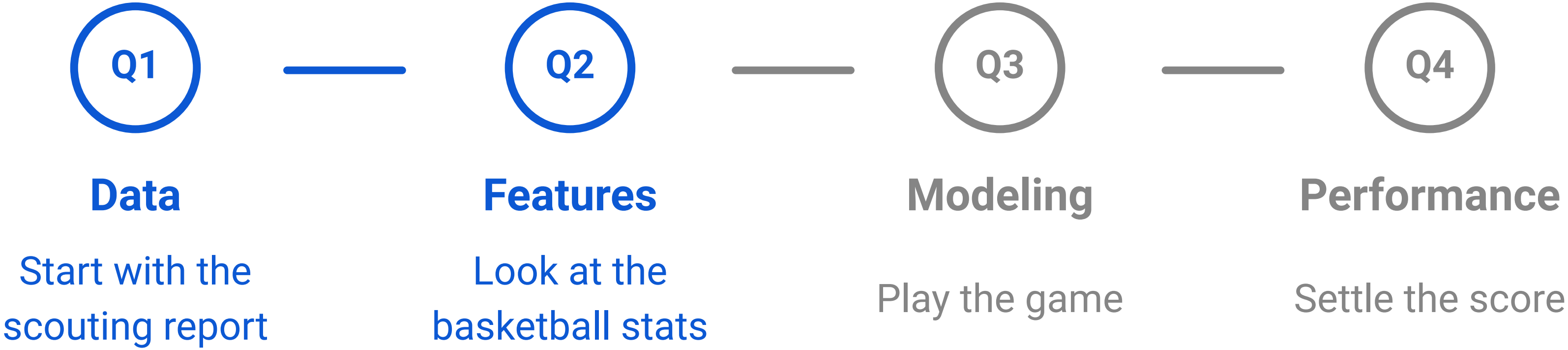
College Basketball

MODEL MADNESS

by Allen Chen

The Game Plan



Q1

Data

Start with the scouting report

Q2

Features

Look at the basketball stats

Q3

Modeling

Play the game

Q4

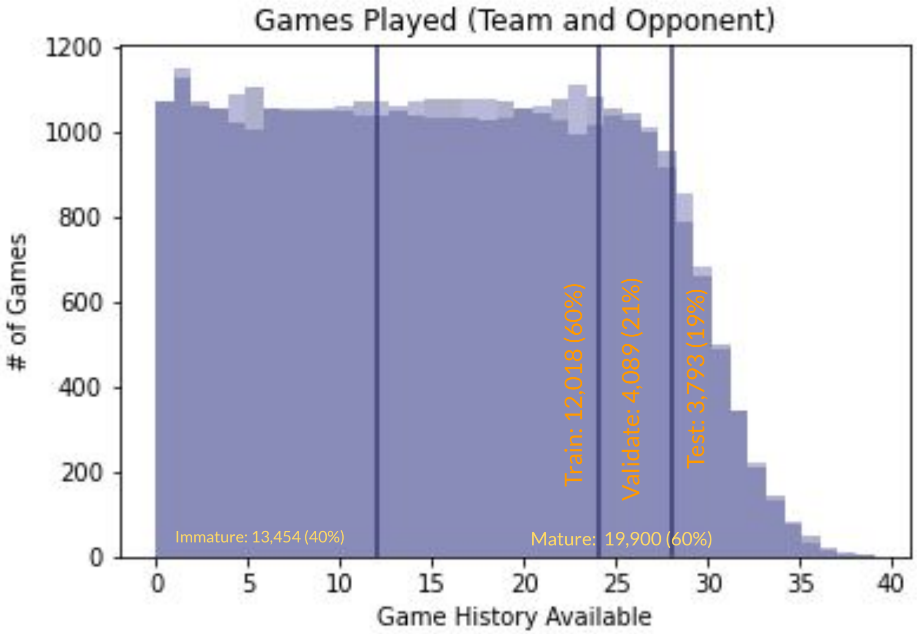
Performance

Settle the score

Scouting Report

Data

- Basketball-reference.com
- Seasons: Six (2014-2019)
- D1 Teams: 353
- Games Played: 33,354



Basketball Statistics

Predictor Variables

Date
Home
Away
GP
Wins

Tm_vs	Opp_vs	**P_S_vs**
FG_vs	FG_O_vs	FG_vs_S
FGA_vs	FGA_O_vs	FGA_vs_S
FG%_vs	FG%_O_vs	FG%_vs_S
2P_vs	2P_O_vs	
2PA_vs	2PA_O_vs	
2P%_vs	2P%_O_vs	
3P_vs	3P_O_vs	3P_vs_S
3PA_vs	3PA_O_vs	3PA_vs_S
3P%_vs	3P%_O_vs	3P%_vs_S
FT_vs	FT_O_vs	FT_vs_S
FTA_vs	FTA_O_vs	FTA_vs_S
FT%_vs	FT%_O_vs	FT%_vs_S
ORB_vs	ORB_O_vs	
DRB_vs	DRB_O_vs	
TRB_vs	TRB_O_vs	TRB_vs_S
AST_vs	AST_O_vs	AST_vs_S
STL_vs	STL_O_vs	STL_vs_S
BLK_vs	BLK_O_vs	BLK_vs_S
TOV_vs	TOV_O_vs	TOV_vs_S
PF_vs	PF_O_vs	PF_vs_S

ORtg_vs	ORtg_O_vs	
DRtg_vs	DRtg_O_vs	
Pace_vs	Pace_O_vs	
FTr_vs	FTr_O_vs	FTr_vs_S
3PAr_vs	3PAr_O_vs	3PAr_vs_S
TS%_vs	TS%_O_vs	TS%_vs_S
TRB%_vs	TRB%_O_vs	TRB%_vs_S
AST%_vs	AST%_O_vs	AST%_vs_S
STL%_vs	STL%_O_vs	STL%_vs_S
BLK%_vs	BLK%_O_vs	BLK%_vs_S
OeFG%_vs	OeFG%_O_vs	
TOV%_vs	TOV%_O_vs	
ORB%_vs	ORB%_O_vs	
OFT/FGA_vs	OFT/FGA_O_vs	
DeFG%_vs	DeFG%_O_vs	
DTOV%_vs	DTOV%_O_vs	
DRB%_vs	DRB%_O_vs	
DFT/FGA_vs	DFT/FGA_O_vs	

Tm	Opp	**P_S**
FG	FG_O	FG_S
FGA	FGA_O	FGA_S
FG%	FG%_O	FG%_S
2P	2P_O	
2PA	2PA_O	
2P%	2P%_O	
3P	3P_O	3P_S
3PA	3PA_O	3PA_S
3P%	3P%_O	3P%_S
FT	FT_O	FT_S
FTA	FTA_O	FTA_S
FT%	FT%_O	FT%_S
ORB	ORB_O	
DRB	DRB_O	
TRB	TRB_O	TRB_S
AST	AST_O	AST_S
STL	STL_O	STL_S
BLK	BLK_O	BLK_S
TOV	TOV_O	TOV_S
PF	PF_O	PF_S

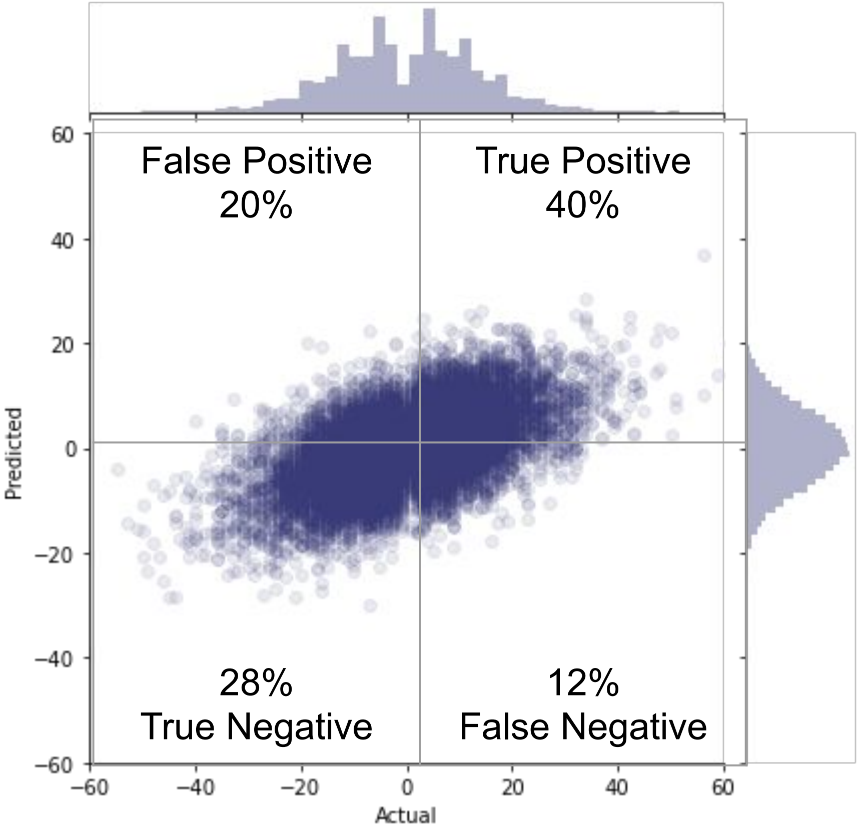
ORtg	ORtg_O	
DRtg	DRtg_O	
Pace	Pace_O	
FTr	FTr_O	FTr_S
3PAr	3PAr_O	3PAr_S
TS%	TS%_O	TS%_S
TRB%	TRB%_O	TRB%_S
AST%	AST%_O	AST%_S
STL%	STL%_O	STL%_S
BLK%	BLK%_O	BLK%_S
OeFG%	OeFG%_O	
TOV%	TOV%_O	
ORB%	ORB%_O	
OFT/FGA	OFT/FGA_O	
DeFG%	DeFG%_O	
DTOV%	DTOV%_O	
DRB%	DRB%_O	
DFT/FGA	DFT/FGA_O	

Date
Home_vs
Away_vs
GP_vs
Wins_vs

The Game

Model: Linear Regression

- Away
- ORtg
- ORtg_vs
- DRtg
- DRtg_vs
- Win%



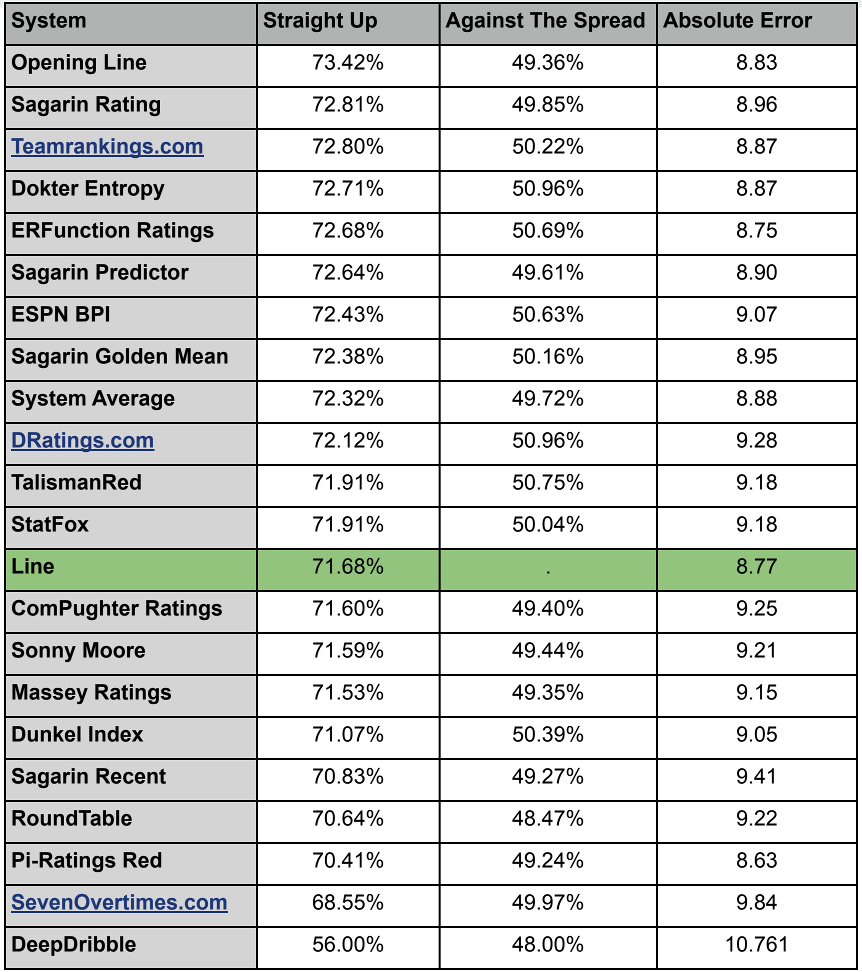
The Score

Model Performance

- Mean Absolute Error - 8.90
  - 7 out of 22 compared to published predictors*
- Straight up - 68%
  - 21 out of 22 compared to published predictors*

*Note: Comparator metrics are based on a different experience period than my model's test data. However, they do provide useful context for measuring overall model performance.

2020 YTD Prediction Tracker
(the prediction tracker.com)



System	Straight Up	Against The Spread	Absolute Error
Opening Line	73.42%	49.36%	8.83
Sagarin Rating	72.81%	49.85%	8.96
Teamrankings.com	72.80%	50.22%	8.87
Dokter Entropy	72.71%	50.96%	8.87
ERFunction Ratings	72.68%	50.69%	8.75
Sagarin Predictor	72.64%	49.61%	8.90
ESPN BPI	72.43%	50.63%	9.07
Sagarin Golden Mean	72.38%	50.16%	8.95
System Average	72.32%	49.72%	8.88
DRatings.com	72.12%	50.96%	9.28
TalismanRed	71.91%	50.75%	9.18
StatFox	71.91%	50.04%	9.18
Line	71.68%	.	8.77
ComPughter Ratings	71.60%	49.40%	9.25
Sonny Moore	71.59%	49.44%	9.21
Massey Ratings	71.53%	49.35%	9.15
Dunkel Index	71.07%	50.39%	9.05
Sagarin Recent	70.83%	49.27%	9.41
RoundTable	70.64%	48.47%	9.22
Pi-Ratings Red	70.41%	49.24%	8.63
SevenOvertimes.com	68.55%	49.97%	9.84
DeepDribble	56.00%	48.00%	10.761

Overtime

Future Work

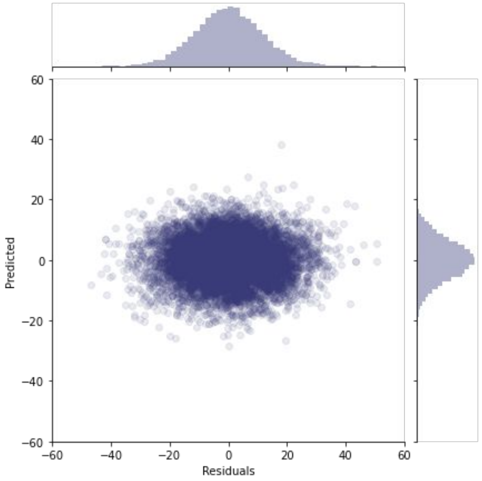
- Compare predictions against "Vegas" and other forecasters on a game by game basis
- Continue model validation using Cross Validation (holding out various seasons for each fold, instead of later games under simple validation)
- Pursue other models
  - Time Series
  - Classification models: Boosted trees
    - This can help with feature selection
- Additional feature modeling
  - Classifying playing styles, and a team's performance against other styles
  - Better incorporating rank/strength of schedule.
    - This can be done by running linear regressions to determine each team's own contribution to the shared stats (e.g. Pace, Offensive Rating)

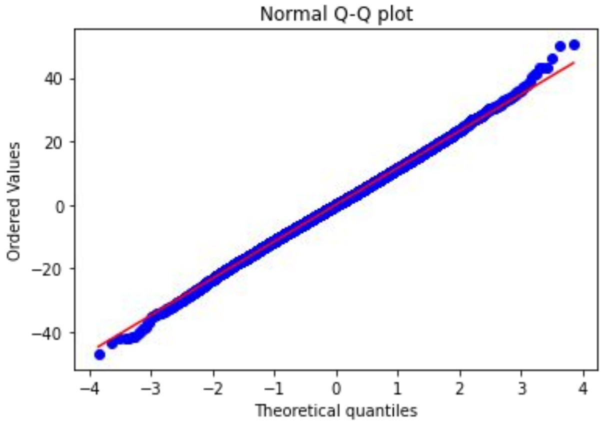
Training Camp

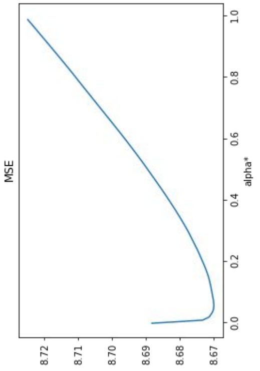
Appendices

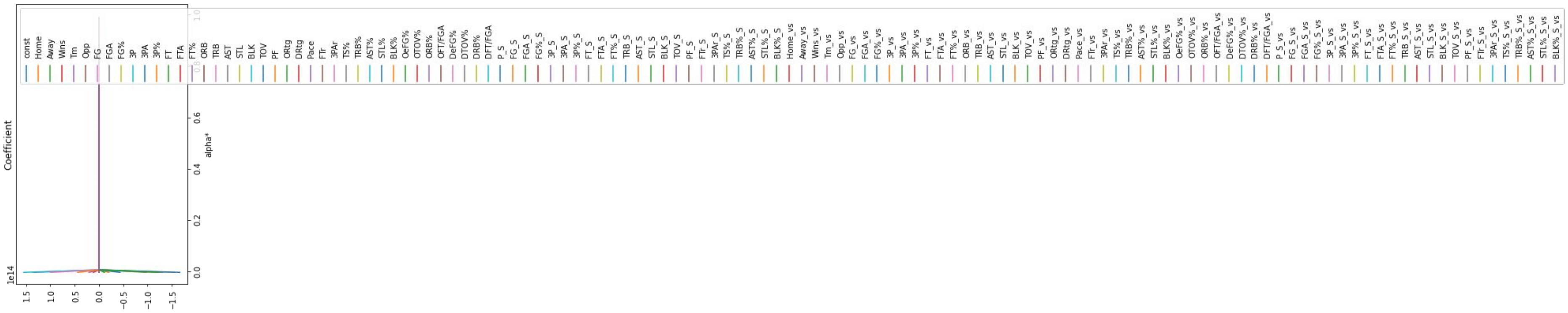
Linear Regression Assumptions

1. Linear in beta coefficients
   a. Reasonable given that residuals look evenly distributed (figure 1); predicted results are symmetric (shown on The Game slide)
2. Errors are normally distributed and has population mean of 0
   a. Reasonable assumption given that QQ plot (figure 2) shows that the errors are mostly normal except at the extremes where there is skewness. This can be attributed to blowout games that are outliers and difficult to predict
3. Homoskedasticity
   a. Errors appear to have constant variance across predictions (The Game slide)
4. Errors are uncorrelated across observations
   a. Durbin-Watson statistic of 1.991
5. Little to no multi-collinearity
   a. This was a major concern with many models that were tested, which exhibited multi-collinearity among variables. For the model shown, the condition number of the exogenous matrix is quite low, at 5.89.









Lasso regression… Just for laughs