NBA Awards Prediction - Part 1

Ryan Drost, Luigi Noto

Methodology

Our goal was to build a model to predict the players who would earn All-NBA awards for any given season using only data available prior to the beginning of that season. Our final dataset included the seasons from 1989 to 2020 because 1989 was the first year that the NBA selected three All-NBA teams; prior to that only two were selected.

We opted to use All-NBA as our response variable, as opposed to MVP Share, since it was more natural to use a purely binary variable for logistic regression and it translated more easily into a probability.[1] Our first step was to evaluate whether the possible predictors in our dataset were significantly different between players who were selected to an All-NBA team and those who weren't. We considered the value of each statistic from one to five seasons prior for that player, and also for the player's career prior to that season. It was not surprising to confirm that All-NBA players were better by essentially every metric, but the degree of significance for every statistic indicated that the sparsity of the All-NBA class would be problematic. We opted not to pursue a multiclass classification because of the sparsity issue even with binary classes.

For the prediction, we started by fitting a logistic regression model, a simple yet powerful model which is also easy to interpret. However, we noticed that the predictive ability of most of the existing statistics was largely already captured in the Prev_1_AllNBA variable, and we weren't able to achieve substantial gains beyond that baseline one-variable model. Because of that we turned our focus to feature engineering in order to improve our model's performance. As an example, age intuitively has predictive ability since very young and very old players are much less likely to win the award, but adding age as a variable actually greatly worsened model performance.

Results

Hypothesis testing

We did permutation tests for 177 predictors using the difference in median or mean values, depending which was more appropriate for the specific statistic, between the two classes (All-NBA: 1, otherwise: 0). We found that, regardless of the length of the window, almost every statistic was very significant even when accounting for the number of hypothesis tests with Bonferroni's correction.[2] In fact, many of the permutation tests resulted in a p-value of 0. An example result is displayed in Figure 1.

[1] As noted before, MVP Share is not properly interpreted as the probability of winning MVP. It measures the percentage of the voting points a single player can receive rather than a percentage of the total points available.

[2] Three-point percentage was a notable exception. It was clearly not significant for any span of time. See appendix for full results.

Logistic Regression - assumptions diagnostics

As a first step in the construction of the logistic regression model, before evaluating its predictive performance on the test set, we investigated whether the logistic regression assumptions were satisfied. The main assumptions are: (1) binary outcome type (which is of course satisfied), (2) linearity of independent variables and log-odds, (3) absence of multicollinearity, (4) independence of observations. In particular, we investigated assumptions 3 and 4.

In order to check for multicollinearity, which to a situation where the data contain highly correlated independent variables, we computed the Variance Inflation Factor (VIF) of each continuous independent variable. The VIF of independent variable $\tilde{x}_j$ is defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination of the linear regression of $\tilde{x}_j$ on the other covariates. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic degree of multicollinearity. In our case, as we might expect given that many variables are the same rate stats computed across a varying number of previous seasons, some of the variables have VIFs in the hundreds. Therefore, there's high multicollinearity, which might be a problem for the coefficient estimates and the associated parametric significance tests, which may not be reliable. Since we are now mainly focusing on prediction, we do not care about this issue very much, but if we want to interpret the coefficients in a later stage, then we would need to discard the variables that cause multicollinearity.

In order to check for independence of observations, we fitted the logistic regression model with all predictors and we produced a residual series plot, which is a plot of the deviance residuals of the logistic regression model against the index numbers of the observations, where the observations are sorted by player and year. In this way, we can assess whether there is some autocorrelation, given the fact that we have multiple observations in different years for each player. The residual series plot is displayed in Figure 2. We can see that the residuals are randomly scattered around zero, therefore we can reasonably assume that the observations are independent.

Prediction

Our baseline was a logistic regression model with one binary predictor variable which indicated whether the player made the All-NBA team the season before. This model simply predicted that any player who earned All-NBA the year before would earn it again, and any other player would not. As we expected, this model performed fairly well and we made comparatively marginal (but important) improvements to it. For our final model, we used logistic regression rather than random forest because it performed better. In particular, the implied probability was the same for all players in the same class in the baseline model, whereas the implied probabilities became much more realistic in our model which is reflected in the reduced log loss. [3]

Metric	Baseline Model	Current Model
AUROC	0.793	0.960
Accuracy	0.974	0.978
F1 Score	0.581	0.645
Precision	0.563	0.625
Recall	0.600	0.667
Log Loss	0.086	0.064

[3] See the appendix for All-NBA predictions on the test set (2021 season).

Analysis

Overall, the model does well in precision. There are few false positives, and many of them are top players who were injured during a season when they otherwise were an All-NBA caliber player. There was a delicate balance between precision and recall, as we were able to improve precision further with small tweaks, but recall was significantly worsened; in fact, it proved challenging to improve recall at all. In the 2021 test set, the model actually did better in recall than precision but that was likely due to randomness, as it struggles to identify players who will make an All-NBA team for the first time. This makes sense since many of the input variables relate to having earned previous All-NBA awards. Figure 3 shows ROC and precision-recall curves.

As noted earlier, this model makes a significant improvement in logistic loss, which we consider the most important metric in evaluating this model. The implied probability of making an All-NBA team ranges from $\approx 3 \times 10^{-6}$ for some fringe NBA players who barely play to 0.97 for Shaquille O'Neal in 2000, the highest in the training dataset. This is far more satisfying and informative as a prediction than showing one percentage for any player who made the previous season's All-NBA team and another for anyone who didn't.

Notably, we know that 15 players earn an All-NBA award every season, which increases the importance of the implied probabilities the model produces. In a practical setting where we wanted to predict which players would make the All-NBA team in an upcoming season, we would likely select the 15 players with the highest probability rather than predicting everyone with a probability over the 0.5 threshold to earn the award. This emphasizes the importance of the relative implied probabilities over the binary classification.

Plan for additional analysis

The main area we are looking to improve is the recall of the model. In our current model there is actually no differentiation among rookies, and they are all predicted to make the All-NBA team with the same probability (< 0.0001). While not predicting any rookies to be on an All-NBA team is reasonable, it's rather unsatisfying that there is no differentiation at all. Beyond that, the model struggles with early-career players in general, as its main predictive power results from the previous All-NBA teams a player has made. We will look into including college or amateur data, or other biographical data that could solve this issue. We may also consider implementing different models based on years of experience in the NBA (though this has the downside of losing consistency and making it more difficult to compare between all players) or expanding our use of random forest.

We will also move to predicting future career All-NBA awards. While it is interesting and potentially helpful to make predictions for an upcoming season, predicting a player's number of future All-NBA appearances would have more utility. We would approach this similarly, looking at the players from our dataset who have a complete career (limiting the dataset to about 1990-2008) and using a linear regression to predict the number of future All-NBA teams for each player at any point in their career.

Lastly, we will use the held-out data from the 2022 season and investigate the calibration of the model on this data, along with presenting the model's final prediction for the ongoing season.

Figures

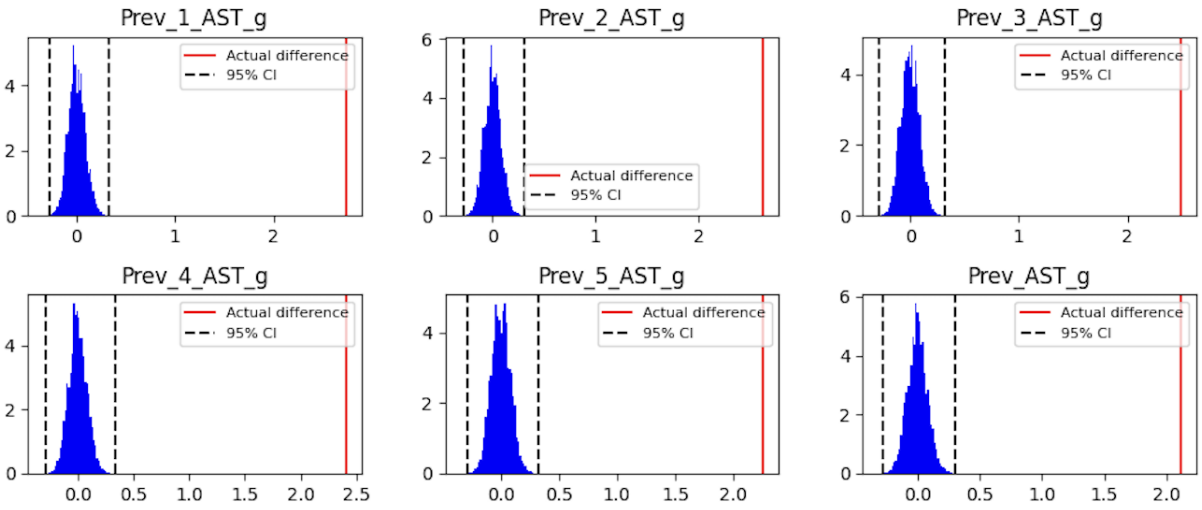

Figure 1: Permutation tests for assists per game for the previous 1, 2, 3, 4, and 5 seasons, and career

Figure 2: Residual series plot

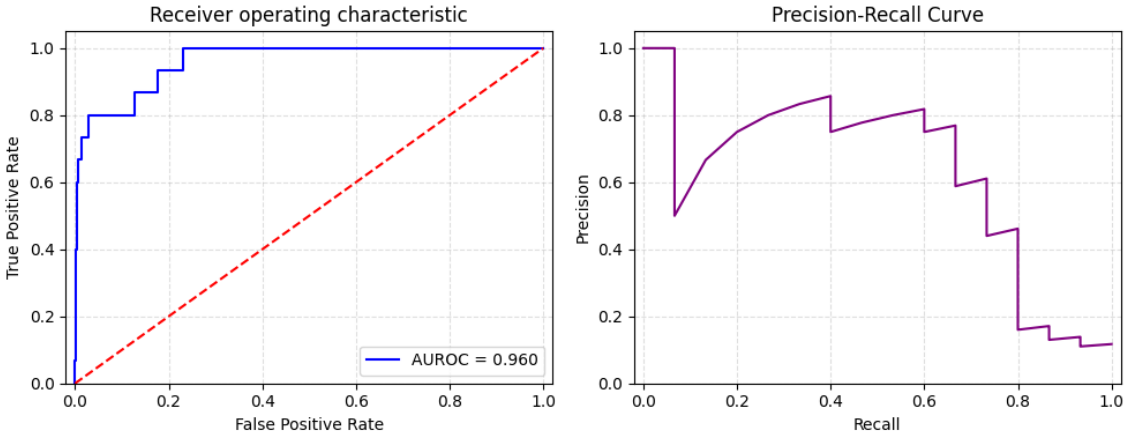

Figure 3: AUROC and Precision-Recall Curve

Appendix

Model Features

Feature	Type	Coefficient	Explanation
Prev_1_AllNBA	Binary	3.16	Made All-NBA Team in previous season
old_first	Binary	-0.82	Age >= 28, Made All-NBA team for the first time in the previous season, but was not on first team
young_star	Binary	0.41	Age < 25, Made All-NBA first or second team in any season previously
prev_rotation	Binary	2.94	Played > 15 minutes per game in previous season
prev_starter	Binary	1.36	Played > 28 minutes per game in previous season
YrsOff	Continuous	-1.62	Number of seasons missed since the player's most recent season
Prev_1_WS_g	Continuous	0.33	Win shares per game in previous season
injured	Binary	1.68	Fewer than 50 games in previous season with more than 15 points per game and Prev_AllNBAV_decay > 10
high_pick	Binary	1.20	Drafted in the top 10, made All-Rookie first team, and fewer than 3 years of experience
star	Binary	1.61	Age <= 30, Prev_3_AllNBAV >= 10, Prev_AllNBAV_decay > 10 (a metric of career AllNBAV with beta decay)

2021 Test Predictions
Players predicted to make All-NBA team in 2021

Player	AllNBA	Prediction	Probability
Giannis Antetokounmpo	1	1	0.845
Anthony Davis	0	1	0.845
Kawhi Leonard	1	1	0.845
Nikola Jokic	1	1	0.845
Damian Lillard	1	1	0.845
Luka Doncic	1	1	0.620
Paul George	1	1	0.555
James Harden	0	1	0.522
Rudy Gobert	1	1	0.522
Jimmy Butler	1	1	0.522
LeBron James	1	1	0.522
Ben Simmons	0	1	0.522
Chris Paul	1	1	0.522
Jayson Tatum	0	1	0.522
Pascal Siakam	0	1	0.522
Russell Westbrook	0	1	0.522

Hypothesis testing for significant predictors

(Significance level based on Bonferroni's correction: $\alpha = 0.05/177 = 0.0003$)

Predictor	**p value**	**Significance**	**95% CI range**		**Actual difference**
Pick	0	Yes	-5	6	-20
Age	0.041	No	-0.703	0.704	0.401
Prev_1_GP	0	Yes	-5.5	5	13
Prev_2_GP	0	Yes	-11.5	8.5	29
Prev_3_GP	0	Yes	-18	16	46
Prev_4_GP	0	Yes	-27.5	22	68.5
Prev_5_GP	0	Yes	-42	35	95
Prev_GP	0	Yes	-64.5	71	165.5
Prev_1_PTS_g	0	Yes	-1.26	1.44	14.04
Prev_2_PTS_g	0	Yes	-1.14	1.23	13.53
Prev_3_PTS_g	0	Yes	-1.17	1.24	13.00
Prev_4_PTS_g	0	Yes	-1.08	1.22	12.60
Prev_5_PTS_g	0	Yes	-1.04	1.26	12.39
Prev_PTS_g	0	Yes	-0.99	1.16	11.45
Prev_1_FG3p	0.707	No	-0.027	0.017	0.002
Prev_2_FG3p	0.698	No	-0.027	0.017	0.002
Prev_3_FG3p	0.863	No	-0.026	0.019	0.001
Prev_4_FG3p	0.763	No	-0.026	0.018	0.002
Prev_5_FG3p	0.778	No	-0.027	0.019	-0.002
Prev_FG3p	0.900	No	-0.028	0.019	0.001
Prev_1_FG2p	0	Yes	-0.011	0.010	0.034
Prev_2_FG2p	0	Yes	-0.010	0.010	0.035
Prev_3_FG2p	0	Yes	-0.009	0.009	0.036
Prev_4_FG2p	0	Yes	-0.009	0.008	0.034
Prev_5_FG2p	0	Yes	-0.008	0.008	0.033
Prev_FG2p	0	Yes	-0.008	0.008	0.031
Prev_1_FTp	0	Yes	-0.021	0.022	0.032
Prev_2_FTp	0	Yes	-0.022	0.02	0.032
Prev_3_FTp	0	Yes	-0.020	0.021	0.028
Prev_4_FTp	0	Yes	-0.019	0.021	0.026
Prev_5_FTp	0	Yes	-0.019	0.018	0.024
Prev_FTp	0.0008	No	-0.018	0.019	0.017
Prev_1_OREB_g	0	Yes	-0.155	0.185	0.760
Prev_2_OREB_g	0	Yes	-0.162	0.18	0.713
Prev_3_OREB_g	0	Yes	-0.165	0.181	0.671
Prev_4_OREB_g	0	Yes	-0.169	0.193	0.691
Prev_5_OREB_g	0	Yes	-0.178	0.187	0.658
Prev_OREB_g	0	Yes	-0.187	0.188	0.710
Prev_1_DREB_g	0	Yes	-0.311	0.364	3.27
Prev_2_DREB_g	0	Yes	-0.297	0.324	3.14
Prev_3_DREB_g	0	Yes	-0.289	0.333	3.01

Predictor	p value	Significance	95% CI range		Actual difference
Prev_4_DREB_g	0	Yes	-0.278	0.324	2.87
Prev_5_DREB_g	0	Yes	-0.291	0.320	2.78
Prev_DREB_g	0	Yes	-0.273	0.329	2.49
Prev_1_AST_g	0	Yes	-0.279	0.332	2.74
Prev_2_AST_g	0	Yes	-0.274	0.314	2.63
Prev_3_AST_g	0	Yes	-0.281	0.322	2.49
Prev_4_AST_g	0	Yes	-0.283	0.339	2.41
Prev_5_AST_g	0	Yes	-0.288	0.326	2.26
Prev_AST_g	0	Yes	-0.272	0.308	2.11
Prev_1_STL_g	0	Yes	-0.086	0.098	0.716
Prev_2_STL_g	0	Yes	-0.086	0.089	0.725
Prev_3_STL_g	0	Yes	-0.084	0.093	0.699
Prev_4_STL_g	0	Yes	-0.083	0.088	0.704
Prev_5_STL_g	0	Yes	-0.084	0.090	0.680
Prev_STL_g	0	Yes	-0.086	0.086	0.650
Prev_1_BLK_g	0	Yes	-0.061	0.074	0.413
Prev_2_BLK_g	0	Yes	-0.060	0.077	0.430
Prev_3_BLK_g	0	Yes	-0.058	0.075	0.427
Prev_4_BLK_g	0	Yes	-0.061	0.076	0.424
Prev_5_BLK_g	0	Yes	-0.063	0.079	0.423
Prev_BLK_g	0	Yes	-0.060	0.080	0.428
Prev_1_PF_g	0	Yes	-0.169	0.171	0.592
Prev_2_PF_g	0	Yes	-0.163	0.162	0.598
Prev_3_PF_g	0	Yes	-0.147	0.160	0.586
Prev_4_PF_g	0	Yes	-0.145	0.160	0.594
Prev_5_PF_g	0	Yes	-0.137	0.155	0.586
Prev_PF_g	0	Yes	-0.147	0.147	0.578
Prev_1_OWS_g	0	Yes	-0.005	0.006	0.072
Prev_2_OWS_g	0	Yes	-0.005	0.006	0.069
Prev_3_OWS_g	0	Yes	-0.005	0.005	0.065
Prev_4_OWS_g	0	Yes	-0.004	0.005	0.062
Prev_5_OWS_g	0	Yes	-0.005	0.005	0.059
Prev_OWS_g	0	Yes	-0.004	0.005	0.054
Prev_1_DWS_g	0	Yes	-0.003	0.003	0.031
Prev_2_DWS_g	0	Yes	-0.003	0.003	0.030
Prev_3_DWS_g	0	Yes	-0.003	0.003	0.029
Prev_4_DWS_g	0	Yes	-0.003	0.003	0.028
Prev_5_DWS_g	0	Yes	-0.003	0.003	0.027
Prev_DWS_g	0	Yes	-0.003	0.003	0.025
Prev_1_EFGp	0	Yes	-0.010	0.011	0.028
Prev_2_EFGp	0	Yes	-0.009	0.009	0.027
Prev_3_EFGp	0	Yes	-0.009	0.009	0.027
Prev_4_EFGp	0	Yes	-0.008	0.009	0.026
Prev_5_EFGp	0	Yes	-0.008	0.008	0.026
Prev_EFGp	0	Yes	-0.007	0.008	0.025

Predictor	p value	Significance	95% CI range		Actual difference
Prev_1_TSp	0	Yes	-0.010	0.010	0.044
Prev_2_TSp	0	Yes	-0.010	0.009	0.043
Prev_3_TSp	0	Yes	-0.009	0.009	0.041
Prev_4_TSp	0	Yes	-0.008	0.008	0.039
Prev_5_TSp	0	Yes	-0.009	0.008	0.037
Prev_TSp	0	Yes	-0.008	0.007	0.034
Prev_1_FTr	0	Yes	-0.028	-0.029	0.109
Prev_2_FTr	0	Yes	-0.027	0.029	0.107
Prev_3_FTr	0	Yes	-0.027	0.026	0.104
Prev_4_FTr	0	Yes	-0.026	0.027	0.101
Prev_5_FTr	0	Yes	-0.025	0.026	0.100
Prev_FTr	0	Yes	-0.025	0.025	0.096
Prev_1_MVP_Share	0	Yes	0	0	0.002
Prev_2_MVP_Share	0	Yes	0	0	0.019
Prev_3_MVP_Share	0	Yes	0	0	0.033
Prev_4_MVP_Share	0	Yes	0	0	0.036
Prev_5_MVP_Share	0	Yes	0	0	0.043
Prev_MVP_Share	0	Yes	0	0	0.046
Prev_1_PWp	0	Yes	-0.037	0.037	0.110
Prev_2_PWp	0	Yes	-0.036	0.030	0.104
Prev_3_PWp	0	Yes	-0.029	0.028	0.098
Prev_4_PWp	0	Yes	-0.028	0.027	0.085
Prev_5_PWp	0	Yes	-0.027	0.027	0.078
Prev_PWp	0	Yes	-0.024	0.024	0.067
Prev_1_AllNBA	0	Yes	-0.649	-0.701	1.982
Prev_2_AllNBA	0	Yes	-0.052	0.063	1.106
Prev_3_AllNBA	0	Yes	-0.071	0.096	1.531
Prev_4_AllNBA	0	Yes	-0.094	0.116	1.871
Prev_5_AllNBA	0	Yes	-0.107	0.137	2.147
Prev_AllNBA	0	Yes	-0.197	0.261	2.885
Prev_1_AllNBA1	0	Yes	-0.014	0.022	0.288
Prev_2_AllNBA1	0	Yes	-0.026	0.039	0.546
Prev_3_AllNBA1	0	Yes	-0.037	0.056	0.771
Prev_4_AllNBA1	0	Yes	-0.045	0.071	0.946
Prev_5_AllNBA1	0	Yes	-0.055	0.086	1.09
Prev_AllNBA1	0	Yes	-0.095	0.159	1.49
Prev_1_AllNBA2	0	Yes	-0.014	0.025	0.210
Prev_2_AllNBA2	0	Yes	-0.026	0.035	0.363
Prev_3_AllNBA2	0	Yes	-0.033	0.047	0.488
Prev_4_AllNBA2	0	Yes	-0.039	0.054	0.585
Prev_5_AllNBA2	0	Yes	-0.048	0.064	0.656
Prev_AllNBA2	0	Yes	-0.076	0.101	0.878

Predictor	p value	Significance	95% CI range		Actual difference
Prev_1_AllNBA3	0	Yes	-0.014	0.023	0.127
Prev_2_AllNBA3	0	Yes	-0.022	0.034	0.213
Prev_3_AllNBA3	0	Yes	-0.030	0.039	0.290
Prev_4_AllNBA3	0	Yes	-0.036	0.048	0.357
Prev_5_AllNBA3	0	Yes	-0.040	-0.053	0.420
Prev_AllNBA3	0	Yes	-0.060	0.078	0.540
Prev_1_AllDef	0	Yes	-0.023	0.029	0.260
Prev_2_AllDef	0	Yes	-0.039	0.054	0.488
Prev_3_AllDef	0	Yes	-0.055	0.075	0.680
Prev_4_AllDef	0	Yes	-0.066	0.096	0.823
Prev_5_AllDef	0	Yes	-0.082	0.110	0.945
Prev_AllDef	0	Yes	-0.132	0.190	1.24
Prev_1_AllDef1	0	Yes	-0.014	0.023	0.183
Prev_2_AllDef1	0	Yes	-0.024	0.038	0.334
Prev_3_AllDef1	0	Yes	-0.034	0.054	0.458
Prev_4_AllDef1	0	Yes	-0.044	0.068	0.559
Prev_5_AllDef1	0	Yes	-0.052	0.082	0.641
Prev_AllDef1	0	Yes	-0.084	0.136	0.831
Prev_1_AllDef2	0	Yes	-0.014	0.023	0.086
Prev_2_AllDef2	0	Yes	-0.024	0.037	0.162
Prev_3_AllDef2	0	Yes	-0.029	0.044	0.230
Prev_4_AllDef2	0	Yes	-0.038	0.051	0.273
Prev_5_AllDef2	0	Yes	-0.041	0.058	0.313
Prev_AllDef2	0	Yes	-0.063	0.086	0.412
Prev_1_AllRook	0.030	No	-0.023	0.029	0.016
Prev_2_AllRook	< 0.0001	Yes	-0.031	0.040	0.051
Prev_3_AllRook	0	Yes	-0.040	0.048	0.105
Prev_4_AllRook	0	Yes	-0.046	0.051	0.153
Prev_5_AllRook	0	Yes	-0.050	0.056	0.203
Prev_AllRook	0	Yes	-0.066	0.072	0.409
Prev_1_AllRook1	< 0.0001	Yes	-0.014	0.023	0.025
Prev_2_AllRook1	0	Yes	-0.022	0.030	0.068
Prev_3_AllRook1	0	Yes	-0.028	0.034	0.117
Prev_4_AllRook1	0	Yes	-0.031	0.040	0.159
Prev_5_AllRook1	0	Yes	-0.036	0.042	0.193
Prev_AllRook1	0	Yes	-0.053	0.057	0.366
Prev_1_AllRook2	1	No	-0.014	0.023	-0.001
Prev_2_AllRook2	0.240	No	-0.022	0.032	-0.009
Prev_3_AllRook2	0.714	No	-0.027	0.033	-0.004
Prev_4_AllRook2	0.826	No	-0.032	0.042	0.003
Prev_5_AllRook2	0.088	No	-0.035	0.045	0.019
Prev_AllRook2	0.0003	No	-0.048	0.054	0.052

Predictor	**p value**	**Significance**	**95% CI range**		**Actual difference**
Prev_1_AS	0	Yes	-0.037	0.045	0.639
Prev_2_AS	0	Yes	-0.066	0.077	1.15
Prev_3_AS	0	Yes	-0.090	0.106	1.57
Prev_4_AS	0	Yes	-0.117	0.138	1.89
Prev_5_AS	0	Yes	-0.134	0.164	2.17
Prev_AS	0	Yes	-0.230	0.291	2.78