Linear Regression Project

Predicting the Masters Golf Tournament

Colin Gallagher

Introduction

- Goal is to predict the ***Average Score Per Round*** for a golfer using performance statistics measured at the tournament

- Model will be useful to the golfers, gamblers, sports analysts, and the PGA

Methodology

- Scraped statistics and results at Masters 2001-2022 from PGA Tour website
- Organized data into one main dataframe
- 1157 rows of data and 37 features to start
- Selected different features and applied different model fittings
- Compared models with each other and then selected the best one
- All done using Python and many of the packages available

Results

- Ended up with 4 final candidate models

Model	Adj R^2	MAE	CV R^2 Avg
Model 1	0.89172	0.41402	0.90
Model 2	0.83768	0.53281	0.83
Model 3	0.88898	0.43373	0.88
Model 4	0.86423	0.44962	0.87

Model 3 Closer Look

- Features were picked using own intuition to start

- Driving Distance, Driving Accuracy, Good Drive Percentage, Greens in Regulation Percentage, Scrambling, 1-Putts per Round, 3-Putts per Round, and Putting Average

- No features that gave slight information/hint at what a player scored on a hole (anything with birdies, pars, bogeys)

- Those stats imply birdie more to score better (too obvious)

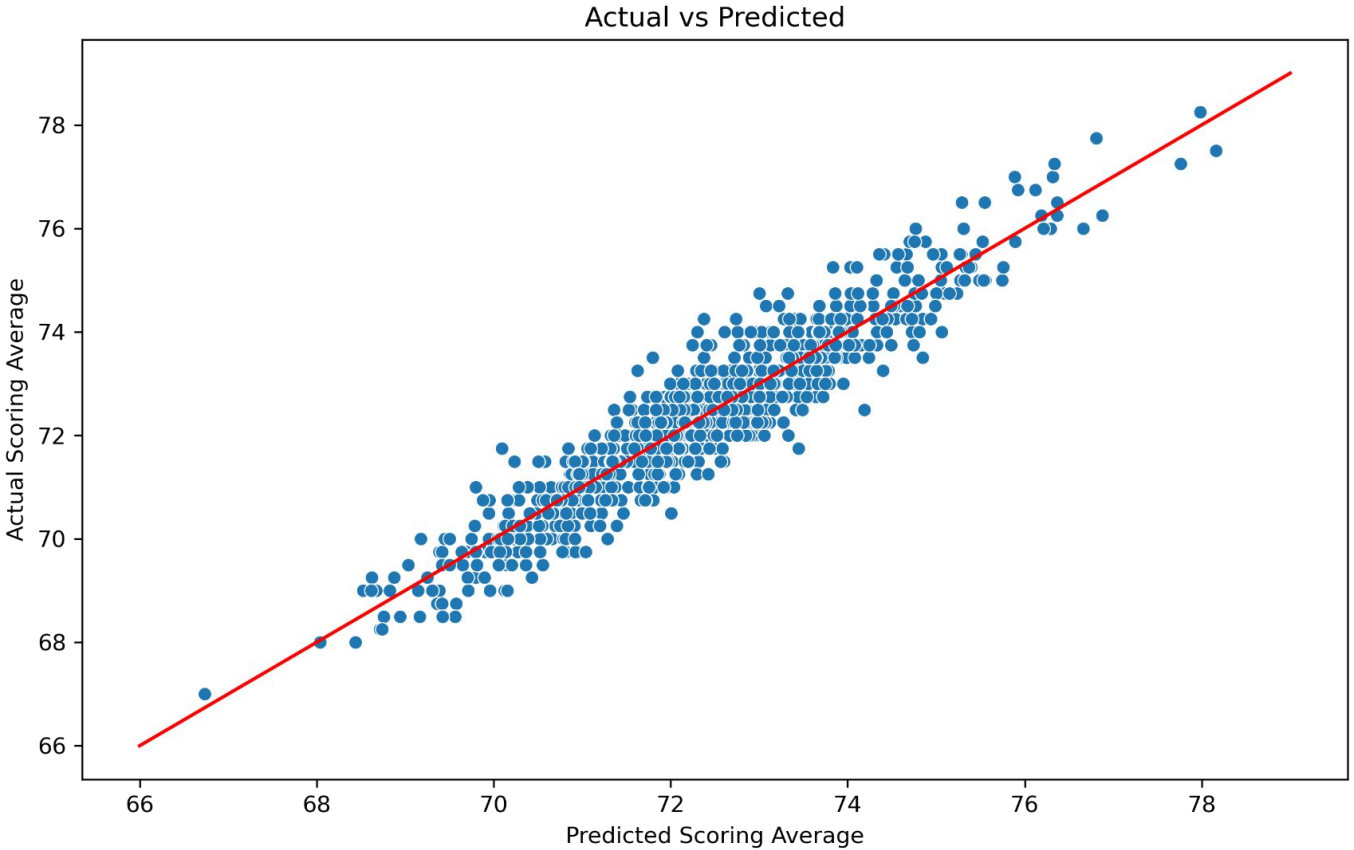
- Looking more for driver better, chip better, putt better, etc.

Model 3 Closer Look

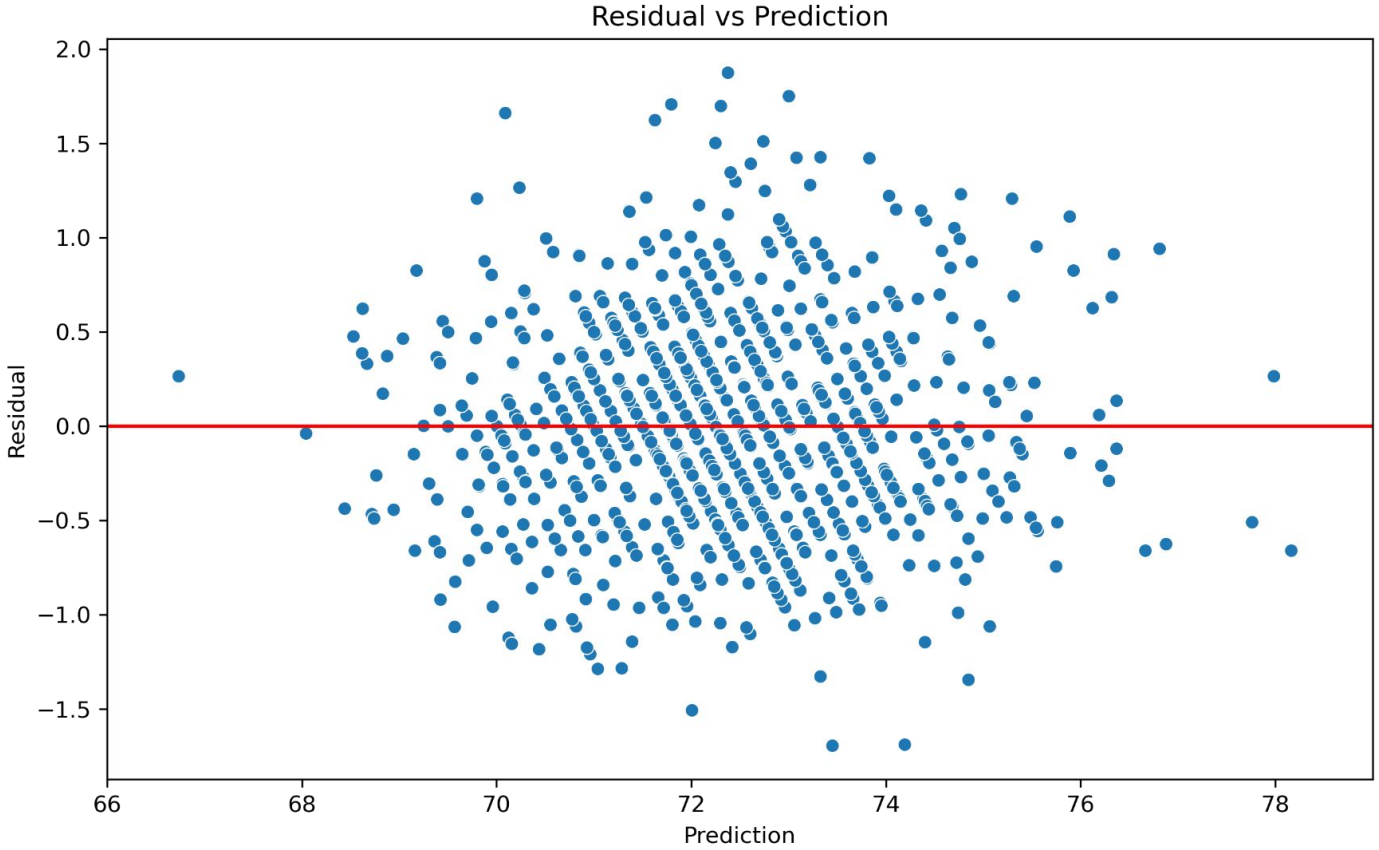
R-squared	0.885
Adj. R-squared	0.884
R-squared Test	0.876
MAE	0.442

Feature	Coefficient	VIF
const	70.9345	-
Driving Distance	-0.0137	1.112
Greens in Regulation Percentage	-0.1352	1.114
Scrambling	-0.0713	1.036
3-Putts per Round	0.2783	1.479
Putting Average	9.7285	1.493

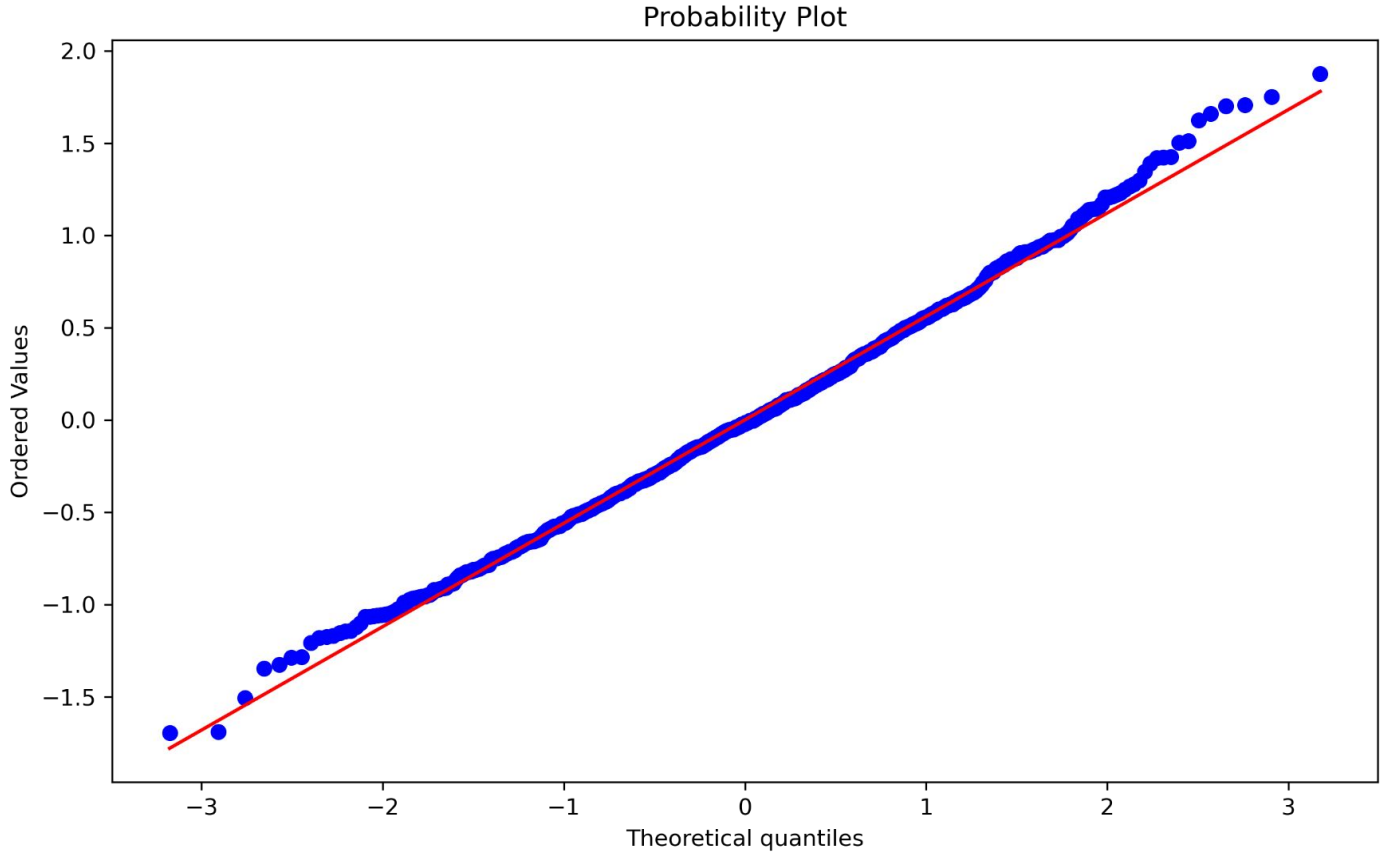
Model 3 Closer Look



Model 3 Closer Look



Model 3 Closer Look



Model 3 Closer Look

- Final Model Equation:

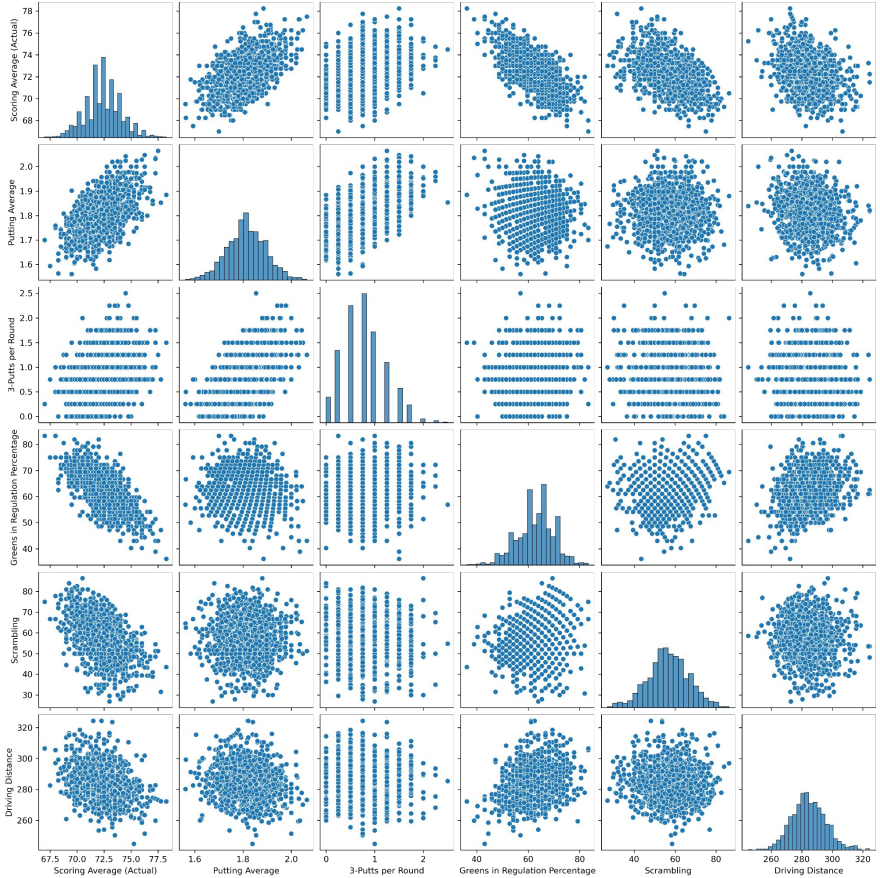
Scoring Average (predicted)=
70.9345
+9.7285*(Putting Average)
+0.2783*(3-Putts per Round)
-0.1352*(Greens in Regulation Percentage)
-0.0713*(Scrambling)
-0.0137*(Driving Distance)

Start with a base score and depending on how a golfer performs by measuring the statistics, they either lose or gain a certain amount of strokes

(1/18)*100=5.56%

-0.1352*5.56%= -0.75 ; everytime a golfer gets a GIR, they reduce their score by 0.75 of a stroke

Model 3 Closer Look



Model 3 Example: Future Prediction

Rory McIlroy:

Putting Average = 1.748

3-Putts per Round = .52

Greens in Regulation Percentage = 69.56%

Scrambling = 63.14%

Driving Distance = 319.4

Masters Predicted Average Round Score = 69.8

2022 Masters Average Score = 70.25

Conclusions

- A Linear Regression Model was able to be created

- Predicts the Scoring Average per Round at the Masters

- Mean Absolute Error of 0.44 strokes

- Model can be used to benefit the golfers themselves, gamblers, sports analysts, and PGA

Future Work

- More tournaments
- Explore different features/statistics
- More evaluations on how well it predicts the future
- More feature engineering to improve accuracy

Questions?

Appendix

- Model 0

- Started with all features initially and ended with:

Scoring Average (pred)=

4*(Par 3 Avg)+10*(Par 4 Avg)+4*(Par 5 Avg)

- Not interesting at all because coefficients are just how many Par 3, Par 4, and Par 5 holes are on the course

- This is what got me thinking about removing “scoring” features

Appendix

- Model 1

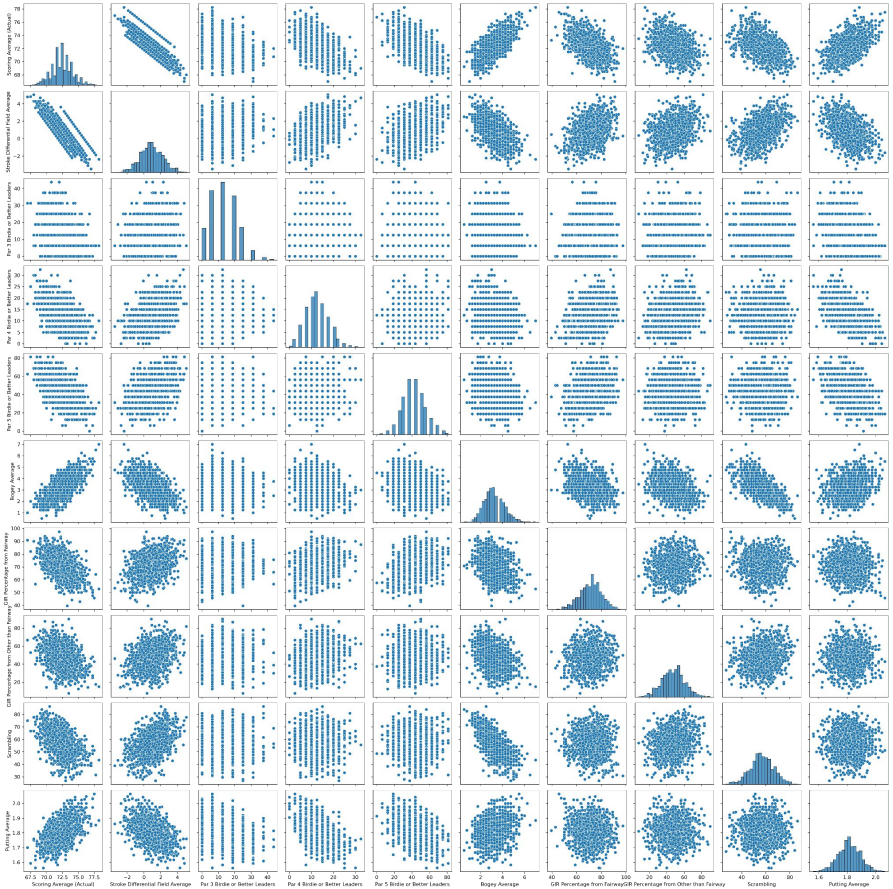
- Started with features that were more correlated to target minus the 3 final features from Model 0 (Par 3, 4, 5 Averages)

- Ended with equation:

Scoring Average (pred)=
75.4475
-0.2648*(Stroke Differential Field Average)
-0.0372*(Par 3 Birdie or Better Leaders)
-0.0622*(Par 4 Birdie or Better Leaders)
-0.0333*(Par 5 Birdie or Better Leaders)
+0.3488(Bogey Average)
-0.0364*(GIR Percentage from Fairway)
-0.0127*(GIR Percentage from Other than Fairway)
-0.0411*(Scrambling)
+2.3505*(Putting Average)

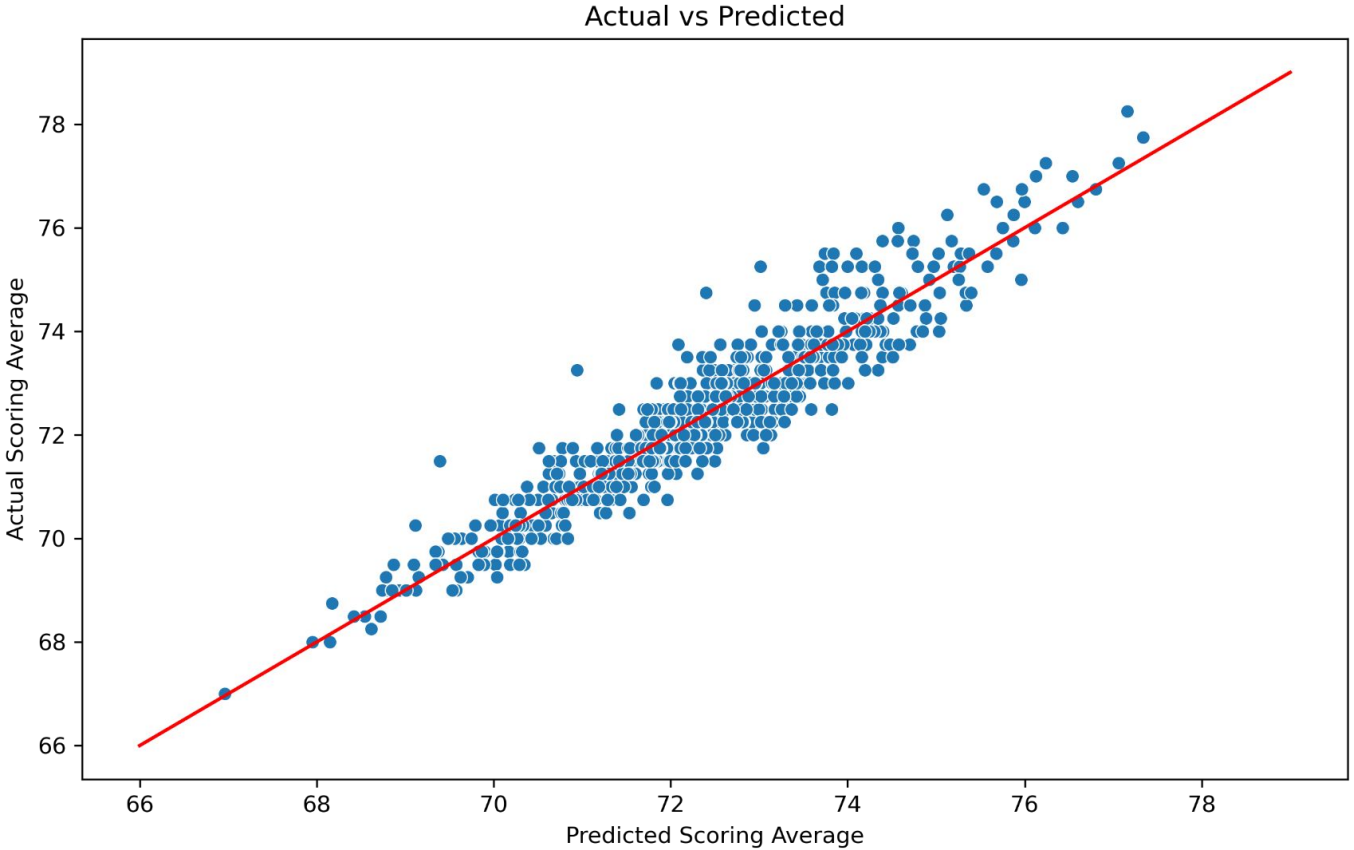
Appendix

- **Model 1**



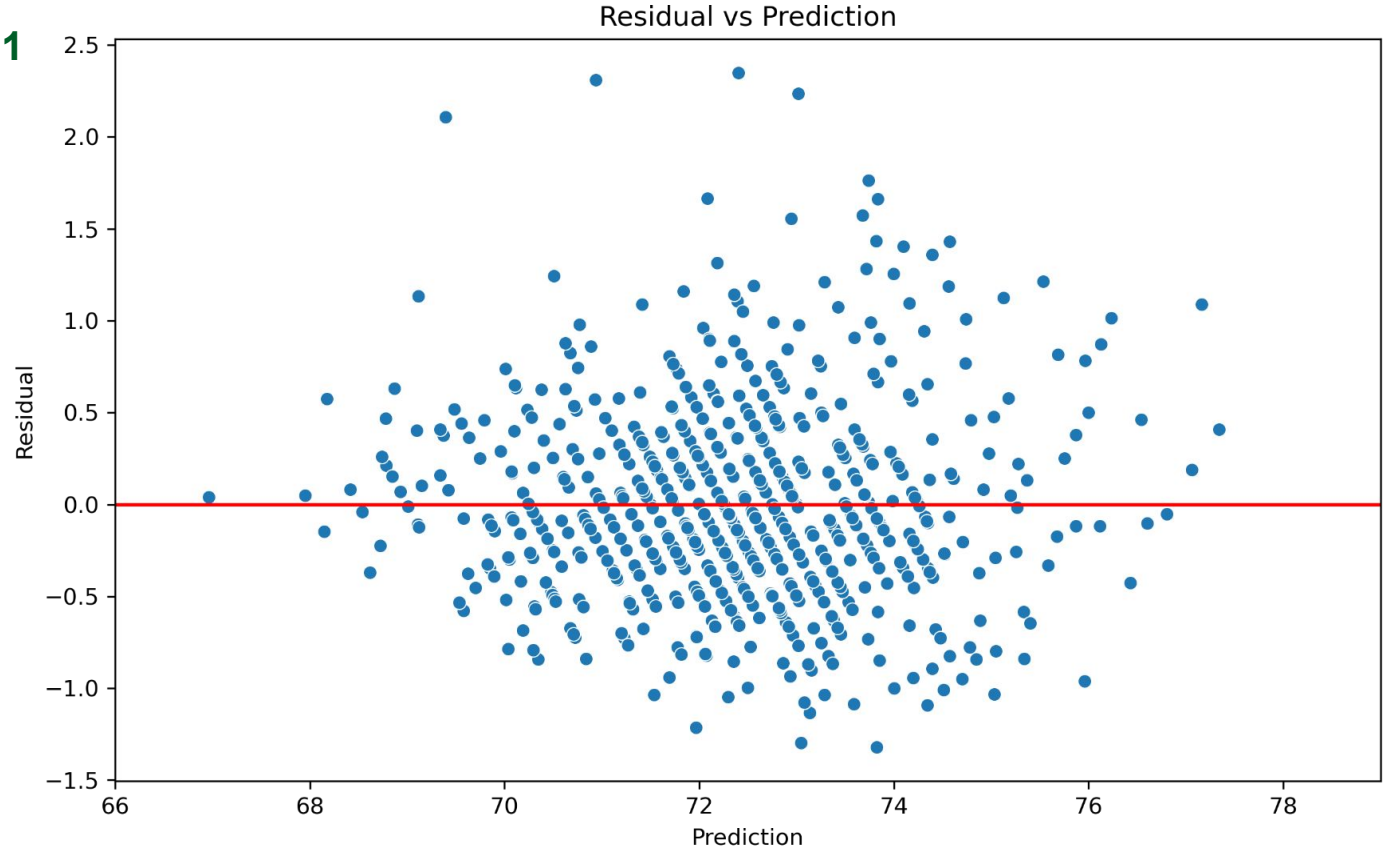
Appendix

- Model 1



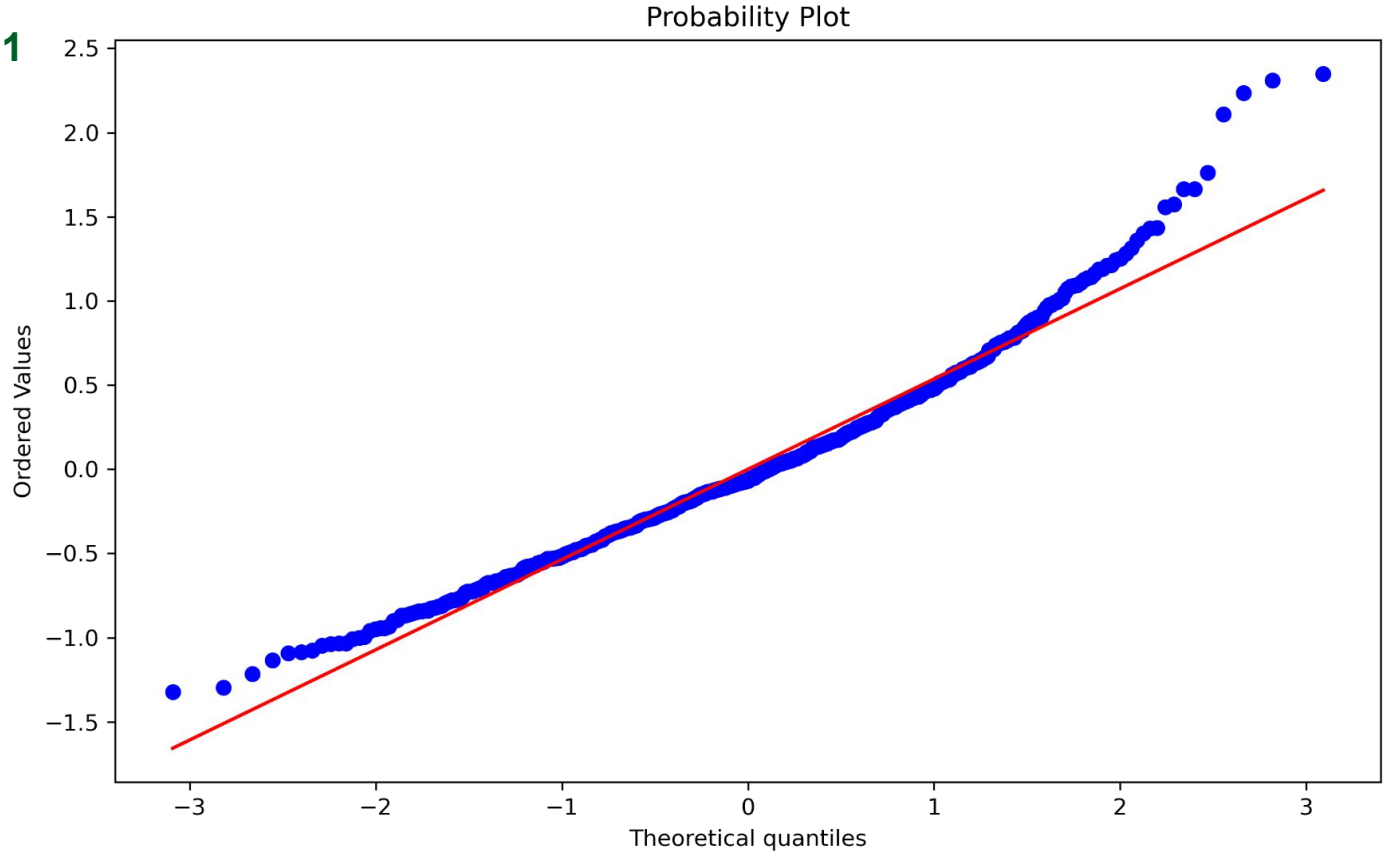
Appendix

- Model 1



Appendix

- Model 1



Appendix

- Model 2

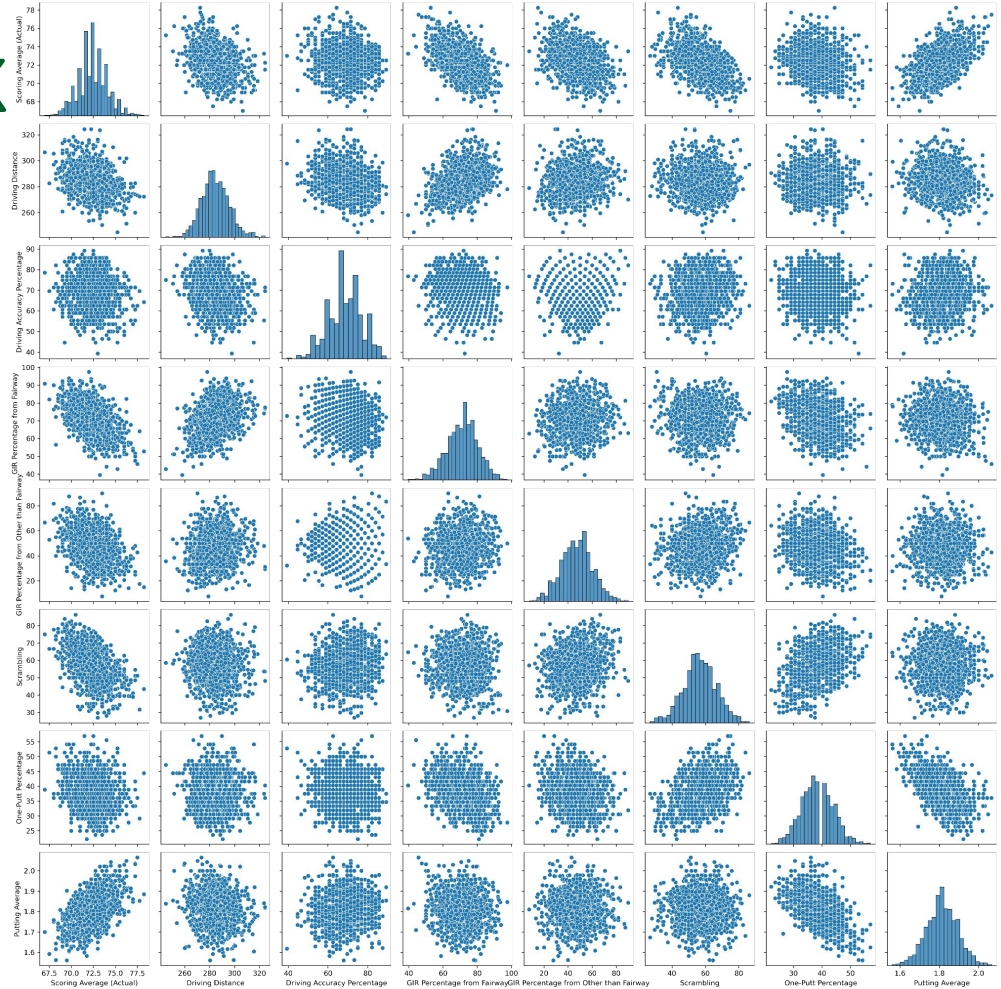
- Started with features that were more correlated to target minus the 3 final features from Model 0 (Par 3, 4, 5 Averages)

- Ended with equation:

66.2557
-0.0178*(Driving Distance)
-0.0285*(Driving Accuracy Percentage)
-0.0695*(GIR Percentage from Fairway)
-0.0265*(GIR Percentage from Other than Fairway)
-0.0888*(Scrambling)
+0.0489*(One-Putt Percentage)
+12.4200*(Putting Average)

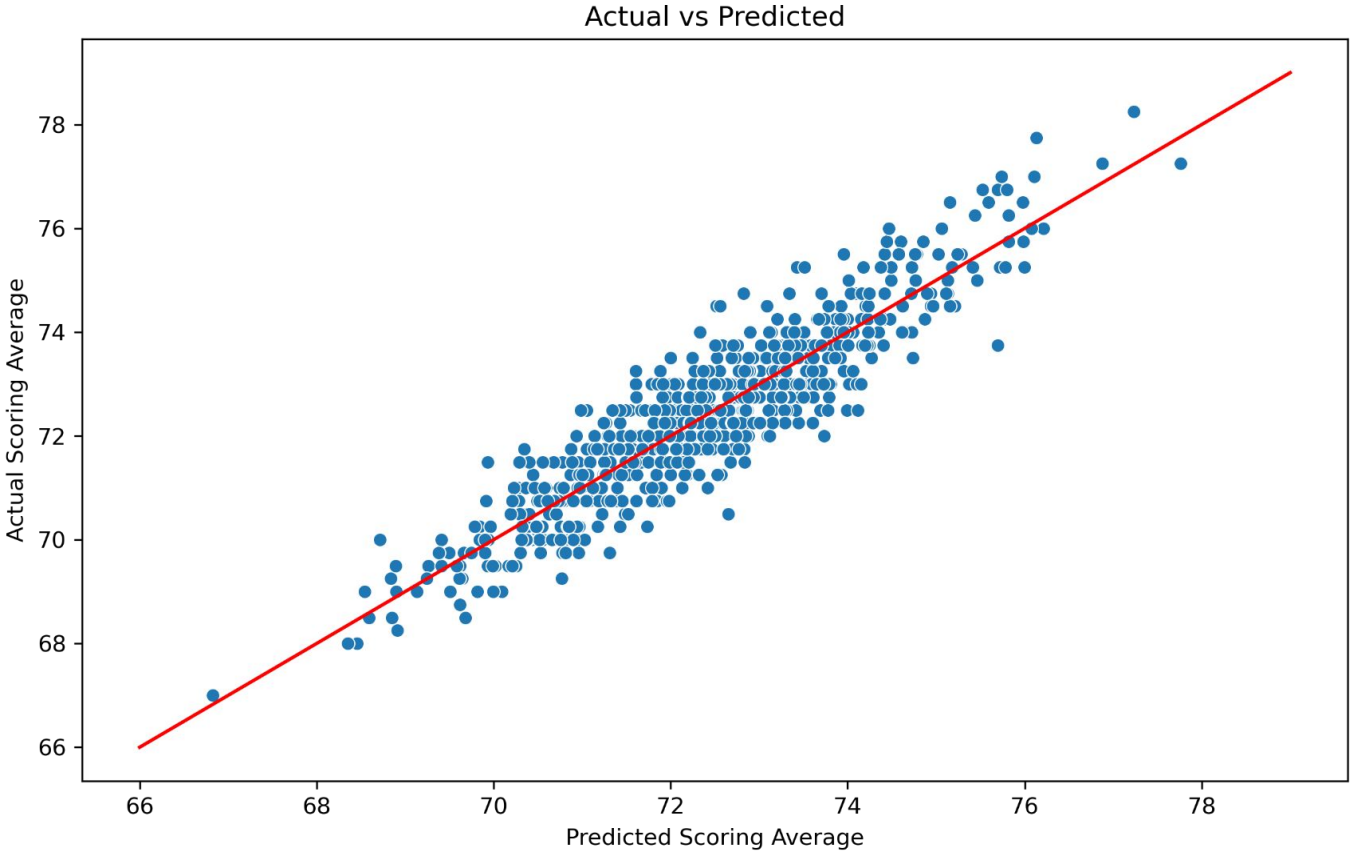
Appendix

- Model 2



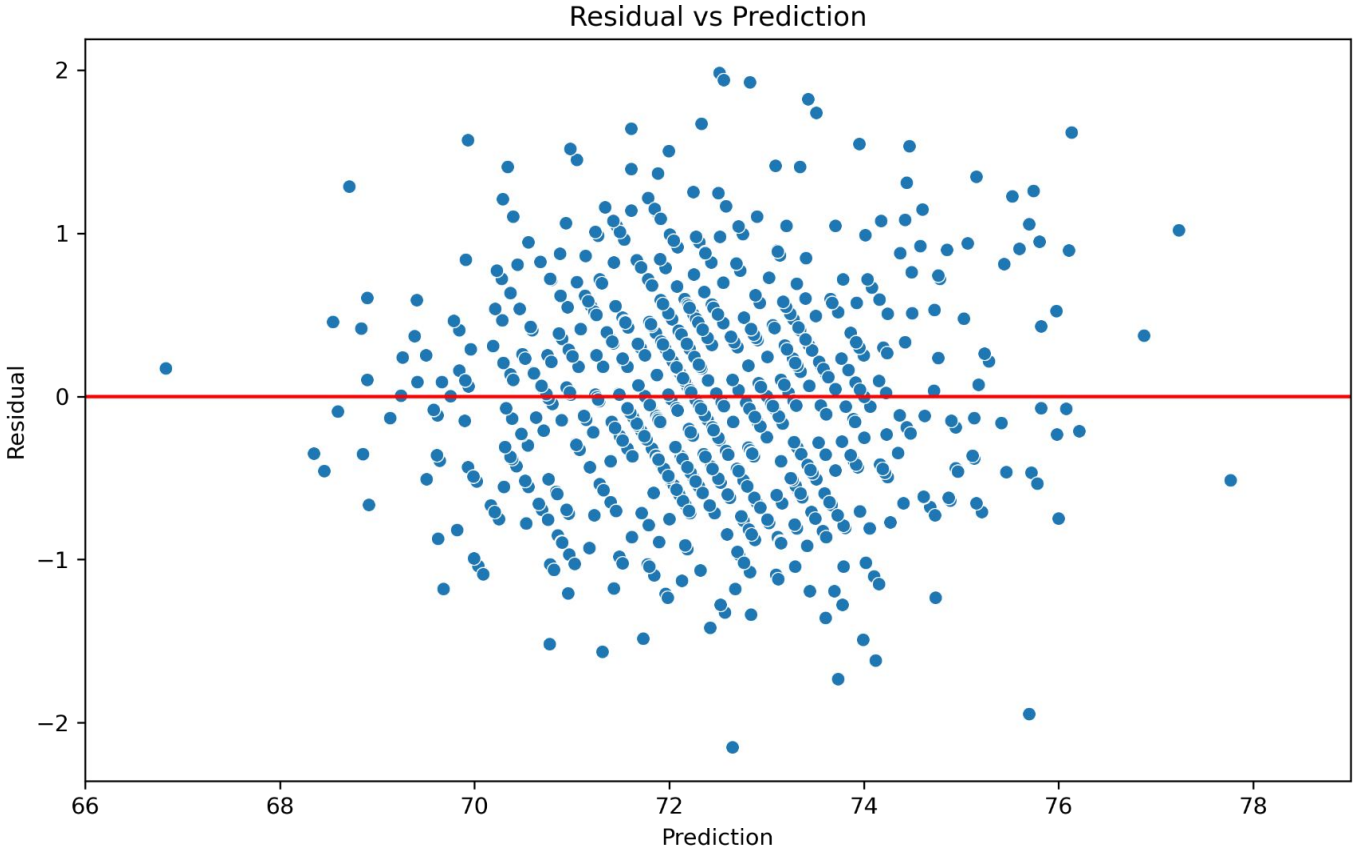
Appendix

- Model 2



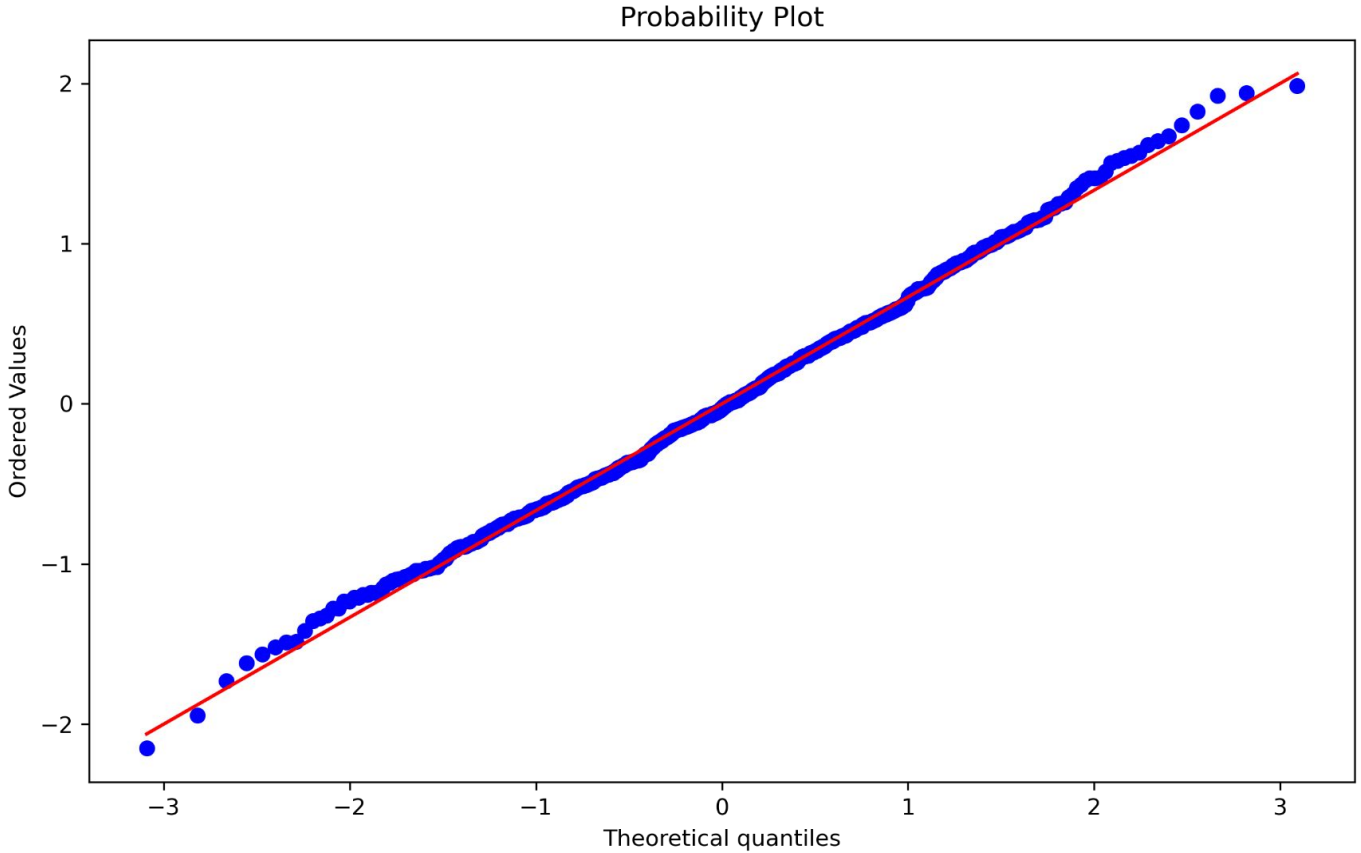
Appendix

- Model 2



Appendix

- Model 2



Appendix

- Model 4

- Slight modification of Model 3, really wanted to keep Good Drive Percentage feature, but it still ended up being dropped out

- Ended with equation:

67.6072

-0.1443*(Greens in Regulation Percentage)

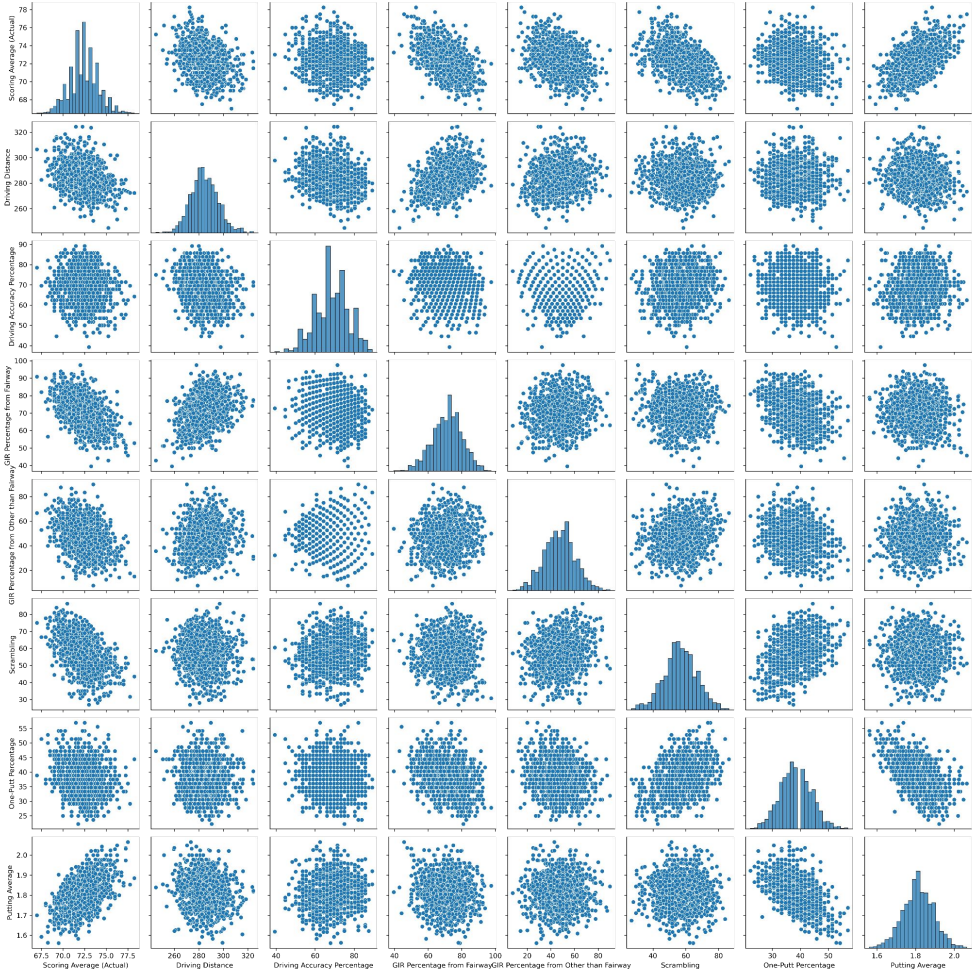
-0.0708*(Scrambling)

+0.3161*(3-Putts per Round)

+9.7000*(Putting Average)

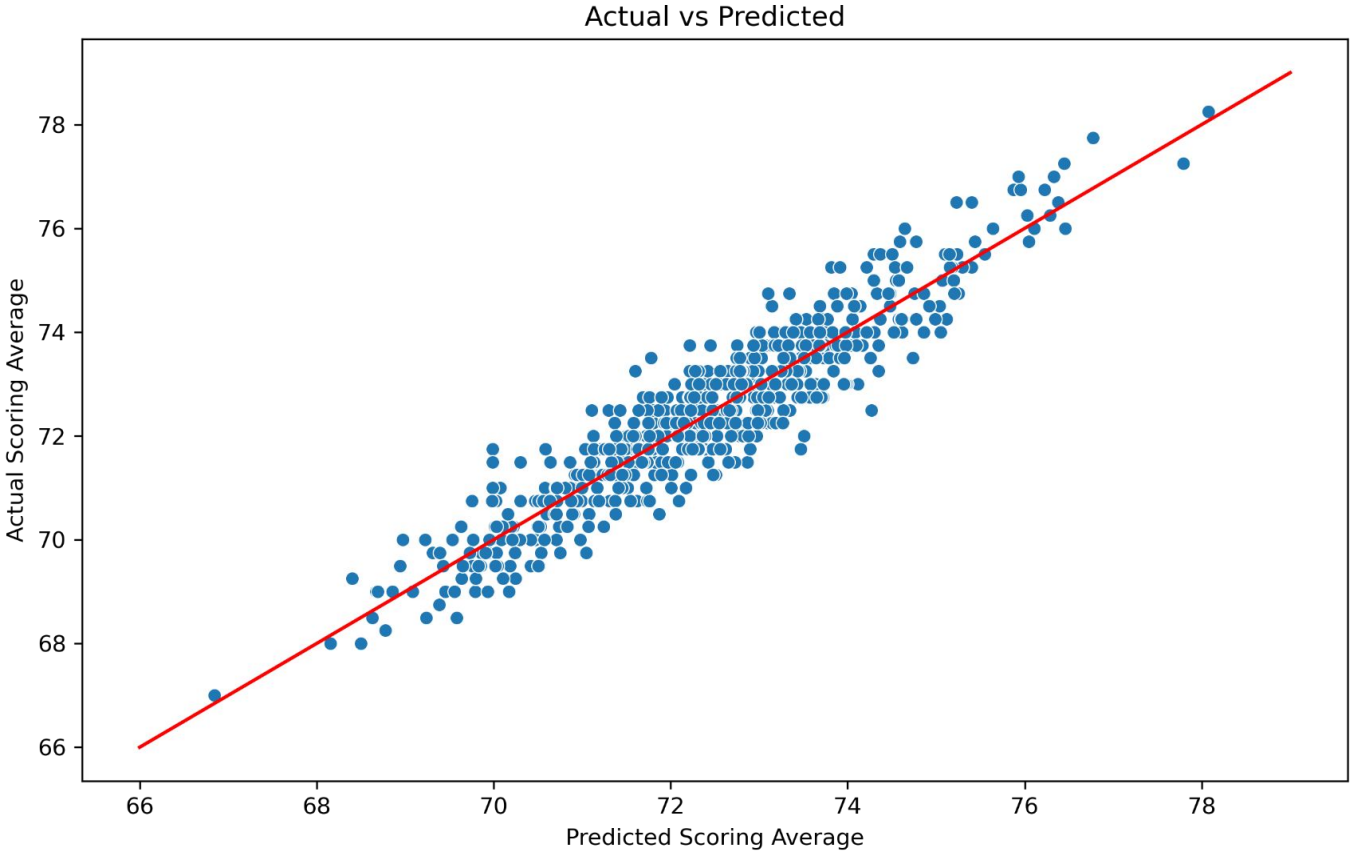
Appendix

- Model 4



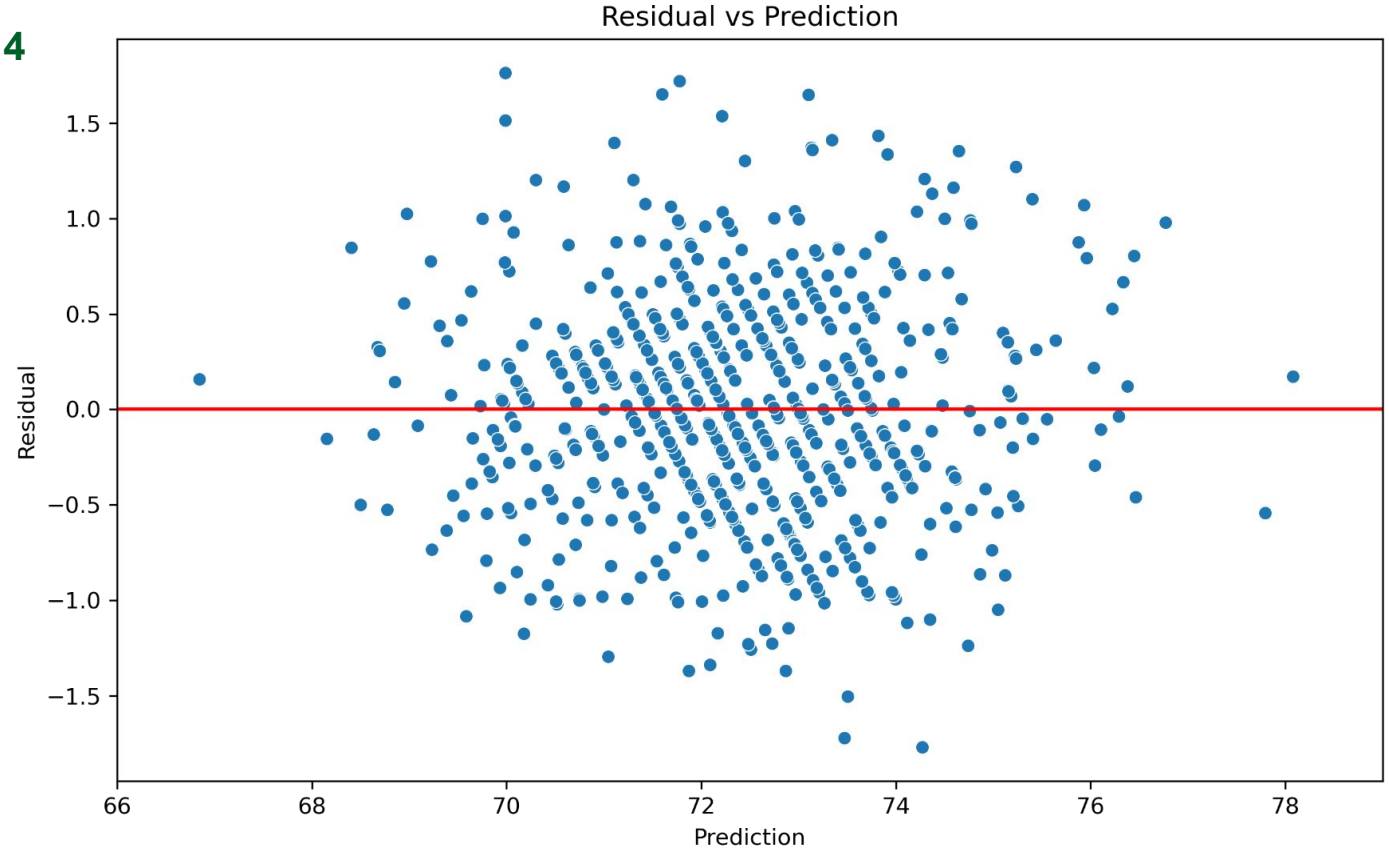
Appendix

- Model 4



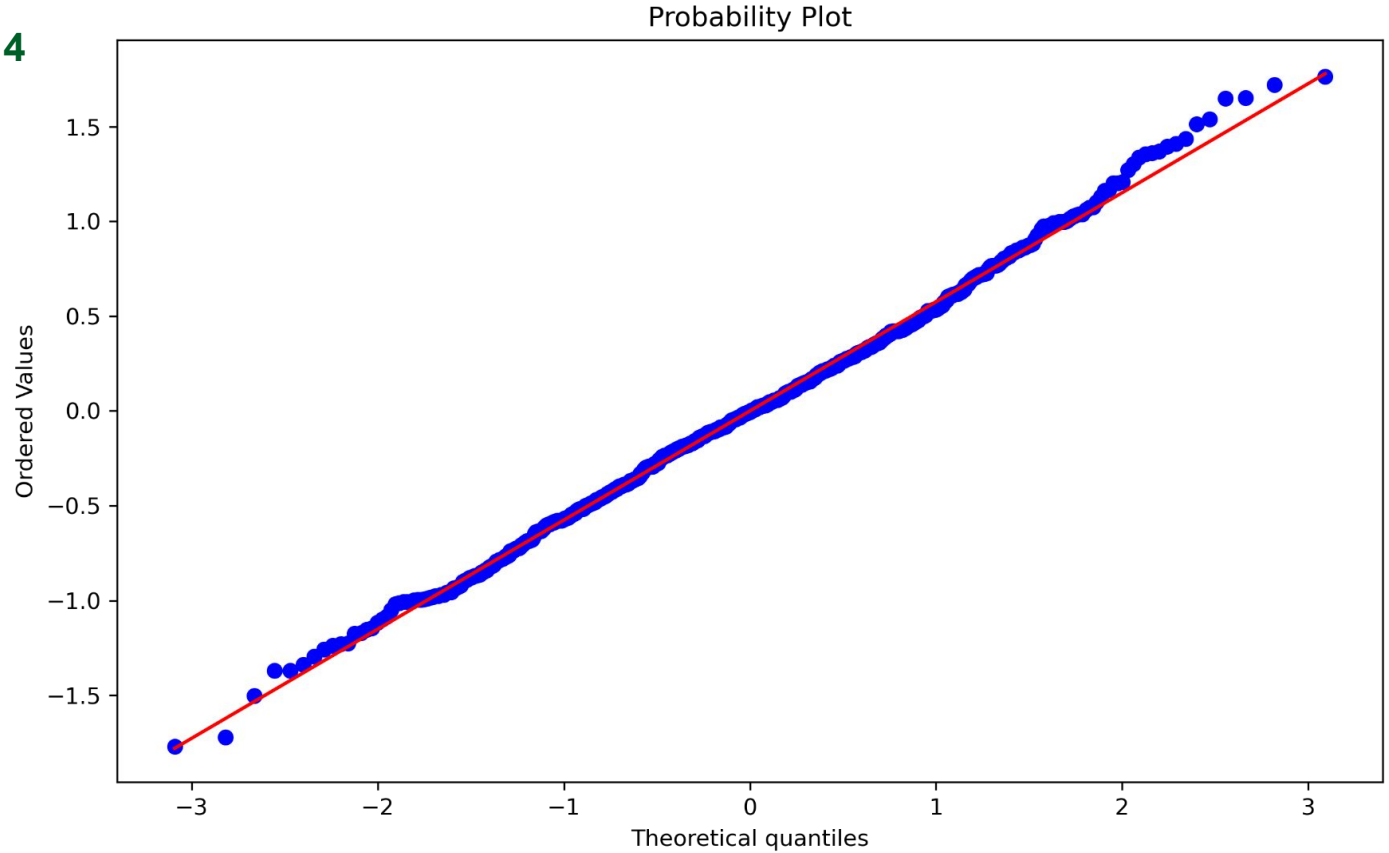
Appendix

- Model 4



Appendix

- Model 4



Appendix

- Model 1 Ridge, Lasso, and Elastic Net

```
Alpha: 4.297004704320839
R^2: 0.8931202382212727
R^2 Val: 0.918705641652148

Alpha: 0.0001
R^2: 0.893130931829825 1
R^2 Val: 0.91829758364512

Alpha: 0.006747544053110693
l1 ratio: 1e-05
R^2: 0.893130931829825 1
R^2 Val: 0.9187414248211012
```

Appendix

- Model 2 Ridge, Lasso, and Elastic Net

```
Alpha: 2.833096101839324
R^2: 0.8392994196131048
R^2 Val: 0.8214131426990146

Alpha: 0.0002300430119772917
R^2: 0.8393248587934845
R^2 Val: 0.8211303764251581

Alpha: 0.004247571552536898 5
l1 ratio: 1e-05
R^2: 0.8393248587934845
R^2 Val: 0.8214253853551688
```

Appendix

- Model 3 Ridge, Lasso, and Elastic Net

```
Alpha: 0.890735463861044
R^2: 0.8899463837572834
R^2 Val: 0.8671365491841347

Alpha: 0.0002002200371815584 5
R^2: 0.8899474872040717
R^2 Val: 0.8671487864934854

Alpha: 0.0012173827277396 62
l1 ratio: 0.1
R^2: 0.8899474872040717
R^2 Val: 0.8671521642559241
```

Appendix

- Model 4 Ridge, Lasso, and Elastic Net

```
Alpha: 0.775259748862946
R^2: 0.8803063067598482
R^2 Val: 0.8642630879533648

Alpha: 0.00037399373024787 98
R^2: 0.880307077160883
R^2 Val: 0.8642540914726357

Alpha: 0.0011357333583431 05
l1 ratio: 0.1
R^2: 0.880307077160883
R^2 Val: 0.8642684291361382
```