# Internet Appendix

1. **Sample Splitting**

This paper studies a few sample splitting methods discussed in the prior literature (West, 2006). The first sample split method is the fixed splits where the data are divided into training, validation, and testing samples. This method trains models with training dataset and validates models with validation dataset to get all models’ accuracy performances. Finally, this method uses the best model with test dataset to make a prediction.

The second sample split method is the rolling method, where the training and validation datasets regularly change before training and validating to contain fresher data and keeps all time periods in each training and validation dataset fixed. For every rolling window, scholars can refit the model from the current training and verification datasets, and track the accuracy of the model in the residual test dataset that is not fitted in the rolling window. The result is a series of accuracy simulations regarding every rolling window. This method shows the advantage of using more up-to-date information to make predictions than fixed solutions.

The third sample splitting method is the recursive sample splitting method. Similar to the rolling method, it regularly contains more up-to-date sample in the training and validation datasets. However, the recursive method usually holds the whole history in the training dataset. Therefore, the size of the window regularly goes up. The rolling and recursive method are quite expensive, especially for more complex models.

In this paper, I adopt the fixed split and split the whole dataset into three sub-datasets, train dataset, validation dataset, and test dataset. I first split the whole dataset into 80% and 20%, and then split the 80% one to 64% and 16%. As a result, the proportion of the train, validation, and test datasets are 64%, 16%, and 20%. The observation number of these three datasets are 9444, 2362, 2952.

1. **Summary Statistics for different datasets.**

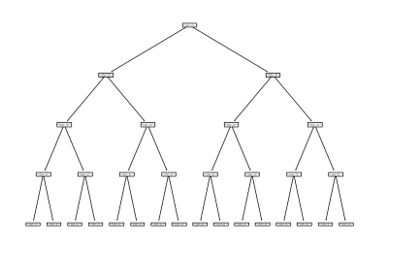
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TABLE 1 B.1**  **Summary Statistics for Original Dataset** | | | | | |
| **Variable** | **Obs** | **Mean** | **Std.Dev.** | **Min** | **Max** |
| *gmDate* | 14758 | 20290.04 | 622.54 | 19296.00 | 21285.00 |
| *pointspread* | 14758 | 0.00 | 13.70 | -61.00 | 61.00 |
| *teamLocx* | 14758 | 0.50 | 0.50 | 0.00 | 1.00 |
| *teamDayOff* | 14758 | 1.94 | 1.07 | 0.00 | 11.00 |
| *teamMinx* | 14758 | 48.36 | 0.52 | 47.50 | 53.00 |
| *teamPTSx* | 14758 | 102.15 | 6.54 | 75.00 | 140.00 |
| *teamASTx* | 14758 | 22.28 | 2.58 | 9.00 | 36.00 |
| *teamTOturN~x* | 14758 | 14.47 | 1.78 | 8.00 | 27.00 |
| *teamSTLx* | 14758 | 7.76 | 1.33 | 2.00 | 16.00 |
| *teamBLKx* | 14758 | 4.90 | 1.19 | 0.00 | 18.00 |
| *teamPFpers~x* | 14758 | 20.28 | 2.12 | 13.00 | 32.00 |
| *teamFGAx* | 14758 | 84.06 | 3.62 | 66.00 | 107.00 |
| *teamFGMx* | 14758 | 38.12 | 2.42 | 24.00 | 53.00 |
| *teamFGx* | 14758 | 0.45 | 0.02 | 0.31 | 0.57 |
| *team2PAx* | 14758 | 60.19 | 5.35 | 36.40 | 84.00 |
| *team2PMx* | 14758 | 29.61 | 2.52 | 17.00 | 44.00 |
| *team2Px* | 14758 | 0.49 | 0.03 | 0.31 | 0.65 |
| *team3PAx* | 14758 | 23.87 | 5.39 | 9.00 | 48.00 |
| *team3PMx* | 14758 | 8.51 | 2.23 | 0.00 | 19.00 |
| *team3Px* | 14758 | 0.35 | 0.04 | 0.00 | 0.68 |
| *teamFTAx* | 14758 | 22.92 | 3.53 | 10.00 | 50.00 |
| *teamFTMx* | 14758 | 17.41 | 2.78 | 5.00 | 34.00 |
| *teamFTx* | 14758 | 0.76 | 0.05 | 0.39 | 1.00 |
| *teamORBx* | 14758 | 10.57 | 1.86 | 2.00 | 26.00 |
| *teamDRBx* | 14758 | 32.61 | 2.48 | 19.00 | 47.00 |
| *teamTRBx* | 14758 | 43.18 | 2.94 | 30.00 | 62.00 |
| *teamPTS1x* | 14758 | 25.59 | 2.50 | 10.00 | 43.00 |
| *teamPTS2x* | 14758 | 25.61 | 2.44 | 7.00 | 37.00 |
| *teamPTS3x* | 14758 | 25.23 | 2.44 | 12.00 | 41.00 |
| *teamPTS4x* | 14758 | 25.01 | 2.32 | 12.00 | 41.00 |
| *teamPTS5x* | 14758 | 0.60 | 0.90 | 0.00 | 17.00 |
| *teamPTS6x* | 14758 | 0.09 | 0.33 | 0.00 | 6.50 |
| *teamPTS7x* | 14758 | 0.01 | 0.12 | 0.00 | 1.88 |
| *teamPTS8x* | 14758 | 0.00 | 0.08 | 0.00 | 2.00 |
| *teamTREBx* | 14758 | 50.00 | 2.32 | 36.67 | 63.33 |
| *teamASSTx* | 14758 | 58.36 | 5.15 | 27.27 | 81.58 |
| *teamTSx* | 14758 | 0.54 | 0.03 | 0.39 | 0.68 |
| *teamEeffec~x* | 14758 | 0.51 | 0.03 | 0.33 | 0.65 |
| *teamOREBx* | 14758 | 24.27 | 3.82 | 5.88 | 46.15 |
| *teamDREBx* | 14758 | 75.73 | 3.22 | 53.85 | 94.12 |
| *teamTOTurn~x* | 14758 | 13.31 | 1.55 | 6.71 | 23.24 |
| *AS* | 14758 | 8.10 | 1.35 | 1.99 | 16.11 |
| *AT* | 14758 | 5.12 | 1.25 | 0.00 | 18.66 |
| *teamBLKRx* | 14758 | 8.28 | 2.12 | 0.00 | 33.33 |
| *teamPPSpoi~x* | 14758 | 1.22 | 0.07 | 0.85 | 1.53 |
| *teamFICflo~x* | 14758 | 75.87 | 8.38 | 29.75 | 115.75 |
| *teamFIC40p~x* | 14758 | 62.78 | 6.96 | 24.69 | 96.46 |
| *teamOrtgo~10* | 14758 | 106.83 | 5.37 | 75.25 | 130.60 |
| *teamDrtgd~10* | 14758 | 106.82 | 4.93 | 75.25 | 130.60 |
| *teamEDiffe~e* | 14758 | 0.01 | 6.75 | -48.14 | 48.14 |
| *teamPlayx* | 14758 | 0.43 | 0.02 | 0.27 | 0.53 |
| *teamARx* | 14758 | 16.94 | 1.62 | 7.59 | 24.24 |
| *teamASTTOx* | 14758 | 1.68 | 0.31 | 0.48 | 3.45 |
| *teamSTLTOx* | 14758 | 57.72 | 11.98 | 14.29 | 155.56 |
| *pointspr~10x* | 14758 | 0.01 | 6.46 | -48.00 | 48.00 |
| *teamMiny* | 14758 | 241.78 | 2.60 | 237.50 | 265.00 |
| *teamDayOffy* | 14758 | 1.89 | 0.41 | 0.00 | 3.20 |
| *teamPTSy* | 14758 | 102.15 | 6.54 | 75.00 | 140.00 |
| *teamASTy* | 14758 | 22.28 | 2.58 | 9.00 | 36.00 |
| *teamTOTurn~y* | 14758 | 14.47 | 1.78 | 8.00 | 27.00 |
| *teamSTLy* | 14758 | 7.76 | 1.33 | 2.00 | 16.00 |
| *teamBLKy* | 14758 | 4.90 | 1.19 | 0.00 | 18.00 |
| *teamPFpers~y* | 14758 | 20.28 | 2.12 | 13.00 | 32.00 |
| *teamFGAy* | 14758 | 84.06 | 3.62 | 66.00 | 107.00 |
| *teamFGMy* | 14758 | 38.12 | 2.42 | 24.00 | 53.00 |
| *teamFGy* | 14758 | 0.45 | 0.02 | 0.31 | 0.57 |
| *team2PAy* | 14758 | 60.19 | 5.35 | 36.40 | 84.00 |
| *team2PMy* | 14758 | 29.61 | 2.52 | 17.00 | 44.00 |
| *team2Py* | 14758 | 0.49 | 0.03 | 0.31 | 0.65 |
| *team3PAy* | 14758 | 23.87 | 5.39 | 9.00 | 48.00 |
| *team3PMy* | 14758 | 8.51 | 2.23 | 0.00 | 19.00 |
| *team3Py* | 14758 | 0.35 | 0.04 | 0.00 | 0.68 |
| *teamFTAy* | 14758 | 22.92 | 3.53 | 10.00 | 50.00 |
| *teamFTMy* | 14758 | 17.41 | 2.78 | 5.00 | 34.00 |
| *teamFTy* | 14758 | 0.76 | 0.05 | 0.39 | 1.00 |
| *teamORBy* | 14758 | 10.57 | 1.86 | 2.00 | 26.00 |
| *teamDRBy* | 14758 | 32.61 | 2.48 | 19.00 | 47.00 |
| *teamTRBy* | 14758 | 43.18 | 2.94 | 30.00 | 62.00 |
| *teamPTS1y* | 14758 | 25.59 | 2.50 | 10.00 | 43.00 |
| *teamPTS2y* | 14758 | 25.61 | 2.44 | 7.00 | 37.00 |
| *teamPTS3y* | 14758 | 25.23 | 2.44 | 12.00 | 41.00 |
| *teamPTS4y* | 14758 | 25.01 | 2.32 | 12.00 | 41.00 |
| *teamPTS5y* | 14758 | 0.60 | 0.90 | 0.00 | 17.00 |
| *teamPTS6y* | 14758 | 0.09 | 0.33 | 0.00 | 6.50 |
| *teamPTS7y* | 14758 | 0.01 | 0.12 | 0.00 | 1.88 |
| *teamPTS8y* | 14758 | 0.00 | 0.08 | 0.00 | 2.00 |
| *teamTREBy* | 14758 | 50.00 | 2.32 | 36.67 | 63.33 |
| *teamASSTy* | 14758 | 58.36 | 5.15 | 27.27 | 81.58 |
| *teamTSy* | 14758 | 0.54 | 0.03 | 0.39 | 0.68 |
| *teamEeffec~y* | 14758 | 0.51 | 0.03 | 0.33 | 0.65 |
| *teamOREBy* | 14758 | 24.27 | 3.82 | 5.88 | 46.15 |
| *teamDREBy* | 14758 | 75.73 | 3.22 | 53.85 | 94.12 |
| *teamTOturn~y* | 14758 | 13.31 | 1.55 | 6.71 | 23.24 |
| *CS* | 14758 | 8.10 | 1.35 | 1.99 | 16.11 |
| *CT* | 14758 | 5.12 | 1.25 | 0.00 | 18.66 |
| *teamBLKRy* | 14758 | 8.28 | 2.12 | 0.00 | 33.33 |
| *teamPPSpoi~y* | 14758 | 1.22 | 0.07 | 0.85 | 1.53 |
| *teamFICflo~y* | 14758 | 75.87 | 8.38 | 29.75 | 115.75 |
| *teamFIC40p~y* | 14758 | 62.78 | 6.96 | 24.69 | 96.46 |
| *CY* | 14758 | 106.83 | 5.37 | 75.25 | 130.60 |
| *CZ* | 14758 | 106.82 | 4.93 | 75.25 | 130.60 |
| *DA* | 14758 | 0.01 | 6.75 | -48.14 | 48.14 |
| *teamPlayy* | 14758 | 0.43 | 0.02 | 0.27 | 0.53 |
| *teamARy* | 14758 | 16.94 | 1.62 | 7.59 | 24.24 |
| *teamASTTOy* | 14758 | 1.68 | 0.31 | 0.48 | 3.45 |
| *teamSTLTOy* | 14758 | 57.72 | 11.98 | 14.29 | 155.56 |
| *pointspr~10y* | 14758 | 0.01 | 6.46 | -48.00 | 48.00 |
| *pointsprea~g* | 14758 | 0.01 | 13.59 | -61.00 | 61.00 |
| *pointsprea~5* | 14758 | 0.00 | 7.81 | -28.80 | 28.80 |
| *season* | 14758 | 2015.50 | 1.71 | 2013.00 | 2018.00 |
| *area* | 14758 | 0.50 | 0.50 | 0.00 | 1.00 |
| Note: This table reports the summary statistics for NBA games observations from season 2013 to season 2018 where numbers of observations, mean, standard deviation, maximum, and minimum values are shown.  All variables in this table are defined in Appendix E.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **TABLE B.2**  **Summary Statistics for Original Train Dataset** | | | | | | | **Variable** | **Obs** | **Mean** | **Std.Dev.** | **Min** | **Max** | | *v1* | 9444 | 7369.39 | 4261.07 | 0.00 | 14755.00 | | *pointspread* | 9444 | 0.23 | 13.53 | -61.00 | 61.00 | | *teamlocx* | 9444 | 0.50 | 0.50 | 0.00 | 1.00 | | *teamdayoff* | 9444 | 1.93 | 1.04 | 0.00 | 11.00 | | *teamminx* | 9444 | 48.36 | 0.53 | 47.60 | 53.00 | | *teamptsx* | 9444 | 102.19 | 6.56 | 75.00 | 131.00 | | *teamastx* | 9444 | 22.28 | 2.57 | 9.00 | 36.00 | | *teamto~pverx* | 9444 | 14.44 | 1.78 | 8.00 | 27.00 | | *teamstlx* | 9444 | 7.77 | 1.34 | 3.00 | 16.00 | | *teamblkx* | 9444 | 4.90 | 1.19 | 0.00 | 17.00 | | *teampfpers~x* | 9444 | 20.28 | 2.10 | 14.00 | 32.00 | | *teamfgax* | 9444 | 84.08 | 3.63 | 70.00 | 104.00 | | *teamfgmx* | 9444 | 38.13 | 2.43 | 24.00 | 49.00 | | *teamfgx* | 9444 | 0.45 | 0.02 | 0.31 | 0.56 | | *team2pax* | 9444 | 60.23 | 5.42 | 36.40 | 84.00 | | *team2pmx* | 9444 | 29.64 | 2.54 | 19.00 | 43.00 | | *team2px* | 9444 | 0.49 | 0.03 | 0.31 | 0.65 | | *team3pax* | 9444 | 23.85 | 5.46 | 9.00 | 48.00 | | *team3pmx* | 9444 | 8.49 | 2.26 | 0.00 | 18.00 | | *team3px* | 9444 | 0.35 | 0.04 | 0.00 | 0.68 | | *teamftax* | 9444 | 22.97 | 3.53 | 10.00 | 50.00 | | *teamftmx* | 9444 | 17.44 | 2.78 | 5.00 | 34.00 | | *teamftx* | 9444 | 0.76 | 0.05 | 0.39 | 1.00 | | *teamorbx* | 9444 | 10.58 | 1.88 | 3.00 | 26.00 | | *teamdrbx* | 9444 | 32.61 | 2.47 | 20.00 | 47.00 | | *teamtrbx* | 9444 | 43.19 | 2.94 | 30.00 | 62.00 | | *teampts1x* | 9444 | 25.59 | 2.51 | 10.00 | 43.00 | | *teampts2x* | 9444 | 25.63 | 2.44 | 14.00 | 37.00 | | *teampts3x* | 9444 | 25.23 | 2.44 | 12.00 | 38.00 | | *teampts4x* | 9444 | 25.01 | 2.33 | 12.00 | 41.00 | | *teampts5x* | 9444 | 0.60 | 0.90 | 0.00 | 17.00 | | *teampts6x* | 9444 | 0.09 | 0.34 | 0.00 | 6.50 | | *teampts7x* | 9444 | 0.01 | 0.11 | 0.00 | 1.70 | | *teampts8x* | 9444 | 0.00 | 0.09 | 0.00 | 2.00 | | *teamtrebx* | 9444 | 50.00 | 2.32 | 36.67 | 63.33 | | *teamasstx* | 9444 | 58.34 | 5.14 | 27.27 | 81.58 | | *teamtsx* | 9444 | 0.54 | 0.03 | 0.39 | 0.68 | | *teameeffec~x* | 9444 | 0.51 | 0.03 | 0.33 | 0.64 | | *teamorebx* | 9444 | 24.29 | 3.84 | 7.32 | 46.15 | | *teamdrebx* | 9444 | 75.71 | 3.23 | 53.85 | 92.50 | | *teamto~overx* | 9444 | 13.29 | 1.55 | 7.28 | 22.43 | | *v42* | 9444 | 8.11 | 1.35 | 3.11 | 15.90 | | *v43* | 9444 | 5.13 | 1.25 | 0.00 | 17.86 | | *teamblkrx* | 9444 | 8.28 | 2.12 | 0.00 | 29.43 | | *teamppspoi~x* | 9444 | 1.22 | 0.07 | 0.85 | 1.51 | | *teamficflo~x* | 9444 | 75.92 | 8.38 | 29.75 | 115.75 | | *teamfic40p~x* | 9444 | 62.82 | 6.97 | 24.69 | 96.46 | | *teamortgof~s* | 9444 | 106.87 | 5.39 | 75.25 | 128.20 | | *teamdrtgde~s* | 9444 | 106.78 | 4.94 | 75.25 | 130.60 | | *teamediffe~x* | 9444 | 0.09 | 6.79 | -48.14 | 48.14 | | *teamplayx* | 9444 | 0.43 | 0.02 | 0.27 | 0.53 | | *teamarx* | 9444 | 16.93 | 1.62 | 7.59 | 24.24 | | *teamasttox* | 9444 | 1.68 | 0.31 | 0.48 | 3.45 | | *teamstltox* | 9444 | 57.86 | 12.02 | 15.79 | 145.45 | | *pointspr~10x* | 9444 | 0.08 | 6.49 | -48.00 | 48.00 | | *teamminy* | 9444 | 241.80 | 2.63 | 238.00 | 265.00 | | *teamdayoffy* | 9444 | 1.90 | 0.41 | 0.00 | 3.10 | | *teamptsy* | 9444 | 102.07 | 6.50 | 75.00 | 140.00 | | *teamasty* | 9444 | 22.27 | 2.55 | 10.00 | 36.00 | | *teamtoturn~y* | 9444 | 14.47 | 1.78 | 8.00 | 27.00 | | *teamstly* | 9444 | 7.77 | 1.33 | 3.00 | 16.00 | | *teamblky* | 9444 | 4.90 | 1.21 | 0.00 | 18.00 | | *teampfpers~y* | 9444 | 20.28 | 2.13 | 13.00 | 32.00 | | *teamfgay* | 9444 | 84.09 | 3.63 | 70.00 | 107.00 | | *teamfgmy* | 9444 | 38.11 | 2.42 | 24.00 | 53.00 | | *teamfgy* | 9444 | 0.45 | 0.02 | 0.31 | 0.57 | | *team2pay* | 9444 | 60.26 | 5.28 | 37.10 | 84.00 | | *team2pmy* | 9444 | 29.62 | 2.51 | 19.00 | 44.00 | | *team2py* | 9444 | 0.49 | 0.03 | 0.31 | 0.65 | | *team3pay* | 9444 | 23.83 | 5.33 | 9.00 | 48.00 | | *team3pmy* | 9444 | 8.48 | 2.21 | 0.00 | 18.00 | | *team3py* | 9444 | 0.35 | 0.04 | 0.00 | 0.59 | | *teamftay* | 9444 | 22.89 | 3.51 | 10.00 | 50.00 | | *teamftmy* | 9444 | 17.38 | 2.75 | 5.00 | 34.00 | | *teamfty* | 9444 | 0.76 | 0.05 | 0.39 | 1.00 | | *teamorby* | 9444 | 10.58 | 1.86 | 3.00 | 26.00 | | *teamdrby* | 9444 | 32.63 | 2.48 | 21.00 | 47.00 | | *teamtrby* | 9444 | 43.20 | 2.93 | 31.00 | 62.00 | | *teampts1y* | 9444 | 25.55 | 2.47 | 10.00 | 43.00 | | *teampts2y* | 9444 | 25.57 | 2.44 | 7.00 | 37.00 | | *teampts3y* | 9444 | 25.21 | 2.43 | 14.00 | 41.00 | | *teampts4y* | 9444 | 25.01 | 2.33 | 12.50 | 41.00 | | *teampts5y* | 9444 | 0.61 | 0.90 | 0.00 | 15.00 | | *teampts6y* | 9444 | 0.09 | 0.33 | 0.00 | 6.50 | | *teampts7y* | 9444 | 0.01 | 0.12 | 0.00 | 1.70 | | *teampts8y* | 9444 | 0.00 | 0.08 | 0.00 | 2.00 | | *teamtreby* | 9444 | 50.00 | 2.32 | 36.67 | 63.33 | | *teamassty* | 9444 | 58.35 | 5.11 | 35.14 | 81.58 | | *teamtsy* | 9444 | 0.54 | 0.03 | 0.39 | 0.68 | | *teameeffec~y* | 9444 | 0.50 | 0.03 | 0.33 | 0.64 | | *teamoreby* | 9444 | 24.28 | 3.84 | 7.89 | 46.15 | | *teamdreby* | 9444 | 75.73 | 3.23 | 53.85 | 92.68 | | *v93* | 9444 | 13.31 | 1.56 | 7.28 | 23.24 | | *v94* | 9444 | 8.11 | 1.35 | 2.94 | 16.11 | | *v95* | 9444 | 5.13 | 1.27 | 0.00 | 18.66 | | *teamblkry* | 9444 | 8.27 | 2.14 | 0.00 | 33.33 | | *teamppspoi~y* | 9444 | 1.22 | 0.07 | 0.85 | 1.53 | | *teamficflo~y* | 9444 | 75.80 | 8.33 | 29.75 | 115.75 | | *teamfic40p~y* | 9444 | 62.72 | 6.92 | 24.69 | 96.46 | | *v100* | 9444 | 106.74 | 5.36 | 75.25 | 130.60 | | *v101* | 9444 | 106.80 | 4.95 | 75.25 | 128.20 | | *teamediffe~y* | 9444 | -0.05 | 6.75 | -48.14 | 48.14 | | *teamplayy* | 9444 | 0.43 | 0.02 | 0.27 | 0.52 | | *teamary* | 9444 | 16.93 | 1.61 | 8.14 | 24.24 | | *teamasttoy* | 9444 | 1.68 | 0.31 | 0.48 | 3.45 | | *teamstltoy* | 9444 | 57.82 | 12.09 | 15.79 | 155.56 | | *pointspr~10y* | 9444 | -0.05 | 6.46 | -48.00 | 48.00 | | *pointsprea~g* | 9444 | 0.14 | 13.63 | -61.00 | 61.00 | | *pointsprea~5* | 9444 | 0.11 | 7.82 | -27.80 | 28.80 | | Note: This table reports the summary statistics for original train dataset from season 2013 to season 2018 where numbers of observations, mean, standard deviation, maximum, and minimum values are shown.  All variables in this table are defined in Appendix E. | | | | | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **TABLE B.3**  **Summary Statistics for Original Validation Dataset** | | | | | | | **Variable** | **Obs** | **Mean** | **Std.Dev.** | **Min** | **Max** | | *v1* | 2362 | 7325.67 | 4260.84 | 4.00 | 14751.00 | | *pointspread* | 2362 | -0.46 | 14.01 | -49.00 | 49.00 | | *teamlocx* | 2362 | 0.49 | 0.50 | 0.00 | 1.00 | | *teamdayoff* | 2362 | 1.94 | 1.10 | 0.00 | 10.00 | | *teamminx* | 2362 | 48.36 | 0.52 | 47.71 | 51.50 | | *teamptsx* | 2362 | 102.11 | 6.50 | 79.00 | 140.00 | | *teamastx* | 2362 | 22.25 | 2.54 | 13.50 | 32.80 | | *teamto~pverx* | 2362 | 14.50 | 1.77 | 8.00 | 24.00 | | *teamstlx* | 2362 | 7.79 | 1.35 | 2.00 | 16.00 | | *teamblkx* | 2362 | 4.88 | 1.19 | 0.00 | 11.00 | | *teampfpers~x* | 2362 | 20.31 | 2.13 | 14.00 | 30.00 | | *teamfgax* | 2362 | 84.05 | 3.58 | 71.00 | 102.00 | | *teamfgmx* | 2362 | 38.11 | 2.38 | 29.00 | 53.00 | | *teamfgx* | 2362 | 0.45 | 0.02 | 0.36 | 0.57 | | *team2pax* | 2362 | 60.17 | 5.19 | 37.50 | 77.33 | | *team2pmx* | 2362 | 29.60 | 2.46 | 17.00 | 44.00 | | *team2px* | 2362 | 0.49 | 0.03 | 0.35 | 0.65 | | *team3pax* | 2362 | 23.88 | 5.31 | 9.00 | 45.70 | | *team3pmx* | 2362 | 8.51 | 2.19 | 2.00 | 19.00 | | *team3px* | 2362 | 0.35 | 0.04 | 0.17 | 0.60 | | *teamftax* | 2362 | 22.87 | 3.54 | 12.00 | 39.13 | | *teamftmx* | 2362 | 17.39 | 2.80 | 10.00 | 32.00 | | *teamftx* | 2362 | 0.76 | 0.05 | 0.52 | 0.95 | | *teamorbx* | 2362 | 10.53 | 1.86 | 2.00 | 21.00 | | *teamdrbx* | 2362 | 32.61 | 2.51 | 21.00 | 42.33 | | *teamtrbx* | 2362 | 43.14 | 2.96 | 31.00 | 55.00 | | *teampts1x* | 2362 | 25.58 | 2.48 | 15.00 | 37.33 | | *teampts2x* | 2362 | 25.62 | 2.42 | 13.00 | 36.00 | | *teampts3x* | 2362 | 25.23 | 2.43 | 16.00 | 41.00 | | *teampts4x* | 2362 | 24.97 | 2.36 | 14.00 | 34.00 | | *teampts5x* | 2362 | 0.62 | 0.92 | 0.00 | 8.50 | | *teampts6x* | 2362 | 0.08 | 0.31 | 0.00 | 2.40 | | *teampts7x* | 2362 | 0.02 | 0.14 | 0.00 | 1.70 | | *teampts8x* | 2362 | 0.00 | 0.03 | 0.00 | 1.70 | | *teamtrebx* | 2362 | 49.96 | 2.36 | 38.30 | 61.11 | | *teamasstx* | 2362 | 58.31 | 5.16 | 39.47 | 75.80 | | *teamtsx* | 2362 | 0.54 | 0.03 | 0.43 | 0.67 | | *teameeffec~x* | 2362 | 0.51 | 0.03 | 0.39 | 0.64 | | *teamorebx* | 2362 | 24.19 | 3.82 | 5.88 | 43.75 | | *teamdrebx* | 2362 | 75.71 | 3.18 | 55.77 | 87.50 | | *teamto~overx* | 2362 | 13.34 | 1.54 | 7.46 | 23.24 | | *v42* | 2362 | 8.12 | 1.36 | 1.99 | 16.11 | | *v43* | 2362 | 5.11 | 1.25 | 0.00 | 11.64 | | *teamblkrx* | 2362 | 8.25 | 2.07 | 0.00 | 16.67 | | *teamppspoi~x* | 2362 | 1.22 | 0.07 | 0.94 | 1.53 | | *teamficflo~x* | 2362 | 75.75 | 8.30 | 39.13 | 113.75 | | *teamfic40p~x* | 2362 | 62.68 | 6.89 | 32.60 | 94.79 | | *teamortgof~s* | 2362 | 106.75 | 5.36 | 81.54 | 130.60 | | *teamdrtgde~s* | 2362 | 106.83 | 4.95 | 87.31 | 125.92 | | *teamediffe~x* | 2362 | -0.08 | 6.73 | -28.89 | 33.31 | | *teamplayx* | 2362 | 0.43 | 0.02 | 0.34 | 0.52 | | *teamarx* | 2362 | 16.92 | 1.61 | 10.69 | 22.61 | | *teamasttox* | 2362 | 1.68 | 0.31 | 0.79 | 3.13 | | *teamstltox* | 2362 | 57.75 | 11.85 | 14.29 | 105.98 | | *pointspr~10x* | 2362 | -0.08 | 6.43 | -29.00 | 33.00 | | *teamminy* | 2362 | 241.77 | 2.54 | 238.80 | 257.50 | | *teamdayoffy* | 2362 | 1.89 | 0.40 | 0.00 | 3.20 | | *teamptsy* | 2362 | 102.35 | 6.64 | 85.00 | 131.00 | | *teamasty* | 2362 | 22.29 | 2.64 | 9.00 | 33.30 | | *teamtoturn~y* | 2362 | 14.44 | 1.78 | 8.00 | 22.14 | | *teamstly* | 2362 | 7.73 | 1.34 | 2.00 | 16.00 | | *teamblky* | 2362 | 4.89 | 1.16 | 1.00 | 11.00 | | *teampfpers~y* | 2362 | 20.31 | 2.10 | 14.20 | 29.50 | | *teamfgay* | 2362 | 84.05 | 3.72 | 66.00 | 102.00 | | *teamfgmy* | 2362 | 38.16 | 2.46 | 30.00 | 49.00 | | *teamfgy* | 2362 | 0.46 | 0.02 | 0.35 | 0.56 | | *team2pay* | 2362 | 60.13 | 5.50 | 36.70 | 81.00 | | *team2pmy* | 2362 | 29.62 | 2.58 | 20.25 | 43.00 | | *team2py* | 2362 | 0.50 | 0.03 | 0.39 | 0.63 | | *team3pay* | 2362 | 23.92 | 5.44 | 10.90 | 45.20 | | *team3pmy* | 2362 | 8.54 | 2.26 | 3.00 | 17.90 | | *team3py* | 2362 | 0.35 | 0.04 | 0.20 | 0.50 | | *teamftay* | 2362 | 23.04 | 3.58 | 12.20 | 46.00 | | *teamftmy* | 2362 | 17.50 | 2.81 | 8.40 | 32.00 | | *teamfty* | 2362 | 0.76 | 0.04 | 0.61 | 0.93 | | *teamorby* | 2362 | 10.57 | 1.83 | 4.00 | 17.50 | | *teamdrby* | 2362 | 32.58 | 2.45 | 24.20 | 42.00 | | *teamtrby* | 2362 | 43.16 | 2.91 | 33.00 | 53.40 | | *teampts1y* | 2362 | 25.67 | 2.52 | 15.50 | 37.33 | | *teampts2y* | 2362 | 25.67 | 2.54 | 11.00 | 36.00 | | *teampts3y* | 2362 | 25.28 | 2.52 | 15.67 | 37.00 | | *teampts4y* | 2362 | 25.01 | 2.34 | 14.67 | 35.00 | | *teampts5y* | 2362 | 0.60 | 0.87 | 0.00 | 6.50 | | *teampts6y* | 2362 | 0.09 | 0.33 | 0.00 | 3.33 | | *teampts7y* | 2362 | 0.01 | 0.11 | 0.00 | 1.88 | | *teampts8y* | 2362 | 0.01 | 0.10 | 0.00 | 2.00 | | *teamtreby* | 2362 | 50.02 | 2.29 | 38.89 | 58.63 | | *teamassty* | 2362 | 58.31 | 5.26 | 27.27 | 75.16 | | *teamtsy* | 2362 | 0.54 | 0.03 | 0.45 | 0.64 | | *teameeffec~y* | 2362 | 0.51 | 0.03 | 0.39 | 0.61 | | *teamoreby* | 2362 | 24.30 | 3.72 | 10.26 | 37.90 | | *teamdreby* | 2362 | 75.77 | 3.28 | 56.25 | 94.12 | | *v93* | 2362 | 13.29 | 1.59 | 6.71 | 21.89 | | *v94* | 2362 | 8.06 | 1.37 | 1.99 | 15.90 | | *v95* | 2362 | 5.12 | 1.22 | 1.08 | 11.64 | | *teamblkry* | 2362 | 8.28 | 2.04 | 1.67 | 16.54 | | *teamppspoi~y* | 2362 | 1.22 | 0.07 | 1.01 | 1.45 | | *teamficflo~y* | 2362 | 75.99 | 8.49 | 44.25 | 107.43 | | *teamfic40p~y* | 2362 | 62.88 | 7.06 | 36.88 | 89.36 | | *v100* | 2362 | 107.04 | 5.40 | 89.23 | 126.40 | | *v101* | 2362 | 106.91 | 4.97 | 81.54 | 130.60 | | *teamediffe~y* | 2362 | 0.13 | 6.91 | -29.36 | 28.89 | | *teamplayy* | 2362 | 0.43 | 0.02 | 0.34 | 0.53 | | *teamary* | 2362 | 16.94 | 1.66 | 7.59 | 23.01 | | *teamasttoy* | 2362 | 1.68 | 0.31 | 0.64 | 3.00 | | *teamstltoy* | 2362 | 57.55 | 12.01 | 14.29 | 125.00 | | *pointspr~10y* | 2362 | 0.13 | 6.60 | -29.00 | 29.00 | | *pointsprea~g* | 2362 | -0.25 | 13.60 | -49.00 | 49.00 | | *pointsprea~5* | 2362 | -0.32 | 7.61 | -22.60 | 24.40 | | Note: This table reports the summary statistics for original validation dataset from season 2013 to season 2018 where numbers of observations, mean, standard deviation, maximum, and minimum values are shown.  All variables in this table are defined in Appendix E. | | | | | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **TABLE B.4**  **Summary Statistics for Original Test Dataset** | | | | | | | | | **Variable** | **Obs** | **Mean** | **Std.Dev.** | **Min** | **Max** | | *v1* | 2952 | 7449.93 | 4258.51 | 9.00 | 14757.00 | | *pointspread* | 2952 | -0.36 | 13.95 | -51.00 | 51.00 | | *teamlocx* | 2952 | 0.52 | 0.50 | 0.00 | 1.00 | | *teamdayoff* | 2952 | 1.96 | 1.13 | 0.00 | 11.00 | | *teamminx* | 2952 | 48.36 | 0.50 | 47.50 | 51.50 | | *teamptsx* | 2952 | 102.08 | 6.51 | 82.00 | 125.00 | | *teamastx* | 2952 | 22.31 | 2.62 | 12.00 | 34.00 | | *teamto~pverx* | 2952 | 14.52 | 1.78 | 8.00 | 26.00 | | *teamstlx* | 2952 | 7.72 | 1.31 | 3.00 | 14.00 | | *teamblkx* | 2952 | 4.90 | 1.20 | 0.00 | 18.00 | | *teampfpers~x* | 2952 | 20.29 | 2.15 | 13.00 | 31.50 | | *teamfgax* | 2952 | 84.02 | 3.60 | 66.00 | 107.00 | | *teamfgmx* | 2952 | 38.10 | 2.42 | 30.20 | 46.90 | | *teamfgx* | 2952 | 0.45 | 0.02 | 0.37 | 0.55 | | *team2pax* | 2952 | 60.09 | 5.22 | 39.30 | 78.00 | | *team2pmx* | 2952 | 29.54 | 2.50 | 19.67 | 40.00 | | *team2px* | 2952 | 0.49 | 0.03 | 0.40 | 0.59 | | *team3pax* | 2952 | 23.94 | 5.26 | 11.00 | 45.40 | | *team3pmx* | 2952 | 8.56 | 2.18 | 2.67 | 17.50 | | *team3px* | 2952 | 0.36 | 0.04 | 0.17 | 0.56 | | *teamftax* | 2952 | 22.80 | 3.50 | 12.00 | 46.00 | | *teamftmx* | 2952 | 17.32 | 2.75 | 8.00 | 32.67 | | *teamftx* | 2952 | 0.76 | 0.05 | 0.53 | 1.00 | | *teamorbx* | 2952 | 10.56 | 1.80 | 6.00 | 17.00 | | *teamdrbx* | 2952 | 32.63 | 2.48 | 19.00 | 42.50 | | *teamtrbx* | 2952 | 43.19 | 2.91 | 31.00 | 59.00 | | *teampts1x* | 2952 | 25.60 | 2.47 | 17.00 | 36.30 | | *teampts2x* | 2952 | 25.55 | 2.47 | 7.00 | 36.00 | | *teampts3x* | 2952 | 25.21 | 2.44 | 15.00 | 39.00 | | *teampts4x* | 2952 | 25.01 | 2.28 | 12.50 | 37.00 | | *teampts5x* | 2952 | 0.60 | 0.86 | 0.00 | 6.50 | | *teampts6x* | 2952 | 0.09 | 0.31 | 0.00 | 2.30 | | *teampts7x* | 2952 | 0.01 | 0.13 | 0.00 | 1.88 | | *teampts8x* | 2952 | 0.00 | 0.07 | 0.00 | 2.00 | | *teamtrebx* | 2952 | 50.05 | 2.31 | 40.40 | 59.34 | | *teamasstx* | 2952 | 58.46 | 5.19 | 32.43 | 75.71 | | *teamtsx* | 2952 | 0.54 | 0.03 | 0.41 | 0.66 | | *teameeffec~x* | 2952 | 0.51 | 0.03 | 0.40 | 0.65 | | *teamorebx* | 2952 | 24.30 | 3.74 | 13.95 | 37.73 | | *teamdrebx* | 2952 | 75.81 | 3.21 | 57.78 | 94.12 | | *teamto~overx* | 2952 | 13.37 | 1.56 | 6.71 | 22.85 | | *v42* | 2952 | 8.05 | 1.33 | 2.99 | 14.14 | | *v43* | 2952 | 5.13 | 1.26 | 0.00 | 18.66 | | *teamblkrx* | 2952 | 8.30 | 2.15 | 0.00 | 33.33 | | *teamppspoi~x* | 2952 | 1.22 | 0.07 | 0.87 | 1.45 | | *teamficflo~x* | 2952 | 75.81 | 8.42 | 42.75 | 114.63 | | *teamfic40p~x* | 2952 | 62.73 | 6.99 | 35.63 | 95.52 | | *teamortgof~s* | 2952 | 106.78 | 5.33 | 83.53 | 125.67 | | *teamdrtgde~s* | 2952 | 106.94 | 4.89 | 86.05 | 122.11 | | *teamediffe~x* | 2952 | -0.15 | 6.65 | -24.91 | 30.08 | | *teamplayx* | 2952 | 0.43 | 0.02 | 0.34 | 0.52 | | *teamarx* | 2952 | 16.96 | 1.64 | 9.14 | 23.43 | | *teamasttox* | 2952 | 1.68 | 0.31 | 0.50 | 2.95 | | *teamstltox* | 2952 | 57.24 | 11.93 | 15.79 | 155.56 | | *pointspr~10x* | 2952 | -0.14 | 6.37 | -25.00 | 31.00 | | *teamminy* | 2952 | 241.74 | 2.55 | 237.50 | 265.00 | | *teamdayoffy* | 2952 | 1.89 | 0.41 | 0.00 | 2.90 | | *teamptsy* | 2952 | 102.26 | 6.58 | 82.00 | 123.30 | | *teamasty* | 2952 | 22.32 | 2.60 | 13.50 | 33.80 | | *teamtoturn~y* | 2952 | 14.48 | 1.75 | 8.00 | 21.50 | | *teamstly* | 2952 | 7.76 | 1.33 | 3.00 | 14.00 | | *teamblky* | 2952 | 4.90 | 1.16 | 0.00 | 12.00 | | *teampfpers~y* | 2952 | 20.27 | 2.09 | 14.00 | 31.00 | | *teamfgay* | 2952 | 83.99 | 3.51 | 71.00 | 100.00 | | *teamfgmy* | 2952 | 38.13 | 2.41 | 29.00 | 46.90 | | *teamfgy* | 2952 | 0.46 | 0.02 | 0.36 | 0.55 | | *team2pay* | 2952 | 60.03 | 5.43 | 36.40 | 80.00 | | *team2pmy* | 2952 | 29.56 | 2.51 | 17.00 | 40.00 | | *team2py* | 2952 | 0.50 | 0.03 | 0.35 | 0.64 | | *team3pay* | 2952 | 23.96 | 5.54 | 10.00 | 45.70 | | *team3pmy* | 2952 | 8.57 | 2.29 | 2.00 | 19.00 | | *team3py* | 2952 | 0.36 | 0.04 | 0.12 | 0.68 | | *teamftay* | 2952 | 22.92 | 3.54 | 10.00 | 40.00 | | *teamftmy* | 2952 | 17.43 | 2.81 | 6.00 | 32.60 | | *teamfty* | 2952 | 0.76 | 0.05 | 0.52 | 1.00 | | *teamorby* | 2952 | 10.53 | 1.86 | 2.00 | 23.00 | | *teamdrby* | 2952 | 32.59 | 2.49 | 19.00 | 46.00 | | *teamtrby* | 2952 | 43.12 | 2.98 | 30.00 | 59.50 | | *teampts1y* | 2952 | 25.63 | 2.56 | 17.00 | 38.00 | | *teampts2y* | 2952 | 25.69 | 2.37 | 15.00 | 33.10 | | *teampts3y* | 2952 | 25.25 | 2.40 | 12.00 | 39.00 | | *teampts4y* | 2952 | 24.97 | 2.30 | 12.00 | 36.00 | | *teampts5y* | 2952 | 0.59 | 0.91 | 0.00 | 17.00 | | *teampts6y* | 2952 | 0.09 | 0.33 | 0.00 | 2.50 | | *teampts7y* | 2952 | 0.01 | 0.12 | 0.00 | 1.70 | | *teampts8y* | 2952 | 0.00 | 0.06 | 0.00 | 1.70 | | *teamtreby* | 2952 | 49.99 | 2.35 | 40.00 | 60.30 | | *teamassty* | 2952 | 58.45 | 5.22 | 36.75 | 79.49 | | *teamtsy* | 2952 | 0.54 | 0.03 | 0.42 | 0.67 | | *teameeffec~y* | 2952 | 0.51 | 0.03 | 0.38 | 0.65 | | *teamoreby* | 2952 | 24.23 | 3.82 | 5.88 | 42.22 | | *teamdreby* | 2952 | 75.72 | 3.14 | 57.14 | 90.19 | | *v93* | 2952 | 13.33 | 1.51 | 7.35 | 19.60 | | *v94* | 2952 | 8.09 | 1.35 | 3.11 | 13.92 | | *v95* | 2952 | 5.13 | 1.21 | 0.00 | 12.66 | | *teamblkry* | 2952 | 8.32 | 2.11 | 0.00 | 22.22 | | *teamppspoi~y* | 2952 | 1.22 | 0.07 | 0.92 | 1.51 | | *teamficflo~y* | 2952 | 76.01 | 8.42 | 47.00 | 111.07 | | *teamfic40p~y* | 2952 | 62.91 | 7.00 | 39.17 | 91.34 | | *v100* | 2952 | 106.95 | 5.40 | 86.05 | 125.88 | | *v101* | 2952 | 106.83 | 4.84 | 83.53 | 121.74 | | *teamediffe~y* | 2952 | 0.13 | 6.63 | -30.34 | 24.63 | | *teamplayy* | 2952 | 0.43 | 0.02 | 0.33 | 0.52 | | *teamary* | 2952 | 16.97 | 1.64 | 9.90 | 23.33 | | *teamasttoy* | 2952 | 1.68 | 0.30 | 0.72 | 2.98 | | *teamstltoy* | 2952 | 57.55 | 11.58 | 20.00 | 120.00 | | *pointspr~10y* | 2952 | 0.12 | 6.35 | -30.67 | 25.00 | | *pointsprea~g* | 2952 | -0.21 | 13.46 | -51.00 | 48.00 | | *pointsprea~5* | 2952 | -0.09 | 7.91 | -28.80 | 28.40 | | Note: This table reports the summary statistics for original test dataset from season 2013 to season 2018 where numbers of observations, mean, standard deviation, maximum, and minimum values are shown.  All variables in this table are defined in Appendix E. | | | | | | | | | | | |

1. **Results for All Models**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE C.1**  **Results for All Models and Comparison between Linear Regressions and Machine Learning Algorithms and Prior Literature Variables** | | | | | | | | | | | | | | |
| **Panel A: Results for Linear Regression** | | | | | | | | | | | | | |  |
| **Target Training Parameter** | **Datasets** | **MAE** | | | | | **RMSE** | | | | **Explained Variance** | | | |
| **Accuracy Measures** |  | **Explained Variance** | **Max Error** | | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** |
| **Normalization** | **All Features** | 0.12 | 51.82 | | 10.94 | -0.02 | 0.12 | 51.82 | 10.94 | -0.02 | 0.12 | 51.82 | 10.94 | -0.02 |
| **Feature Selected** | - | + | | + | - | - | + | + | - | - | + | + | - |
| **PCA** | 0.14 | 51.22 | | 13.07 | -0.39 | 0.14 | 51.22 | 13.07 | -0.39 | 0.14 | 51.22 | 13.07 | -0.39 |
| **Standardization** | **All Features** | 0.13 | 56.72 | | 9.97 | 0.13 | 0.13 | 56.72 | 9.97 | 0.13 | 0.13 | 56.72 | 9.97 | 0.13 |
| **Feature Selected** | - | + | | + | - | - | + | + | - | - | + | + | - |
| **PCA** | 0.17 | 55.10 | | 9.77 | 0.16 | 0.17 | 55.10 | 9.77 | 0.16 | 0.17 | 55.10 | 9.77 | 0.16 |
| **Panel B: Results for Decision Tree** | | | | | | | | | | | | | | |
| **Target Training Parameter** | **Datasets** | **MAE** | | | | | **RMSE** | | | | **Explained Variance** | | | |
| **Accuracy Measures** |  | **Explained Variance** | | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** |
| **Normalization** | **All Features** | 0.10 | 50.69 | | 12.00 | -0.21 | 0.10 | 50.69 | 12.00 | -0.21 | 0.10 | 50.69 | 12.00 | -0.21 |
| **Feature Selected** | 0.10 | 50.69 | | 12.00 | -0.21 | 0.10 | 50.69 | 12.00 | -0.21 | 0.10 | 50.69 | 12.00 | -0.21 |
| **PCA** | 0.08 | 52.67 | | 12.20 | -0.26 | 0.08 | 52.67 | 12.20 | -0.26 | 0.08 | 52.67 | 12.20 | -0.26 |
| **Standardization** | **All Features** | 0.14 | 55.60 | | 9.88 | 0.14 | 0.13 | 54.13 | 9.98 | 0.13 | 0.14 | 55.60 | 9.88 | 0.14 |
| **Feature Selected** | 0.14 | 55.60 | | 9.88 | 0.14 | 0.14 | 55.60 | 9.88 | 0.14 | 0.14 | 55.60 | 9.88 | 0.14 |
| **PCA** | 0.10 | 55.07 | | 10.08 | 0.10 | 0.10 | 55.07 | 10.08 | 0.10 | 0.10 | 55.07 | 10.08 | 0.10 |
| **Prior Literature** | **All** | 0.08 | 67.52 | | 10.64 | 0.08 |  |  |  |  |  |  |  |  |
| **Panel C: Results for Random Forest** | | | | | | | | | | | | | | |
| **Target Training Parameter** | **Datasets** | **MAE** | | | | | **RMSE** | | | | **Explained Variance** | | | |
| **Accuracy Measures** |  | **Explained Variance** | **Max Error** | | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** |
| **Normalization** | **All Features** | 0.14 | 50.77 | | 11.07 | -0.06 | 0.15 | 50.02 | 11.10 | -0.06 | 0.14 | 51.20 | 11.13 | -0.07 |
| **Feature Selected** | 0.07 | 53.68 | | 10.89 | -0.05 | 0.07 | 54.30 | 10.70 | -0.01 | 0.07 | 54.29 | 10.72 | -0.02 |
| **PCA** | 0.07 | 54.24 | | 11.41 | -0.14 | 0.07 | 54.01 | 11.42 | -0.14 | 0.07 | 54.12 | 11.36 | -0.13 |
| **Standardization** | **All Features** | 0.16 | 54.11 | | 9.77 | 0.16 | 0.16 | 54.86 | 9.77 | 0.16 | 0.16 | 54.20 | 9.75 | 0.16 |
| **Feature Selected** | 0.16 | 55.42 | | 9.74 | 0.15 | 0.15 | 55.35 | 9.77 | 0.15 | 0.13 | 59.79 | 9.91 | 0.13 |
| **PCA** | 0.11 | 54.95 | | 10.08 | 0.11 | 0.11 | 54.99 | 10.06 | 0.11 | 0.11 | 54.77 | 10.07 | 0.11 |
| **Prior Literature** | **All** | 0.12 | 61.38 | | 10.19 | 0.12 |  |  |  |  |  |  |  |  |
| **Panel D: Results for XGBoost** | | | | | | | | | | | | | | |
| **Target Training Parameter** | **Datasets** | **MAE** | | | | | **RMSE** | | | | **Explained Variance** | | | |
| **Accuracy Measures** |  | **Explained Variance** | **Max Error** | | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** |
| **Normalization** | **All Features** | 0.15 | 51.01 | | 10.64 | 0.02 | 0.14 | 52.74 | 10.57 | 0.03 | 0.15 | 50.84 | 10.64 | 0.02 |
| **Feature Selected** | 0.14 | 54.03 | | 10.47 | 0.04 | 0.16 | 51.62 | 10.32 | 0.07 | 0.14 | 49.91 | 10.44 | 0.04 |
| **PCA** | 0.13 | 52.69 | | 11.58 | -0.14 | 0.12 | 52.76 | 11.69 | -0.17 | 0.14 | 54.66 | 13.36 | -0.43 |
| **Standardization** | **All Features** | 0.15 | 53.87 | | 9.84 | 0.15 | 0.17 | 54.28 | 9.71 | 0.16 | 0.16 | 54.26 | 9.84 | 0.16 |
| **Feature Selected** | 0.15 | 55.79 | | 9.82 | 0.15 | 0.15 | 55.61 | 9.86 | 0.15 | 0.16 | 55.40 | 9.84 | 0.15 |
| **PCA** | 0.14 | 54.16 | | 9.93 | 0.14 | 0.16 | 53.79 | 9.86 | 0.15 | 0.14 | 51.89 | 10.27 | 0.08 |
| **Prior Literature** | **All** | 0.11 | 59.87 | | 10.19 | 0.11 |  |  |  |  |  |  |  |  |
| **Panel E: Results for Support Vector Machine** | | | | | | | | | | | | | | |
| **Target Training Parameter** | **Datasets** | **MAE** | | | | | **RMSE** | | | | **Explained Variance** | | | |
| **Accuracy Measures** |  | **Explained Variance** | **Max Error** | | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** |
| **Normalization** | **All Features** | 0.16 | 52.25 | | 10.24 | 0.08 | 0.16 | 52.25 | 10.24 | 0.08 | 0.16 | 52.25 | 10.24 | 0.08 |
| **Feature Selected** | 0.16 | 53.62 | | 10.11 | 0.10 | 0.16 | 53.62 | 10.11 | 0.10 | 0.16 | 53.62 | 10.11 | 0.10 |
| **PCA** | 0.12 | 53.17 | | 13.44 | -0.45 | 0.15 | 50.93 | 12.22 | -0.23 | 0.15 | 50.93 | 12.22 | -0.23 |
| **Standardization** | **All Features** | 0.17 | 55.81 | | 9.72 | 0.17 | 0.17 | 55.81 | 9.72 | 0.17 | 0.17 | 55.81 | 9.72 | 0.17 |
| **Feature Selected** | 0.17 | 56.83 | | 9.73 | 0.16 | 0.17 | 56.83 | 9.73 | 0.16 | 0.17 | 56.83 | 9.73 | 0.16 |
| **PCA** | 0.16 | 55.47 | | 9.77 | 0.16 | 0.16 | 55.47 | 9.77 | 0.16 | 0.16 | 55.47 | 9.77 | 0.16 |
| **Prior Literature** | **All** | 0.03 | 62.88 | | 11.13 | 0.03 |  |  |  |  |  |  |  |  |
| **Panel F: Results for Light GBM** | | | | | | | | | | | | | | |
| **Target Training Parameter** | **Datasets** | **MAE** | | | | | **RMSE** | | | | **Explained Variance** | | | |
| **Accuracy Measures** |  | **Explained Variance** | **Max Error** | | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** | **Explained Variance** | **Max Error** | **MAE** | **R2** |
| **Normalization** | **All Features** | 0.16 | 52.23 | | 10.40 | 0.06 | 0.15 | 51.95 | 10.31 | 0.07 | 0.15 | 51.95 | 10.31 | 0.07 |
| **Feature Selected** | 0.16 | 53.05 | | 10.38 | 0.06 | 0.15 | 52.03 | 10.41 | 0.05 | 0.15 | 52.03 | 10.41 | 0.05 |
| **PCA** | 0.12 | 52.55 | | 11.68 | -0.17 | 0.12 | 52.55 | 11.68 | -0.17 | 0.12 | 52.55 | 11.68 | -0.17 |
| **Standardization** | **All Features** | 0.16 | 55.52 | | 9.81 | 0.16 | 0.16 | 55.70 | 9.77 | 0.16 | 0.16 | 55.70 | 9.77 | 0.16 |
| **Feature Selected** | 0.15 | 58.65 | | 9.84 | 0.15 | 0.15 | 58.65 | 9.84 | 0.15 | 0.15 | 58.65 | 9.84 | 0.15 |
| **PCA** | 0.15 | 55.01 | | 9.83 | 0.15 | 0.15 | 55.01 | 9.83 | 0.15 | 0.15 | 55.01 | 9.83 | 0.15 |
| **Prior Literature** | **All** | 0.12 | 58.33 | | 11.78 | 0.12 |  |  |  |  |  |  |  |  |
| **Panel G: Results for Light DNN** | | | | | | | | | | | | | | |
| **Accuracy Measures** | **Datasets** | **Max Error** | **MAE** | | **RMSE** | **R2** |  |  |  |  |  |  |  |  |
| **Normalization** | **All Features** | 55.23 | 9.84 | | 12.61 | 0.15 |  |  |  |  |  |  |  |  |
| **Feature Selected** | 54.67 | 9.82 | | 12.61 | 0.15 |  |  |  |  |  |  |  |  |
| **PCA** | 56.10 | 9.86 | | 12.71 | 0.13 |  |  |  |  |  |  |  |  |
| **Prior Literature** | **All** | 59.32 | 10.32 | | 13.28 | 0.12 |  |  |  |  |  |  |  |  |
| Table 3 reports all models’ results and compares linear regressions and ML algorithms. Panel A reports the accuracy results for the linear regressions. Panel A reports the accuracy results for the DTR. Panel C reports the accuracy results for the RFR. Panel D reports the accuracy results for the XGBoost. Panel E reports the accuracy results for the SVM. Panel F reports the accuracy results for the LightGBM. Panel G reports the accuracy results for the DNN. “+” and “-” mean a very big positive or a very big negative number. | | | | | | | | | | | | | | |

1. **Best Models’ Visualization**

**Tree Based Models (Decision Tree, Random Forest, XGBoost, and LightGBM)**



**Deep Neural Network**

**Diagram

Description automatically generated**

**Chart, radar chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

1. **Additional Tables and Figures**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE E.1**  **Detail of the Characteristics** | | | | | | | | |
| **NO.** | **Acronym** | **Team Characteristics** | | **Calculation or Explanation** | | | | **Avg Last 10 Games or Not** |
| 1 | *gmDate* | Game date | | the day that this game started on | | | | no |
| 2 | *teamAbbr* | Team name | |  | | | | no |
| 3 | *teamID* | Team ID number | |  | | | | no |
| 4 | *opptAbbr* | Opponent team name | |  | | | | no |
| 5 | *pointspread* | Point spread | | team's total points minus opponent’s total points | | | | no |
| 6 | *teamLoc.x* | Team location | | home or away based on where this game played | | | | no |
| 7 | *teamDayOff* | Team day off | | days team rested before the coming game | | | | no |
| 8 | *teamMin.x* | Team minutes | | how long the team played for one game | | | | yes |
| 9 | *teamPTS.x* | Team points | | number of points the team get for one game | | | | yes |
| 10 | *teamAST.x* | Team assists | | number of assists the team get for one game | | | | yes |
| 11 | *teamTO.turNOPver..x* | Team turnovers | | number of turnovers the team get for one game | | | | yes |
| 12 | *teamSTL.x* | Team steal | | number of steals the team get for one game | | | | yes |
| 13 | *teamBLK.x* | Team Block | | number of Blocks the team get for one game | | | | yes |
| 14 | *teamPF.personal.fouls..x* | Team personal fouls | | number of personal fouls the team get for a game | | | | yes |
| 15 | *teamFGA.x* | Team field goal attempts | | number of field goal attempts the team have | | | | yes |
| 16 | *teamFGM.x* | Team field goal makes | | number of field goal makes the team have | | | | yes |
| 17 | *teamFG..x* | Team field goal percentage | | team field goal makes/ team field goal attempts | | | | yes |
| 18 | *team2PA.x* | 2 points attempts | | number of 2 points attempts the team have | | | | yes |
| 19 | *team2PM.x* | 2 points makes | | number of 2 points makes the team have | | | | yes |
| 20 | *team2P..x* | 2 points percentage | | team 2 points makes/ team 2 points attempts | | | | yes |
| 21 | *team3PA.x* | 3 points attempts | | number of 3 points attempts the team have | | | | yes |
| 22 | *team3PM.x* | 3 points makes | | number of 3 points makes the team have | | | | yes |
| 23 | *team3P..x* | 3 points percentage | | team 3 points makes/ team 3 points attempts | | | | yes |
| 24 | *teamFTA.x* | Free throw attempts | | number of free throws attempts the team have | | | | yes |
| 25 | *teamFTM.x* | Free throw makes | | number of free throws makes the team have | | | | yes |
| 26 | *teamFT..x* | Free throw percentage | | team free throw makes/ team free throw attempts | | | | yes |
| 27 | *teamORB.x* | Offensive rebound | | number of rebounds the team get from opponent’s side | | | | yes |
| 28 | *teamDRB.x* | Defensive rebound | | number of rebounds the team get from their own side | | | | yes |
| 29 | *teamTRB.x* | Total rebound | | total number of rebounds | | | | yes |
| 30 | *teamPTS1.x* | Team points first quarter | |  | | | | yes |
| 31 | *teamPTS2.x* | Team points second quarter | |  | | | | yes |
| 32 | *teamPTS3.x* | Team points third quarter | |  | | | | yes |
| 33 | *teamPTS4.x* | Team points fourth quarter | |  | | | | yes |
| 34 | *teamPTS5.x* | Team points over time one | |  | | | | yes |
| 35 | *teamPTS6.x* | Team points over time two | |  | | | | yes |
| 36 | *teamPTS7.x* | Team points over time three | |  | | | | yes |
| 37 | *teamPTS8.x* | Team points over time four | |  | | | | yes |
| 38 | *teamTREB%.x* | Team total rebounds percentage | | teamTRB.x/(teamTRB.x + teamTRB.y) | | | | yes |
| 39 | *teamASST%.x* | Team assists percentage | | teamAST.x/(teamAST.x +teamAST.y) | | | | yes |
| 40 | *teamTS%.x* | Team Steal percentage | | teamSTL.x/(teamSTL.x + teamSTL.y) | | | | yes |
| 41 | *teamE.effective.FG%.x* | Effective field goal percentage | | (team2PM.x + 1.5\*team3PM.x) / teamFGA.x | | | | yes |
| 42 | *teamOREB%.x* | Offensive rebound percentage | | teamORB.x/(teamORB.x +teamDRB.y) | | | | yes |
| 43 | *teamDREB%.x* | Defensive rebound percentage | | teamDRB.x/(teamDRB.x +teamORB.y) | | | | yes |
| 44 | *teamTO(TurnOver)%.x* | Turnover percentage | | 100 \* team.TO.x/(team.FGA.x + 0.44\*team.FTA.x + team.TO.x) | | | | yes |
| 45 | *teamSTL%.x* | Steal percentage | | 100 \* (team.STL.x \* (teamMin.x / 5)) / (team.Min.x \* Possessions.y) | | | | yes |
| 46 | *teamBLK%.x* | team Block percentage | | 100 \* (team.BLK.x \* (teamMin.x / 5)) / (teamMin.x \* (team.FGA.y - team.3PA.y)) | | | | yes |
| 47 | *teamBLKR.x* | team Block ratio | | team.BLK.x/teamBLK.x + teamBLK.y | | | | yes |
| 48 | *teamPPS.points.per.shot..x* | points per shot | | (3\*team3PM.x + 2\*team2PM.x)/team.FGA.x | | | | yes |
| 49 | *teamFIC.floor.impact.counter..x* | floor impact counter | | (team.PTS.x + team.ORB.x + 0.75 team.DRB.x + team.AST.x + team.STL.x + team.BLK.x –0.75 team.FGA.x – 0.375 team.FTA.x – team.TO.x – 0.5 team.PF.x) | | | | yes |
| 50 | *teamFIC40.per.40.minuts..x* | floor impact counter per 40 minutes | | using 40 minutes data in a game to calculate floor impact counter | | | | yes |
| 51 | *teamOrtg.offensive.rating.per.100.possesion..x* | offensive rating per 100 possessions | | points scored per 100 possessions | | | | yes |
| 52 | *teamDrtg(defensive.rating.per.100.possesion).x* | defensive rating per 100 possessions | | points allowed per 100 possessions | | | | yes |
| 53 | *teamEDiff(efficiency.difference).x* | efficiency difference | | teamOrtg.x -teamDrtg.x | | | | yes |
| 54 | *teamPlay%.x* | team play percentage | | time period that the team hold the ball | | | | yes |
| 55 | *teamAR.x* | average rebounds | | number of average rebounds | | | | yes |
| 56 | *teamAST/TO.x* | assist turnover ratio | | team.AST.x/team.TO.x | | | | yes |
| 57 | *teamSTL/TO.x* | steal turnover ratio | | team.STL.x/team.TO.x | | | | yes |
| 58 | *pointspread.l10.x* | point spread last 10 games | | the average number of the point spread last 10 games | | | | yes |
| 59 | *teamMin.y* | opponent minutes | | how long the opponent played for one game | | | | yes |
| 60 | *teamDayOff.y* | Team day off | | days opponent rested before the coming game | | | | yes |
| 61 | *teamPTS.y* | opponent points | | number of points the opponent get for one game | | | | yes |
| 62 | *teamAST.y* | opponent assists | | number of assists the opponent get for one game | | | | yes |
| 63 | *teamTO(TurnOver).y* | opponent turnovers | | number of turnovers the opponent get for one game | | | | yes |
| 64 | *teamSTL.y* | opponent steal | | number of steals the opponent get for one game | | | | yes |
| 65 | *teamBLK.y* | opponent Block | | number of Blocks the opponent get for one game | | | | yes |
| 66 | *teamPF(personal.fouls).y* | opponent personal fouls | | number of personal fouls the opponent get for a game | | | | yes |
| 67 | *teamFGA.y* | opponent field goal attempts | | number of field goal attempts the opponent have | | | | yes |
| 68 | *teamFGM.y* | opponent field goal makes | | number of field goal makes the opponent have | | | | yes |
| 69 | *teamFG..y* | opponent field goal percentage | | opponent field goal makes/ opponent field goal attempts | | | | yes |
| 70 | *team2PA.y* | 2 points attempts | | number of 2 points attempts the opponent have | | | | yes |
| 71 | *team2PM.y* | 2 points makes | | number of 2 points makes the opponent have | | | | yes |
| 72 | *team2P..y* | 2 points percentage | | opponent 2 points makes/ opponent 2 points attempts | | | | yes |
| 73 | *team3PA.y* | 3 points attempts | | number of 3 points attempts the opponent have | | | | yes |
| 74 | *team3PM.y* | 3 points makes | | number of 3 points makes the opponent have | | | | yes |
| 75 | *team3P..y* | 3 points percentage | | opponent 3 points makes/ opponent 3 points attempts | | | | yes |
| 76 | *teamFTA.y* | Free throw attempts | | number of free throws attempts the opponent have | | | | yes |
| 77 | *teamFTM.y* | Free throw makes | | number of free throws makes the opponent have | | | | yes |
| 78 | *teamFT..y* | Free throw percentage | | opponent free throw makes/ opponent free throw attempts | | | | yes |
| 79 | *teamORB.y* | Offensive rebound | | number of rebounds the opponent get from opponent’s side | | | | yes |
| 80 | *teamDRB.y* | Defensive rebound | | number of rebounds the opponent get from their own side | | | | yes |
| 81 | *teamTRB.y* | Total rebound | | total number of rebounds | | | | yes |
| 82 | *teamPTS1.y* | Team points first quarter | |  | | | | yes |
| 83 | *teamPTS2.y* | Team points second quarter | |  | | | | yes |
| 84 | *teamPTS3.y* | Team points third quarter | |  | | | | yes |
| 85 | *teamPTS4.y* | Team points fourth quarter | |  | | | | yes |
| 86 | *teamPTS5.y* | Team points over time one | |  | | | | yes |
| 87 | *teamPTS6.y* | Team points over time two | |  | | | | yes |
| 88 | *teamPTS7.y* | Team points over time three | |  | | | | yes |
| 89 | *teamPTS8.y* | Team points over time four | |  | | | | yes |
| 90 | *teamTREB%.y* | Team total rebounds percentage | | teamTRB.y/(teamTRB.y + teamTRB.y) | | | | yes |
| 91 | *teamASST%.y* | Team assists percentage | | teamAST.y/(teamAST.y +teamAST.y) | | | | yes |
| 92 | *teamTS%.y* | Team Steal percentage | | teamSTL.y/(teamSTL.y + teamSTL.y) | | | | yes |
|  | | |  | |  |  |  | |
| 93 | *teamE(effective)FG%.y* | Effective field goal percentage | | (team2PM.y + 1.5\*team3PM.y) / teamFGA.y | | | | yes |
| 94 | *teamOREB%.y* | Offensive rebound percentage | | teamORB.y/(teamORB.y +teamDRB.y) | | | | yes |
| 95 | *teamDREB%.y* | Defensive rebound percentage | | teamDRB.y/(teamDRB.y +teamORB.y) | | | | yes |
| 96 | *teamTO(turnOver)%.y* | Turnover percentage | | 100 \* team.TO.y/(team.FGA.y + 0.44\*team.FTA.y + team.TO.y) | | | | yes |
| 97 | *teamSTL%.y* | Steal percentage | | 100 \* (team.STL.y \* (teamMin.y / 5)) / (team.Min.y \* Possessions.y) | | | | yes |
| 98 | *teamBLK%.y* | team Block percentage | | 100 \* (team.BLK.y \* (teamMin.y / 5)) / (teamMin.y \* (team.FGA.y - team.3PA.y)) | | | | yes |
| 99 | *teamBLKR.y* | team Block ratio | | team.BLK.y/teamBLK.y + teamBLK.y | | | | yes |
| 100 | *teamPPS(points.per.shot).y* | points per shot | | (3\*team3PM.y + 2\*team2PM.y)/team.FGA.y | | | | yes |
| 101 | *teamFIC(floor.impact.counter).y* | floor impact counter | | (team.PTS.y + team.ORB.y + 0.75 team.DRB.y + team.AST.y + team.STL.y + team.BLK.y –0.75 team.FGA.y – 0.375 team.FTA.y – team.TO.y – 0.5 team.PF.y) | | | | yes |
| 102 | *teamFIC40(per.40.minuts).y* | floor impact counter per 40 minutes | | using 40 minutes data in a game to calculate floor impact counter | | | | yes |
| 103 | *teamOrtg(offensive.rating.per.100.possesion).y* | offensive rating per 100 possessions | | points scored per 100 possessions | | | | yes |
| 104 | *teamDrtg(defensive.rating.per.100.possesion).y* | defensive rating per 100 possessions | | points allowed per 100 possessions | | | | yes |
| 105 | *teamEDiff(efficiency.difference).y* | efficiency difference | | teamOrtg.y -teamDrtg.y | | | | yes |
| 106 | *teamPlay%.y* | team play percentage | | time period that the team hold the ball | | | | yes |
| 107 | *teamAR.y* | average rebounds | | number of average rebounds | | | | yes |
| 108 | *teamAST/TO.y* | assist turnover ratio | | team.AST.y/team.TO.y | | | | yes |
| 109 | *teamSTL/TO.y* | steal turnover ratio | | team.STL.y/team.TO.y | | | | yes |
| 110 | *pointspread.l10.y* | point spread last 10 games | | the average number of points spread last 10 games | | | | yes |
| 111 | *SPAT\_H* | SumPlus Averaged Team Ability for Home team | | SumPlus measures one player’s ability in a game. *SPAT\_H* sums all players’ averaged SumPlus, who will participate to the next game. The averaged SumPlus for one player is calculated as averaged one player’s last ten games’ SumPlus. If he does not participate ten games before, just averaged his SumPlus that he has so far. | | | | yes |
| 112 | *SPAT\_A* | SumPlus Averaged Team Ability for Away team | |  |
| 113 | *pointspread.lag* | point spread lag | | last game's point spread for the team | | | | no |
| 114 | *pointspreadl5* | point spread average for last five games the team vs this opponent | |  | | | | no |
| 115 | *season* | NBA regular season | |  | | | | no |
| 116 | *area* | team's location belongs | | eastern or western | | | | no |
| 117 | *spread1* | One type of point spread | | If team1’s points add spread1 is higher than team2’s points, then team1 wins from bookmakers’ perspective. Else versa. | | | | no |
| 118 | *spread2* | One type of point spread | | If team2’s points add spread2 is higher than team1’s points, then team2 wins from bookmakers’ perspective. Else versa. | | | | no |
| 119 | *price1* | One type of price | | How much money you need to pay to get 100 dollars pay-out if you bet team1 and team1 wins. | | | | no |
| 120 | *price2* | One type of price | | How much money you need to pay to get 100 dollars pay-out if you bet team2 and team2 wins. | | | | no |
| 121 | *odds1* | Odds | | Odds for team1 | | | | no |
| 122 | *odds2* | Odds | | Odds for team2 | | | | no |
| Note: This table reports details of characteristics. | | | | | | | | |
|  | | | | | | | | |  |  |  |  |

1. **Feature Importance Results**

Normalization dataset

1. Linear regression
2. scores

feature\_importance

team2PA.x 8.134863e+08

teamFGA.x 6.823920e+08

team3PA.x 6.272048e+08

teamFIC(floor.impact.counter).y 5.286506e+08

teamFIC.floor.impact.counter..x 4.114809e+08

teamPTS.y 4.114571e+08

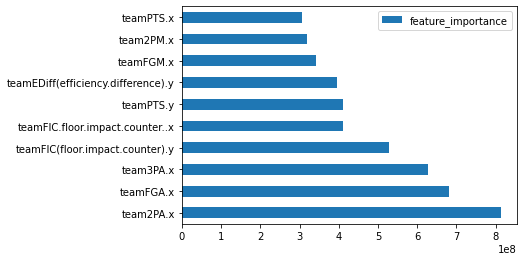
teamEDiff(efficiency.difference).y 3.966545e+08

teamFGM.x 3.415315e+08

team2PM.x 3.179791e+08

teamPTS.x 3.062201e+08

1. plot



1. DecisionTreeRegressor
2. Scores

feature\_importance

teamEDiff(efficiency.difference).x 0.106979

teamEDiff(efficiency.difference).y 0.075535

pointspreadl5 0.025703

SPAT\_H 0.021940

SPAT\_A 0.016607

teamDREB%.x 0.015401

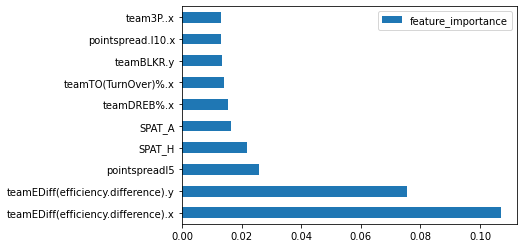
teamTO(TurnOver)%.x 0.014032

teamBLKR.y 0.013511

pointspread.l10.x 0.013170

team3P..x 0.013117

1. Plot



RandomForestRegressor

1. Scores

feature\_importance

teamEDiff(efficiency.difference).x 0.068635

teamEDiff(efficiency.difference).y 0.053793

pointspread.l10.x 0.040174

pointspread.l10.y 0.031437

pointspreadl5 0.024143

SPAT\_H 0.020595

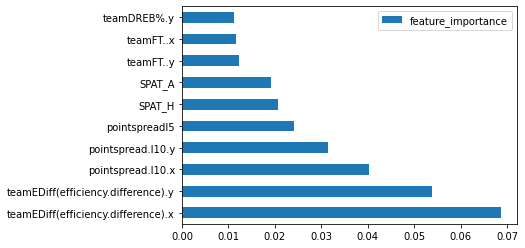
SPAT\_A 0.019255

teamFT..y 0.012239

teamFT..x 0.011537

teamDREB%.y 0.011203

1. Plot



XGBRegressor

1. Scores

feature\_importance

teamEDiff(efficiency.difference).x 0.077497

teamEDiff(efficiency.difference).y 0.059128

pointspread.l10.x 0.047216

pointspread.l10.y 0.046649

pointspreadl5 0.018418

teamFIC40(per.40.minuts).y 0.015727

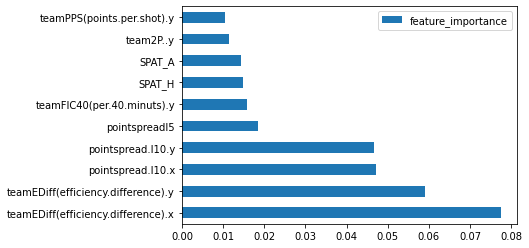
SPAT\_H 0.014941

SPAT\_A 0.014388

team2P..y 0.011429

teamPPS(points.per.shot).y 0.010525

1. Plot



SVR

1. Scores

feature\_importance

pointspreadl5 10.374906

teamDrtg(defensive.rating.per.100.possesion).y 8.064559

teamEDiff(efficiency.difference).x 7.593304

teamDrtg(defensive.rating.per.100.possesion).x 7.465870

pointspread.l10.x 7.205861

SPAT\_H 6.885843

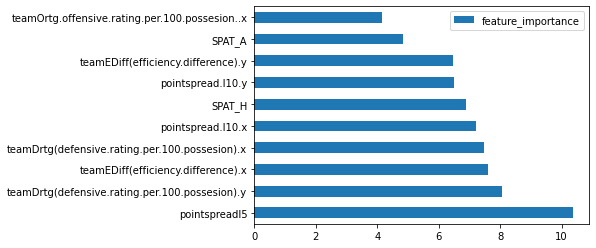
pointspread.l10.y 6.487634

teamEDiff(efficiency.difference).y 6.463246

SPAT\_A 4.836793

teamOrtg.offensive.rating.per.100.possesion..x 4.155316

1. Plot



LightGBM

1. Scores

feature\_importance

pointspreadl5 78

SPAT\_H 61

teamDREB%.y 55

teamEDiff(efficiency.difference).y 52

pointspread.l10.y 49

SPAT\_A 49

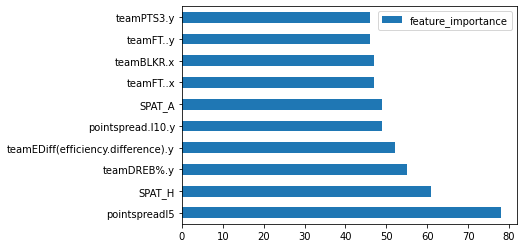
teamFT..x 47

teamBLKR.x 47

teamFT..y 46

teamPTS3.y 46

1. Plot



Permutation

feature\_importance

teamSTL.x 1.278015

teamE(effective)FG%.y 1.135124

teamFG..y 1.102324

pointspread.lag 1.064089

teamSTL%.x 0.931730

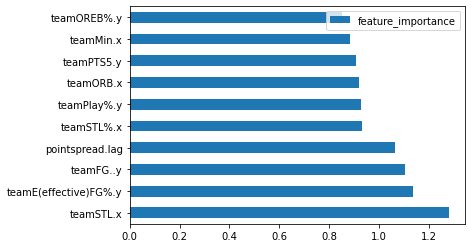
teamPlay%.y 0.927310

teamORB.x 0.917904

teamPTS5.y 0.907654

teamMin.x 0.882251

teamOREB%.y 0.850311



Standardization dataset

1. Linear regression
2. Score

feature\_importance

teamTRB.y 4.441238e+08

team3PA.y 4.241022e+08

team2PA.y 4.217652e+08

teamDRB.y 4.158042e+08

teamFGA.y 3.421623e+08

teamORB.y 3.219606e+08

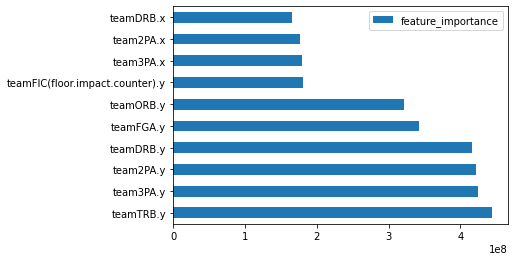
teamFIC(floor.impact.counter).y 1.809595e+08

team3PA.x 1.793690e+08

team2PA.x 1.772246e+08

teamDRB.x 1.657382e+08

1. Plot



1. DTR
2. Score

feature\_importance

teamEDiff(efficiency.difference).x 0.104284

teamEDiff(efficiency.difference).y 0.076315

pointspreadl5 0.025626

SPAT\_H 0.021719

SPAT\_A 0.016544

teamDREB%.x 0.015795

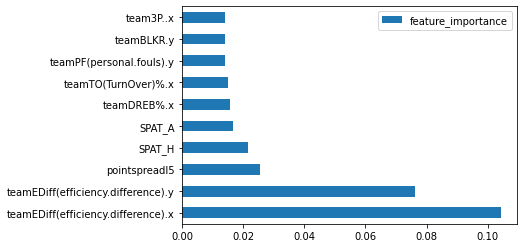
teamTO(TurnOver)%.x 0.015011

teamPF(personal.fouls).y 0.014050

teamBLKR.y 0.013933

team3P..x 0.013917

1. Plot



1. RFR
2. Score

feature\_importance

teamEDiff(efficiency.difference).x 0.069812

teamEDiff(efficiency.difference).y 0.055170

pointspread.l10.x 0.040588

pointspread.l10.y 0.030625

pointspreadl5 0.023239

SPAT\_H 0.019361

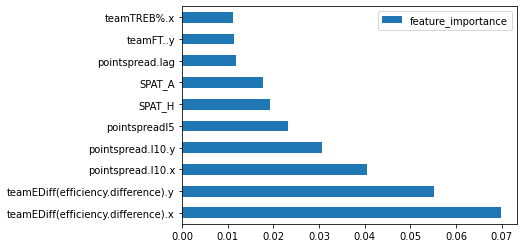
SPAT\_A 0.017838

pointspread.lag 0.011727

teamFT..y 0.011442

teamTREB%.x 0.011071

1. Plot



XGB

1. Score

feature\_importance

teamEDiff(efficiency.difference).x 0.077497

teamEDiff(efficiency.difference).y 0.059128

pointspread.l10.x 0.047216

pointspread.l10.y 0.046649

pointspreadl5 0.018418

teamFIC40(per.40.minuts).y 0.015727

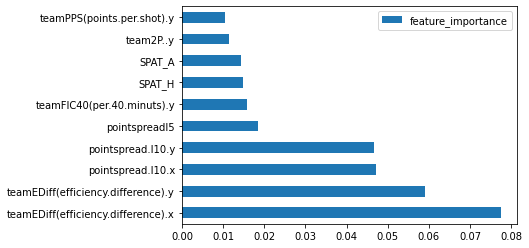
SPAT\_H 0.014941

SPAT\_A 0.014388

team2P..y 0.011429

teamPPS(points.per.shot).y 0.010525

1. Plot



SVR

1. Score

feature\_importance

team2P..y 2.917532

teamAR.y 2.300794

teamSTL%.x 2.107224

teamASST%.y 2.096641

teamPPS.points.per.shot..x 1.937383

teamE(effective)FG%.y 1.757093

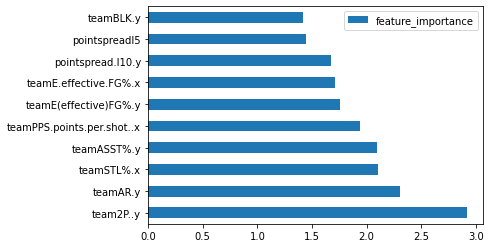
teamE.effective.FG%.x 1.706945

pointspread.l10.y 1.672402

pointspreadl5 1.442920

teamBLK.y 1.415923

1. Plot



LGB

1. Score

feature\_importance

pointspreadl5 75

SPAT\_H 64

teamFT..x 61

teamPTS1.x 50

teamDREB%.y 50

pointspread.l10.y 49

SPAT\_A 49

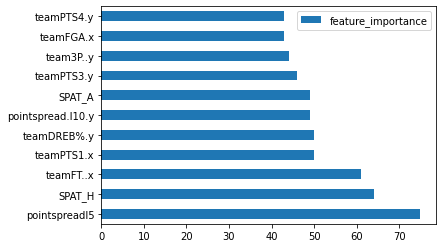
teamPTS3.y 46

team3P..y 44

teamFGA.x 43

teamPTS4.y 43

1. Plot



7) permutation

feature\_importance

teamFT..y 1.056648

teamTREB%.x 1.022956

pointspreadl5 0.767542

teamTRB.x 0.752871

SPAT\_H 0.751874

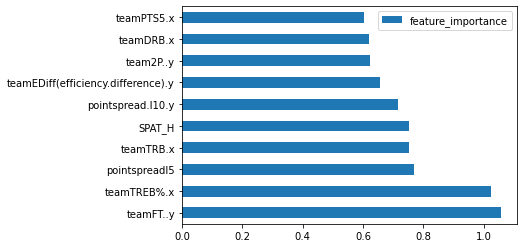
pointspread.l10.y 0.714671

teamEDiff(efficiency.difference).y 0.654805

team2P..y 0.623529

teamDRB.x 0.618481

teamPTS5.x 0.602847



1. **comparison between our paper and prior literature variable used.**

|  |  |
| --- | --- |
| **TABLE G.1:**  **Comparison between our paper and prior literature variables used** | |
| **Team-level variable our paper used** | **Prior literature used** |
| *gmDate* |  |
| *teamAbbr* |  |
| *teamID* |  |
| *opptAbbr* |  |
| *pointspread* | True |
| *teamLoc.x* | True |
| *teamDayOff* |  |
| *teamMin.x* |  |
| *teamPTS.x* | True |
| *teamAST.x* | True |
| *teamTO.turNOPver..x* | True |
| *teamSTL.x* | True |
| *teamBLK.x* | True |
| *teamPF.personal.fouls..x* | True |
| *teamFGA.x* | True |
| *teamFGM.x* | True |
| *teamFG..x* |  |
| *team2PA.x* |  |
| *team2PM.x* |  |
| *team2P..x* |  |
| *team3PA.x* |  |
| *team3PM.x* |  |
| *team3P..x* | True |
| *teamFTA.x* |  |
| *teamFTM.x* |  |
| *teamFT..x* | True |
| *teamORB.x* | True |
| *teamDRB.x* | True |
| *teamTRB.x* | True |
| *teamPTS1.x* |  |
| *teamPTS2.x* |  |
| *teamPTS3.x* |  |
| *teamPTS4.x* |  |
| *teamPTS5.x* |  |
| *teamPTS6.x* |  |
| *teamPTS7.x* |  |
| *teamPTS8.x* |  |
| *teamTREB%.x* |  |
| *teamASST%.x* |  |
| *teamTS%.x* |  |
| *teamE.effective.FG%.x* |  |
| *teamOREB%.x* |  |
| *teamDREB%.x* |  |
| *teamTO(TurnOver)%.x* |  |
| *teamSTL%.x* |  |
| *teamBLK%.x* |  |
| *teamBLKR.x* |  |
| *teamPPS.points.per.shot..x* | True |
| *teamFIC.floor.impact.counter..x* |  |
| *teamFIC40.per.40.minuts..x* |  |
| *teamOrtg.offensive.rating.per.100.possesion..x* |  |
| *teamDrtg(defensive.rating.per.100.possesion).x* |  |
|  |  |
| *teamEDiff(efficiency.difference).x* |  |
| *teamPlay%.x* |  |
| *teamAR.x* |  |
| *teamAST/TO.x* |  |
| *teamSTL/TO.x* |  |
| *pointspread.l10.x* |  |
| *pointspread.lag* |  |
| *pointspreadl5* |  |
| *season* |  |
| *area* |  |

1. **New Features**

Table H.1 and H.2 show the comparison of different ways of feature engineering methods of team-level features. Table H.1 shows the top 30 Team-level features of RF univariate feature important results. The max error and the MAE are 52.88 and 12.28. Although these two numbers are relatively high, the main purpose of this process is to select important features. The results show that most of the top 30 features are from *Avg82* and *Avg41* subsets. These two subsets mean rolling averaged datasets for the last 82 and 41 games. We chose 82 games because one NBA season has 82 games. Therefore, we can get a team’s average performance against all other teams, which simulates one team’s ability. we chose 41 games because we want a team’s performance against all other teams and more recent performance. Results suggest that rolling averaged features contribute most to the point spread prediction, which supports our assumption that rolling averaged features perform better than lags because rolling averaged data simulate a team’s ability against all teams but lags are biased. Table H.2 shows the number of features we selected by 0.001 cut off and relative importance ranks for each subset of features. Results indicate that although lags have more features selected, their features importance is worse than rolling averaged features. That is, lags are recent data, so more features were selected from lags. However, because of bias, lags features are not as important as rolling averaged features. Therefore, if we can find a way to drop the bias for lags features, they will probably perform much better.

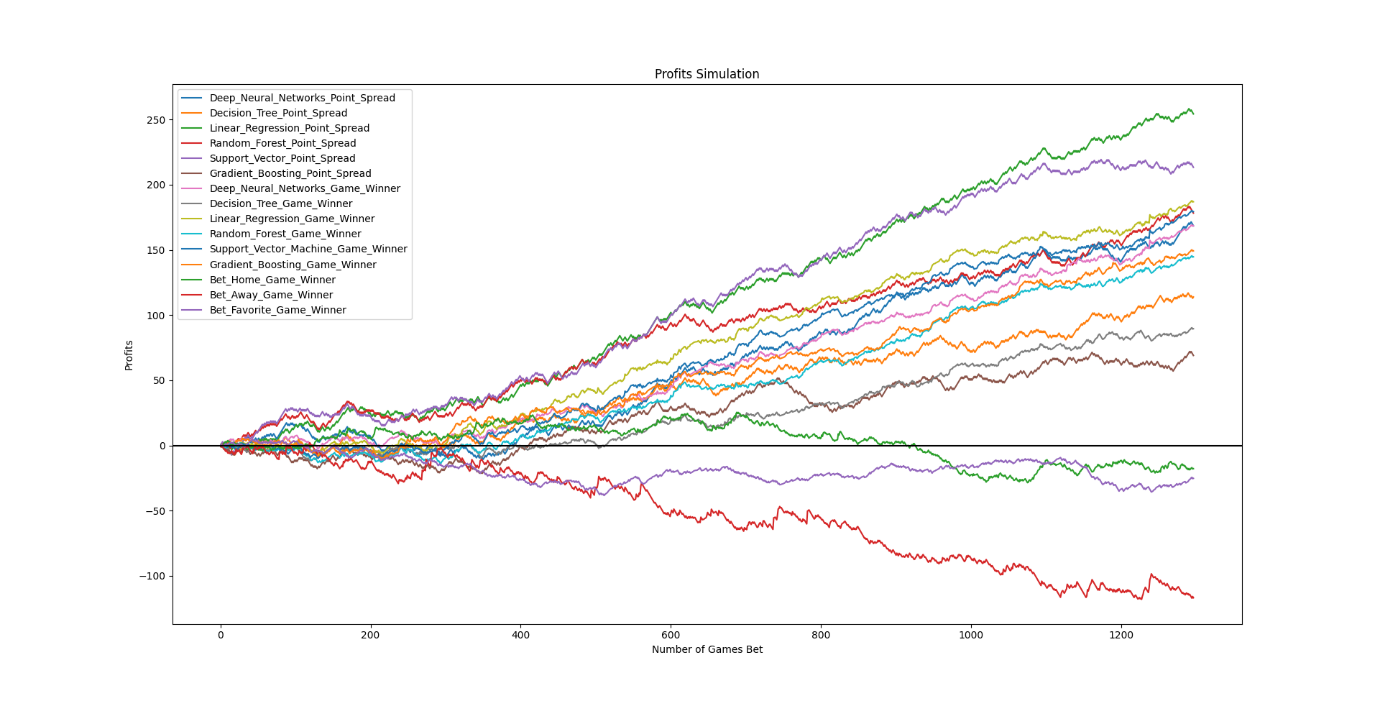
Table H.3 shows the top 30 Player-level features of RF univariate feature important results. The max error and the MAE are 61.74 and 11.27. Although these two numbers are relatively high, the main purpose of this process is to select important features. The results show that most of the top 30 features are from *OppoAvg30* and *OppoAvg60* subsets. These two subsets mean rolling averaged features for one player’s opponent’s last 30 and 60 games. Results suggest that rolling averaged features of the opponent contribute most to the prediction of players’ performance next game (*PLUS\_MINUS*), which is the target variable for the player dataset. Results also supportmy idea that Lags features are biased. Table H.4 shows the number of features we selected by 0.001 cut off and relative importance ranks for each subset features. Results indicate that although lags have more features selected, their features importance is worse than rolling averaged features. That is, lags are recent data, so more features were selected from lags. However, because of bias, lags features are not as important as rolling averaged features.

|  |  |  |
| --- | --- | --- |
| **TABLE H.1**  **Top 30 Team-level Features** | | |
| **RF max error** | | 52.88 |
| **RF mean absolute error** | | 12.28046 |
|  | | **RF features importance** |
| *teamLoc* | | 0.037459 |
| *Avg41teampointspread* | | 0.030589 |
| *Avg41oppopointspread* | | 0.028007 |
| *Avg82teamopptEDiff* | | 0.025065 |
| *Avg82oppoteamEDiff(efficiency difference)* | | 0.024993 |
| *Avg82oppoopptEDiff* | | 0.024679 |
| *Avg82teamteamEDiff(efficiency difference)* | | 0.022651 |
| *Avg82oppopointspread* | | 0.00476 |
| *Avg82teampointspread* | | 0.004037 |
| *Lag1BteamDrtg(defensive rating per 100 possesion)* | | 0.003427 |
| *AvsBLag2team3P%* | | 0.002991 |
| *AvsBLag2teamSTL%* | | 0.002744 |
| *Lag1Ateam2P%* | | 0.002474 |
| *Lag3ATeamLoc\*teamTO(turnover)%* | | 0.002415 |
| *Lag2BteamSTL/TO* | | 0.00239 |
| *AvsBLag2oppt3P%* | | 0.00238 |
| *Lag2Bteam2P%* | | 0.002358 |
| *Lag2BteamBLK%* | | 0.002342 |
| *AvsBLag2opptASST%* | | 0.002328 |
| *Lag1AteamDREB%* | | 0.00232 |
| *Lag3AteamPTS4* | | 0.002271 |
| *Avg41teamteamBLKR* | | 0.002252 |
| *Lag1AteamOrtg(offensive rating per 100 possesion)* | | 0.002247 |
| *Avg41oppoteamBLKR* | | 0.002187 |
| *Avg41teamopptBLKR* | | 0.002167 |
| *Lag2AteamSTL/TO* | | 0.002166 |
| *Lag3BteamFT%* | | 0.00216 |
| *Avg41teamopptTRB* | | 0.002118 |
| *Lag1AteamSTL/TO* | | 0.002116 |
| *Avg82teamteamSTL%* | | 0.002102 |
| Note: this table records feature importance for top 30 features for Team-level data. | | |
| |  |  |  | | --- | --- | --- | | **TABLE H.2**  **Team-level Subset Features Comparison** | | | |  | **Num\_selected** | **Rank Level** | | *Avg82* | 14\*4 = 56 | High Rank | | *Avg41* | 20\*4 = 80 | Medium High Rank | | *AvsB* | 21\*4 = 84 | Medium Rank | | *Lag123* | 29\*6 = 174 | Low Rank | | total | 394 |  | | Note: This table summarises Team-level subset features comparison. Avg82 means rolling average 82 games features; Avg41 means rolling average 41 games features; AvsB means last two times these two teams played against each other’s features; Lag123 means these two teams’ last three games but might not play with the same opponent.Num\_selected means the number of features selected from different subsets. Rank level means the rank of feature importance of these subsets on average. | | |  |  |  | | --- | --- | | **TABLE H.3**  **Top 30 Team-level Features** | | | **RF max error** | 61.74 | | **RF mean absolute error** | 11.27834979 | |  | **RF Feature Importance** | | *Avg60PLUS\_MINUS* | 0.04555286 | | *OppoAvg60PLUS\_MINUS* | 0.027528777 | | *OppoAvg60pointspread* | 0.02384224 | | *LOC* | 0.022215372 | | *OppoAvg30PLUS\_MINUS* | 0.017280338 | | *Avg30PLUS\_MINUS* | 0.015639225 | | *OppoAvg301G3\_PCT\_away* | 0.015086818 | | *OppoAvg301T\_PCT\_home* | 0.014387532 | | *OppoAvg601T\_PCT\_away* | 0.014298168 | | *OppoAvg601G3\_PCT\_away* | 0.014041858 | | *OppoAvg601T\_PCT\_home* | 0.01376509 | | *OppoAvg301T\_PCT\_away* | 0.013679472 | | *OppoAvg601G3\_PCT\_home* | 0.013297229 | | *OppoAvg301G\_PCT\_away* | 0.013267728 | | *OppoAvg601G\_PCT\_away* | 0.01298525 | | *OppoAvg301G3\_PCT\_home* | 0.012756403 | | *OppoAvg30AST\_home* | 0.0122292 | | *OppoAvg30REB\_home* | 0.011850565 | | *OppoAvg60AST\_home* | 0.011822495 | | *OppoAvg30AST\_away* | 0.011814652 | | *OppoAvg60AST\_away* | 0.011573609 | | *OppoAvg601G\_PCT\_home* | 0.011363207 | | *OppoAvg60REB\_away* | 0.01133446 | | *OppoAvg301G\_PCT\_home* | 0.011147495 | | *OppoAvg60REB\_home* | 0.010842658 | | *OppoAvg30REB\_away* | 0.010475469 | | *START\_POSITION* | 0.010174045 | | *Lag1PLUS\_MINUS* | 0.010145831 | | *Lag3PLUS\_MINUS* | 0.009804804 | | *Lag2PLUS\_MINUS* | 0.009678647 | | Note: this table records feature importance for top 30 features for Player-level data | |  |  |  |  | | --- | --- | --- | | **TABLE H.4**  **Team-level Subset Features Comparison** | | | |  | **Num\_selected** | **Rank Level** | | *Avg60* | 16 | Medium Rank | | *Avg30* | 16 | Medium Rank | | *AvgPlayerVsOppo* | 12 | Medium Rank | | *Lag123* | 21 | Low Rank | | *OppoAvg60* | 14 | High Rank | | *OppoAvg30* | 14 | High Rank | | total | 93 |  | | Note: This table summarises Team-level subset features comparison. Avg60 means player’s rolling average 60 games features; Avg30 means player’s rolling average 30 games features; *AvgPlayerVsOppo* means averaged features a player played against the next opponent; Lag123 means these players’ features last three games but might not play with the same opponent. *OppoAvg60* means the rolling average of 60 games’ opponent’s team-level features. *OppoAvg30* means the rolling average of 30 games’ opponent’s team-level features. Num\_selected means the number of features selected from different subsets. Rank level means the rank of feature importance of these subsets on average. | | | | | |

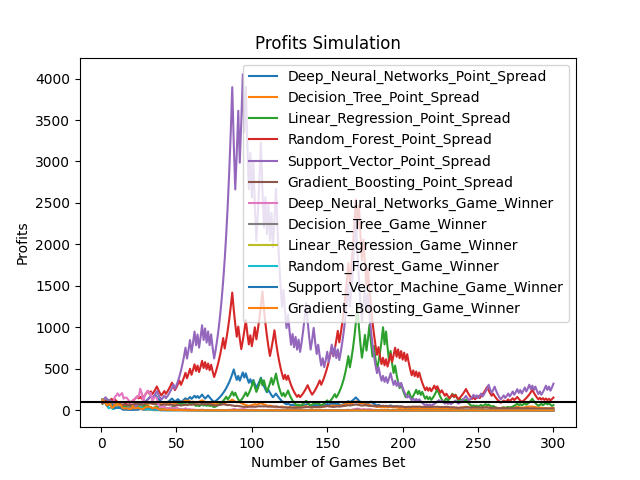
1. **New Results and Profits Simulation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table I.1 All model comparisons | | | | | |
| **Algorithms** | | **Max Error** | **MAE** | **RMSE** | **R2** |
| **OLS** | | 56.86 | 8.96 | 11.70 | 0.31 |
| **DT** | | 58.86 | 9.40 | 12.16 | 0.26 |
| **SVM** | | 55.77 | 9.04 | 11.81 | 0.30 |
| **RF** | | 55.69 | 9.18 | 11.98 | 0.28 |
| **XGBoost** | | 57.89 | 9.36 | 12.22 | 0.25 |
| **DNN** | | 58.10 | 9.17 | 11.98 | 0.28 |
|  |

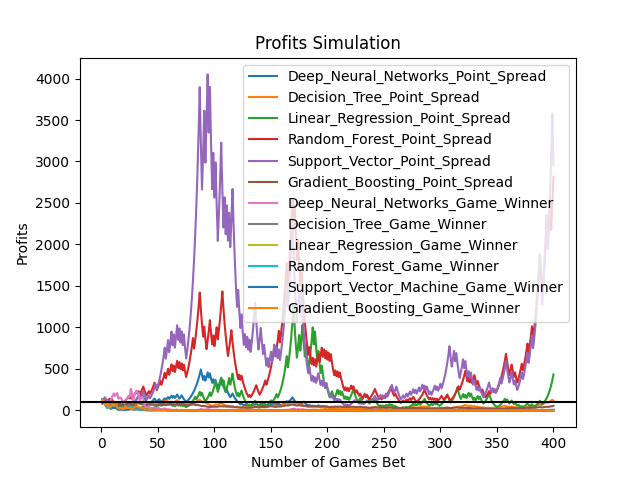
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table I.2Percentage of correct predictions. | | | | | | | | | |
|  | **DNN** | **DT** | **XGBoost** | **Linear Regression** | **RF** | **SVM** | **Home** | **Favorite** | **Away** |
| **Bet on all test dataset** | 58.00% | 55.80% | 54.01% | 61.34% | 58.33% | 59.72% | 59.72% | 68.90% | 40.28% |
| **One bet per day** | 58.75% | 52.50% | 56.25% | 57.50% | 60.00% | 58.75% |  |  |  |

Here are some profits or returns simulation figures:

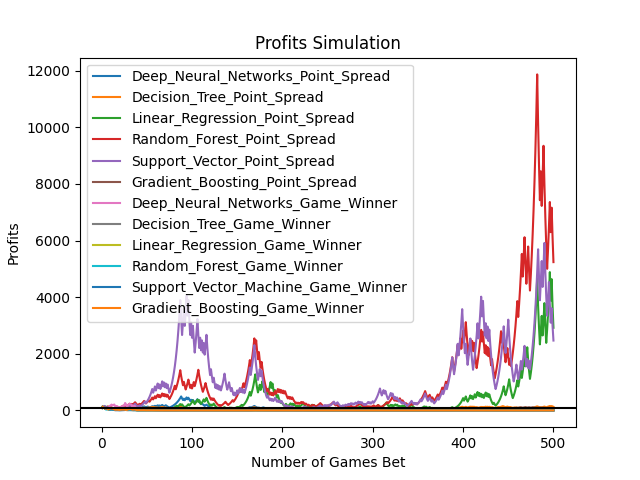
1. Simple betting rule for all games in 2018.



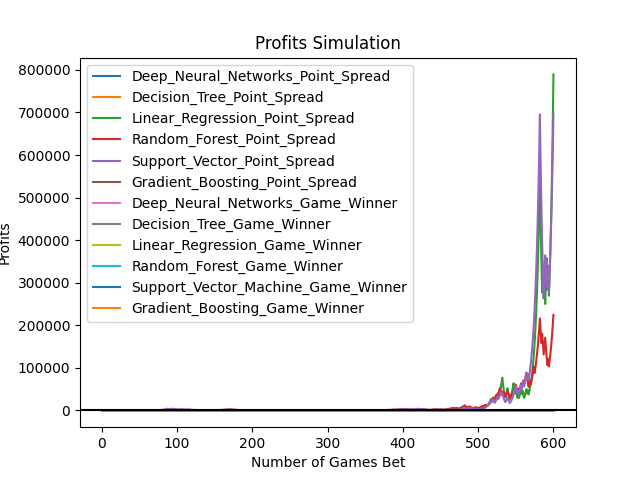
1. First 300 games of Kelly rule in 2018. We want to show first 300 games because if we show all games, you cannot see it clearly. The gaps between good models and bad models are too much when we present all games. You can see it later.



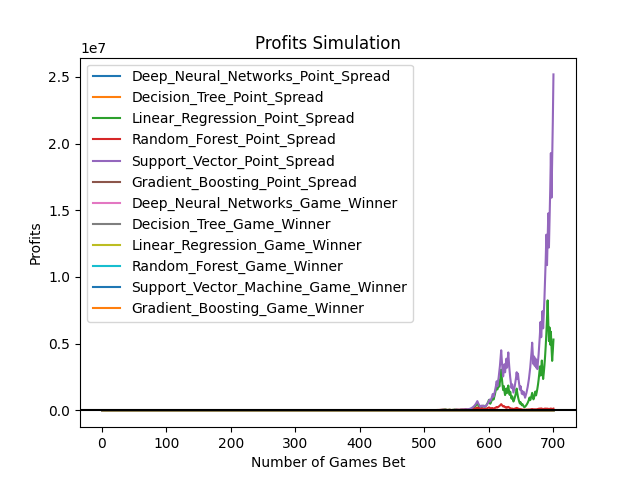
1. First 400 games of Kelly rule in 2018.



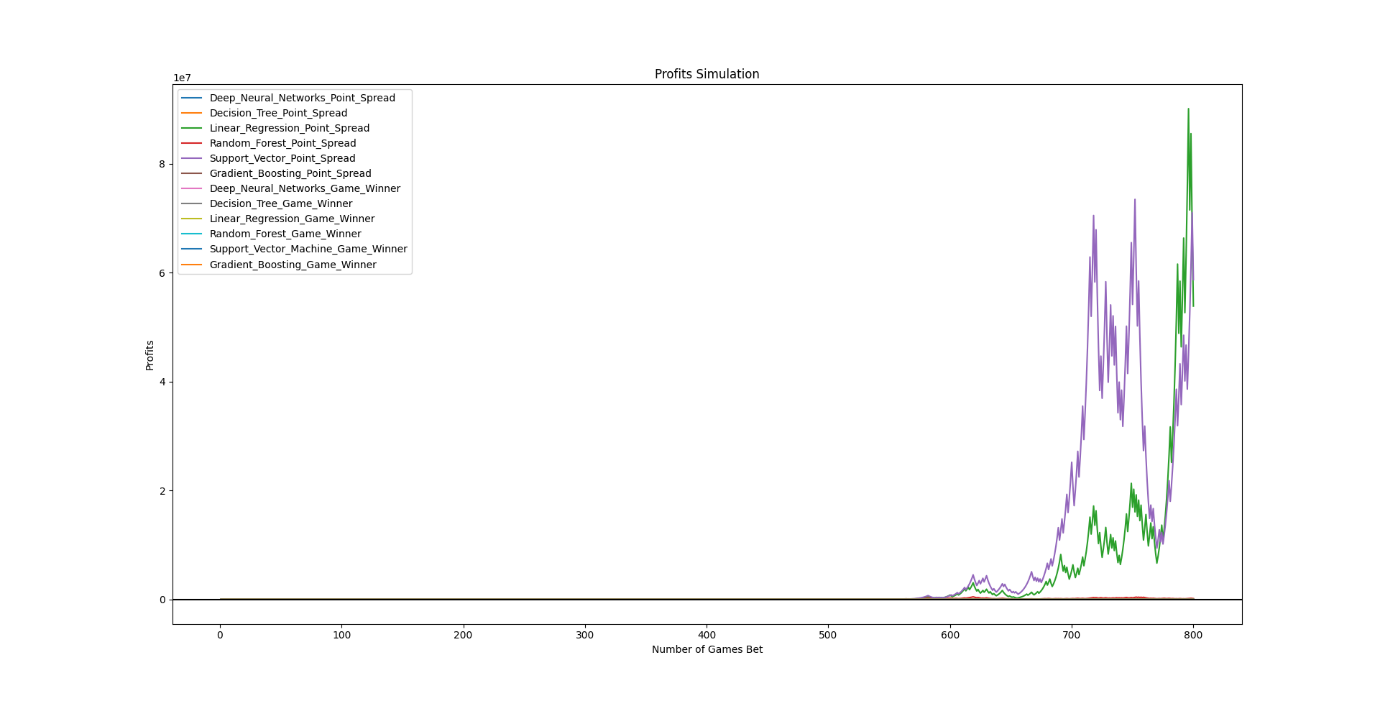
1. First 500 games of Kelly rule in 2018.



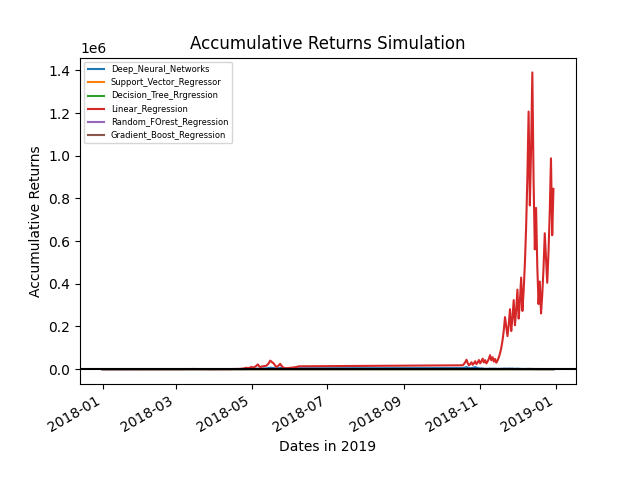
1. First 600 games of Kelly rule in 2018.



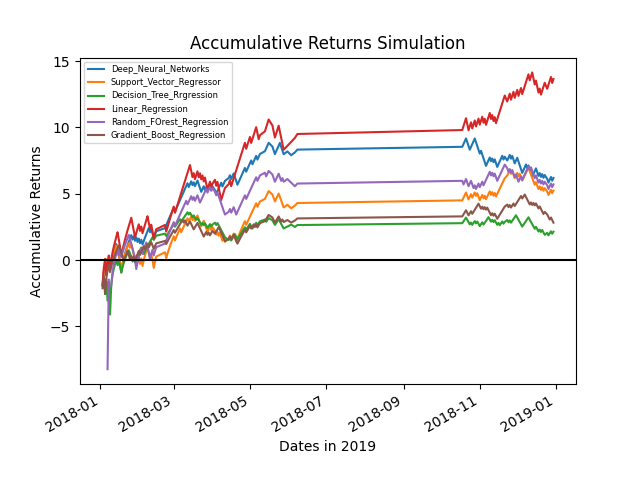
1. First 700 games of Kelly rule in 2018.



1. First 800 games of Kelly rule in 2018. Now you see…



1. Bet one game per day Kelly rule no log transformation in 2018.



1. Bet one game per day Kelly rule with log transformation in 2018.
2. **Extensive Literature review.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Papers** | **Variable of Interest** | **Data Used and Number of Variables** | **Feature Engineering** | **Method** | **Accuracy** | **Limitations** |
| Miljković et al. (2010) | Game-Winner and Point spread | 18 team-level basketball variables and 14 team standing variables | Doesn’t mention, but can see that some of features are season averaged | Classification: Decision Tree, K nearest neighbours, Naïve Bayes, and SVM.  Regression: not mentioned. | 65% for classification but no results for regression. | Only team-level data and the number are limited.  Season averaged data are not accessible.  Accuracy is not high enough for profitability.  Test dataset is randomly split not at the end of the period. |
| Rodrıguez (2019) | Game-Winner | 18 Player-level variables | Use top 3 players stats as features | Logistic Regression, 3 layers NN | 59.8% | Use one game features to predict same game’s outcome.  No team-level stats.  Other players also contribute to game outcome.  Only two algorithms.  Not accurate.  Test dataset is randomly split not at the end of the period. |
| Jones (2016) | Game-Winner | Select 144 games from 3 NBA seasons 2008-2011. 33 team-level variables. 50 games as test dataset. | No feature engineering | Linear Regression | 88-94%  R-squared 0.91 | Small sample size.  No player-level data.  Use one game features to predict same game’s outcome.  Only applied linear regression.  Size of datasets are small.  Test dataset is randomly split not at the end of the period. |
| Jain and Kaur (2017) | Game-Winner | Not specified | Not specified | SVM and HFSVM | 60-65% | Not profitable against the odds.  Only two algorithms.  Low accuracy.  Test dataset is randomly split not at the end of the period. |
| Ryan and Alameda-Basora (2019) | Game-winner | 2014-2018 NBA season data. 54 In-game team-level data from end of each quarter. | In game data at the end of each quarter. | Bayesian Network | 51.85% to 78.26% | Test datasets sizes are small, only 97 to 100 games, leading to not stable as can be seen from accuracy range.  In-game odds change quickly leading to hard to bet with prediction because prediction also takes time.  No player-level data.  Only one algorithm, cannot compare with others.  Cannot find how they get in-game data.  Test dataset is randomly split not at the end of the period. |
| Manner (2015) | Point spread | Not mentioned | Not mentioned | Linear Regression | MAE around 9 | Cannot find variable used or feature engineering techniques.  Only one algorithm.  Test dataset is randomly split not at the end of the period. |
| Kayhan and Watkins (2019) | Point spread | 2009-2016 NBA season data. In-game data. | In game data at the end of each quarter. | Snapshot approach with  LSTM | MAE 11 before game starts. | Low accuracy  Cannot find In-game data from provided website: <http://stats.nba.com>  Only one algorithm.  Error is big.  Test dataset is randomly split not at the end of the period. |
| Torres (2013) | Game-winner | 2007-2013 NBA regular seasons. 8 Past win-loss variables. | Past 8 games. | Linear regression,  Maximum Likelihood Classifier,  Multi-Layer Perceptron | 63.98% to 68.44% | Could have more variables.  No player-level data.  Test dataset is randomly split not at the end of the period. |
| Lin, Short, and Sundaresan (2014) | Game-winner | 92-98 NBA seasons.  17 team-level variables. | Recent games. | Logistic regression, SVM, AdaBoost, Random Forest. | 60% to 65% | Data is old.  No player-level data.  Team-level variables could be more.  Not accurate to be profitable.  Test dataset is randomly split not at the end of the period. |
| Cheng et al. (2013) | Spread Over/Under classification | 02-12 NBA seasons game-level and player-level data. | Not mentioned | SVM, Naïve Bayes. | 52.78% | Data is old.  Only two algorithms.  Not accurate enough to be profitable.  Test dataset is randomly split not at the end of the period. |
| Shi and Song (2021) | Point spread | 13-17 NBA seasons in-game data. | In game data, use data in every minute to feed forward. | Markov chain | MAE of 10.8 before game starts. | Only team-level data.  Only one algorithm.  Error is relatively large.  Test dataset is randomly split not at the end of the period. |
| Cai et al. (2019) | Game-winner | 380 games of 20 teams from 16-17 CBA regular season. 12 team-level variables | Last 5 games | SVM with Bagging, Naïve Bayes, Logistic Regression, Neural Networks. | 80% | only use 12 variables, didn't consider player-level variables. Sample size is too small, only 280 in total, so their accuracy can be biased. Test dataset is randomly split not at the end of the period. |
| Safer et al. (2018) | Game-winner | 07-17 NBA seasons 13 team-level in-game variables. | In game data at the end of each quarter. | Logistic regression,  Naïve Bayes, SVM, Neural Networks, Random Forest, Model stacking, Adaboost. | 60% to 70% of each season. | Number of variables could be more.  In-game odds change fast.  Model predictability decreases with year increases.  Test dataset is randomly split not at the end of the period. |
| Zhang et al. (2021) | Point spread | 17-21 NBA seasons, regular team-level data with injuries and salaries of player-level data. | Not mentioned | Linear regression | 13.1 RMSE  65% game winner accuracy | only linear regression.  only limited team-level data.  No test dataset. |

The above table summarizes recent literature that predicts NBA game outcomes. We have excluded this table and the subsequent discussion from the main paper due to brevity and included them in the internet appendix. The below explanations further highlight how our study differ from prior literature, issues with using some of those data sets, limitations and the contribution of our study and significance.

## Output variables of Interest

The NBA sports betting market has three prevalent betting mechanisms popular among bettors: 1. which team will win a game? 2. is the total points score of both teams over or under a specific points target provided by the betting firms? 3. is the point spread over or under specific points provided by the betting firms? Therefore, sports betting literature primarily focus on predicting these three targets.

One of the significant limitations of predicting the game-winner is limited profitability. In more detail, some teams would obviously outperform their opponents consistently resulting in majority of bettors betting on that particular outperforming team. As a result, the odds for that favourite team would be over-valued. Therefore, bettors who predict game-winners would always falls into betting on over-valued odds leading to limited profitability. Another issue that arises with predicting the game-winner is the wide selection of games to choose from on a given day. For example, on the one hand, if there are 10 games a day, bettors cannot choose just a single game to bet because the predictions are all going to be which team win or losses but not by how much they win. But, on the other hand if bettors know how much a game will be won by, they can select the game that is farthest from the line to have a safe bet and higher profits. In addition, majority of prior literature uses a single games’ data to predict that same games’ result. Thus, compared to game-winner predictions, point spread prediction can aid to avoid betting on the over-valued odds and helps bettors select under-valued odds leading to considerable profits. Moreover, we adopt point spread prediction since spreads can be further explained by the difference of each teams’ abilities as well identified by the amalgamation of both player and team level data sets.

## 2. Alternative Datasets

With regard to various data sets, four primary data sets are used at various levels by prior literature. They are player-level data, in-game-level data, team-level data, player and team-level data. However, significant limitations remain using solely player or team-level data and ignoring other relevant data. Two common limitations inherent for the in-game dataset would be intensity and high cost. Once a game begins, bettors would have to pay close attention to the game and adjust inputs continuously as the game proceeds to fully make use of an extensive array of in-game data for their models. Furthermore, timely accurate and precise in-game data is required leading to intensity, cost and accessibility issues, especially for average bettors. Moreover, odds can change in quick succession for in-game betting, but predictions take time (machine learning models take a considerable amount of time to run even with exceptional hardware and other resources). Hence, when bettors subsequently obtain game outcome predictions from their models, the odds may have already changed. Finally, prior literature uses player and team-level data only for top 3 players to measure their ability which is a limited representation of very rich features and their interactions that are available that contribute to game outcomes. However, our study includes both team and player-level data for all players.

## 3. Number of variables and sample size in original dataset

As can be observed from Table 7, our study has a considerable array of features both at team and player level that contributes to game outcomes leading to higher accuracy, compared to prior literature. Hence, our features, results and machine learning models would be of use even to professional bettors.

Sample size is another issue with regard to data sets used by prior literature. Several studies only use a single seasons’ data or historical data going back to 1990s to predict game outcomes which limits their ability and relevance. Our study in contrast, uses the most recent six seasons’ data for prediction purposes.

## 4. Feature engineering techniques

Over one third of prior literature do not detail any feature engineering technique used for their analysis. The rest of the prior literature mainly use three types of techniques:

1. use a single game’s features to predict the same game’s outcome. For example, they use one game’s Rebound, Steal, and Field Goal Percentage to predict the same game’s point spread or outcome. The issue with this type of feature engineering technique is that it’s difficult to ascertain what the value of these features are before the end of a game. Therefore, one cannot predict the outcome before a game end in real time for out of sample predictions. 2. use a single seasons’ averaged game features to predict a single game’s outcome in that season. For example, the NBA season starts from October and ends in April the following year. Prior literature uses all games in that period and averaged feature to predict a single game’s outcome for that particular season. The issue in this case is that one cannot identify the features for a particular game in that season and the games at the end of the same season. Thus, those predictions seem to assume that bettors would ex-ante have access to data/features not available prior to that game but even of games that would follow the game to predicted which is problematic. In contrast, our study only uses features to predict a game that would only be available ex-ante a game as opposed to ex-post. 3. Prior literature further use past 5 or 8 games to create averaged features which could lead to opponent bias. We first tried to address this issue by using a rolling average of the most recent past 10 games. However, to address reviewer 3 concerns we explored other feature engineering methods discussed in Table 9 and 11. We also implemented additional machine learning models including LSTM suggested by reviewer 3, which are still running and results would be available in the internet appendix for brevity.

## 5. Prior Literature Prediction Models

Our study as opposed to prior literature implements a wide array of standard and ensemble algorithms such as SVM, DT regression, Ensemble method (averaged model of Elastic Net, Gboost, KRR, and Lasso; XGBoost; LightGBM; Random Forest), and DNN. Newly implemented models following the second round of revisions include Ridge, Decision Trees, AdaBoost, Neural Networks, LSTM and PCA with the same models. Another common limitation that needs to be addressed regarding majority of prior literature is how to quantify ‘home team advantage’ where team A plays a game in team A’s own city with its supportive spectators. As a result, spectators would encourage team A when team A is on offence and defence, leading to significant mental pressure on the ‘away’ team. Previous studies always quantify the home team advantage as an interception when they implement regression and target point spreads because they believe home advantage’s effect on point spread does not change over time (Manner, 2016; Miljković et al., 2010; Jones, 2016; Kayhan and Watkins, 2019; Cheng et al., 2013; Shi and Song, 2019). However, from our perspective, different teams’ home advantages’ effects on the point spread might be time variant and not so simple to quantify. It could depend on the number of spectators/intensity but also other geographic factors as well. For example, Utah, which is a mountain city above sea level would make unaccustomed basketball players to having breathing difficulties. But Utah Jazz’s (Utah home team) players can handle these circumstances better due to familiarity with these conditions. In addition, a single team’s home advantage might change depending on each opponent they play against as well. For instance, the Los Angeles Lakers (LAL) always beat the Los Angeles Clippers (LAC) in LAL home court, while the LAL always loses to Houston Rockets (HOU) in LAL home court. However, LAC always beats HOU in LAC home court. Hence, our study treats the home team advantage as an independent dummy variable instead of an interception variable.

All the new results and extensive literature review tables are made available in the internet appendix to stay within the page limits for the main paper.

**A.11. Comparison between game-winner and point spread odds.**

Predicting point spread has significant economic implications as we can calculate by how much a team is going to win by instead a single outcome. For example, the figure one is a snapshot of Money Line Betting (game-winner betting) in oddsportal.com for 7 Mar 2022. There are 10 games on a single day that you can bet on. But by predicting merely the game winner or loser cannot aid in the decision making of which game to bet upon in this case. Most of the time your winner versus loser predictions would suggest betting on the favourites (small odd in each game). However, persistently betting on the small odds is not profitable as these odds are over-valued. The reason is that most bettors would choose to bet on the favourites. Moskowitz (2021) argued about this phenomenon as well. However, if you predict point spreads, you will get 10 spread predictions for these ten games and an individual can choose the farthest line’s from the prediction as a safe option.

