Statistical Learning Applied to NCAA Basketball

Jake Urban

March 18, 2018

Introduction

About the Data

Each row in the data represents a NCAA Basksetball game during the 2017 season. The features shown in both tables are a subset of all features. To clarify some potentially cryptic variable names, $GameID$, $T1$, and $T2$ are identifiers for the game and each team, WLR is an abbreviation for $WinLossRatio$, and $Tournament$ is a binary indicator of the row being a tournament game. Every variable (that is not an identifer) is the average difference for the season between $T1$ and $T2$. Positive values imply $T1$ performed beter on average than $T2$, and vice versa. For Example, a value of 2 for FG would mean that $T1$ scored 2 more field goals than $T2$ on average this season up to this game. Our dataset has 5535 rows.

	FG	FT	Assists	Blocks	WLR
1486	0.90	5.74	2.82	0.69	0.54
2095	-0.50	-4.70	1.60	-1.40	-0.30
3412	1.50	8.17	2.67	0.17	0.33
4527	-1.08	7.50	-1.38	-0.99	0.23
3987	5.64	6.82	3.04	-1.18	0.69

Table 1: Subset of Game Stats Data

	T1	T1.Score	T2	T2.Score	Tournament
1486	519	84	3065	68	0
2095	3273	78	3334	81	0
3412	3196	81	3255	94	0
4527	3067	75	3106	79	0
3987	3198	78	476	70	0

Table 2: Subset of Game Meta Data

Formulating Questions

Given the data we have, what are some valuable questions we could reasonably build models for? Obviously, the simplest question to ask is which team will win. To get a more accuarate evaluation of the model we could also try to predict point spread, i.e. the difference in points scored between $T1$ and $T2$, or have a muticategorical response variable for intervals of the spread.

Regardless of the response variable's form, we will focus our modeling efforts on this question, as well as relevant meta-questions about the quality of our model relative to other models and variables.

Analysis

Linear Regression

We have two different choices for response variables: continuous (spread) and binary (winner). Fitting a linear model, **LM1**, to predict spread using all variables in game stats, we get an R-squared of 0.26623 and a RMSE of 12.97. The R-squared is much lower than expected, but what it truly telling is that **LM1**'s predictions of spread are 12.97 points off on average, easily enough points to change the outcome of a game. We also have a number of variables that are not significant at the 1% level.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.8333	0.1751	-4.76	0.0000
FG	-0.0278	0.0773	-0.36	0.7189
Assists	0.5291	0.0785	6.74	0.0000
Blocks	0.5716	0.1114	5.13	0.0000
D.Reb	0.1265	0.0658	1.92	0.0547
FT	0.0452	0.0489	0.92	0.3557
O.Reb	0.3514	0.0693	5.07	0.0000
PF	-0.1302	0.0637	-2.05	0.0409
Steals	0.1154	0.0993	1.16	0.2451
Turnovers	-0.5524	0.0797	-6.93	0.0000
WLR	17.4348	0.8216	21.22	0.0000

Table 3: Summary of All-Variable Linear Model Fit

These results are a bit of a reality check. Given **LM1**'s R-squared, it is apparent that we simply do not have the infomation needed to predict the spread with linear models accurately. We will continue with different models, but another focus will be looking more closely at our data to understand what other variables may improve our prediction quality.

Before we move on to logistic regression, lets get a good subset of our current predictors, since it is clear that some are not useful. We'll use a forward stepwise function for this.

	FG	Assists	Blocks	D.Reb	FT	O.Reb	PF	Steals	Turnovers	WLR
1 (1)										*
2 (1)		*								*
3 (1)		*							*	*
4 (1)		*	*						*	*
5 (1)		*	*			*			*	*
6 (1)		*	*			*	*		*	*
7 (1)		*	*	*		*	*		*	*
8 (1)		*	*	*		*	*	*	*	*
9 (1)		*	*	*	*	*	*	*	*	*
10 (1)	*	*	*	*	*	*	*	*	*	*

Table 4: Forward Stepwise Selection

Now that we have an ordering to pick our variables, how many should we keep? We can perform F tests to determine if a model is significantly better than a subset of its variables. Below, the index of each row is the number of variables in the linear model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	5533	962949.24				
2	5532	950247.08	1	12702.15	75.36	0.0000
3	5531	943695.01	1	6552.07	38.87	0.0000
4	5530	936984.15	1	6710.86	39.82	0.0000
5	5529	932636.02	1	4348.13	25.80	0.0000
6	5528	931915.39	1	720.63	4.28	0.0387
7	5527	931428.79	1	486.60	2.89	0.0894
8	5526	931214.10	1	214.69	1.27	0.2591
9	5525	931066.53	1	147.56	0.88	0.3495
10	5524	931044.69	1	21.84	0.13	0.7189

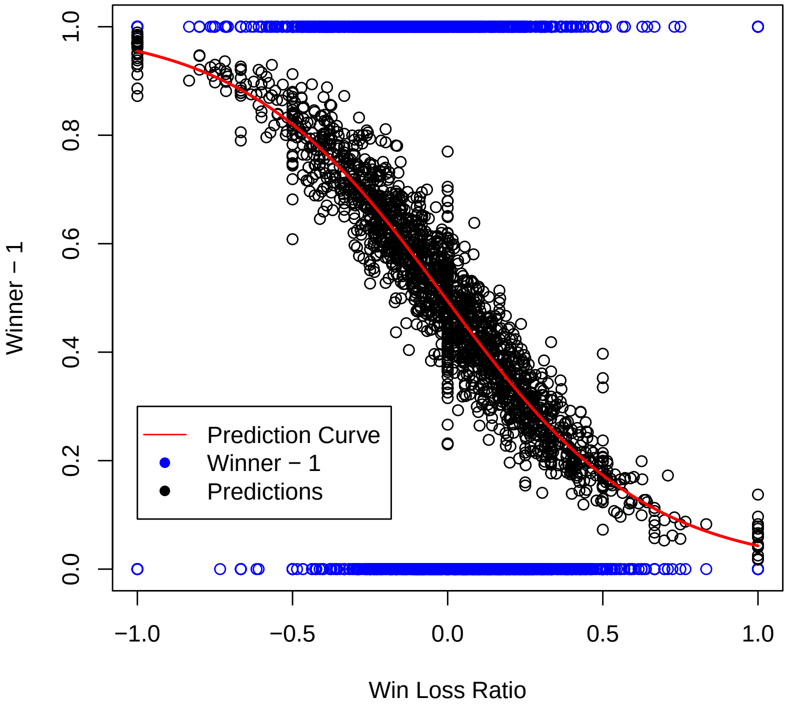
Table 5: Analysis of Variance

Looking at **Table 5**'s p-values, we see that the model with 6 variables is not significantly better than the model with 5 variables at the 1% level. Therefore, we will continue on with only WLR, $Assists$, $Turnovers$, $Blocks$, and $O.Reb$.

Logisitic Regression

	0	1
1	641	306
2	283	615

Table 6: Winner v. Predictions Confusion Matrix

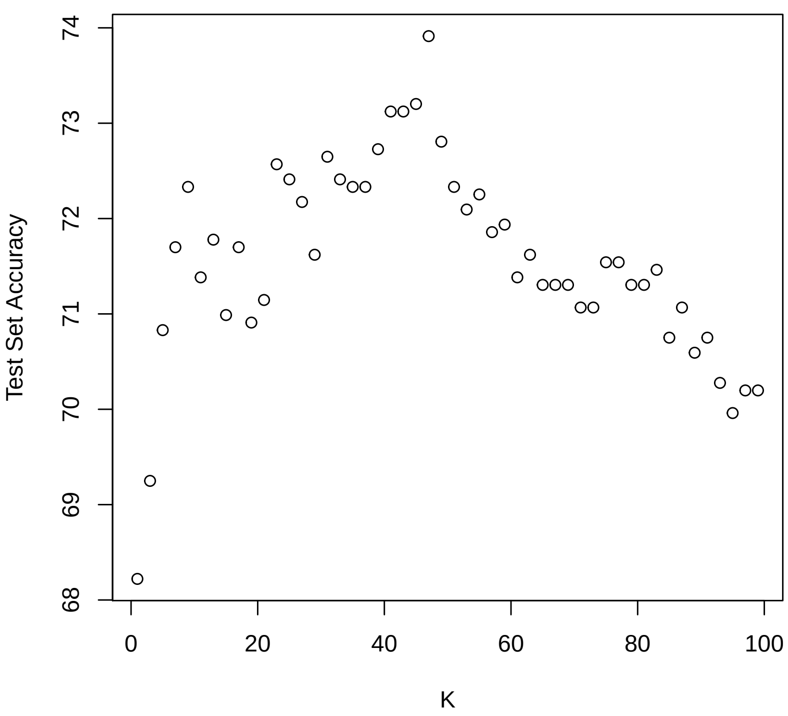


Using the predictors selected in the previous section, our logistic regression model was trained on a 66% split, and got a prediction accuracy of 68.08% on our test set. Lets explore some other classification techniques and see if we get better results.

K-Nearest Neighbor

Compared to linear models, KNN is more intuitive. Given a game, lets get the previous games that are most similar and predict the outcome based on the similar games. There are 143 days this season, so lets split the season into 95 and 48 days for our training and test set, respectively. We also need to find the optimal value for K, so we'll train multiple KNN models with differing K's and select the model with the highest accuracy.

Figure 1. Finding the Optimal K



Out of K values 1 to 100, $K = 47$ was found to be optimal, with 74% accuracy. Looking at **Figure 1.**, it seems as though our optimal K is slightly higher due to our sample of data, and the real test accuracy is around 73%.

Conclusions

Based on the results of fitting and optimizing our first nonlinear model, other nonlinear models may also be promising. However, linear and logisitic regression did not acheive great results, and after visualizing our data it became apparent that our data was not linearly separable. The models' R-squared's were also indicative of the weak predictive power each of our predictors had on the response.

Polynomial regression is one model we did not explore, since none of our predictors appeared to have a polynomial relationship with our response variable. Smoothing spines and other continuous nonlinear functions would not be effective for the same reason.

The reality of the data is that it is so dense and nonlinear that the models we build must have many more useful predictors in order to be effective above the 90th percentile. Specfically, we need predictors that better describe the relative difference in a team's skill level compared to their opponent. For example, "Field Goals Per Game" only describes a team's skill relative to teams they've played in the past, not teams they play in the future.

The same statistic also assumes that each team has had past opponents of equal average difficulty, which is certainly not true. A statistic that takes into account the skill difference between each team's past opponents would be a great predictor to have, since it would tell us more about how the two teams in the game may compare to each other. For example, if team A plays the two best teams in the league and losses, and team B plays the two worst team and wins, the models we built in this analysis would almost certainly predict team B to win the game, even though team A may end up being the 3rd best and team B may end up being the 3rd worst.

While we did not build a competitive model, we explored and tested various types of models, performed feature selection, and visualized our data effectively. We also have a better understanding our data and the underlying assumptions we make when we build models with its predictors. As I continue to add to this dataset, I am optimistic that the quality of models such as KNN will continue to improve.