로지스틱 회귀분석을 이용한 한국 프로야구 승/패 예측모형

응용통계학과

201552043 전태양

목차

1. 서론

야구는 이닝마다 공격과 수비를 번갈아가며 진행하고, 투수와 타자가 1대 1로 겨룬다. 그렇기에 다른 스포츠들과 달리 정지되는 순간도 많고, 원인과 결과가 분명하다. 따라서 경기 중 발생하는 모든 상황과 플레이를 끊어서 턴마다 분석이 가능하다. 이러한 야구의 특성 덕분에 야구는 과거부터 수많은 통계 자료를 쌓아 놓을 수 있었으며, 해당 자료들을 기반으로 여러가지 수리적 지표를 동원해 야구 선수를 객관적으로 평가하는 세이버메트릭스가 다른 스포츠에 비해 빠르게 활성화될 수 있었다.

경기로그

이닝	투수	타자	P	결과	이전상황	이후상황	LEV	REs	REa	WPe	WPa
1초	뷰캐넌	1 정수빈	7(2-2)	1루수 땅볼	무사 0:0	1사 0:0	0.87	0.555	-0.258	52.2%	-0.022
	뷰캐넌	2 페르난데스	1(0-0)	좌익수 뜬공	1사 0:0	2사 0:0	0.62	0.297	-0.180	53.8%	-0.016
	뷰캐넌	3 박건우	7(2-2)	3루수 땅볼	2사 0:0	이닝종료 0:0	0.40	0.117	-0.117	54.8%	-0.010
1말	최원준	1 박해민	2(1-0)	2루수 땅볼	무사 0:0	1사 0:0	0.87	0.555	-0.258	52.6%	-0.022

[그림1. 야구는 경기 중 발생하는 모든 상황을 기록한다.]

데이터를 활용하는 가장 큰 목적은 결국 야구 경기의 승리를 위해서다. 객관적으로 다루는 지표를 활용하면 팀의 약점이 무엇인지, 강점이 무엇인지 판단하기 용이하고 상대팀의 지표와 비교가 가능해 전략적인 경기를 설계하기에 용이하기 때문이다.

"타율 3할 5푼을 치는 타자와 처음 상대해도, 데이터 분석에서 나온 대로 던지기만 하면 거의 100% 삼진을 잡을 수 있다"[1]라는 류현진 선수의 말처럼, 현대 프로야구에서 데이터는 무시할 수 없을 정도로 발전했다.

앞으로도 국내 야구 리그의 규모가 커지고 관중의 규모가 커질수록 매경기마다 쏟아내는 다양한 데이터들을 활용한 분석은 새로운 지표들을 생성해내고 승리와 패배 원인을 분석하기 위한 다양한 방식으로 활용될 것이다.

이렇듯 본 분석에서는 야구데이터 기록사이트인 스탯티즈[2]에서 2017년부터 2020년까지 4년의 경기 데이터를 수집하여 팀 타율, 홈런 개수, 삼진 개수와 같은 지표들 중 어떤 지표가 팀 승리에 크게 기여하는지 분석하고 그에 따른 다른 경기들의 승/패 예측을 진행하고자 한다.

---

[1] https://sports.v.daum.net/v/20190301175502754 기사 참조

[2] http://www.statiz.co.kr

2. 데이터 설명

GDAY_DS	T_ID	VS_TID	TB_SC	PA	...	LOB	OBP	OOO	Win
20170331	HH	OB	T	33	...	11	0.219	0.138	0
20170324	OB	HH	B	33	...	15	0.242	0.148	1
20170331	KT	HT	T	42	...	12	0.357	0.206	1
20170331	HT	KT	B	33	...	8	0.25	0.226	0

[표1. 데이터 형태]

변수 명	변수 타입	변수 설명	SCALE
GDAY_DS	Int	경기 일자	
T_ID	Chr	홈 팀	Ex)한화(HH)
VS_TID	Chr	상대 팀	Ex)기아(HT)
TB_SC	Chr	홈/어웨이 여부	T:어웨이 B:홈
PA	Int	타석: 타자가 타석에 선 모든 횟수	17 ~ 65
AB	Int	타수: 타석에서 볼넷, 몸에 맞는 볼, 희생플라이를 제외한 횟수	15 ~ 54
RBI	Int	타점: 타자가 타격을 통해 낸 점수	0 ~ 26
RUN	Int	득점: 출루한 선수가 직접 홈 플레이트로 돌아와 낸 점수	0 ~ 26
HIT	Int	안타를 친 횟수	0 ~ 25
H2	Int	2루타를 친 횟수	0 ~ 16
H3	Int	3루타를 친 횟수	0 ~ 6
HR	Int	홈런을 친 수	0 ~ 8
SB	Int	도루를 성공한 횟수	0 ~ 14
CS	Int	도루를 실패한 횟수	0 ~ 4
SF	Int	희생플라이: 타자가 아웃된 플라이 타구에 주자가 홈인 했을 경우	0 ~ 4
BB	Int	볼넷: 볼 4개를 얻어 출루한 횟수	0 ~ 16
HP	Int	몸에 맞은 공으로 출루한 횟수	0 ~ 5
KK	Int	삼진을 당한 횟수	0 ~ 18
GD	Int	병살: 더블 아웃을 당한 횟수	0 ~ 5
LOB	Int	잔루율: 투수가 루상에 주자를 남긴 채로 이닝을 끝낸 수치	1 ~ 42
OBP	Float	팀 출루율: 타자가 타석에서 1루를 밟은 수치	0.068 ~ 0.6
OOO	Float	팀 타율	0 ~ 0.533
Win	Int	홈 승/패	0:패배 1:승리

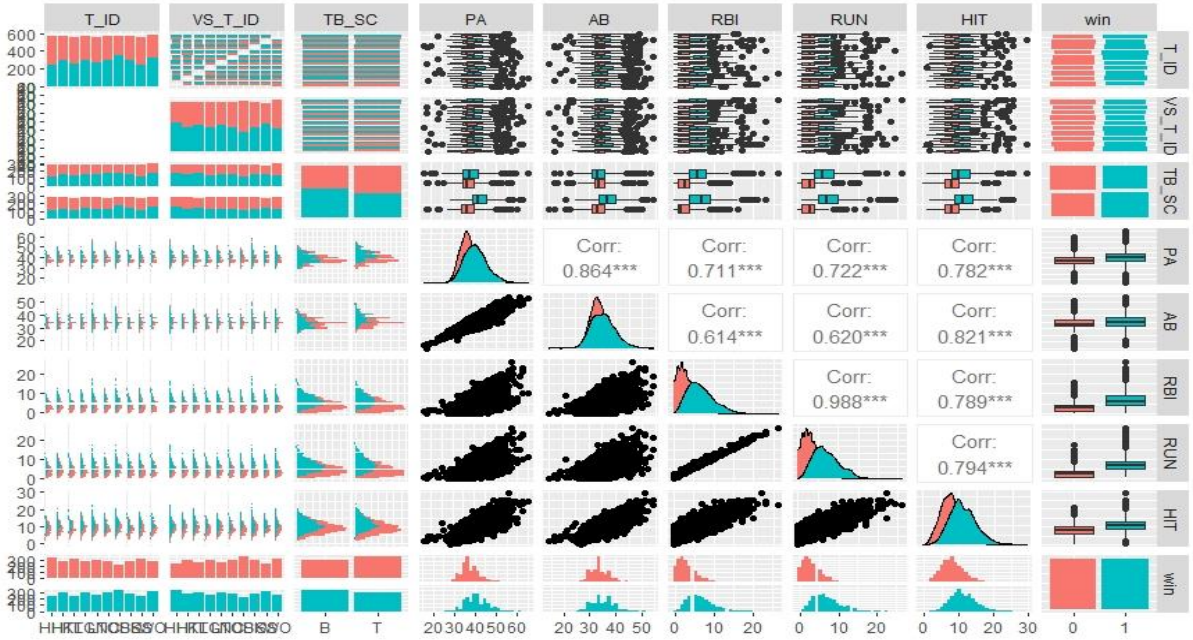
[표2. 변수 설명]

본 분석에서는 2017년 3월 31일부터 2020년 8월 18일까지 총 5756경기의 한국 프로야구 공식 경기 정보 데이터를 이용했으며, 해당 데이터는 경기일자와 홈/어웨이 팀과 같은 경기 외적인 정보가 담긴 변수 4개와 경기 내용이 담긴 18개의 세이버메트릭스 지표를 이용해 승패예측을 진행했다.

3. EDA

3-1. 산점도 행렬

[그림2. 산점도 행렬 1]



먼저 T_ID부터 HIT(안타)까지의 설명변수와 반응변수 Win과의 산점도 행렬을 살펴본 결과, 전반적으로 타석(PA), 타수(AB), 타점(RBI), 득점(RUN), 안타(HIT)의 분포가 강한 상관 관계를 보였으며 특히 타석과 타수, 득점과 타점이 강한 상관관계 분포를 보였다.

경기 외적인 변수와 승패(Win)과의 관계를 살펴보면 홈일 때 승리가 더 많았으며 4년간 승리를 많이 한 팀은 두산 베어스(OB)와 키움 히어로즈(WO)였고, 패배가 가장 많은 팀은 한화 이글스(HH)였다.

득점과 타점의 분포가 우측으로 긴 꼬리의 형태를 가지고 있음을 확인할 수 있었다.

상관계수 행렬에서도 전반적으로 강한 양의 상관관계임을 확인할 수 있다.

[그림3. 상관계수 행렬 1]

[그림4. 산점도 행렬 2]

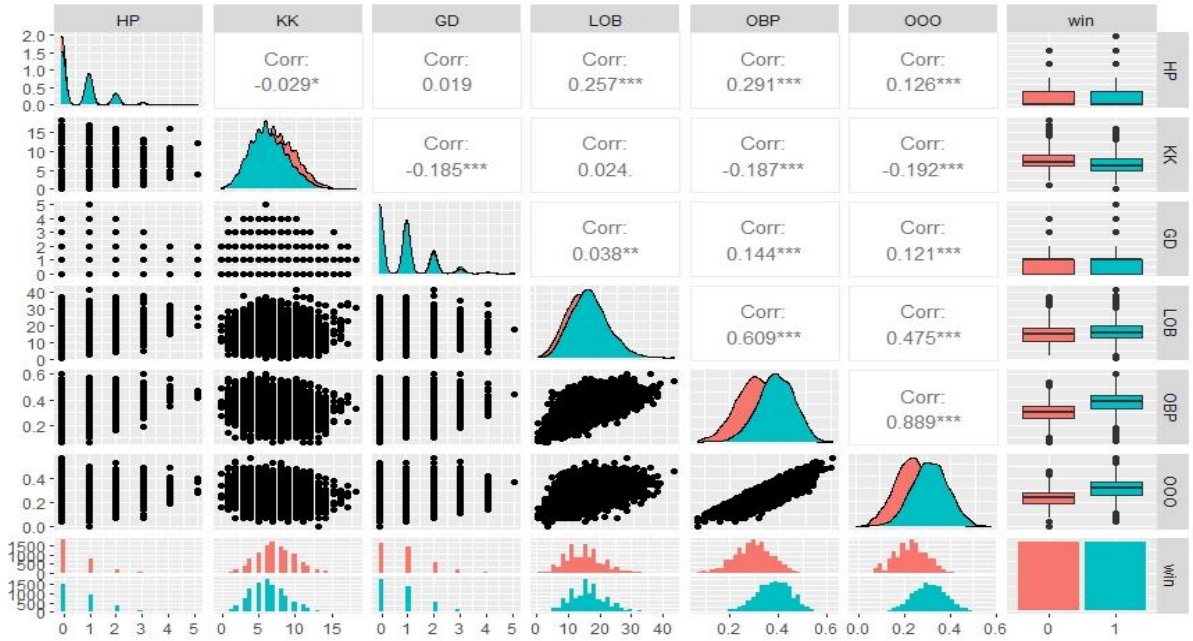
2루타와 3루타를 친 횟수(H2, H3)나 홈런을 친 횟수(HR), 도루 성공, 실패 횟수(SB, CS), 희생플라이(SF), 볼넷(BB) 횟수는 변수 간 상관관계가 있다고 보기 어려우며 승패와 큰 상관성을 보이지 않는다.

3루타와 몸에 맞는 볼, 도루 실패, 희생플라이 횟수는 0의 빈도가 높았으며, 홈런 수와 볼넷 수도 좌측에 치우쳐 있는 것을 확인했다.

[그림5. 상관계수 행렬 2]

마찬가지로 상관계수 행렬에서도 서로 강한 상관관계를 보이지 않았다.

[그림6. 산점도 행렬 3]



출루율(OBP)과 팀타율(OOO)은 강한 양의 상관관계를 띄고 있으며, 잔루율(LOB)과 출루율, 팀타율은 약한 양의 상관관계를 띄는 것을 확인할 수 있다.

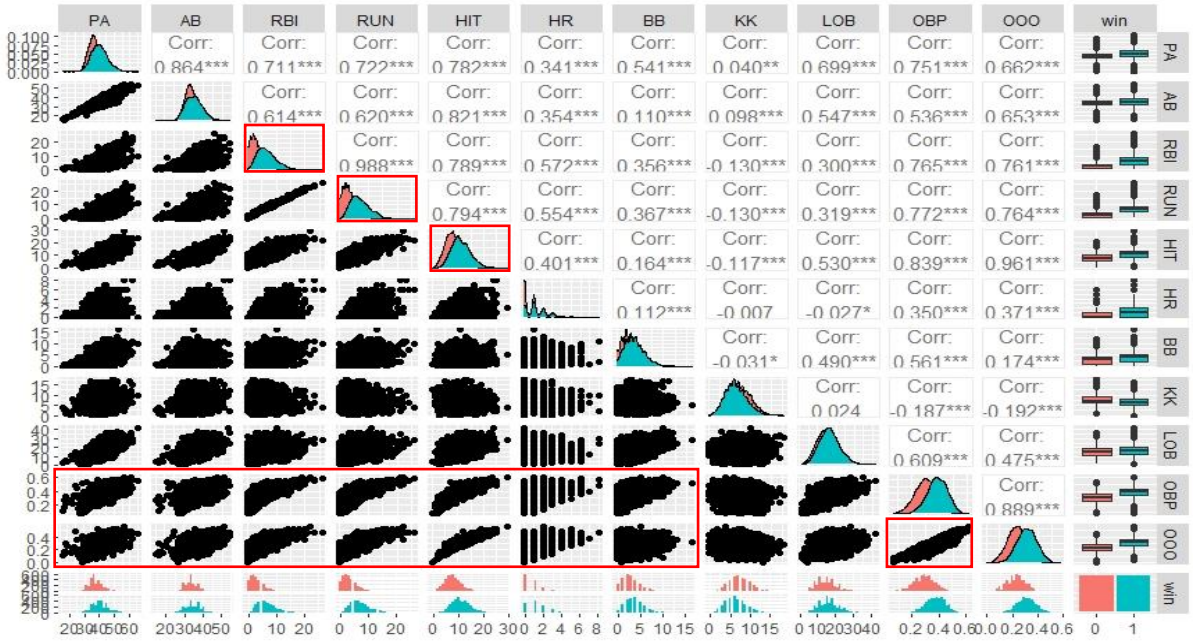
몸에 맞는 볼 횟수(HP)와 삼진 수(KK), 병살 수(GD)는 승패에 큰 영향을 미치지 못하는 것으로 확인되며 출루율과 팀타율이 높을수록 승리에 분포되어 있는 것을 확인할 수 있다.

[그림7. 상관계수 행렬 3]

상관계수 행렬에서도 출루율과 팀타율, 잔루율 지표가 양의 상관관계임을 확인할 수 있다.

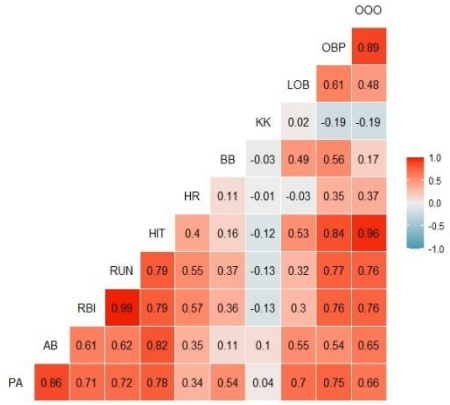
지금까지 설명변수가 많아 산점도 행렬을 분할하여 살펴보았다. 그러나 지금과 같은 방식은 분할한 변수들의 분포만을 확인할 수 있기에 연속적이거나 데이터에 0이 적은 변수들을 골라 재차 확인하는 작업이 필요하다고 판단했으며 아래와 같이 진행했다.

[그림8. 산점도 행렬 4]



추가적으로 재차 확인한 결과, 팀 타율과 팀 출루율은 공격과 관련된 지표와 대체적으로 양의 상관관계로 나타났다.

당연하게도 팀 타율과 팀 출루율이 높을수록, 안타 횟수와 득점과 타점이 많을수록 승리하는 것을 확인할 수 있다.



[그림9. 상관계수 행렬 4]

4. 로지스틱 회귀 분석

4-1. 자료 분리

먼저 데이터셋을 학습시킬 Train 데이터 80%와 예측에 쓰일 Test 데이터 20%로 분리했다. 이때, 각 팀의 비율을 고르게 하기위해 T_ID를 기준으로 층화 추출하였다.

	HH	HT	KT	LG	LT	NC	OB	SK	SS	WO	총 데이터 수	Target 분포
Train (80%)	453	452	445	450	448	452	457	455	441	470	4523	0: 2251 1: 2271
Test (20%)	113	113	111	112	112	112	114	113	110	117	1127	0: 574 1: 553

[표3. 분리한 Train Test 셋]

추가적으로 경기 일자를 나타내는 GDAY_DS 변수가 팀/홈/원정 별로 총 10번씩 반복되므로 데이터셋에서 제외하였다.

4-2. 회귀모형 적합

분리한 Train 데이터를 기반으로 로지스틱 회귀 분석을 실시하였다. 로지스틱 회귀 모형은 패배(0)와 승리(1)에 해당하는 이항변수를 선형으로 근사화시킨다.

Formula : Win ~ T_IDHT + … + OOO -GDAY_DS(경기일자를 제외한 모든 변수)	
유의한 변수	T_ID, VS_T_ID, TB_SC, PA, AB, RBI, RUN, HIT, H3, SB, CS, BB, HP, KK, GD, OBP, OOO
AIC	3717.938
BIC	3961.782

회귀모형의 적합도 검정

Resid. DF	Resid. DEV	DF	Deviance	P-value
4522	6270.1			
4485	3641.9	37	2628.2	<2. 2e-16

[표4. 회귀분석 결과]

로지스틱 회귀에서는 F검정이 없어진 대신 Deviance(이탈도)라는 척도를 가지고 모형이 적합한지에 대해 판단한다. 적합도 검정 결과 P-value가 <2. 2e-16이므로 '모형의 모든 변수의 회귀계수는 0이다'라는 귀무가설을 기각한다. 즉, 변수가 없이 상수항만 있는 모형 대비 통계적으로 유의한 모형이라고 볼 수 있다.

설명 변수의 효과 분석

로지스틱 회귀에서 설명 변수가 반응 변수를 효과적으로 설명하는지에 대해 판별하기 위해서는 오즈비(Odds Ratio)를 사용한다. 오즈비는 나머지 변수를 고정시킨 상태에서 한 변수를 1만큼 증가시켰을 때 변화하는 Odds의 비율[3]이다.

변수 명	Odds Ratio	변수 명	Odds Ratio
(T_ID) HT	1.1512	**PA**	1.6988
KT	0.9952	**AB**	0.4732
LG	1.4382	**RBI**	0.8594
LT	1.32	**RUN**	2.08993
NC	1.3314	**HIT**	0.7266
OB	1.6631	**H2**	1.0256
SK	1.5436	**H3**	1.111
SS	1.0415	**HR**	0.964
WO	1.6593	**SB**	1.159
(VS_T_ID) HT	0.5541	**CS**	1.148
KT	0.8049	**SF**	0.4848
LG	0.7111	**BB**	0.4587
LT	0.8253	**HP**	0.4546
NC	0.5859	**KK**	0.9466
OB	0.536	**GD**	0.6386
SK	0.6564	**LOB**	0.9956
SS	0.9628	**OBP**	4.6339e+05
WO	0.6224	**OOO**	6.1124e+04
(TB_SC) T	1.216		

[표5. 회귀 모형 오즈비]

범주형 변수는 하나의 기준 대비 다른 범주들의 차이를 비교한다. 따라서 T_ID와 VS_T_ID는 한화 이글스(HH)를 기준으로, TB_SC는 홈(B)를 기준으로 승리 확률을 비교할 수 있다.

예를 들어 홈팀이 기아(HT)일 경우엔 한화일 때 보다 승리 확률이 1.15배 증가한다고 볼 수 있으며, 상대팀이 기아(HT)일 경우엔 한화일 때보다 승리 확률이 (1 – 0.5541) * 100 = 44.59% 감소한다고 볼 수 있다.

타석과 득점, 출루율과 팀타율이 늘어날수록 승률이 크게 늘어났으나, 반대로 상관관계가 강했던 타수와 타점이 늘어날수록 승률이 감소하는 현상을 보여 다중 공선성이 의심된다.

[3] Odds : 사건이 발생할 확률(P)를 사건이 발생하지 않을 확률(1-P)로 나눈 비율이다.

4-3. 변수 선택

지금까지 모든 설명변수를 활용하여 회귀모형을 설정하고, 오즈비를 활용하여 설명변수들의 효과를 해석해봤다. 그 결과, 타석이 많아질수록 승리 확률이 올라가는데 타수가 많아지면 승리 확률이 오히려 감소하는 등, 변수들 간에 강한 상관관계로 인해 다중 공선성 문제가 발생했음을 판단했고, 그에 근거하여 변수 선택을 진행했다.

변수 선택을 하는 기준은 모형에 변수를 추가할수록 페널티를 부여해 모형의 품질을 평가하는 AIC와 BIC를 사용했다.

모형 선택 및 단계적 회귀	선택된 변수
AIC에 의한 전진 선택 (M1)	T_ID + VS_T_ID + TB_SC + PA + AB + RBI + RUN + HIT + H2 + H3 + SB + CS + SF + BB + HP + KK + GD + OBP + OOO
AIC에 의한 후진 소거	전진 선택 모형과 동일
BIC에 의한 전진 선택(M2)	TB_SC + AB + RUN + SB + SF + BB + HP + KK + GD + OBP
BIC에 의한 후진 소거(M3)	PA + AB + RUN + SB + SF + BB + HP + KK + GD + OBP

[표6. 변수 선택]

변수 선택 결과 3가지 모형이 채택되었고 3개의 모형의 AIC와 BIC를 비교했으며 VIF를 활용하여 다중 공선성 문제가 존재하는지 확인했다.

AIC	BIC	VIF
M1: 3714.424	M1: 3945.433	M1: VIF 최대 10.4(OOO, HIT)
M2: 3779.182	M2: 3836.934	M2: VIF 최대 2.81(AB)
M3: 3736.297	M3: 3806.883	M3: VIF 최대 64(AB, PA)

[표7. 3가지 모형 결과 값]

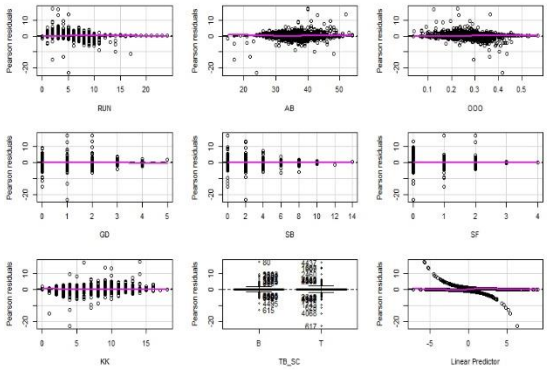
3가지 모형의 기준이 큰 차이가 없었으나 M1과 M3에서 VIF의 값이 10을 넘는 설명변수들이 존재했다. 따라서 M2를 추정 모형으로 선택했다.

4-4. 변수 선택을 통해 선정한 추정 모형

$$log\left(\frac{\textit{승리 확률}(y=1|x)}{\textit{패배 확률}(y=1|x)}\right) \sim 6.37 + (-0.358) * AB + 0.568 * RUN + \cdots + (-0.051) * KK$$

[표8. 추정 모형식]

모형 진단

잔차 산점도	변수 명	P-value
	AB	2.488e-05
	RUN	2.869e-10
	OOO	0.397468
	GD	0.59
	SB	0.133
	SK	0.9133
	KK	0.0021
	TB_SC	

[표9. 모형 진단]

잔차 산점도로 모형을 진단한 결과, 비모수 회귀 곡선인 마젠타 곡선이 전반적으로 수평선을 유지하는 것을 확인했다. 그러나 2차항 유의성 검정에서는 AB와 RUN, KK가 유의하게 나왔기에 3가지 변수에 대해 2차항을 추가한 모형을 재차 회귀 진단했고, 모든 변수가 유의하다는 결과를 확인했다. (부록 참고)

2차항 추가 모형

$$log\left(\frac{\textit{승리 확률}(y=1|x)}{\textit{패배 확률}(y=1|x)}\right) \sim 6.37 + (-0.358) * AB + 0.568 * RUN + \cdots + (-0.028) * RUN^2 + 0.016 * AB^2 + 0.012 * KK^2$$

[표10. 2차항 추가 모형식]

AIC	BIC
기존 모형 : 3779.182 2차항을 추가한 모형 : 3692.784	기존 모형 : 3836.934 2차항을 추가한 모형 : 3769.787

[표11. 모형 비교]

2차항을 추가한 모형이 AIC와 BIC가 더 낮게 나왔지만 큰 차이를 드러냈다고 보기 어려워 분류 성능으로 최종 모형을 선택했다.

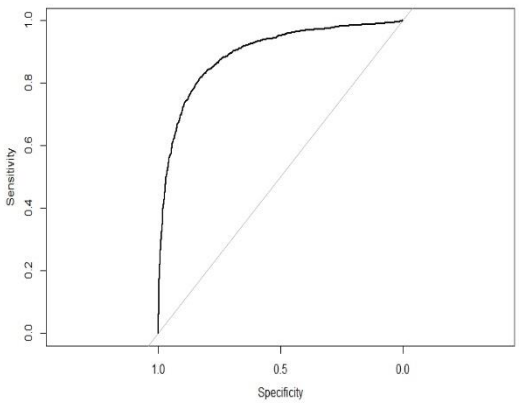
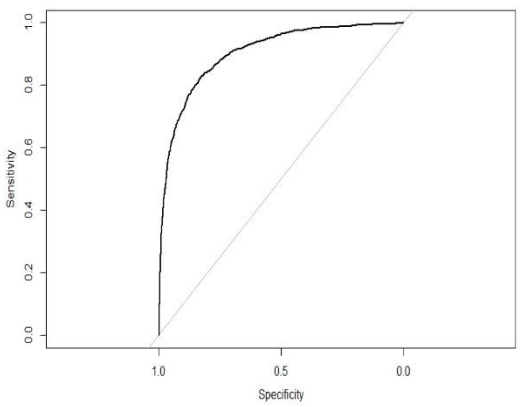
모형의 분류 성능 평가

추정 모형		
Prediction (예측)	**Condition (실제)**	
	승(Win = 1)	**패(Win = 0)**
승(Win = 0)	1834	372
패(Win = 0)	438	1879

Accuracy : 0.8209
Sensitivity : 0.8072
Specificity : 0.8347
F1 - Score : 0.8191

2차항 추가 모형		
Prediction (예측)	**Condition (실제)**	
	승(Win = 1)	**패(Win = 0)**
승(Win = 1)	1879	393
패(Win = 0)	394	1857

Accuracy : 0.826
Sensitivity : 0.827
Specificity : 0.825
F1 - Score : 0.8268

추정 모형 ROC CURVE	**2차항 추가 모형 ROC CURVE**
AUC : 0.8938	**AUC : 0.9014**
	

[표12. 모형들의 분류 성능 평가]

모형의 분류 성능 평가 결과, 2차항 추가 모형의 정확도와 F1 – Score, AUC가 모두 기존의 추정 모형보다 높게 나왔으며, 분류 성능 평가를 기반으로 2차항 추가 모형을 최종 모형으로 채택하였다.

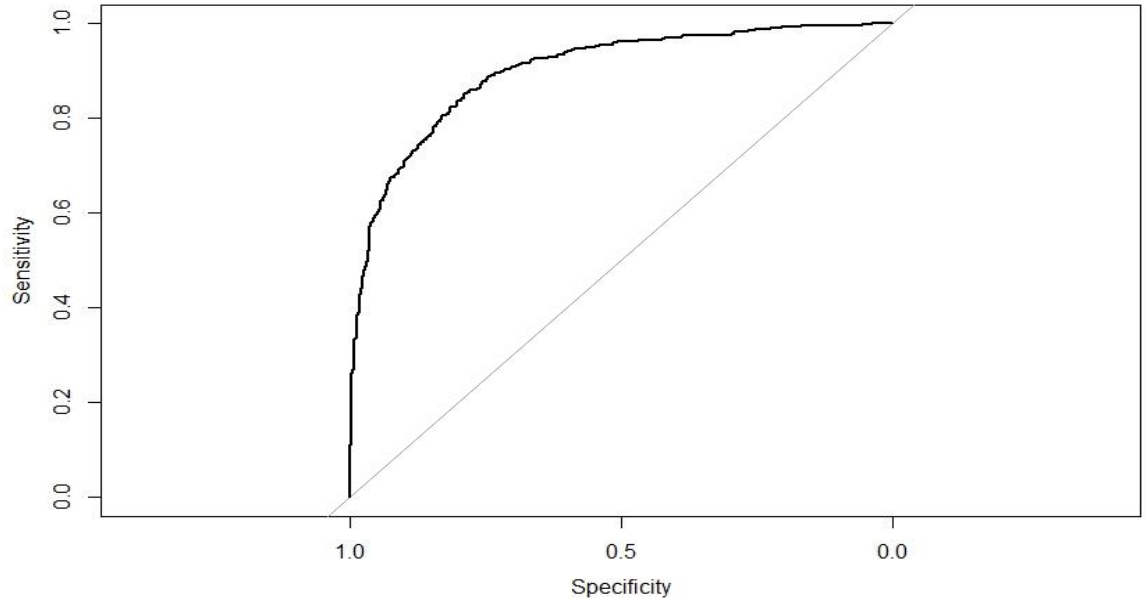
5. 최종 모형으로 Test data 예측

앞서 채택한 최종 모형(표10.)을 활용하여 사전에 분리한 Test data의 승패를 예측했다.

Prediction (예측)	Condition (실제)	
	승(Win = 1)	패(Win = 0)
승(Win = 1)	455	111
패(Win = 0)	98	463

Accuracy : 0.8146
Sensitivity : 0.8228
Specificity : 0.8066
F1 - Score : 0.8132

TEST DATA ROC Curve
AUC : 0.899



[표13. Test data 분류 성능 평가]

최종 모형을 활용하여 Test data를 예측한 결과 분류 정확도는 81%가 나왔으며 민감도와 특이도는 각각 82%와 80.6%로 나타났다.

6. 결론

지금까지 2017 ~ 2020 시즌 한국 프로야구 공식 경기 데이터를 활용하여 승패 예측을 진행하였다.
홈/어웨이 여부, 타수와 득점, 도루를 성공한 횟수, 희생플라이 횟수, 더블아웃을 당한 횟수, 삼진을 잡은 횟수, 팀 출루율이 승리 여부에 영향을 끼치는 것으로 나타났으며, 해당 지표를 기반으로 승패를 예측한 결과, 전체 분류 정확도는 81%로 나타났다.

본 분석에서는 22개의 지표를 사용했지만 공격 지표가 대부분을 차지한다는 아쉬움이 있다.
추후, Fpct[4]과 같은 수비 지표와 방어율과 같은 개인 선수 별 지표를 추가하여 분석한다면 지금보다 다양한 방면으로 승패에 끼치는 영향력을 확인할 수 있을 것이며 해당사항을 기반으로 지금보다 높은 성능을 가진 분류기를 생성할 수 있을 것이다.

[4] Fileding Percentage: 수비율, 수비수가 수비 기회 중 아웃 처리에 성공한 비율

부록

1. 활용한 데이터셋 정보

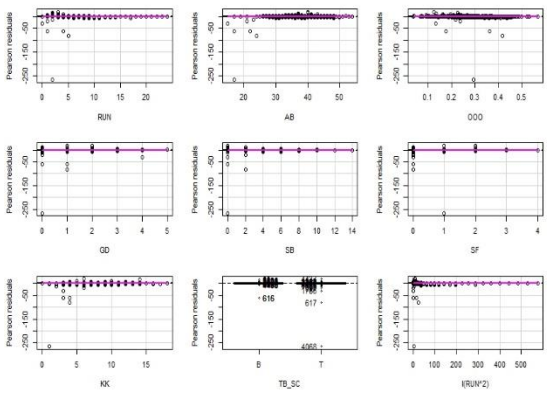
Datasets : Korea Baseball Datasets(KBO 2015 ~ 2020)

- **Sabermetrics of Korea Baseball Pro league from 'statiz' site**

Link :

- **https://www.kaggle.com/park123/korea-baseball-datasetkbo-20152020?select=baseball_2020.csv**

2. 2차항 추가 모형 진단 결과

잔차 산점도	변수 명	P-value
	AB	< 2e-16
	RUN	< 2e-16
	OOO	< 2e-16
	GD	< 2e-16
	SB	5.17e – 10
	SF	1.20e-05
	KK	0.00104
	TB_SC	2.30e-05
	I(RUN^2)	2.26e-16
	I(AB^2)	5.66e-15
	I(KK^2)	0.00896

[표14. 2차항을 추가한 모형 진단]