대학 농구 승패 요인 분석 및 **2024** 정기연고전 농구경기 결과 예측

데이터분석 1조 : 강수민, 권해찬, 유선호, 전유하, 조영인

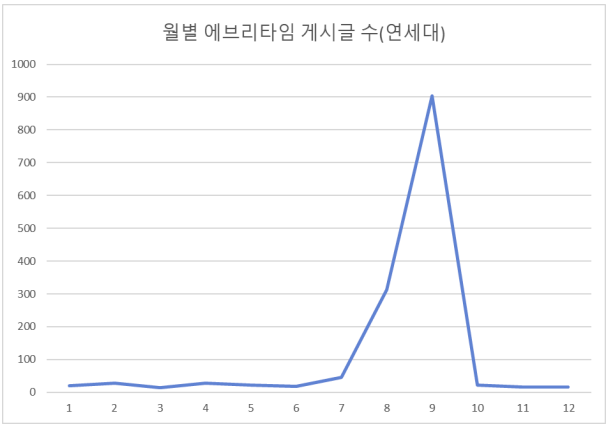
1. 연구배경

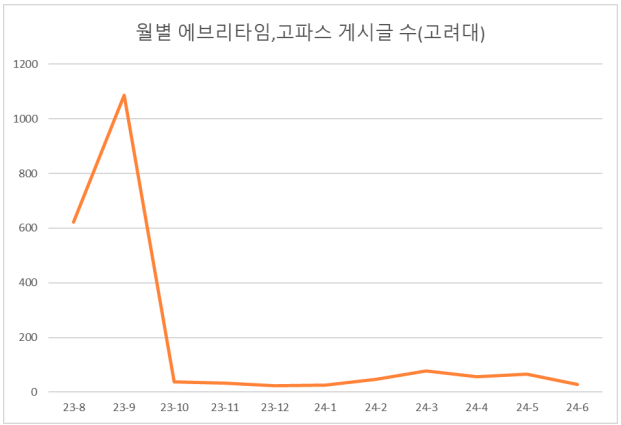
MZ세대 대학생들이 대학생활을 더욱 풍성하게 즐기기 위해 어떤 정보를 필요로 하는지에 대해 알아보고자 두 학교의 에브리타임, 유튜브, 그리고 고려대학교 자체 커뮤니티 고파스를 조사해보았다. 전반적으로 학업과 진로에 대한 질문이나 고민보다는 대학 생활이나 학교 행사에 관한 게시글이 많이 존재했고, 그 중에서도 '연고전/고연전'은 두 학교에서 가장 인기많은 행사이다.

설문조사 결과에 따르면 고려대학교 학생들의 85%, 연세대학교 학생들의 83%가 학교 행사에 적극적으로 참여하며, 특히 연고전과 같은 대규모 행사에 대해 큰 기대감을 가지고 있음을 확인할 수 있다. 또한, 소셜 미디어 분석 결과에서도 연고전 관련 게시물이 두 학교 학생들 사이에서 폭발적인 반응을 얻고 있는 것을 알 수 있다.

고려대, 연세대학교 학생들 뿐만 아니라 타 대학교에 재학중인 학생들 또한 연고전에 관심이 많다. 유명 동영상 크리에이터 연고티비의 동영상 평균 조회수는 266,098회인데, 이 중 연고전 관련 컨텐츠는 조회수가 1,075,382회로 상당히 높음을 알 수 있다.

1년간 연고전에 관련하여 홍보성 글을 제외한 게시물이 연세대학교 에브리타임에 1446개, 고려대학교 에브리타임과 고파스에 2139개 존재했다. 이중 약 80%는 정기전 시즌인 8,9월에 게시되었으며, 비시즌에도 월평균 20-30개의 게시글이 올라오는 것을 확인할 수 있다.





(연세대: 23년 1-12월 데이터, 고려대: 23년8월 - 24년 7월 데이터)

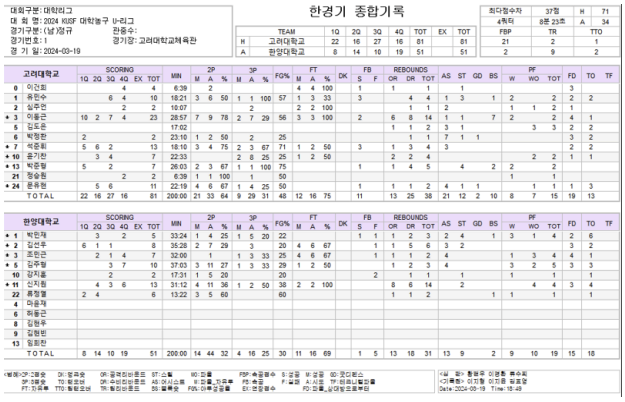
따라서, 본 연구에서는 연고전의 흥미로운 정보를 제공함으로써 학생들의 참여를 촉진하고, 그들이 원하는 정보를 적시에 제공하는 것을 목표로 한다. 특히, 연고전 관련 데이터 분석, 팀별 전력 비교, 과거 경기 기록 등 다양한 데이터 소스를 통해 많은 학생들이 궁금해하는 연고전 승부 예측 결과를 제공한다. 이를 통해 학생들은 경기 전부터 경기 결과에 대한 다양한 예측과 논의를 할 수 있으며, 더욱 재미있고 흥미로운 방식으로 연고전을 즐길 수 있을 것이다. 또한, 연고전 승부 예측 결과 제공은 단순히 정보 제공에 그치지 않고, 학생들 간의 소통과 교류를 촉진하는 역할도 할 것이다. 많은 학생들이 연고전에 대해 더 깊이

이해하고, 이를 통해 학교에 대한 자부심과 소속감을 한층 더 강화할 수 있을 것으로 기대된다. 이와 같은 시도를 통해 고려대학교와 연세대학교 학생들이 더욱 활기차고 의미 있는 대학 생활을 즐길 수 있도록 도울 것이다.

2. Data & Methodology

2.1. 활용 데이터셋

본 연구에서는 2022, 2023, 2024년 KUSF 대학농구 U-리그의 고연전 농구 경기 데이터를 사용하였다. 데이터셋은 각 팀의 경기 결과와 관련된 다양한 피처들로 구성되어 있으며, 여기에는 각 쿼터별 스코어, 공격리바운드, 수비리바운드, 2점슛, 3점슛, 자유투 등 총 33개의 피처가 포함된다. 이러한 데이터는 한국대학농구연맹(http://m.kubf.or.kr/schedule/league.php) 사이트의 경기 기록지를 통해 추출되었으며, 총 299개의 경기를 분석 대상으로 하였다. 각 경기에서 양 팀의 데이터를 각각 활용하여 총 598개의 데이터로 연구를 진행하였다.



한경기 종합기록

대회구분: 대학리그
대 회 명: 2024 KUSF 대학농구 U-리그
경기구분: (남)정규 관중수:
경기번호: 1 경기장: 고려대학교체육관
경 기 일: 2024-03-19

	TEAM	1Q	2Q	3Q	4Q	TOT	EX	TOT
H	고려대학교	22	16	27	16	81		81
A	한양대학교	8	14	10	19	51		51

최다점수차	37점	H	71
4쿼터	8분 23초	A	34

FBP	TR	TTO
21	2	1
2	9	2

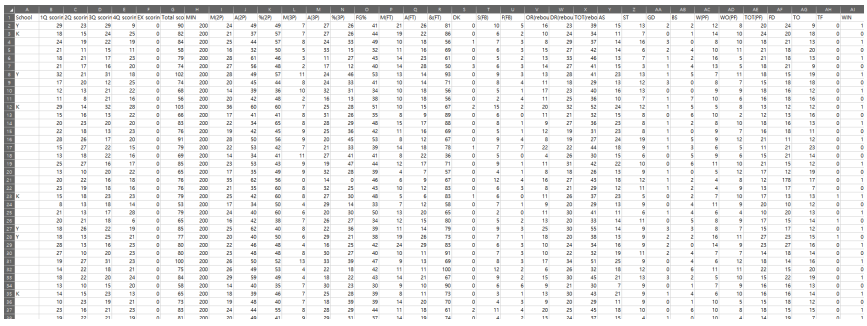
고려대학교

		SCORING						MIN	2P			3P			FG%	FT			FB	REBOUNDS				AS	ST	GD	BS	PF			FD	TO	TF
		1Q	2Q	3Q	4Q	EX	TOT		M	A	%	M	A	%		M	A	%	DK	S	F	OR	DR	TOT				W	WO	TOT			
0	이건희				4		4	6:39		2						4	4	100		1		1		1		1					3		
1	유민수			6	4		10	18:21	3	6	50	1	1	100	57	1	3	33		3			4	4	1	3		1	2		2	2	2
2	심우인				2		2	10:07					2			2	2	100					1	1	2					1	2	1	
+ 3	이동근	10	2	7	4		23	28:57	7	9	78	2	7	29	56	3	3	100		2		6	8	14	1	1	7	2		2	4	1	
5	김도윤							17:02														1	1	2	3	1			3	3	2	2	
6	박정환	2					2	23:10	1	2	50		2		25							1	1	7	1	1					3	2	
+ 7	석준희	5	6	2			13	18:10	3	4	75	2	3	67	71	1	2	50		3		1	3	4	3						2	2
+ 10	윤기찬		3	4			7	22:33				2	8	25	25	1	2	50				2	2	4					2	2	1	1
+ 13	박준형	5		2			7	26:03	2	3	67	1	1	100	75					1		1	4	5		4	2	2		2			
21	정승환				2		2	6:39	1	1	100		1		50														1		1		
+ 24	문유현		5	6			11	22:19	4	6	67	1	4	25	50					1		1	1	2	4	1	1			1	1	1	3
	TOTAL	22	16	27	16		81	200:00	21	33	64	9	29	31	48	12	16	75		11		13	25	38	21	12	2	10	8	7	15	19	13

한양대학교

		SCORING						MIN	2P			3P			FG%	FT			DK	FB	REBOUNDS			AS	ST	GD	BS	PF			FD	TO	TF
		1Q	2Q	3Q	4Q	EX	TOT		M	A	%	M	A	%		M	A	%		S	F	OR	DR	TOT				W	WO	TOT			
+ 1	박한재		3		2		5	33:24	1	4	25	1	5	20	22					1	1	1	2	3	2	4		1	3	1	4	2	6
+ 2	김선우		6	1	1		8	35:28	2	7	29		3		20	4	6	67		1	1	1	5	6	3	2						3	2
+ 3	조민근		2	1	4		7	32:00		1		1	3	33	25	4	6	67		1	1	1	1	2	4				1	3	4	4	1
+ 5	김주형			3	7		10	37:03	3	11	27	1	3	33	29	1	2	50				1	2	3	4				3	2	5	3	3
10	김지훈			2			2	17:31	1	5	20				20					2		1	1		1				1	1		1	
+ 11	신지형		4	3	6		13	31:12	4	11	36	1	2	50	38	2	2	100				8	6	14		2				4	4	3	4
22	홍정웅	2	4				6	13:22	3	5	60				60							1	1	2			1	1			1		1
4	마윤재																																
6	허동근																																
8	김현우																																
9	김현빈																																
13	임희찬																																
	TOTAL	8	14	10	19		51	200:00	14	44	32	4	16	25	30	11	16	69		1	5	13	18	31	13	9	2	9	10	19	15	18	

경기 기록지 예시



활용한 데이터셋 구조

2.2. 실험 구조

데이터 전처리 및 기본적인 분석을 통해 다양한 피처들로 구성된 원시 데이터를 정리하고 불필요한 값을 제거하여 분석에 적합한 형태로 가공하였다. 그 다음, 여러 분류 모델을 적용하여 승리 여부와 관련된 피처들을 추출하였다. 이러한 과정을 통해 경기 결과에 중요한 영향을 미치는 주요 요소들을 식별할 수

있었다. 이후, 선별된 피처들을 바탕으로 고연전 경기 결과를 예측하는 모델을 구축하였다. 이 예측 모델은 다양한 머신러닝 기법을 활용하여 성능을 최적화하였으며, 이를 통해 경기 결과를 보다 정확하게 예측할 수 있었다.

2.3. 방법론

2.3.1. 피처 추출

본 연구에서는 경기 결과에 영향을 미치는 피처를 추출하기 위해 8개의 다른 분류 모델을 활용하였다. 특히 고려대와 연세대의 경기는 가중치를 부여하여 보다 정확한 분석을 진행하였다. 데이터셋은 훈련 데이터와 테스트 데이터의 비율을 8:2로 나누어 사용하였으며, 하이퍼파라미터 선정에는 그리드 서치 기법을 활용하였다. 그리드 서치에서는 param_grid_gb, scoring='accuracy', cv=5, n_jobs=-1의 설정을 사용하였다.

연구에 사용된 방법론으로는 상관관계 분석, 주성분 분석(PCA), 특이값 분해(SVD), 로지스틱 회귀, 랜덤 포레스트, 그래디언트 부스팅(Gradient Boosting), 라쏘 회귀, 인공신경망 등이 있다. 각 분류 모델은 다양한 하이퍼파라미터를 적용하고 교차검증을 통해 최적의 파라미터를 선정하여 모델의 성능을 최적화하였다. 이러한 과정을 통해 경기 결과 예측의 정확성을 높일 수 있었다.

2.3.2. 하이퍼파라미터 후보

본 연구에서는 경기 승패에 관련된 요인 분석을 위해 다양한 모델의 하이퍼파라미터 후보를 선정하였다.

로지스틱 회귀에서는 penalty로 L1, L2, elasticnet, none을, C 값으로 0.01, 0.1, 1, 10, 100을 고려하였다. solver는 lbfgs, liblinear, saga를 사용하였으며, class_weight는 고려대와 연세대 경기 가중치로 1.0, 1.5, 2.0을 적용하였다. 또한 l1_ratio는 0, 0.5, 1을 후보로 하였다.

랜덤 포레스트에서는 n_estimators를 100, 200으로 설정하고, max_features는 auto를 고려하였다. max_depth는 None, 10으로 설정하고, min_samples_split은 2, 5을, min_samples_leaf는 1, 2로 하였다. bootstrap은 True와 False를, class_weight는 1.0, 1.5, 2.0을 후보로 삼았다.

Gradient Boosting에서는 n_estimators를 100, 200, 300으로, learning_rate는 0.01, 0.1, 0.2로 설정하였다. subsample은 0.8과 1.0을, max_depth는 3, 5, 7을, min_samples_split은 2, 5, 10을, min_samples_leaf는 1, 2, 4를, max_features는 auto, sqrt, log2를 후보로 선정하였다.

라쏘 회귀에서는 alpha를 0.001, 0.01, 0.1, 1.0으로, tol은 0.0001, 0.001, 0.01로 설정하였고, selection은 cyclic과 random을 후보로 삼았다.

인공신경망에서는 hidden_layer_sizes를 (50, 50), (100, 50)으로 설정하고, activation은 tanh, relu를 고려하였다. solver는 sgd와 adam을, alpha는 0.0001, 0.001을, learning_rate는 constant를, learning_rate_init은 0.001과 0.01을 후보로 선정하였다.

이러한 하이퍼파라미터 후보들을 통해 각 모델의 최적 파라미터를 선정하고, 고연전 경기 결과 예측 모델의 성능을 극대화하였다.

2.3.3. 예측 모델 구축

고연전 경기 결과를 예측하기 위해 다양한 머신러닝 기법을 활용하여 모델을 구축하였다. 사용된 기법으로는 로지스틱 회귀, 랜덤 포레스트, Gradient Boosting, 다층 퍼셉트론(MLP), 서포트 벡터 머신(SVM)이 있다. 이들 예측 모델에는 피처 중요도가 높은 8개의 피처 세트를 각각 적용하여 총 40개의 모델을 구성하였다. 데이터셋은 훈련 데이터와 테스트 데이터의 비율을 8:2로 나누어 사용하였으며, 하이퍼파라미터 선정에는 그리드 서치 기법을 활용하였다. 그리드 서치 설정은 param_grid_gb, scoring='accuracy', cv=5, n_jobs=-1로 진행되었다. 이러한 과정을 통해 각 모델의 성능을 최적화하고 경기 결과 예측의 정확성을 높이는 데 주력하였다.

2.3.4. 예측 모델에서 활용한 하이퍼파라미터 목록

본 연구에서는 고연전 경기 결과를 예측하기 위해 다양한 모델의 하이퍼파라미터 후보를 선정하였다.

로지스틱 회귀 모델에서는 penalty를 l1, l2, elasticnet으로 설정하고, C 값으로는 0.1, 1, 10을 고려하였다. solver는 lbfgs를 사용하였으며, max_iter는 100과 200을 설정하였다. 또한, l1_ratio는 0, 0.5, 1을 후보로 하였다.

랜덤 포레스트 모델에서는 n_estimators를 100, 200으로 설정하고, max_features는 auto로 고려하였다. max_depth는 None과 10으로 설정하고, min_samples_split은 2와 5를, min_samples_leaf는 1과 2를 후보로 삼았다. bootstrap은 True로 설정하였다.

Gradient Boosting 모델에서는 learning_rate를 0.1, 0.2로, n_estimators는 100으로 설정하였다. max_depth는 3과 5를, subsample은 0.8과 1.0을, min_samples_split은 2와 5를 후보로 선정하였다.

MLP 모델에서는 hidden_layer_sizes를 (50, 50), (100, 50), (100, 100)으로 설정하고, activation은 tanh와 relu를 고려하였다. solver는 sgd와 adam을, alpha는 0.001과 0.01을, learning_rate는 constant와 adaptive를 후보로 삼았다.

SVM 모델에서는 C 값을 0.01, 0.1, 1, 10으로 설정하고, kernel은 linear, poly, sigmoid를 고려하였다. gamma는 scale과 auto를 후보로 삼았으며, probability는 True로 설정하였다.

이러한 하이퍼파라미터 후보들을 통해 각 모델의 최적 파라미터를 선정하고, 고연전 경기 결과 예측 모델의 성능을 극대화하였다.

3. Results

3.1. 분류 모델 최적의 파라미터

본 연구에서는 경기 승패에 관련된 요인 분석을 위해 다양한 머신러닝 모델을 활용하여 최적의 하이퍼파라미터를 선정하였다. 각 모델의 최적 하이퍼파라미터와 교차 검증 점수는 다음과 같다.

로지스틱 회귀 모델에서는 C=1, class_weight={0: 1.5, 1: 1.0}, l1_ratio=0.5, penalty=elasticnet, solver=saga가 최적의 하이퍼파라미터로 선정되었으며, 교차 검증 점수는 0.864로 나타났다.

랜덤 포레스트 모델에서는 bootstrap=False, class_weight={0: 1.0, 1: 1.0}, max_depth=None,

max_features=auto, min_samples_leaf=1, min_samples_split=2, n_estimators=200이 최적의 하이퍼파라미터로 선정되었으며, 교차 검증 점수는 0.816이었다.

Gradient Boosting 모델에서는 learning_rate=0.2, max_depth=7, max_features=auto, min_samples_leaf=1, min_samples_split=5, n_estimators=200, subsample=0.8이 최적의 하이퍼파라미터로 선정되었으며, 교차 검증 점수는 0.843이었다.

라쏘 회귀 모델에서는 alpha=0.001, selection=random, tol=0.01이 최적의 하이퍼파라미터로 선정되었으며, 교차 검증 점수는 0.502로 나타났다.

인공신경망 모델에서는 activation=relu, alpha=0.0001, hidden_layer_sizes=(100, 50), learning_rate=constant, learning_rate_init=0.001, solver=adam이 최적의 하이퍼파라미터로 선정되었으며, 교차 검증 점수는 0.858이었다.

모델	교차검증 점수
로지스틱 회귀	0.864
랜덤 포레스트	0.816
Gradient Boosting	0.843
라쏘 회귀	0.502
인공신경망	0.858

3.2. 피처 중요도 결과

	Feature	Correlation	PCA	SVD	Logistic Regression	Random Forest	Gradient Boosting	Lasso Regression	MLP Classifier	Average Importance
13	TOT(rebounds)	0.194083	0.078093	0.078093	2.379297	0.138798	0.133276	0.233902	0.210000	0.430693
7	%(2P)	0.356203	0.087607	0.087607	1.194045	0.051741	0.030349	0.102003	0.034167	0.242965
8	M(3P)	0.346312	0.110487	0.110487	1.059113	0.023633	0.007360	0.128229	-0.006667	0.222369
4	Total scoring	0.439936	0.125424	0.125424	0.541677	0.141801	0.226053	0.119505	0.005000	0.215602
15	ST	0.094996	0.036641	0.036641	1.187563	0.044555	0.047789	0.115939	0.076667	0.205099
11	M(FT)	0.237373	0.061221	0.061221	0.791753	0.024338	0.014927	0.076954	0.012500	0.160036
5	M(2P)	0.404681	0.132808	0.132808	0.000000	0.109672	0.132577	-0.000000	0.021667	0.116777
0	1Q scoring	0.569224	0.075284	0.075284	0.107974	0.053618	0.041401	-0.005195	0.015000	0.116574
1	2Q scoring	0.514916	0.073881	0.073881	0.168957	0.046259	0.021546	0.000000	0.004167	0.112951
14	AS	0.146054	0.112410	0.112410	0.323753	0.067394	0.042820	0.018613	0.004167	0.103453
16	GD	0.085474	0.032974	0.032974	0.483452	0.012608	0.006766	0.053607	-0.001667	0.088273
2	3Q scoring	0.505725	0.079852	0.079852	-0.094533	0.033438	0.021879	-0.035879	0.005000	0.074417
10	%(3P)	0.241370	0.099477	0.099477	0.000000	0.041882	0.054709	-0.027071	0.011667	0.065189
3	4Q scoring	0.448311	0.072228	0.072228	-0.254156	0.037841	0.043109	-0.053810	-0.001667	0.045510
17	BS	0.082023	0.036416	0.036416	0.112555	0.028205	0.030695	0.012092	-0.015833	0.040321
6	A(2P)	0.363094	0.111601	0.111601	-0.825402	0.038159	0.021998	-0.086536	0.008333	-0.032144
12	A(FT)	0.229255	0.056763	0.056763	-1.060275	0.026406	0.027501	-0.113352	0.019167	-0.094722
18	TO	0.051558	0.029839	0.029839	-1.507028	0.047967	0.064636	-0.139436	0.066667	-0.169495
9	A(3P)	0.294267	0.067648	0.067648	-1.775072	0.031685	0.030610	-0.206758	0.044167	-0.180726

3.3. 각 분류 모델 별 평균 이상의 중요도를 가진 피처 모음

본 연구에서는 각 분류 모델별로 평균 이상의 중요도를 가진 피처들을 식별하였다. 이를 통해 예측 모델의

성능을 높일 수 있는 주요 피처들을 도출하였다.

상관관계 분석을 통해 중요한 피처로는 총 득점(Total scoring), 2점슛 성공 개수(M(2P)), 총 리바운드(TOT(rebounds)), 어시스트(AS), 1쿼터 득점(1Q scoring), 2점슛 성공률(%(2P)), 2쿼터 득점(2Q scoring), 3쿼터 득점(3Q scoring), 2점슛 시도 개수(A(2P))가 선정되었다.

PCA와 SVD 분석에서는 2점슛 성공 개수(M(2P)), 총 득점(Total scoring), 어시스트(AS), 2점슛 시도 개수(A(2P)), 3점슛 성공 개수(M(3P)), 3점슛 성공률(%(3P)), 2점슛 성공률(%(2P)), 3쿼터 득점(3Q scoring), 총 리바운드(TOT(rebounds))가 중요한 피처로 나타났다.

로지스틱 회귀 모델에서는 총 리바운드(TOT(rebounds)), 2점슛 성공률(%(2P)), 스틸(ST), 3점슛 성공 개수(M(3P)), 자유투 성공 개수(M(FT)), 총 득점(Total scoring), 득실차(GD), 어시스트(AS), 2쿼터 득점(2Q scoring)이 중요한 피처로 선정되었다.

랜덤 포레스트 모델에서는 총 득점(Total scoring), 총 리바운드(TOT(rebounds)), 2점슛 성공 개수(M(2P)), 어시스트(AS), 1쿼터 득점(1Q scoring)이 주요 피처로 나타났다.

Gradient Boosting 모델에서는 총 득점(Total scoring), 총 리바운드(TOT(rebounds)), 2점슛 성공 개수(M(2P)), 턴오버(TO), 3점슛 성공률(%(3P))이 중요한 피처로 선정되었다.

라쏘 회귀 모델에서는 총 리바운드(TOT(rebounds)), 3점슛 성공 개수(M(3P)), 총 득점(Total scoring), 스틸(ST), 2점슛 성공률(%(2P)), 자유투 성공 개수(M(FT)), 득실차(GD), 어시스트(AS), 블록슛(BS)이 중요한 피처로 나타났다.

MLP 분류 모델에서는 총 리바운드(TOT(rebounds)), 스틸(ST), 턴오버(TO), 3점슛 시도 개수(A(3P)), 2점슛 성공률(%(2P))이 주요 피처로 선정되었다.

3.4. 각 예측 모델 별 최종 파라미터 값 및 교차검증 정확도

본 연구에서는 고연전 경기 결과 예측을 위해 다양한 머신러닝 모델과 피처 세트를 활용하여 최적의 하이퍼파라미터를 선정하고, 교차 검증을 통해 모델의 성능을 평가하였다.

로지스틱 회귀 모델에서는 다양한 피처 세트에 대해 동일한 최적의 하이퍼파라미터(C=0.1, max_iter=200, penalty=l2, solver=lbfgs)가 도출되었으며, 교차 검증 점수는 대부분의 경우 0.9056을 기록하였다. 단, MLP Classifier 피처 세트의 경우 C=10이 최적의 하이퍼파라미터로 선정되었고, 교차 검증 점수는 0.9500이었다.

랜덤 포레스트 모델에서는 피처 세트에 따라 최적의 하이퍼파라미터가 조금씩 달라졌다. 대부분의 경우 bootstrap=True, max_depth=None, max_features=auto, min_samples_leaf=1, min_samples_split=2, n_estimators=100 또는 200이 최적의 하이퍼파라미터로 선정되었다. 교차 검증 점수는 대부분 0.9056을 기록하였으나, Gradient Boosting 피처 세트와 Lasso Regression 피처 세트의 경우 0.9306을 기록하였다.

Gradient Boosting 모델에서는 피처 세트에 따라 learning_rate와 max_depth, min_samples_split, subsample 등이 최적의 하이퍼파라미터로 선정되었다. 교차 검증 점수는 대부분 0.8583에서 0.9306 사이였으며, MLP Classifier 피처 세트의 경우 0.9528을 기록하였다.

MLP 모델에서는 hidden_layer_sizes, activation, solver 등의 하이퍼파라미터가 최적화되었으며, 교차 검증

점수는 대부분 0.9056에서 0.9306 사이였다. 특히 MLP Classifier 피처 세트의 경우 0.9750을 기록하였다.

SVM 모델에서는 대부분의 피처 세트에 대해 C=0.01, gamma=scale, kernel=linear, probability=True가 최적의 하이퍼파라미터로 선정되었으며, 교차 검증 점수는 0.9056을 기록하였다. MLP Classifier 피처 세트의 경우 C=10이 최적의 하이퍼파라미터로 선정되었고, 교차 검증 점수는 0.9750이었다.

다음은 각 모델의 최적 교차 검증 점수를 정리한 표이다.

	Correlation	PCA	SVD	Logistic Regression	Random Forest	Gradient Boosting	Lasso Regression	MLP Classifier
Logistic Regression	0.9056	0.9056	0.9056	0.9056	0.9056	0.9056	0.9056	0.95
Random Forest	0.9056	0.9056	0.9056	0.9056	0.8833	0.9306	0.9306	0.9306
Gradient Boosting	0.8611	0.8583	0.8583	0.9306	0.8833	0.8806	0.9306	0.9528
MLP	0.9056	0.925	0.925	0.8833	0.9056	0.9306	0.8833	0.975
SVM	0.9056	0.9056	0.9056	0.9056	0.9056	0.9056	0.9056	0.975

3.5. 경기 결과 예측

각 모델별 고연전 경기 결과는 다음과 같다.

승리팀		Feature Set						
		Corr	PCA, SVD	LR	RF	GB	Lasso	MLP
Model	LR	Korea	Korea	Korea	Korea	Yonsei	Korea	Korea
	RF	Korea	Korea	Korea	Korea	Yonsei	Korea	Korea
	GB	Korea	Korea	Korea	Korea	Yonsei	Korea	Yonsei
	MLP	Yonsei	Yonsei	Korea	Yonsei	Yonsei	Korea	Korea
	SVM	Korea	Korea	Korea	Korea	Yonsei	Korea	Korea

본 연구에서는 고연전 경기 결과 예측을 위해 다양한 머신러닝 모델과 피처 세트를 활용하였다. 특히, PCA와 SVD의 피처 중요도가 동일하여 중복을 제외하고 총 35개의 예측 모델의 결과를 활용하였다. 이 모델들은 각기 다른 피처 세트를 기반으로 최적의 하이퍼파라미터를 선정하고, 교차 검증을 통해 성능을 평가하였다.

총 35개의 예측 모델 중 9개의 모델은 연세대학교의 승리를 예측하였으며, 나머지 26개의 모델은 고려대학교의 승리를 예측하였다. 이러한 결과는 각 모델이 피처 세트와 하이퍼파라미터에 따라 다른 예측 결과를 제공함을 보여준다.

4. Conclusion

4.1. 기대효과 및 활용방안

본 연구에서는 고연전 경기 결과 예측을 위해 다양한 머신러닝 모델과 피처 세트를 활용하여 최적의 예측 모델을 구축하였다. 이 연구의 기대효과는 다음과 같다.

첫째, 경기 결과 예측의 정확성을 높임으로써 팀의 전략적 의사결정에 도움을 줄 수 있다. 예측 모델을 통해 경기 전에 승패를 예측하고, 이를 바탕으로 선수 기용 및 전술을 조정할 수 있다.

둘째, 선수들의 개별 성과를 평가하고 개선할 수 있는 기초 자료를 제공할 수 있다. 주요 피처들의 중요도를 분석함으로써 어떤 요소들이 경기 결과에 중요한 영향을 미치는지 파악할 수 있다. 이를 통해 선수들의 훈련 및 경기를 보다 효과적으로 계획할 수 있다.

셋째, 대학 농구 리그의 데이터 분석 및 활용 수준을 한 단계 끌어올릴 수 있다. 본 연구의 방법론과 결과는 다른 대학 농구 팀이나 리그에서도 적용 가능하며, 스포츠 데이터 분석의 발전에 기여할 수 있다.

4.2. 연구 한계점

본 연구에는 몇 가지 한계점이 존재한다.

첫째, 데이터의 제한성이다. 2022, 2023, 2024년의 경기 데이터만을 사용하였으며, 더 많은 연도의 데이터를 포함하지 못했다. 이는 모델의 일반화 성능에 영향을 미칠 수 있다.

둘째, 피처 선택의 한계이다. 본 연구에서는 특정 피처 세트에 집중하였으나, 다른 잠재적으로 중요한 피처들을 배제할 가능성이 있다. 다양한 피처를 고려한 추가적인 연구가 필요하다.

셋째, 모델의 복잡성과 해석 가능성이다. 일부 모델은 높은 예측 성능을 보이지만, 해석 가능성이 낮아 실제 적용 시 이해하기 어려울 수 있다. 이를 보완하기 위해 해석 가능한 모델의 활용과 설명 방법이 필요하다.

4.3. 제언

앞으로의 연구에서는 다음과 같은 방향으로 확장할 것을 제언한다.

첫째, 더 많은 데이터를 포함한 분석이 필요하다. 추가적인 연도와 다양한 경기 데이터를 포함하여 모델의 일반화 성능을 높일 수 있을 것이다.

둘째, 다양한 피처를 고려한 모델링이 필요하다. 경기 결과에 영향을 미칠 수 있는 다양한 피처를 추가적으로 발굴하고, 이를 포함한 모델을 구축함으로써 예측 성능을 향상시킬 수 있을 것이다.

셋째, 해석 가능한 모델의 개발이 필요하다. 높은 예측 성능과 함께 모델의 결과를 쉽게 이해하고 해석할 수 있는 방법론을 개발함으로써 실제 경기 전략 수립에 유용하게 활용할 수 있을 것이다.

넷째, 실시간 데이터 분석 및 예측 시스템의 구축을 제언한다. 경기 중 실시간 데이터를 기반으로 예측을 수행하고, 이를 통해 경기 도중에도 전략적 결정을 지원할 수 있는 시스템을 개발하는 것이 유용할 것이다.

5. References

장성용, 김현수, 구승환. (2009). 국내 남자 프로농구 승패 예측 모형 비교 연구. 체육과학연구, 20(4), 704-711.

Yi, J. H., & Lee, S. W. (2020). Prediction of English Premier League Game Using an Ensemble Technique. KIPS Transactions on Software and Data Engineering, 9(5), 161–168. https://doi.org/10.3745/KTSDE.2020.9.5.161

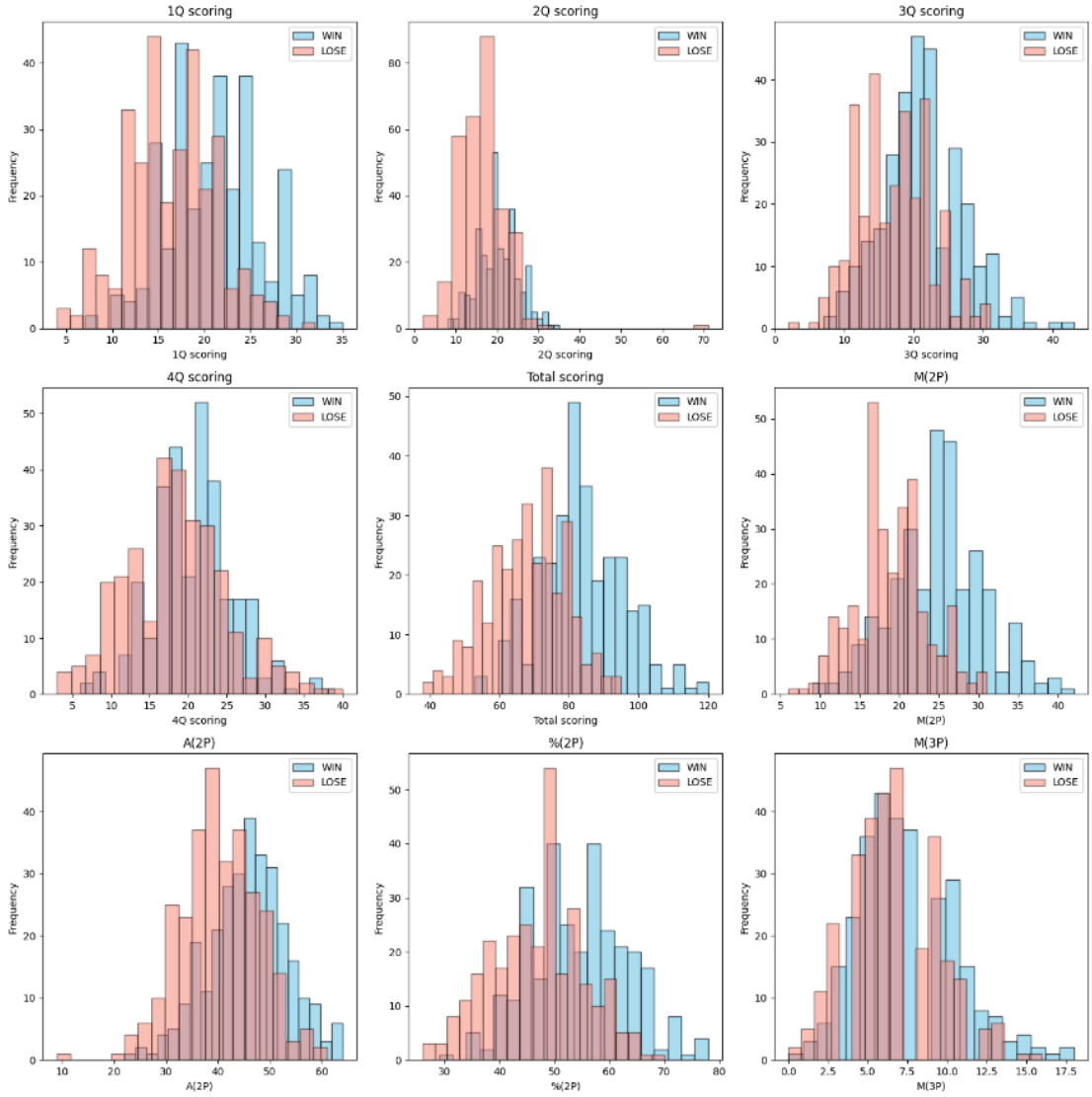
박제영. (2008). 2006-2007시즌 한국프로 농구경기의 승・패 요인 분석. 한국체육과학회지, 17(2), 129-138.

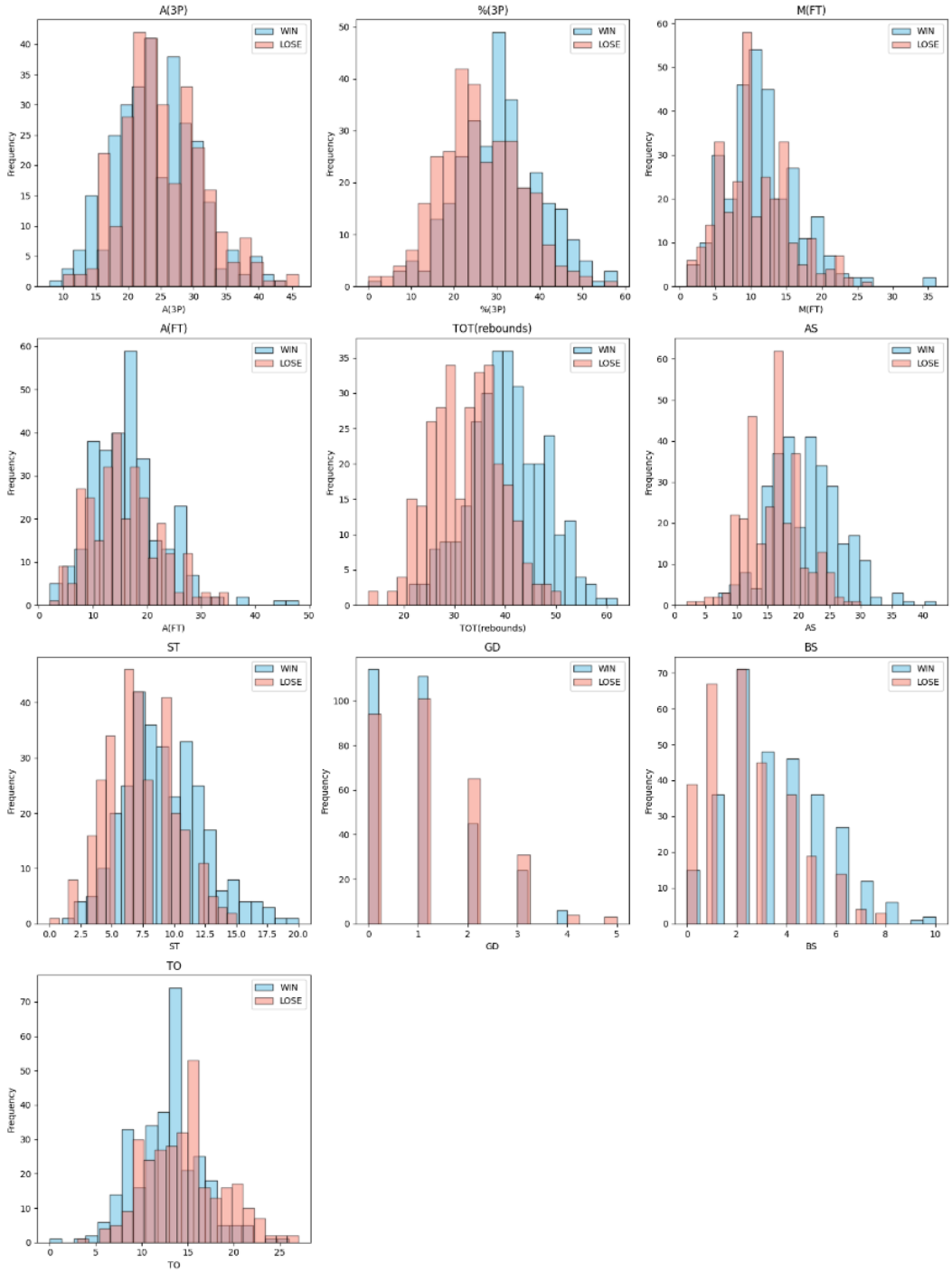
김다희, 신진이, 김현승, 홍성봉. (2022). 한국대학농구 경기력 결정 요인 분석. 한국스포츠학회, 20(3), 689-699.

예원진, 이성노. (2022). 2022 FIBA 남자농구 아시안컵 경기결과를 활용한 머신러닝 분류 모형의 예측 성능 비교. 한국체육측정평가학회지, 24(3), 53-69.

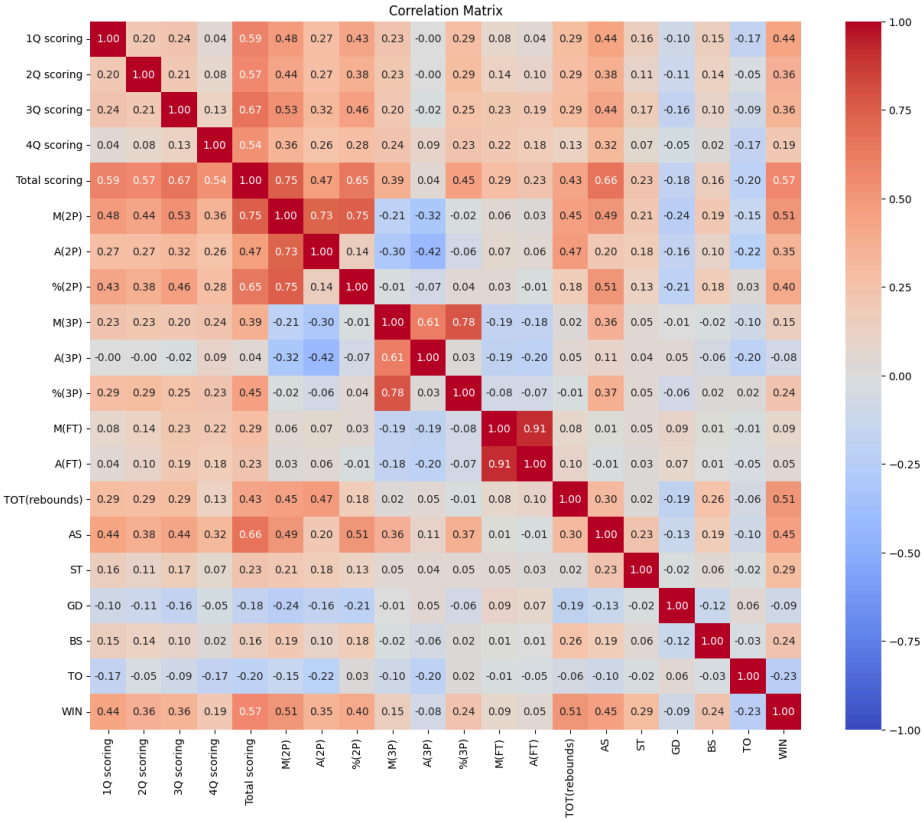
6. 부록

각 피처별 분포 히스토그램

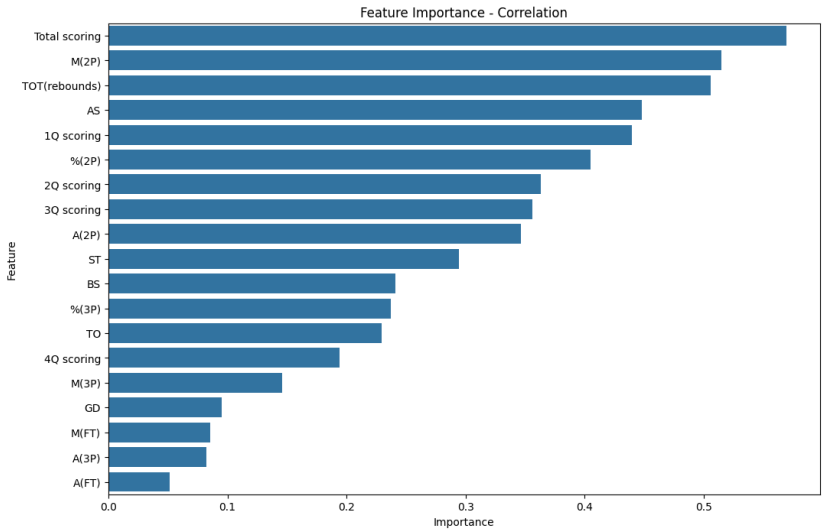

1Q scoring
2Q scoring
3Q scoring
4Q scoring
Total scoring
M(2P)
A(2P)
%(2P)
M(3P)
WIN
LOSE
Frequency

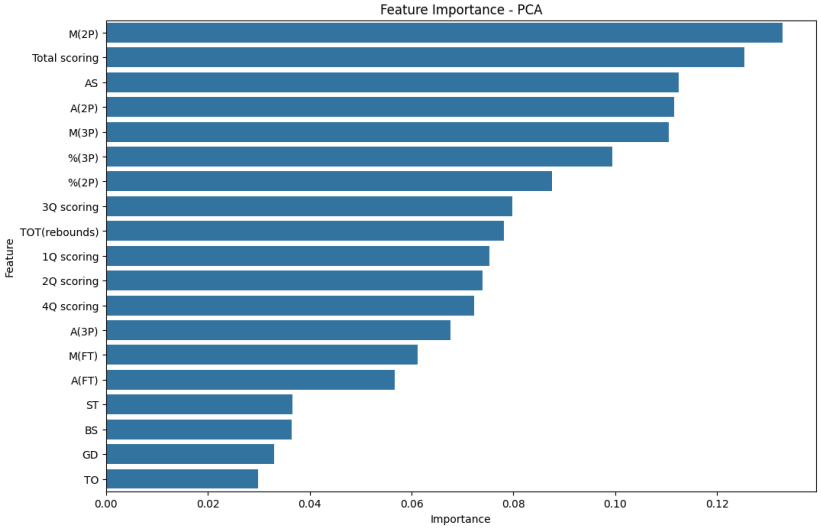


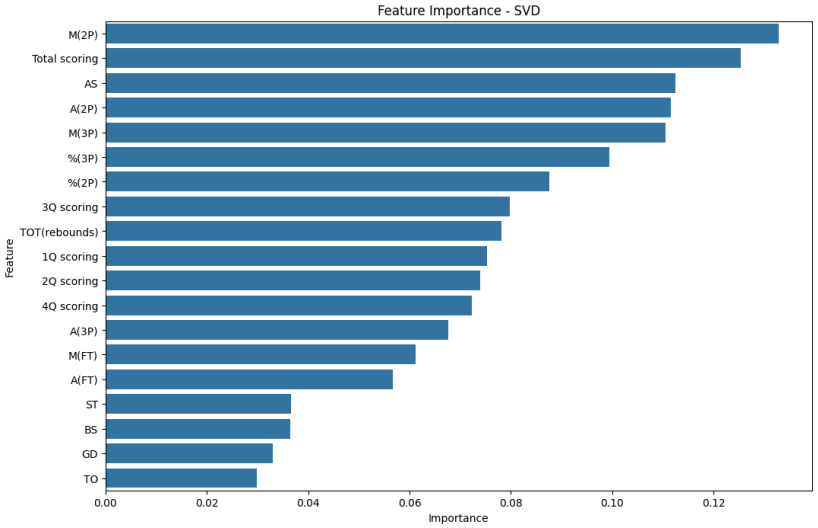
각 피처별 상관관계 분석 그래프

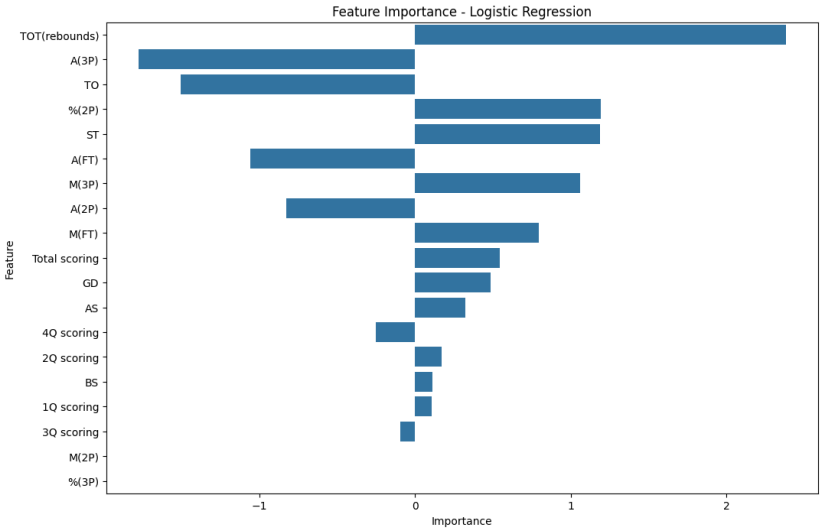


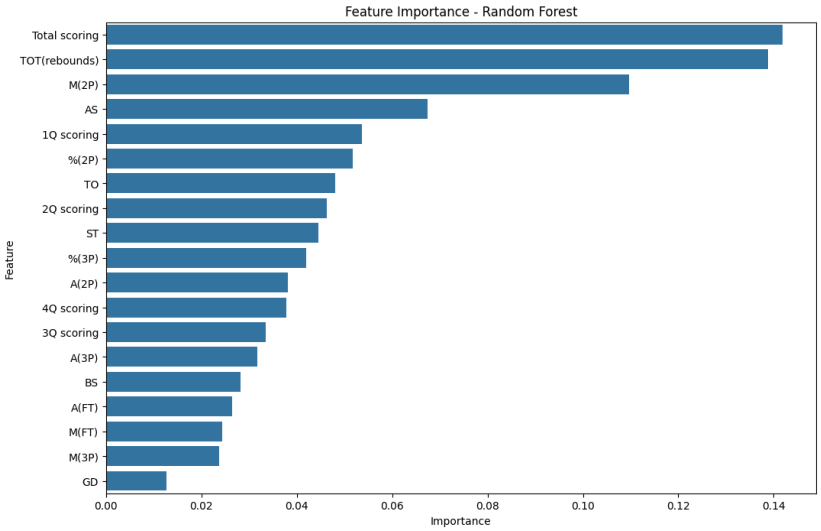
분류 모델별 피처 중요도 시각화

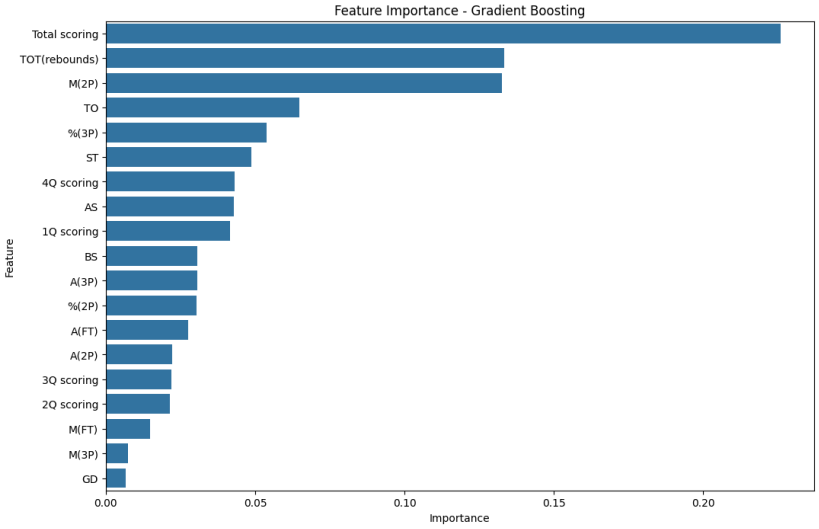


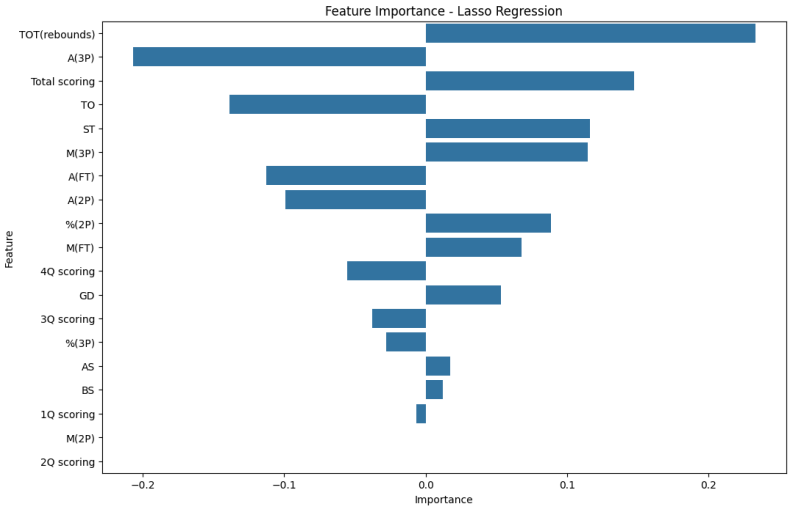

Feature Importance - PCA
M(2P)
Total scoring
AS
A(2P)
M(3P)
%(3P)
%(2P)
3Q scoring
TOT(rebounds)
1Q scoring
2Q scoring
4Q scoring
A(3P)
M(FT)
A(FT)
ST
BS
GD
TO
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance

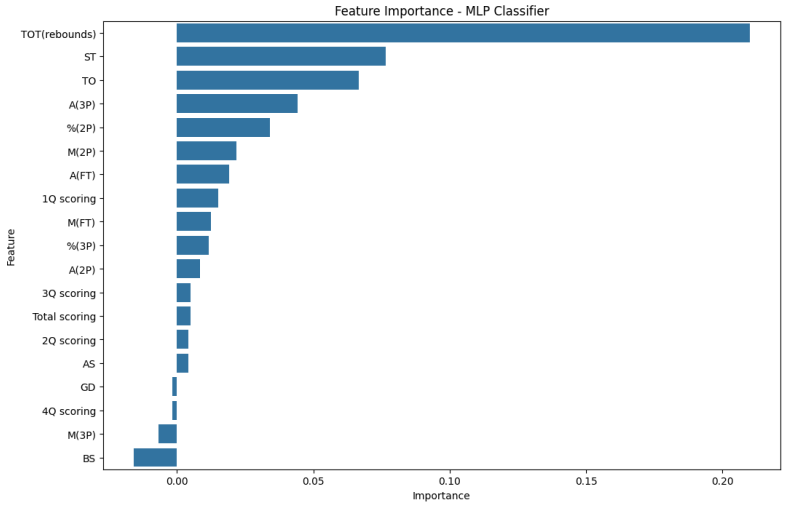

Feature Importance - SVD
M(2P)
Total scoring
AS
A(2P)
M(3P)
%(3P)
%(2P)
3Q scoring
TOT(rebounds)
1Q scoring
2Q scoring
4Q scoring
A(3P)
M(FT)
A(FT)
ST
BS
GD
TO
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance


Feature Importance - Logistic Regression
TOT(rebounds)
A(3P)
TO
%(2P)
ST
A(FT)
M(3P)
A(2P)
M(FT)
Total scoring
GD
AS
4Q scoring
2Q scoring
BS
1Q scoring
3Q scoring
M(2P)
%(3P)
Feature
−1
0
1
2
Importance


Feature Importance - Random Forest
Total scoring
TOT(rebounds)
M(2P)
AS
1Q scoring
%(2P)
TO
2Q scoring
ST
%(3P)
A(2P)
4Q scoring
3Q scoring
A(3P)
BS
A(FT)
M(FT)
M(3P)
GD
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Importance


Feature Importance - Gradient Boosting
Total scoring
TOT(rebounds)
M(2P)
TO
%(3P)
ST
4Q scoring
AS
1Q scoring
BS
A(3P)
%(2P)
A(FT)
A(2P)
3Q scoring
2Q scoring
M(FT)
M(3P)
GD
Feature
0.00
0.05
0.10
0.15
0.20
Importance


Feature Importance - Lasso Regression
TOT(rebounds)
A(3P)
Total scoring
TO
ST
M(3P)
A(FT)
A(2P)
%(2P)
M(FT)
4Q scoring
GD
3Q scoring
%(3P)
AS
BS
1Q scoring
M(2P)
2Q scoring
Feature
−0.2
−0.1
0.0
0.1
0.2
Importance


Feature Importance - MLP Classifier
TOT(rebounds)
ST
TO
A(3P)
%(2P)
M(2P)
A(FT)
1Q scoring
M(FT)
%(3P)
A(2P)
3Q scoring
Total scoring
2Q scoring
AS
GD
4Q scoring
M(3P)
BS
Feature
0.00
0.05
0.10
0.15
0.20
Importance