Revisiting PSM Analysis of College Athletic Success with Machine Learning: LASSO Regression & Gradient Boosting

Gerard Tetegan, Jacob Lara, Daniel Pate, and Lane Whitehead
Texas A&M University
April 29, 2024

Abstract

This study PSM to determine the causal effects of college football success on donations and applications. We replicated their analysis and then used LASSO regression & (possibly) gradient boosting aiming to improve the robustness of the propensity scores.

Introduction

The study uses bookmaker spreads to calculate the propensity scores and found that football success leads to better outcomes. We use LASSO and gradient boosting to aiming improve the PSM estimates and variable selection to see their effects.

Literature Review

- Pairing PSM and gradient boosting significantly enhances predictive accuracy (Kim et al., 2023)
- Gradient boosting is more robust and preferred over other ML methods in finding average treatment effects (Yang, Chuang, Kuan, 2020)
- LASSO regressions have demonstrated their usefulness in subset selection especially in high-dimensional settings (Tibshirani, 1996).
- While LASSO is not as powerful in comparison to other ML methods, combined with PSM it demonstrates extreme robustness (Pirracchio, Petersen, van der Laan, 2015)

Findings

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      lead2_exp_wins_wk11   R-squared:                       0.174
Model:                              OLS   Adj. R-squared:                  0.093
Method:                   Least Squares   F-statistic:                       nan
Date:                  Mon, 29 Apr 2024   Prob (F-statistic):                nan
Time:                          18:15:03   Log-Likelihood:                -2.9759
No. Observations:                   393   AIC:                             77.95
Df Residuals:                       357   BIC:                             221.0
Df Model:                            35
Covariance Type:                cluster
==============================================================================
                                         coef    std err          z
------------------------------------------------------------------------------
lead2_pscore_wk11_group_1              0.1529      0.048      3.182
lead2_pscore_wk11_group_2             -0.0803      0.309     -0.260
lead2_pscore_wk11_group_3              1.6379      1.016      1.613
lead2_pscore_wk11_group_4              0.6030      0.810      0.744
lead2_pscore_wk11_group_5              0.5882      2.390      0.246
lead2_pscore_wk11_group_6              2.1827      0.586      3.722
lead2_pscore_wk11_group_7              0.6534      1.576      0.415
lead2_pscore_wk11_group_8              0.0269      1.503      0.018
lead2_pscore_wk11_group_9             -4.1408      2.235     -1.853
lead2_pscore_wk11_group_10             1.9579      1.815      1.078
lead2_pscore_wk11_group_11            -0.4022      1.408     -0.286
lead2_pscore_wk11_group_12            -0.6552      0.701     -0.935
```

Figure: Simple OLS Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      lead2_exp_wins_wk10   R-squared:                       0.122
Model:                              OLS   Adj. R-squared:                  0.112
Method:                   Least Squares   F-statistic:                     13.71
Date:                  Mon, 29 Apr 2024   Prob (F-statistic):           2.91e-13
Time:                          18:15:03   Log-Likelihood:                -579.68
No. Observations:                  1136   AIC:                             1185.
Df Residuals:                      1123   BIC:                             1251.
Df Model:                            12
Covariance Type:                cluster
==============================================================================
                                         coef    std err          z
------------------------------------------------------------------------------
const                                  0.6200      0.030     20.782
lead2_pscore_wk10_group_1             -0.2114      0.037     -5.694
lead2_pscore_wk10_group_2             -0.0790      0.046     -1.723
lead2_pscore_wk10_group_3             -0.0775      0.053     -1.461
lead2_pscore_wk10_group_5              0.0691      0.050      1.393
lead2_pscore_wk10_group_9              0.0441      0.094      0.471
lead2_pscore_wk10_group_10             0.0313      0.075      0.416
lead2_pscore_wk10_group_12             0.1960      0.099      1.987
```

Figure: LASSO Regression

Machine Learning

While using the machine learning technique of a LASSO regression does yield smaller confidence intervals with less variables, narrowing the subset, it is not statistically different from the original data. This could be explained by less variables around the same mean. More statistical analysis can help explain whether this is an accurate representation.

Conclusions

The study finds that success in college football positively affects alumni donations, application rates, and academic reputation, while also increasing in-state enrollment and SAT scores of incoming students, especially in elite conferences. Our machine learning technique shows that with LASSO while some variables are omitted in the earlier stages it does not change the factors that are overall chosen.

Appendix

Figure: Appendix/Github