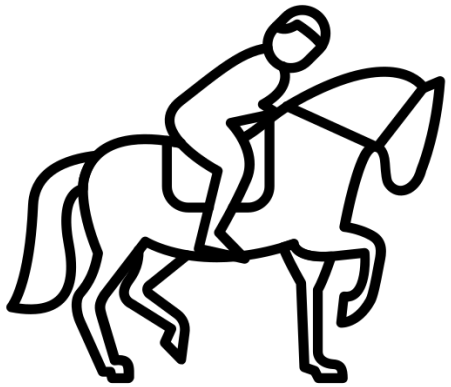
경마 경기 우승마 예측 프로젝트

경마 우승에 영향을 미치는 변수 탐색 및 우승마 예측 모델 구축



이름	학번	역할
권수현	2020122006	데이터 수집 및 결합 및 전처리, 선형회귀분석 모델링 및 디벨롭, 발표자료 제작
김채성	2020122037	데이터 수집 및 결합 및 전처리, 랜덤포레스트 모델링 및 선형회귀분석 디벨롭, 발표자료 제작, 발표
이준서	2020122055	데이터 수집 및 결합 및 전처리, 인공신경망 모델링 및 선형회귀분석 디벨롭, 발표자료 제작

주제 및 목표

“경마 경기 우승마 예측”

5마리 말 선택 후 해당 말 중 1, 2, 3등이 모두 포함되도록 예측

경마 우승에
영향을 미치는 변수?

경마 우승마 예측 모델

데이터 수집

선행연구 조사를 기반으로 하여 말, 경기, 기수, 조교사로 범주를 나누어 경마 예측에 사용할 데이터를 수집함

데이터 출처: 공공데이터포털(https://www.data.go.kr/index.do)

데이터 수집기간: 2019.01~2023.06 (2022.06까지의 데이터를 train data, 이후 데이터는 test data로 활용)

데이터셋	말 변수	경기 변수	기수 변수	조교사 변수
한국마사회_경주성적정보	마명, 마번, 연령, 성별, 마체중, 마체중 증감	경주일자, 경주명, 경주번호, 출주번호, 등급조건, 레이팅(등급), 경주거리, 부담중량	기수명, 기수번호	조교사명, 조교사 번호
한국마사회_일별훈련_상세정보	경기전2주간/1달간훈련시간총합			
한국마사회_경마시행당일_경주결과종합	최근1년간1착/2착비율, 국산외산구분		기수 연령, 기수 경력, 기수최근1년간1착/2착비율	조교사 경력, 조교사최근1년간1착/2착비율
한국마사회_마필진료_정보	최근1개월간진료/질병진단횟수			
한국마사회_경주로정보		주로, 함수율		

데이터 결합 및 전처리

변수 범주	Same unit
말 변수	마명, 마번
경기 변수	경주일자, 경주번호, 경기장
기수 변수	기수명, 기수 번호
조교사 변수	조교사명, 조교사 번호

결합

각 변수 범주별로 unique한 특성을 가지는 변수를 same unit으로 활용하여 데이터를 결합함

전처리

meet(경마장) 전처리

: 부산경남, 부경 등 범주 항목 통일 X → 서울/부산경남/제주 범주로 통일

jkNo(기수번호), trNo(조교사번호) 전처리

: 80482, '080482' 등 str과 int가 섞여서 존재 → 모두 str로 통일

파생변수 생성

y_speed(속력)

: 경주거리(rcDist) / 경주기록(rcTime)

말/기수/조교사 최근 1년간 1착/2착 비율

: (말/기수/조교사 최근 1년간 1착/2착 횟수) / (말/기수/조교사 최근 1년간 출전횟수)

데이터 요약

각 행은 특정 말의 특정 경기에서의 정보

	hrNo	rcDate	rcNo	age	rating	sex_수	sex_암	prdName_외산	trTime_14	recent_ill	...	wgHr_var	wgBudam	chulNo	rcDist	waterPercent	jkCareer	jkord1Rate	trCareer	trord1Rate	y_speed
0	38647	20190104	10	3	43	1	0	0	1380.0	0	...	-2.0	55.0	6	1600.0	3	12	0.037383	19.0	0.130435	15.779093
1	37077	20190104	5	4	0	1	0	0	660.0	0	...	5.0	56.5	5	1300.0	3	12	0.037383	19.0	0.090278	15.476190
2	38954	20190104	5	3	0	1	0	0	1320.0	0	...	-9.0	55.0	10	1300.0	3	8	0.128463	19.0	0.130435	15.494636
3	37279	20190104	5	4	0	1	0	0	1560.0	0	...	14.0	56.5	11	1300.0	3	14	0.000000	19.0	0.090278	15.550239
4	38935	20190104	5	3	0	1	0	0	1620.0	0	...	-3.0	56.0	3	1300.0	3	15	0.074074	19.0	0.085366	15.587530
5	37526	20190104	5	4	0	0	0	0	1860.0	0	...	-7.0	55.5	9	1300.0	3	21	0.181250	8.0	0.075472	15.738499
6	41595	20190104	4	3	39	1	0	1	1920.0	0	...	-3.0	52.0	1	1000.0	3	15	0.074074	9.0	0.125523	15.723270
7	37340	20190104	5	4	0	0	1	0	1500.0	0	...	2.0	54.5	7	1300.0	3	16	0.144681	18.0	0.000000	15.457788
8	31916	20190104	4	7	41	0	0	0	1740.0	0	...	-6.0	53.5	3	1000.0	3	12	0.037383	5.0	0.070248	15.822785
9	36605	20190104	4	5	36	1	0	1	1680.0	0	...	-4.0	52.0	8	1000.0	3	16	0.088305	1.0	0.000000	16.103060

말 변수

나이, 성별, 레이팅(등급), 국산/외산 등

경기 변수

출주번호, 함수율 등

기수 변수

최근1년간1착비율, 기수 경력 등

조교사 변수

최근1년간1착비율, 조교사 경력 등

속력

모델링 전략

한 말에 대해 여러 경기 기록 정보 포함

	hrNo	rcDate	rcNo	age	rating	sex_수	sex_암	prdName_외산	trTime_14	recent_ill	...	wgHr_var	wgBudam	chulNo	rcDist	waterPercent	jkCareer	jkord1Rate	trCareer	trord1Rate	y_speed
0	38647	20190104	10	3	43	1	0	0	1380.0	0	...	-2.0	55.0	6	1600.0	3	12	0.037383	19.0	0.130435	15.779093
1783	38647	20190125	10	3	52	1	0	0	12720.0	0	...	-3.0	53.5	12	1800.0	3	12	0.037383	19.0	0.130435	15.025042
4559	38647	20190301	9	3	52	1	0	0	11160.0	0	...	3.0	51.0	11	1800.0	3	12	0.037383	19.0	0.130435	15.228426
8437	38647	20190414	5	3	52	1	0	0	11220.0	0	...	6.0	57.0	5	1400.0	10	12	0.037383	19.0	0.130435	15.801354
9259	38647	20190426	9	3	51	1	0	0	9360.0	0	...	-9.0	52.0	1	1800.0	20	15	0.064024	19.0	0.130435	15.530630
11939	38647	20190526	6	3	51	1	0	0	11220.0	0	...	-7.0	52.0	1	1600.0	3	8	0.000000	19.0	0.130435	15.503876
13957	38647	20190623	5	3	51	1	0	0	12900.0	0	...	-1.0	53.0	5	1800.0	3	8	0.000000	19.0	0.130435	15.280136
14783	38647	20190705	8	3	51	1	0	0	9720.0	0	...	1.0	53.0	5	1800.0	4	7	0.107143	19.0	0.130435	15.189873
16973	38647	20190803	7	3	51	1	0	0	11760.0	0	...	-6.0	52.0	7	1400.0	3	8	0.000000	19.0	0.130435	16.091954
18859	38647	20190830	8	3	53	1	0	0	13020.0	0	...	5.0	52.0	3	1800.0	11	12	0.037383	19.0	0.130435	15.062762

하지만 경기별로 말의 정보는 다름!

말의 특성/경기 정보 등을 이용하여 해당 말의 ' 속력 ' 을 예측하자!

모델링 전략

말	예측 속력
말1	15.779
말2	15.748
말3	15.518
말4	15.385
말5	15.266
...	
말8	14.565
말9	14.486
말10	14.236

예측 속력 top5

예측 성공

말	실제 순위
말1	1
말2	2
말3	3
말4	4
말5	5

예측 실패

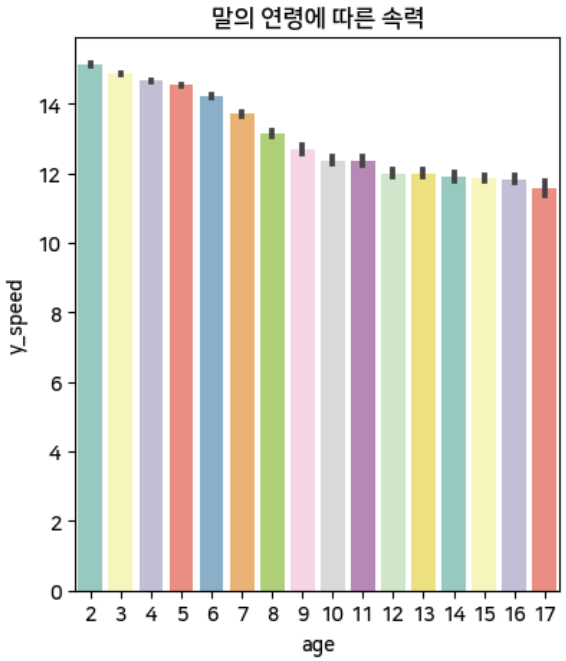
말	실제 순위
말1	7
말2	2
말3	4
말4	6
말5	5

Test set

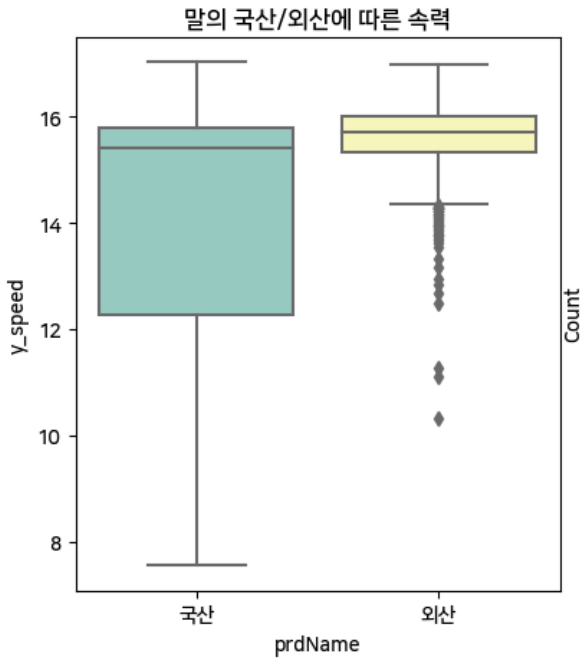
경기	예측 성공 여부
1	1
2	1
3	0
4	0
5	0
...	
2412	0
2413	0
2414	1

예측 성공률
= (예측 성공 경기 수) / (전체 경기 수)

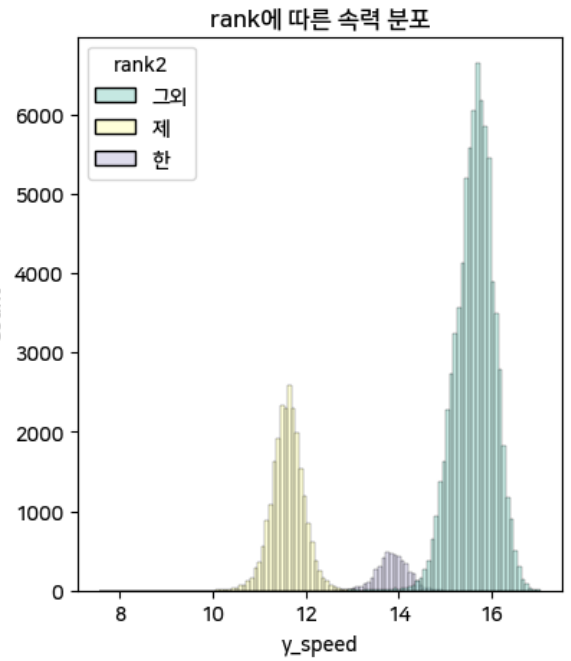
EDA



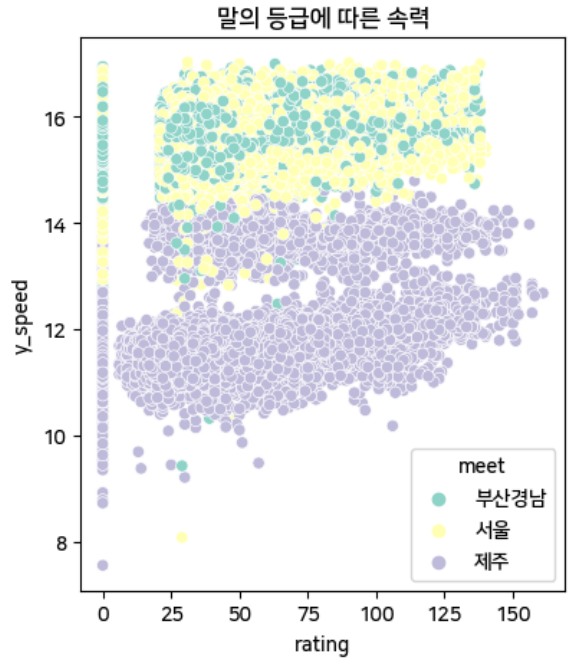
연령이 증가할수록
속력 감소



외산마의 속력이
국산마보다 약간 높음

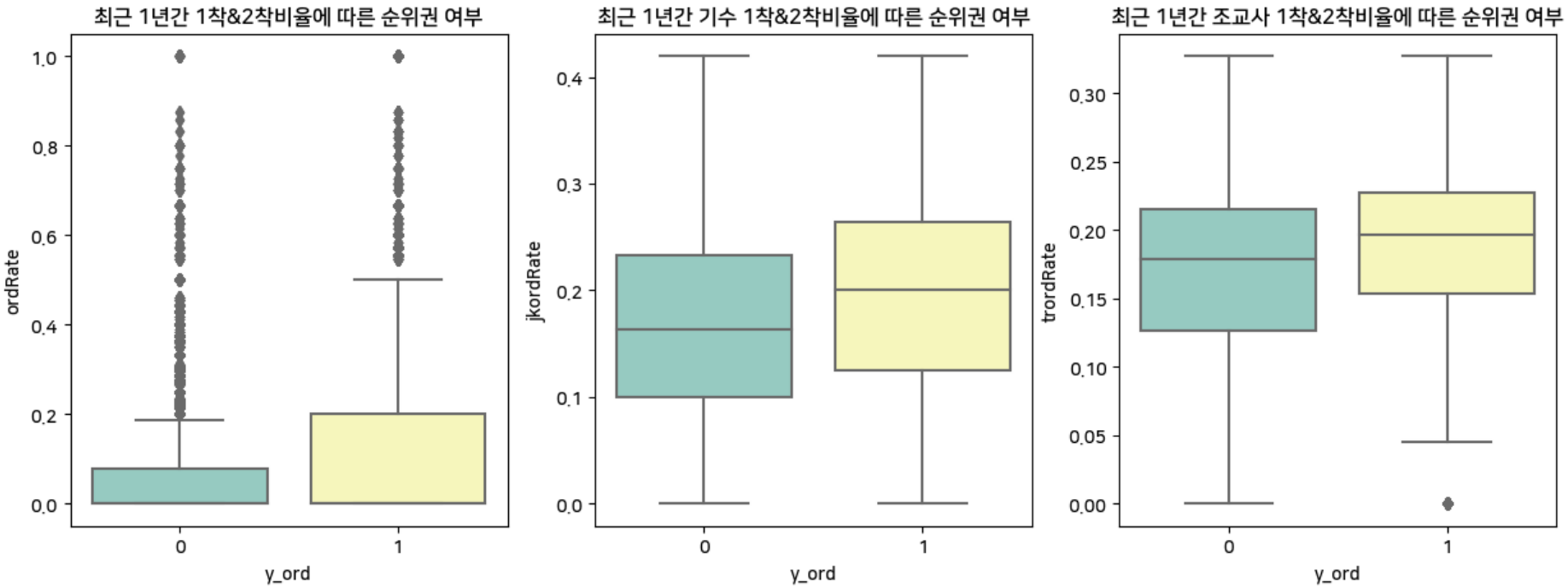


제주마, 한라마, 그외 말의
속력 분포가 다름



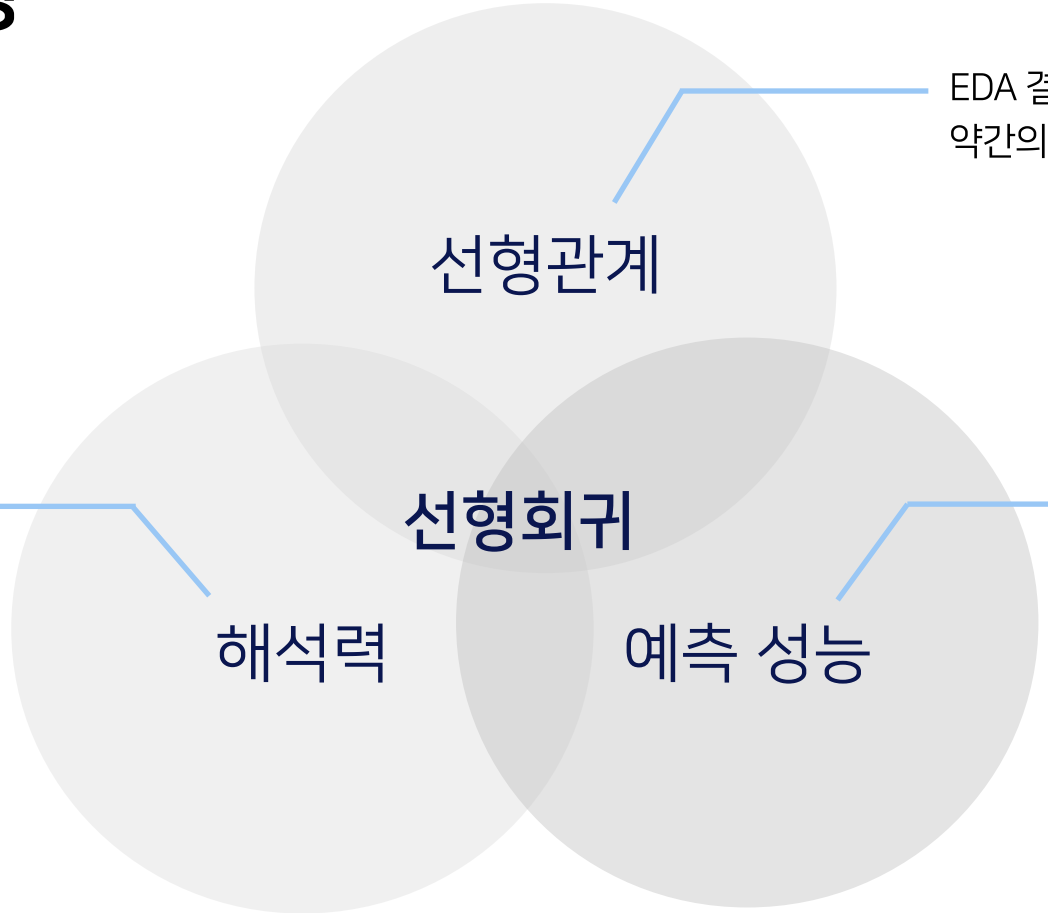
레이팅이 증가할수록
속력이 증가함

EDA



순위권에 든 말의 말/기수/조교사의
1착&2착비율이 더 높음

Base model 선정



EDA 결과, 유의미한 변수들은 속력과 약간의 선형 관계를 가지고 있었음

모델 해석을 통해 속력에 영향을 미치는 변수를 알아보고자 함

Base model 선정을 위해, 모든 변수를 이용하여 선형회귀모형, 랜덤포레스트, 인공신경망을 학습시킨 결과, 선형회귀 모형의 예측 성공률이 가장 높았음

Base Model

Model1: Full Model, 모든 변수 이용

변수 범주	변수
말	연령, 레이팅(등급), 마체중, 마체중 증감, 경기전 1달간 훈련시간 총합, 최근 1개월간 질병진단 횟수, 성별, 외산/국산, 제주마/한라마/그외, 말의 최근 1년간 1, 2착 비율
경기	출주번호, 경주거리, 부담중량, 함수율
기수	기수 경력, 기수의 최근 1년간 1, 2착 비율
조교사	조교사 나이, 조교사 경력, 조교사의 최근 1년간 1, 2착 비율

Adj-R2: 0.967
예측 성공률: 0.342

그러나
다중공선성 존재

Model Develop

Model2: 다중공선성 해결 시도
- VIF가 10 이상인 변수 제거
- Backward Selection을 통한 유의한 변수 선택

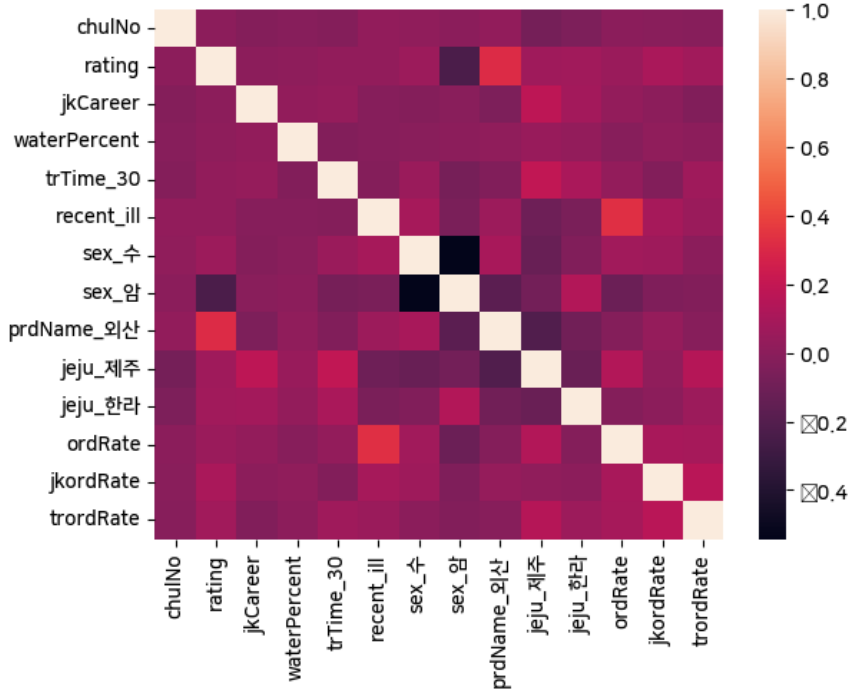
변수 범주	변수
제거 변수	연령, 경주거리, 부담중량, 마체중, 마체중 증감, 기수 나이, 기수 경력, 조교사 경력
말	레이팅(등급), 경기전 1달간 훈련시간 총합, 최근 1개월간 질병진단 횟수, 성별, 외산/국산, 제주마/한라마/그외, 말의 최근 1년간 1, 2착 비율
경기	출주번호, 함수율
기수	기수의 최근 1년간 1, 2착 비율
조교사	조교사의 최근 1년간 1, 2착 비율

Model Develop

Model2: 다중공선성 해결 시도
- VIF가 10 이상인 변수 제거
- Backward Selection을 통한 유의한 변수 선택

Adj-R2: 0.945
예측 성공률: 0.357

Variable	VIF
chulNo	3.787234
rating	2.658026
jkCareer	3.312280
waterPercent	3.200971
trTime_30	6.191925
recent_ill	1.252443
sex_수	2.066692
sex_암	2.250715
prdName_외산	1.425365
jeju_제주	1.686643
jeju_한라	1.175025
ordRate	1.364310
jkordRate	3.569062
trordRate	5.171724



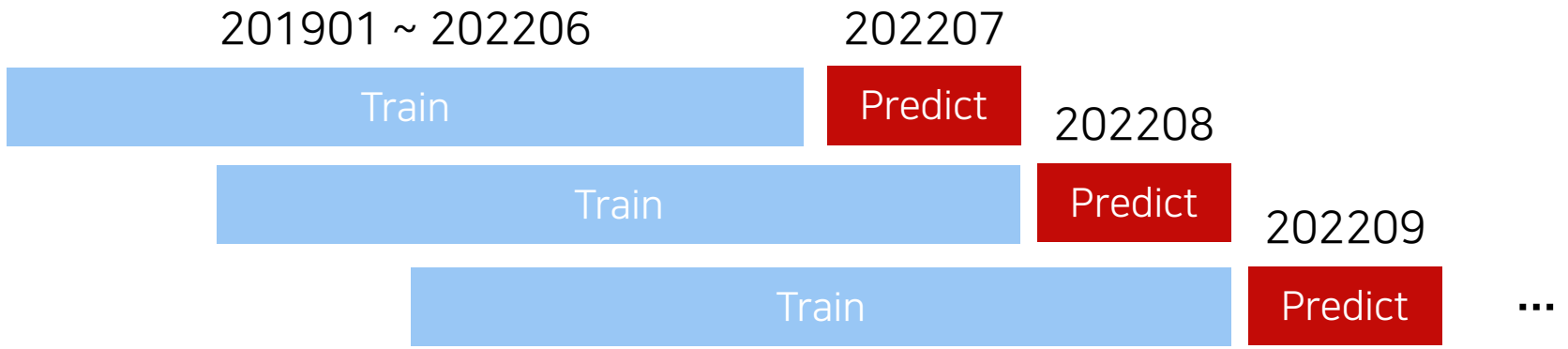
Model Develop

Model3: 선형 회귀 모형 + 규제
- Ridge, Lasso Regression 이용

	Ridge	Lasso
이용 변수	Model2와 동일	
R2	0.960	0.960
예측 성공률	0.357	0.363

Model Develop

Model4: rolling 이용하여 학습 및 예측



월	202207	202208	202209	202210	202211	202212	202301	202302	202303	202304	202305	202306	Var
Ridge	0.262	0.252	0.287	0.393	0.393	0.392	0.38	0.311	0.355	0.463	0.425	0.344	0.004
Lasso	0.284	0.194	0.299	0.389	0.373	0.417	0.387	0.325	0.351	0.454	0.446	0.330	0.005

Final result

최종 모델	Ridge
이용 변수	Model2와 동일
R2	0.960
예측 성공률	0.357
Rolling 평균 예측 성공률	0.355

	0	adj_coef
chulNo	-0.0014	-0.000712
rating	0.0026	0.000152
jkCareer	-0.0002	-0.000040
waterPercent	0.0063	0.002029
trTime_30	0.0000	0.000000
recent_ill	0.0047	0.003784
sex_수	0.0194	0.071458
sex_암	0.0536	0.189059
prdName_외산	-0.0252	-0.135250
jeju_제주	-4.0582	-16.713016
jeju_한라	-1.8185	-15.815634
ordRate	0.3618	4.018367
jkordRate	0.3900	7.087850
trordRate	0.2724	6.564536

경기전 한달간 훈련시간의 총합 중요도 ↓

제주마, 한라마의 경우 속력이 낮음

말/기수/조교사의 최근 1년간 1, 2착 비율의 중요도 ↑

결론 및 한계점

결론

- 우리의 base model인 선형회귀분석을 통한 경마 우승마 예측 모델을 최대한 발전시키고자 하였음
- 경마 우승과 관련된 변수들을 찾아낼 수 있었음
- 설명력 있고 유의한 회귀 모형을 찾아가는 과정에서, 주어진 예측 task 성능을 증진할 뿐 아니라 말의 속력에 대해 어느정도 설명력 있는 모델을 만들어 냄

한계점

- EDA로 파악한 변수 관계와 모델 해석 결과가 다름 → 비선형 모델로 분석 및 성능 개선의 여지가 있음
- 말의 실력과 관련이 있는 변수가 충분하지 않았음
- 선형성이라는 강한 가정으로 인해 예측에 제약이 존재했음
- 평균 속력은 경주 거리의 영향을 받으나, 본 모델에서는 이를 고려하지 않음

Reference

[1] 최혜민, et al. "서울 경마 경기 우승마 예측 모형 연구." 응용통계연구 28.6 (2015): 1133-1146.

[2] 유선경, and 박흥선. "로지스틱 회귀를 통한 경마의 입상확률모형." 응용통계연구 13.1 (2000): 35-43.

[3] 김필수, 이상현, and 전성삼. "머신러닝을 적용한 경륜 경기 순위 예측 및 평가에 관한 연구: 2016~ 2022 년 출주표 정보 및 경주 결과 활용." 한국스포츠산업경영학회지 28.2 (2023): 76-94.

[4] 박가희, 박리라, and 송종우. "통계적 예측모형을 활용한 경륜 경기 순위 분석." 응용통계연구 30.1 (2017): 25-39.