인(보크), 투구수와 ERA 관계 확인 등

-> 피드백

: 운적인 요소가 포함된 변수 (BK) 삭제하기

3. 보고서 스토리라인 기획

1) LGBM

방향성 : ' 왜 lgbm을 쓸 수밖에 없는가'

회귀모형의 한계 (정규성 만족x) -> Tree계열 모델 사용할수밖에 없었음.

Cf. 보고서에서 lgbm의 분량을 줄이는 것이 좋을 듯

Lgbm 파라미터 소개방식으로

보고서에는 VAR을 중점적으로 기재하기

2) VAR

정상성만족 x -> 시계열 모델 써도 된다고 판단

+ 더 고민해보기

4. 다음 시간 회의

1) 역할 분담

상대오빠 : 모델링 코드 취합, 실험코드 작성

창건오빠 : 단변량 LSTM 코드 완성

동현오빠 : 보고서용 EDA

민영언니 : LGBM 승률

윤정 : 스토리라인 짜기(수상자 보고서 확인)

2) 승률 고민