СРАВНЕНИЕ ТРАДИЦИОННЫХ И ПРОДВИНУТЫХ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ В ОПРЕДЕЛЕНИИ УСПЕХА КОМАНД НАЦИОНАЛЬНОЙ БАСКЕТБОЛЬНОЙ АССОЦИАЦИИ

ВВЕДЕНИЕ

Цель исследования

Определить, какой подход (традиционный или современный) к статистическому анализу данных в баскетболе лучше, на основе сравнения регрессионных моделей

Задачи

- Отобрать подходящие предикторы среди традиционных и продвинутых баскетбольных метрик
- Построить 2 модели линейной регрессии на основе выявленных данных
- Оценить качество получившихся моделей и сравнить их
- Сделать конструктивные выводы по применению разных статистических метрик в баскетболе

Методология

База данных - *basketball-reference.com*

База исследования - *Регулярный чемпионат НБА сезона 2018/2019 гг.*

Традиционные метрики (среднее за игру)		*Продвинутые метрики*	
W/L% – процент побед за сезон (целевая переменная)			
G	количество игр	**Age**	средний возраст команды
MP	сыгранные минуты	**W**	количество побед за сезон
FG	реализованные броски с игры	**L**	количество поражений за сезон
FGA	бросковые попытки с игры	**PW / PL**	ожидаемое количество побед / поражений
FG%	процент реализации бросков	**MOV**	средняя разница очков
3P	реализованные трехочковые	**SOS**	сложность расписания
3PA	трехочковые попытки	**SRS**	рейтинг на основе разницы очков и сложности расписания
3P%	процент реализации трехочковых	**ORtg**	количество набранных очков на 100 владений
2P	реализованные двухочковые броски	**DRtg**	количество пропущенных очков на 100 владений
2PA	двухочковые попытки	**NRtg**	разница очков на 100 владений
2P%	процент реализации двухочковых	**Pace**	темп (количество владений на 48 минут)
FT	реализованные штрафные броски	**FTr**	количество попыток штрафных бросков на количество бросковых попыток с игры
FTA	попытки штрафных бросков	**3PAr**	процент трехочковых попыток
FT%	процент реализации штрафных	**TS%**	true shooting
ORB	подборы в атаке	**eFG%**	effective FG%
DRB	подборы в обороне	**TOV%**	количество потерь на 100 владений
TRB	общее количество подборов	**ORB%**	процент собранных подборов в атаке из всех доступных
AST	результативные передачи	**FT/FGA**	количество реализованных штрафных на количество бросковых попыток с игры
STL	количество перехватов	**OppeFG%**	effective FG% оппонента
BLK	количество блокшотов	**OppTOV%**	количество потерь оппонента на 100 владений
TOV	количество потерь	**DRB%**	процент собранных подборов в защите из всех доступных
PF	количество персональных фолов	**OppFT/FGA**	количество реализованных штрафных к количеству бросковых попыток оппонента
PTS	среднее количество очков за игру	**Arena**	название домашней арены
		Attend.	общая посещаемость игр за сезон
		Attend./G	средняя посещаемость домашних матчей за игру

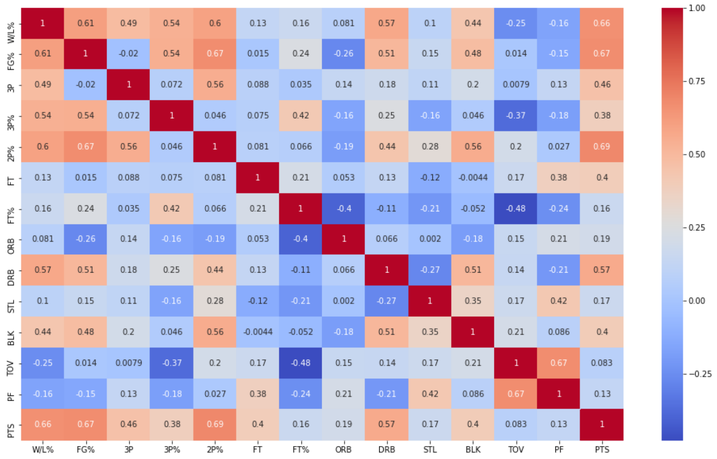
Offensive Four Factors

Defensive Four Factors

$$eFG\% = \frac{100 \times (FG + 0.5 \times 3P)}{FGA}$$

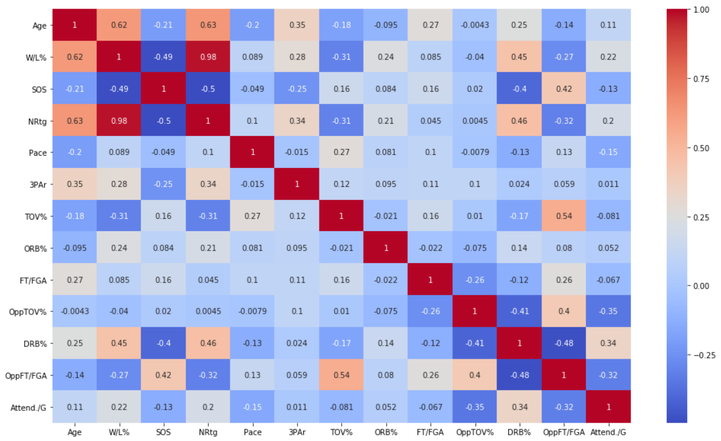
$$TS\% = \frac{100 \times PTS}{2 \times (FGA + 0.44 \times FTA)}$$

ОБРАБОТКА МУЛЬТИКОЛЛИНЕАРНОСТИ



В ходе обработки мультиколлинеарности мы избавились от взаимокррелирующих факторов, убрав те из них, которые слабее влияли на целевую переменную в традиционных метриках:

- FGA и FG (искажали влияние FG%);
- 2PA и 2P (искажали 2P%);
- 3PA и FTA (3P% и FT% соответственно);
- TRB (влияние DRB оказалось сильнее);
- AST (коррелировал с FG%);



Для набора продвинутых метрик такими показателями стали:

- W и L, PW и PL (от них напрямую зависит процент побед);
- OffRtg и DefRtg,
- MOV и SRS,
- TS%, eFG% и OppeFG% (дублировали NRtg);
- FTr и Attend. (производные от более сильных FT/FGA и Attend./G соответственно).

ПОСТРОЕНИЕ РЕГРЕССИОННОЙ МОДЕЛИ

	Традиционные метрики	vs.	Продвинутые метрики
Ошибка на обучении	≈0.0007	>	≈0.0004
Ошибка на валидации	≈0.021	>	≈0.002
Разница	≈0.02	>	≈0.002

Оценив полученные модели линейной регрессии, можно прийти к следующим выводам:

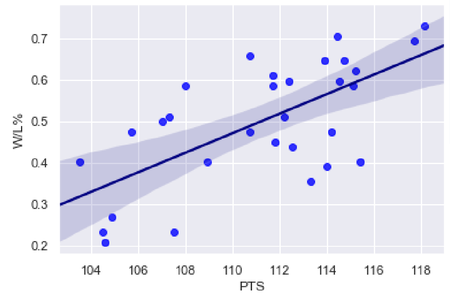
- Наборы как традиционных, так и продвинутых статистических показателей довольно точно определяют целевую переменную (MSE в обоих случаях даже не достигает единицы), что лишний раз объясняет популярность использования статистики в современном баскетболе, который становится все более "цифровым";
- Регрессионная модель на продвинутых метриках чуть более успешна по всем оценкам качества, что говорит о точности и значимости новых показателей, постоянно создаваемых и улучшаемых с течением времени.

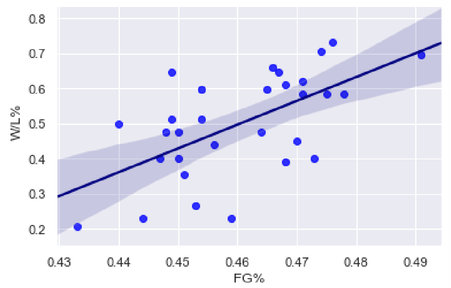
Наиболее сильными предикторами по прямой корреляции для традиционного датасета оказались

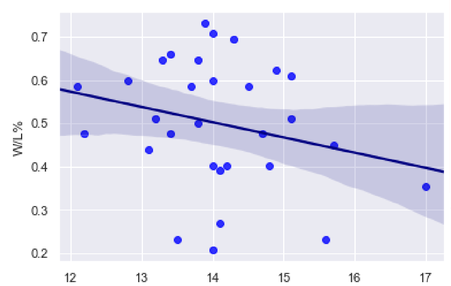
- количество очков (PTS / корреляция: 0.66)
- процент реализации бросков (FG% /корреляция: 0.61).

Следовательно, чем результативнее играет команда и чем точнее ее игроки (причем вес реализации двухочковых больше, чем трехочковых!), тем успешнее она выступает в чемпионате.

- *Визуализация регрессии по традиционным метрикам*

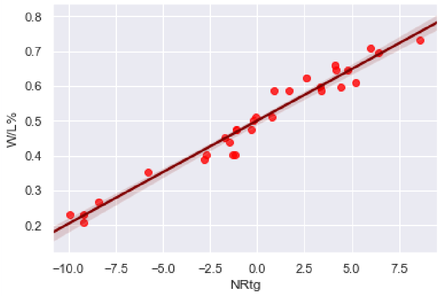


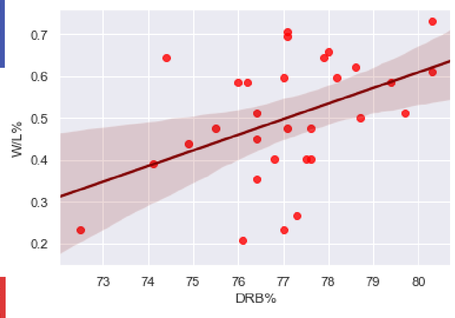


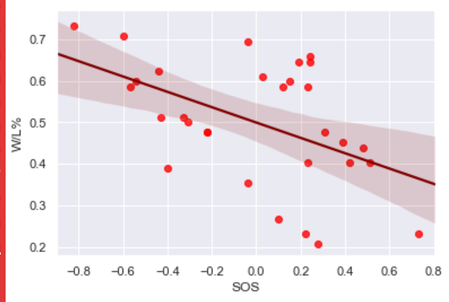


В обратную сторону, в свою очередь, наиболее сильно влияет количество потерь (TOV / -0.25), что говорит о крайней важности технической обученности игроков и качества обращения с мячом.

- *Визуализация регрессии по продвинутым метрикам*







В датасете продвинутых статистик ключевое место занимает NRtg (корреляция - 0.98) - определяющий показатель, показывающий насколько успешно команда переигрывает оппонентов на 100 владений. Немаловажное значение имеет процент собранных подборов в защите (DRB% / 0.45) - не даром говорят, что именно "защита выигрывает чемпионаты".

Обратная корреляция наиболее высока для сложности расписания (SOS / -0.49), которое в настоящее время все явнее требует качественных преобразований и на данный момент является одной из наиболее обсуждаемых проблем в руководстве Ассоциации.

Авторство проекта:

- Мека Василий
- Соляр Евгений

НИУ ВШЭ Факультет Мировой Экономики и Мировой Политики, ОП "Международные отношения", группа БМО-183