Introduzione

Il campionato di pallacanestro NCAA Division I rappresenta il livello più alto del basket collegiale negli USA. Le università che partecipano alla Division I sono circa 350 e sono raggruppate in 32 Conference (leghe formate da un numero variabili di squadre, di solito da 7 a 16). Le Conference non vengono stabilite dalla NCAA ma sono leghe indipendenti formate con accordi e contratti fra le singole università. Nel 1939 è stata introdotta una postseason a livello nazionale, istituendo il torneo NCAA. Il numero delle squadre invitate a farvi parte è cresciuto negli anni: oltre alle 32 vincitrici delle rispettive conference, vi sono squadre invitate in base ai meriti sportivi conseguiti durante il campionato, oggi il numero di squadre partecipanti è 68. La scelta degli inviti viene fatta da un comitato nella cosiddetta "Selection Sunday" nella quale vengono stabilite anche le teste di serie. Solitamente le major conference (ACC, Big East, Big 10, Big 12, PAC-12, SEC) hanno un numero di squadre variabile ma sempre superiore a 1. Il torneo NCAA si svolge durante il mese di marzo e viene chiamato quindi "March Madness", il weekend finale della programmazione del torneo è dedicato alle Final Four, che si compongono in semifinali e finale.
Ovviamente le scopo che si è prefissato è alquanto ambizioso data la grande numerosità delle università partecipanti e anche dal fatto che i criteri di ammissione alla March Madness non sono sempre oggettivi. E', inoltre, importante chiarire che il lavoro qui presentato si basa su dati parziali dal momento che la stagione non si è ancora conclusa, quindi si può valutare come una previsione delle quattro squadre che si giocheranno la vittoria finale se la stagione si fosse conclusa il 9/1/20.

Panoramica e Motivazione

La motivazione che mi spinge ad aver intrapreso questo progetto è che esso riesce ad accumunare due mie grandi passioni: il basket e l'analisi dei dati.
L'idea di base è stata quella di prevedere le semifinaliste dell'anno corrente basandosi sui dati relativi ai cinque anni precedenti fornendo anche un'indagine sul campionato NCAA nel suo complesso.

Dati

La collezione dei dati è stata effettuata reperendo informazioni da più fonti: da sports-reference sono stati presi i dati statici relativi alle squadre, mentre quelli dinamici (2015-2019 e dati parziali del 2020), relativi alle statistiche avanzate, provengono da bart-torvik. Qui sotto sono elencate e spiegate le variabili sulle quali si è lavorato.

VARIABILI QUALITATIVE				
TEAM	Nome Università	Character	347 Università	
CONF	Conference di appartenenza	Factor	32 modalità	"A10" , "ACC", "AE", "Amer", "ASun", "B10", "B12", "BE", "BSky", "BSth","Ivy "BW", "CAA, "CUSA", "Horz","MAAC", MAC, "MEAC", "MVC", "MWC", "NEC", "OVC", "P12", "Pat", "SB", "SC", "SEC", "Slnd", "Sum", "SWAC", "WAC", "WCC"
POSTESEASON	TARGET - ci dice se la squadra in questione partecipa alla postseason	Factor	9 modalità	"Champions", "Second", "F4", "E8", "S16", "R32", "R64", "R68"
post	TARGET - ci dice il piazzamento findale della squadra	Factor	2 modalità	"yes", "no"

VARIABILI QUANTITATIVE DINAMICHE	
G	Numero di partie giocate in stagione
W	Numero di vittorie durante la stagione
L	Numero di sconfitte durante la stagione
ADJOE	"Adjusted Offense Efficiency", punti segnati ogni 100 pessessi
ADJDE	"Adjusted Defense Efficiency", punti subiti ogni 100 pessessi
pp100p_ratio	ADJOE / ADJDE
BARTHAG	Power Rating, possbilità di vincere contro una squadra di medio livello della Divsion1
EFG_O	Percentuale reale dal campo offensiva
EFG_D	Percentuale reale dal campo difensiva
EFG_ratio	EFG_O / EFG_D
TOR	Turnover Rate, stima del numero di palle perse ogni 100 giochi
TORD	Steal Rate, stima del numero di palle recuperate ogni 100 giochi
TOR_ratio	TOR / TORD
ORB	"Offensive Rebound Percentage", percentuale di rimbalzi offensivi catturati
DRB	"Defensive Rebound Percentage", percentuale di rimbalzi difensivi catturati
FTR	"Free Throw Rate", stima quanto spesso una squadra ha a disposizione un tiro libero FTR = FTA / FGA (tiri liberi tentati su totali tiri tentati)
FTRD	"Free Throw Rate Defense", stima quanto spesso una squadra concede un tiro libero
FTR_ratio	FTR / FTRD
2P_O	Percentuale tiri da 2 punti
2P_D	Percentuale concessa tiri da 2 punti
3P_O	Percentuale tiri da 3 punti
3P_D	Percentuale concessa tiri da 3 punti
ADJ_T	"Adjusted Tempo", stima del "pace" di una squadra
WAB	"Wins Above Bubble", vittorie/punti al di sopra della bolla per bolla ci si riferisce al cutoff che dividi che partecipa alla postseason e chi no

VARIABILI QUANTITATIVE STATICHE	
first_year	PrimO anno di partecipazione ad un campionato NCAA in Division I
age	Età della squadra (anni di partecipazione al campionato di Division I)
all_games_played	Numero di partite giocate in Division I
all_WL_perc	Percentruale di vittorie in tutte le partite giocate in Division I
MM	Numero di partecipazioni alla March Madness
F4	Numero di partecipazioni alle Final Four
CHAMP	Numero di titoli vinti

Metodi statistici

Per entrambe le previsioni si è fatta una media aritmetica delle previsioni elaborate tramite l'allenamento di modelli boosting sui dati dal 2015 al 2019 (nc15, nc16, nc17, nc18, nc19).

Per fare la prima previsione si è usato come target "post". In fase di modellazione oltre al metodo boosting si sono allenati modelli di Random Forest, GLM, regressione Ridge e Lasso, questi ultimi 3 si sono dimostrati molto deboli in confronto agli altri due soprattutto la tecnica lasso che mostrava coefficienti diversi da 0 per 3 o 4 variabili a seconda dell'anno. Si è scelto di usare il metodo boosting perché forniva prestazioni migliori, sia in termini di accuratezza che di AUC, del metodo RF quando applicato a dataset di anni differenti che avevano quindi il ruolo di validation set. Qui sotto si mostra la V.I. dei modelli boosting.

Variable Importance over Years									
2015		2016		2017		2018		2019	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
WAB	50.874	WAB	58.747	WAB	52.321	WAB	53.789	WAB	69.015
W	14.998	WL_perc	14.166	W	6.419	W	7.41	W	8.431
BARTHAG	9.241	BARTHAG	9.127	G	5.711	WL_perc	5.545	G	2.892
WL_perc	5.632	TOR_ratio	4.256	BARTHAG	4.392	ADJ_T	2.6	BARTHAG	2.527
pp100p_ratio	2.406	all_WL_perc	2.733	WL_perc	2.846	FTR	2.437	WL_perc	2.264

Per fare la seconda previsione si è usata la variabile POSTSEASON come risposta che è stata trasformata in dicotomica mettendo le squadre dal primo al 16-esimo posto come “S16” (Sweet 16) e le rimanenti come “no_S16”. Si sono provati gli stessi modelli della fase precedente trovando dei risultati simili. In questo caso però la differenza tra boosting e random forest non è stata evidente. Come scritto anche ne blog-post si è seguita la regola: una squadra ha la possibilità di andare alle F4 solo se viene prevista positivamente in almeno 3 dei 5 modelli utilizzati. Qui sotto viene mostrato il grafico dell’influenza relativa delle esplicative sulla risposta dell’anno 2018. Si nota come WAB non sia più le variabili con più importanza, in questo modello quel ruolo lo ha BARTHAG.