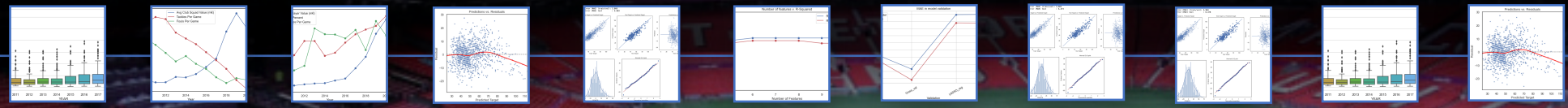
REGRESSION:
PREDICTING EUROPEAN SOCCER LEAGUE POINTS



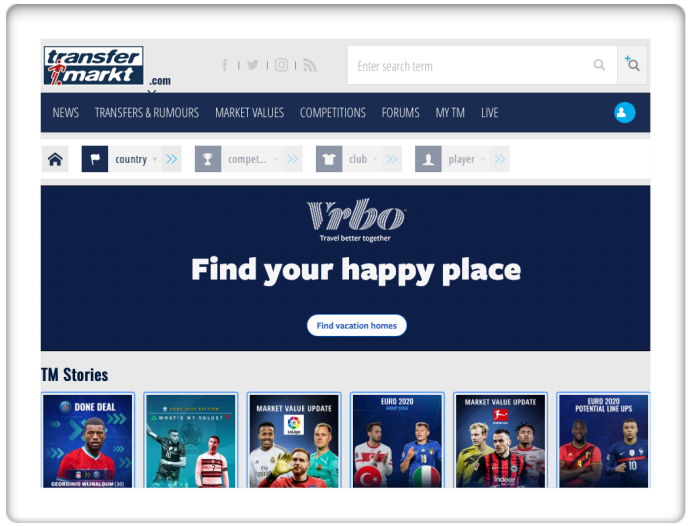
METIS BOOTCAMP | June 11, 2021

Presented by: Kalu Uga

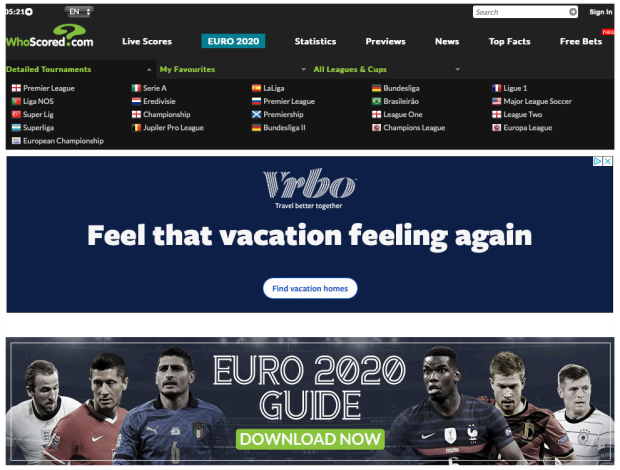
OUTLINE

- Web Scraping with Selenium
- Backward Elimination with R_squared_adj
- Feature Engineering
- Error Analysis
- Result from Model

WEB SCRAPING WITH SELENIUM



- transfermarkt.com
- One page
- Two tables



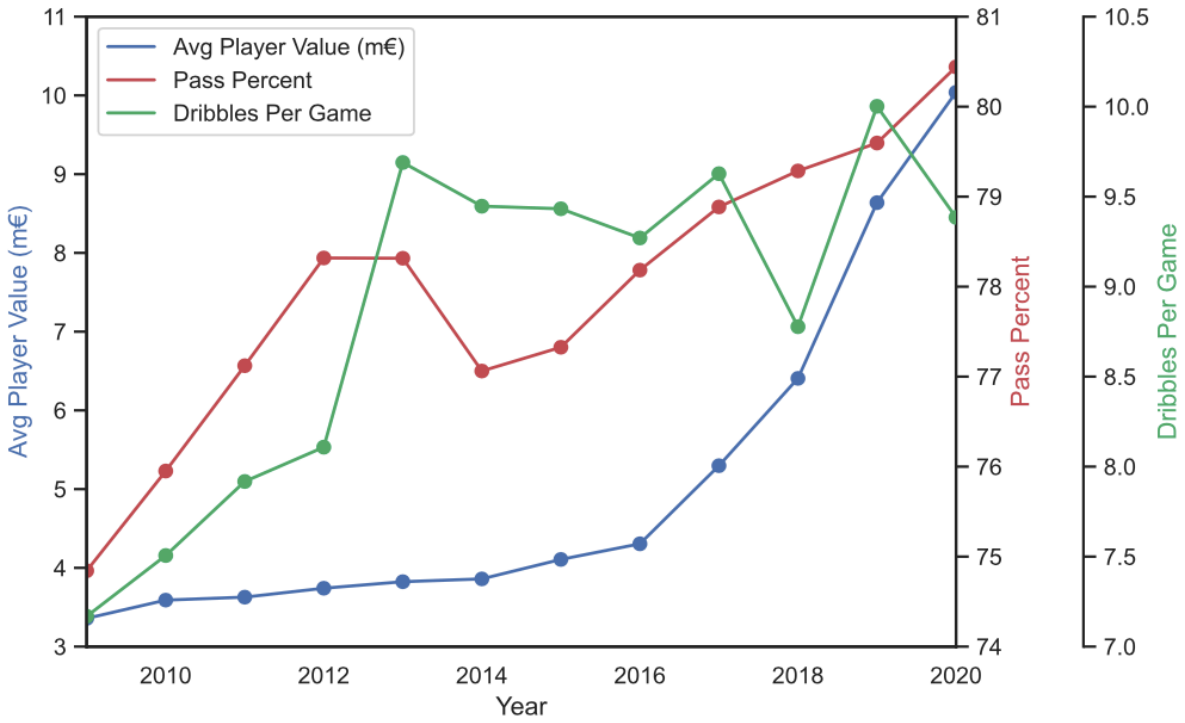
- whoscored.com
- Two pages and three tabs
- Four tables

Top 5 European Leagues	
Country	**Team Count**
England	20
Spain	20
Germany	18
Italy	20
France	20

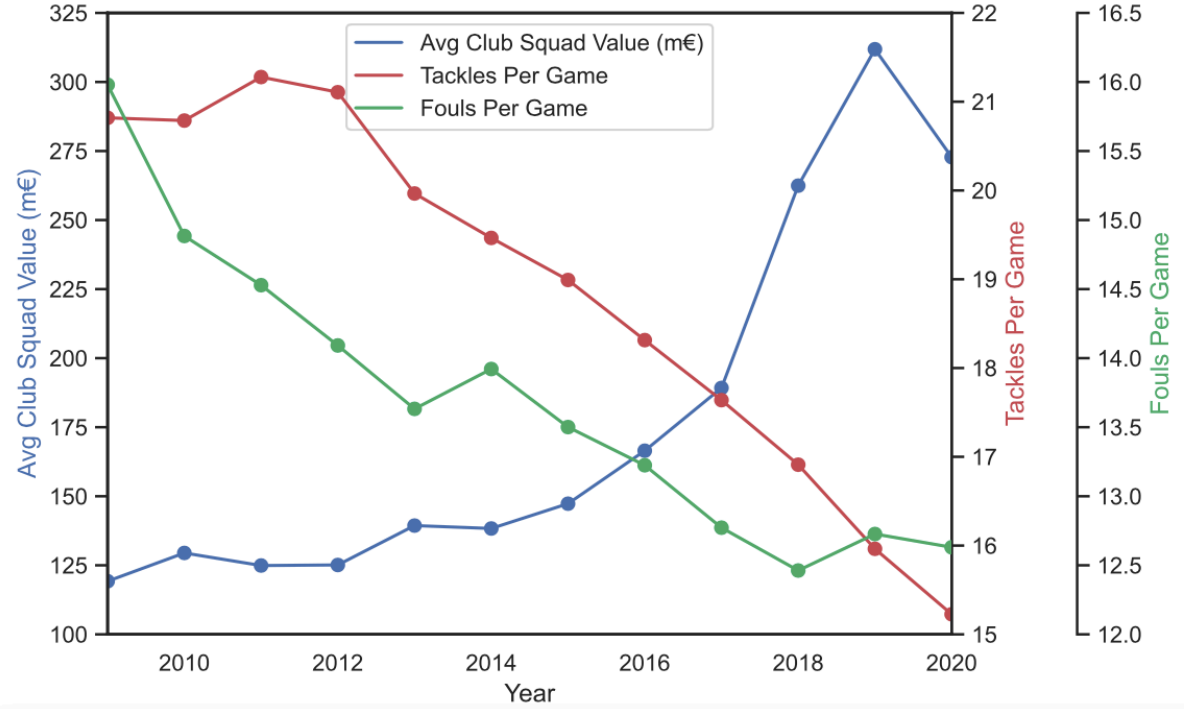
- 12 Seasons (2009 - 2020)
- Table size: 1176 x 31

EDA - TREND

Offensive Trend



Defensive Trend



BACKWARD ELIMINATION WITH R_SQUARED_ADJ

Features

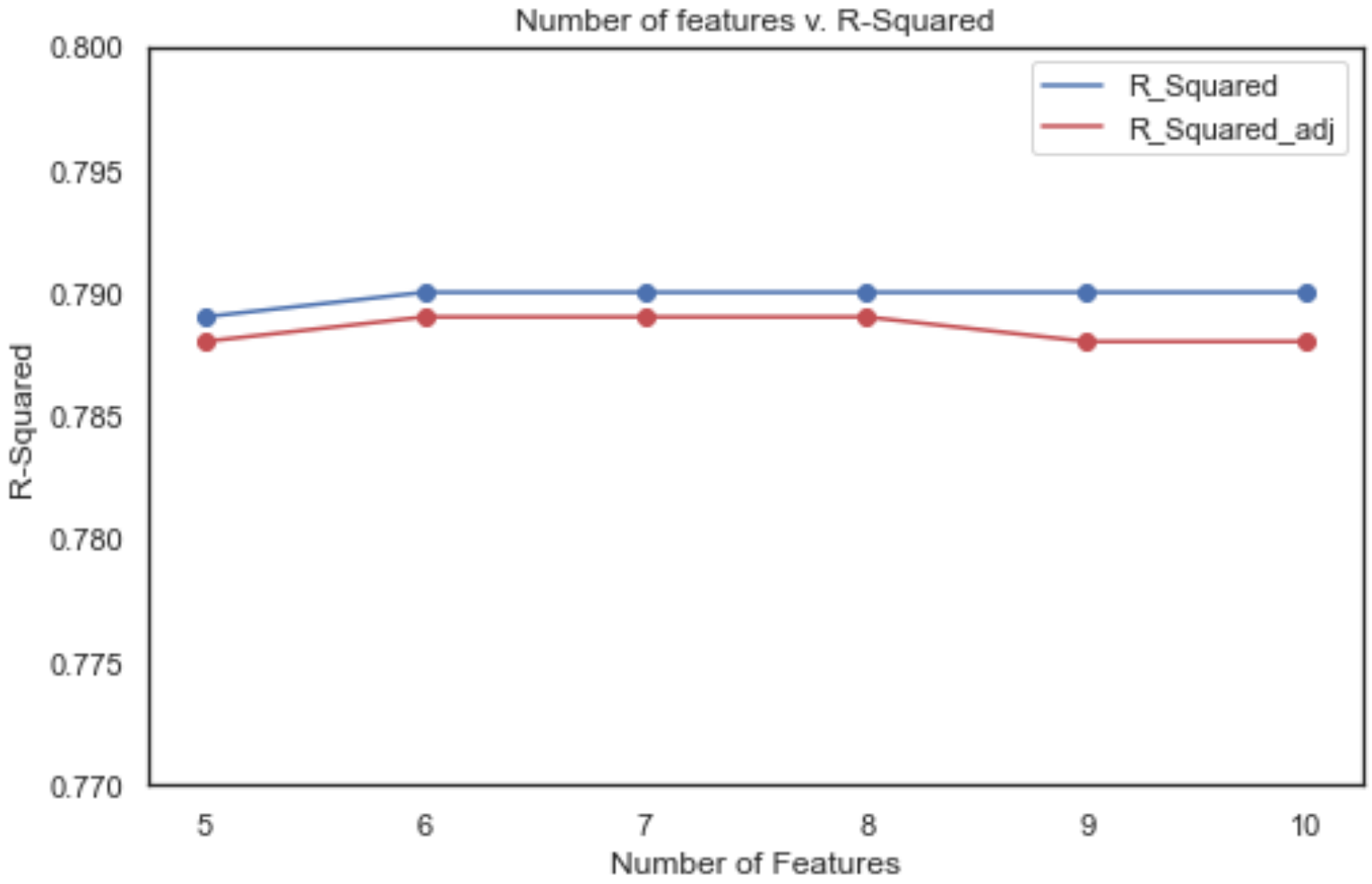
Goals For
Possession
Shots per game
Total Market Value (m€)
Tackles per game
Interceptions …

Target

Total Points

Regression

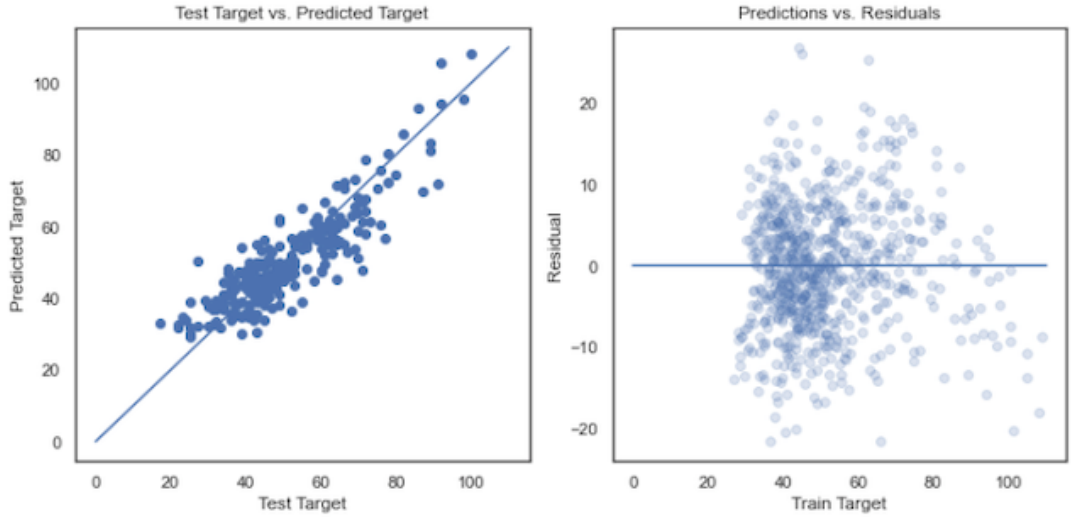
Y = Bx + c



FEATURE ENGINEERING

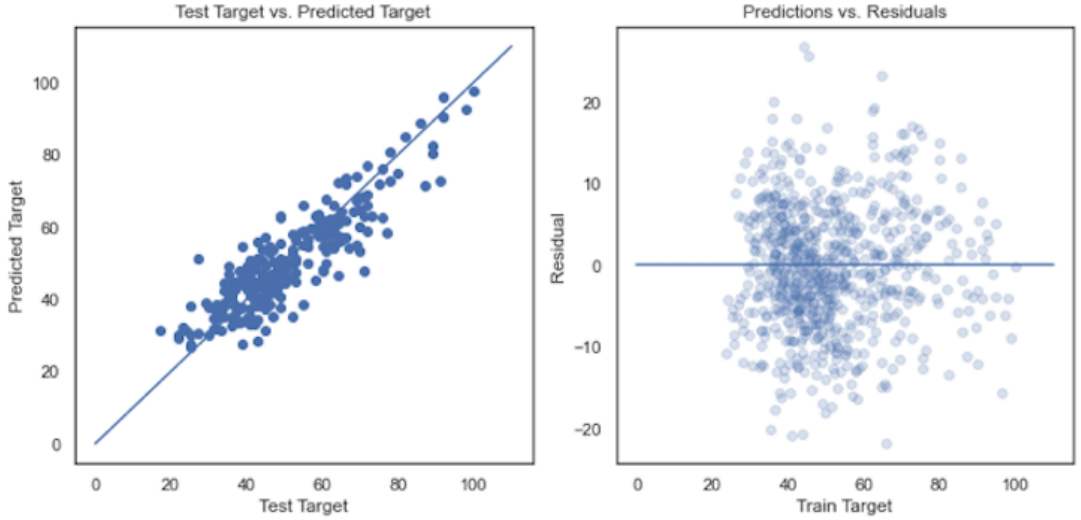
Initial Features = 6

INTERCEPT: -16.535
GF: 0.7284
POSSESSION: 0.5103
SHOTS PG: -0.2471
TOTAL MARKET VALUE (m€): 0.0099
TACKLES PG: 0.2629
INTERCEPTIONS PG: 0.1016

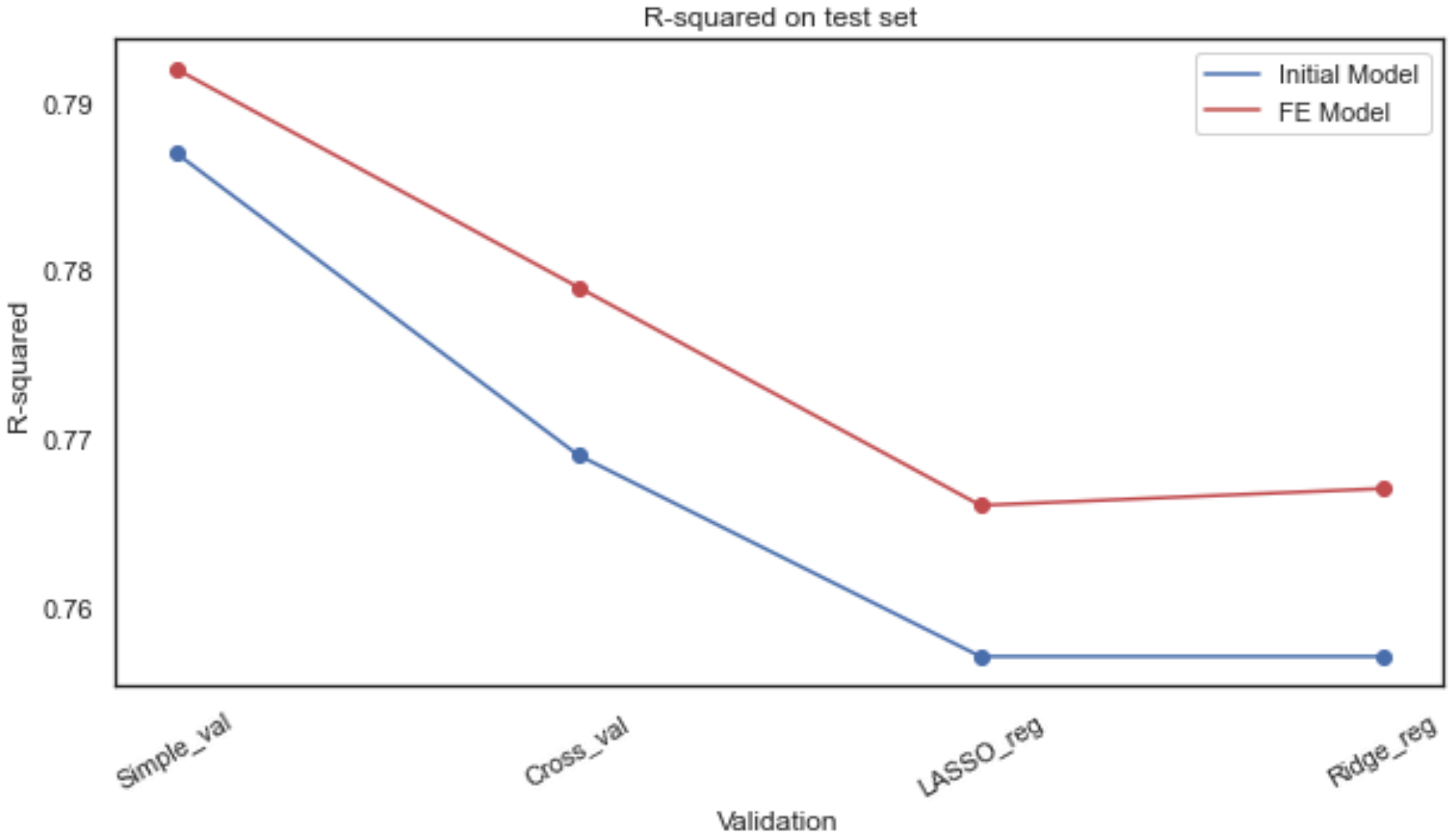


Feature Engineering
Added features = 3

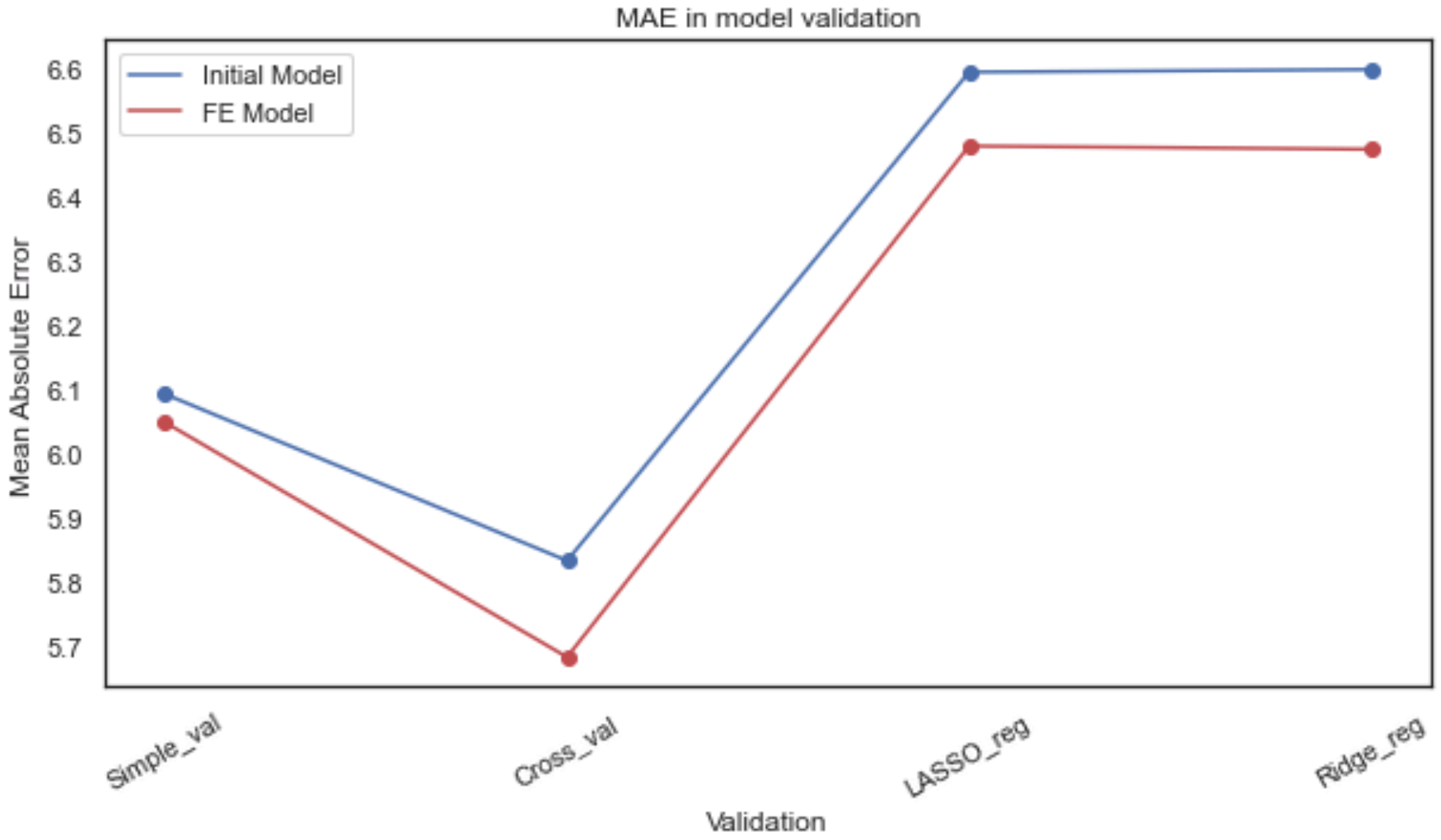
INTERCEPT: -23.523
GF: 1.142
POSSESSION: 0.456
SHOTS PG: -1.235
TOTAL MARKET VALUE (m€): 0.009
TACKLES PG: 0.3
INTERCEPTIONS PG: 0.136
GF2: -0.004
SHOTS PG 2: 0.0035
TOTAL MARKET VALUE_LOG: 2.089



R-SQUARED PLOT

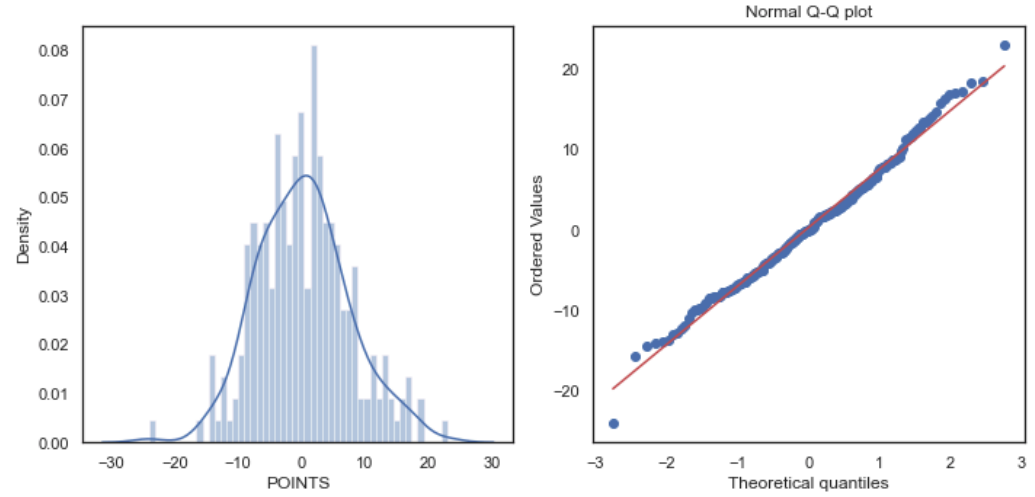


ERROR ANALYSIS



CROSS VALIDATION - RIDGE

```
Mean Absolute Error (MAE) (train/val): 5.969
Mean Absolute Error (MAE) test         : 5.683
```



PREDICTION

- Manchester City (2009/2010), Points = 67

  Calculated Points (FE Model) = 64
  Calculated Points (initial Model) = 71

- Investing an extra 1.0 million Euros

  Point Advantage is 0.009
  The true investment is seen in other features that rank high in the model

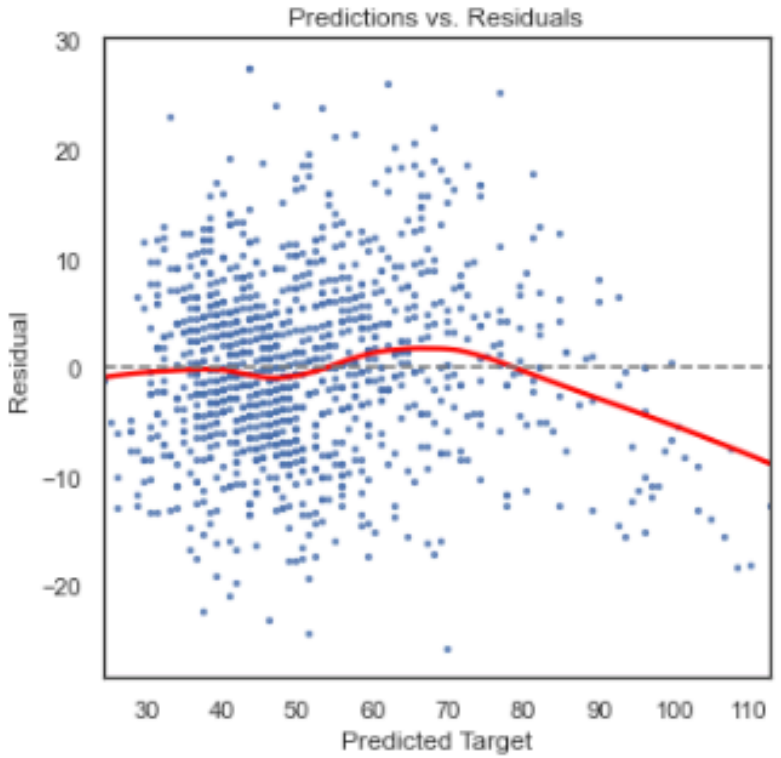
INTERCEPT: -23.523
GF: 1.142
POSSESSION: 0.456
SHOTS PG: -1.235
TOTAL MARKET VALUE (m€): 0.009
TACKLES PG: 0.3
INTERCEPTIONS PG: 0.136
GF2: -0.004
SHOTS PG 2: 0.0035
TOTAL MARKET VALUE_LOG: 2.089

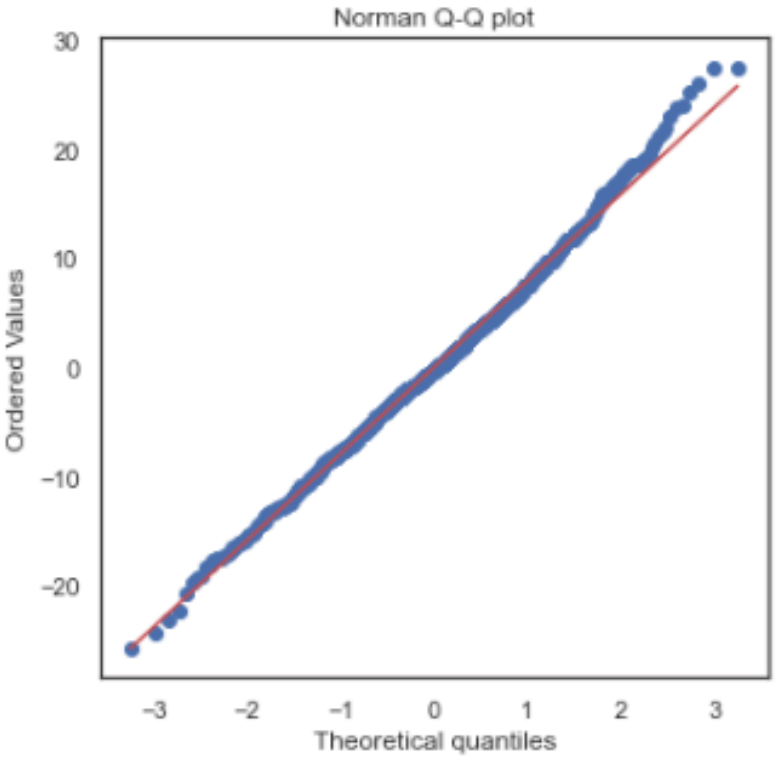
FUTURE WORK

- Predict total points on a league/season basis using a Time Series Model.
- Use Power BI to create interactive dashboard for data visualization

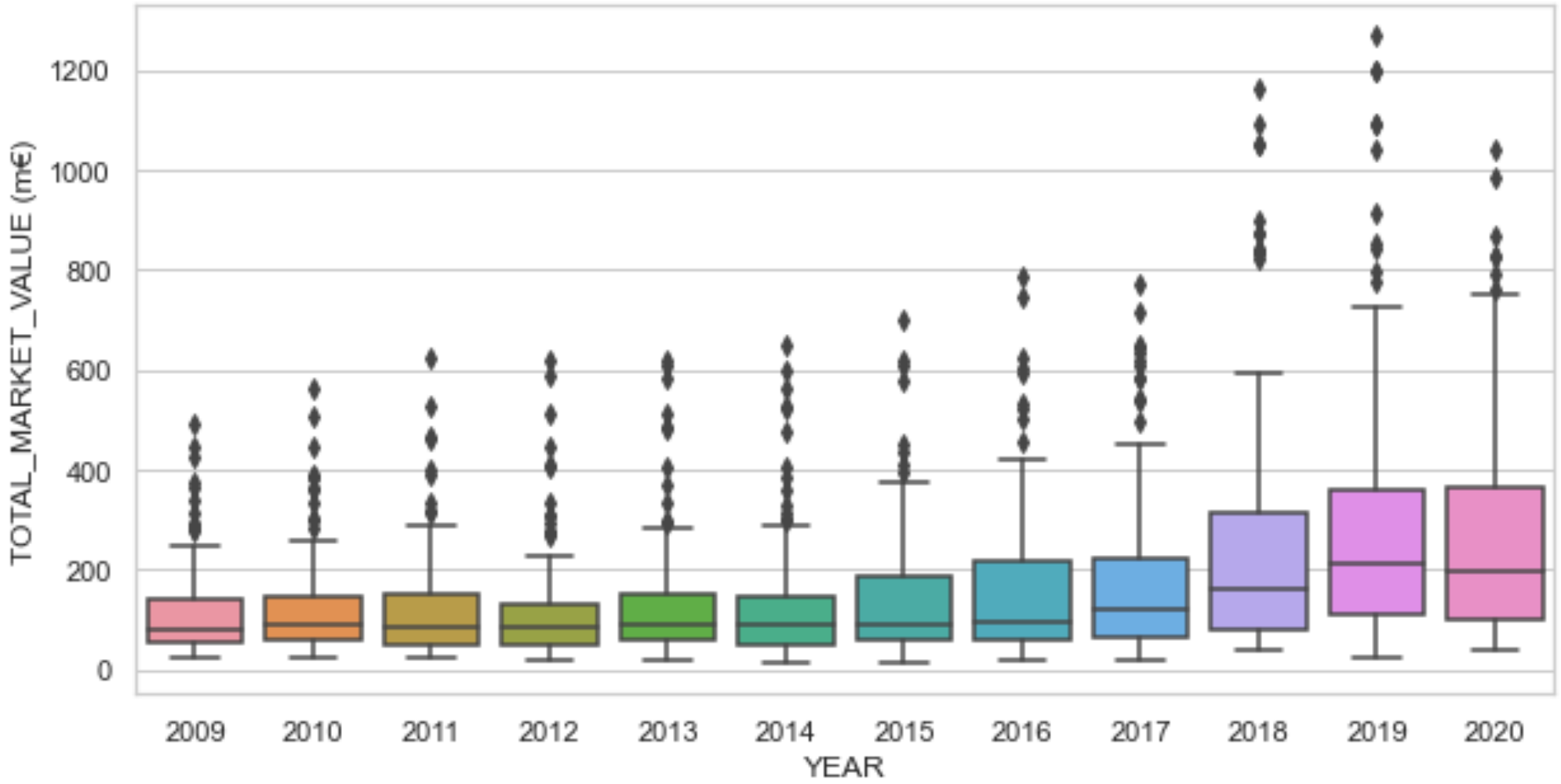
WHY CHOOSE REGRESSION FOR THIS DATA

Diagnostics plot of GOAL_FOR



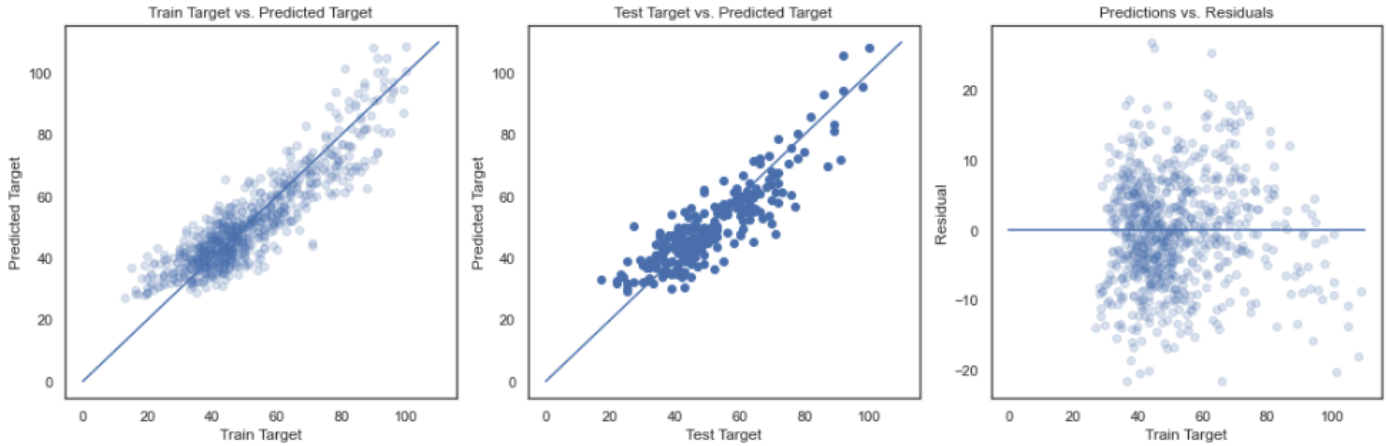


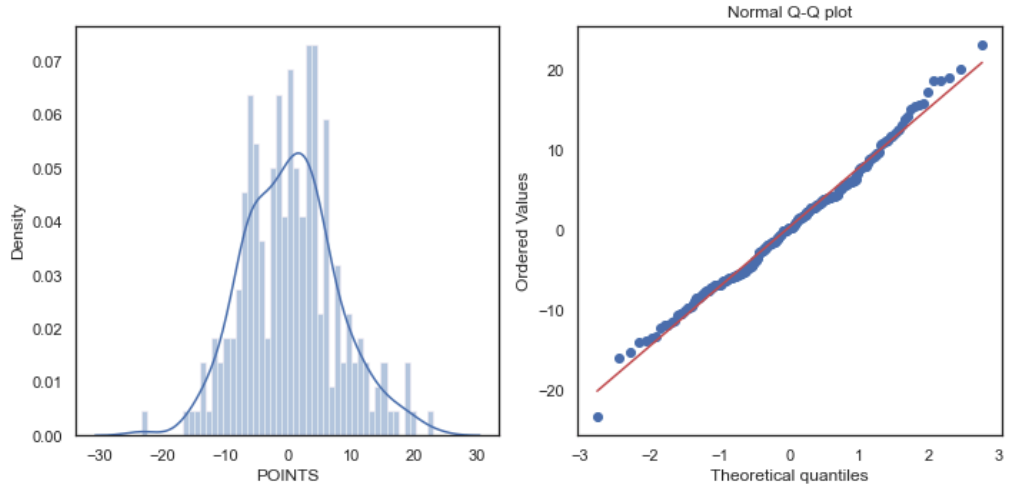
TOTAL MARKET VALUE TREND



CV RIDGE - INITIAL MODEL

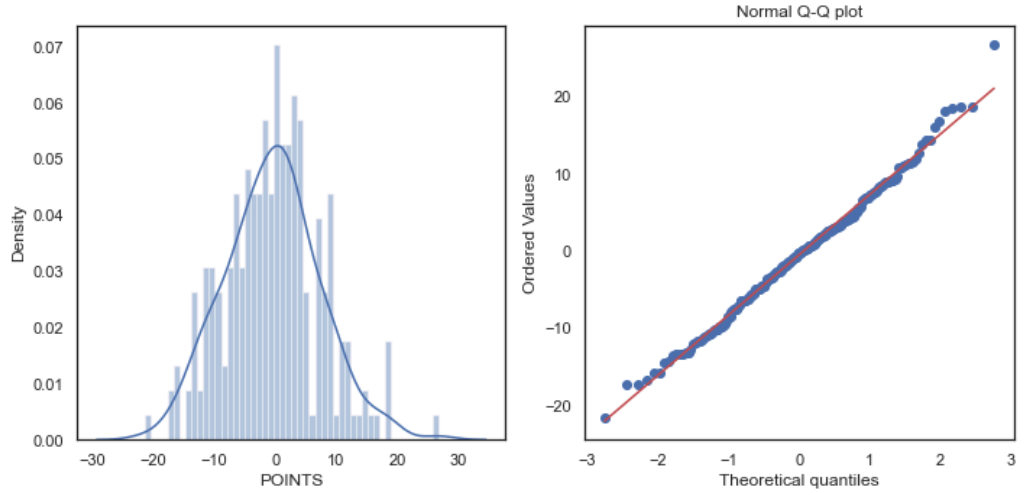
```
Mean Absolute Error (MAE) (train/val): 6.054
Mean Absolute Error (MAE) test       : 5.833
```





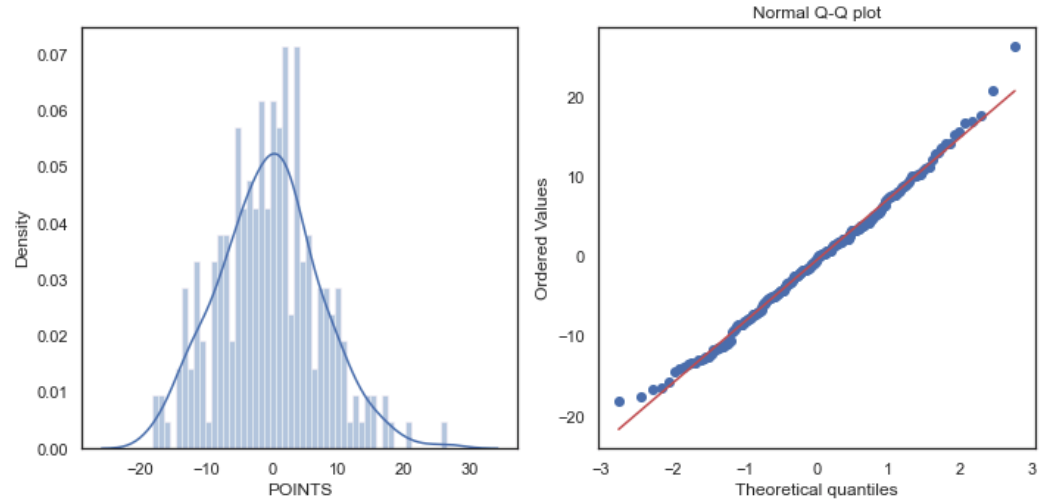
SIMPLE VALIDATION - INITIAL MODEL

```
Mean Absolute Error (MAE) (train/val): 5.992
Mean Absolute Error (MAE) test       : 6.093
```



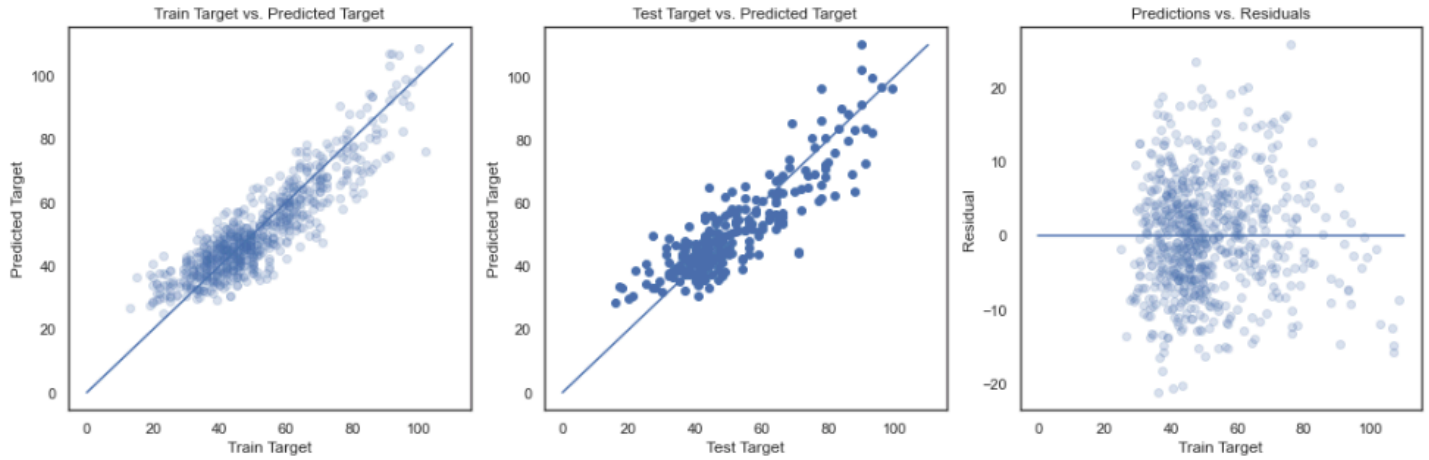
SIMPLE VALIDATION - FEATURE ENG. MODEL

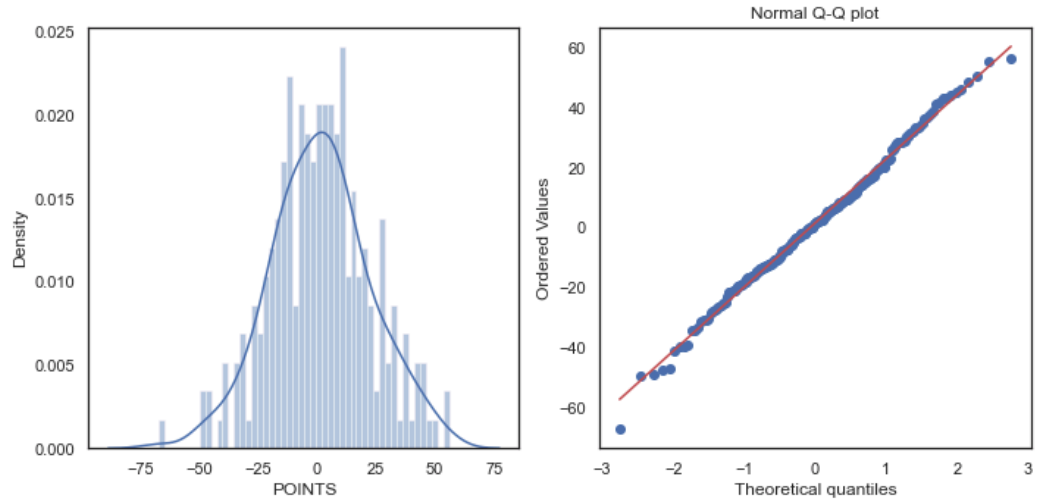
```
Mean Absolute Error (MAE) (train/val): 5.870
Mean Absolute Error (MAE) test       : 6.048
```



LASSO REGULARIZATION - INITIAL MODEL

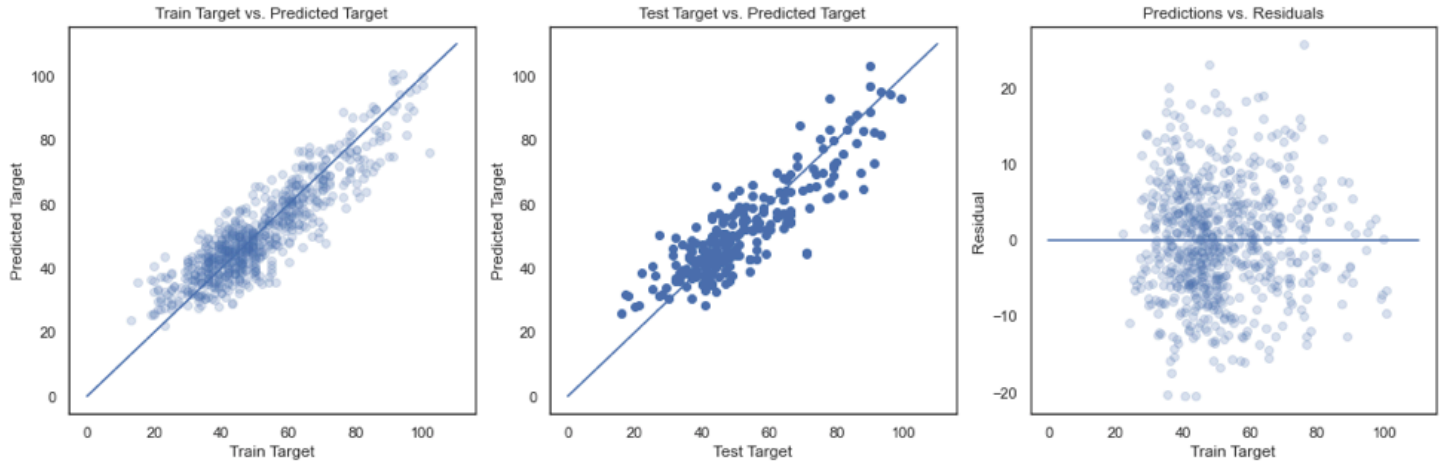
```
Mean Absolute Error (MAE) (train/val): 5.970
Mean Absolute Error (MAE) test         : 6.593
```

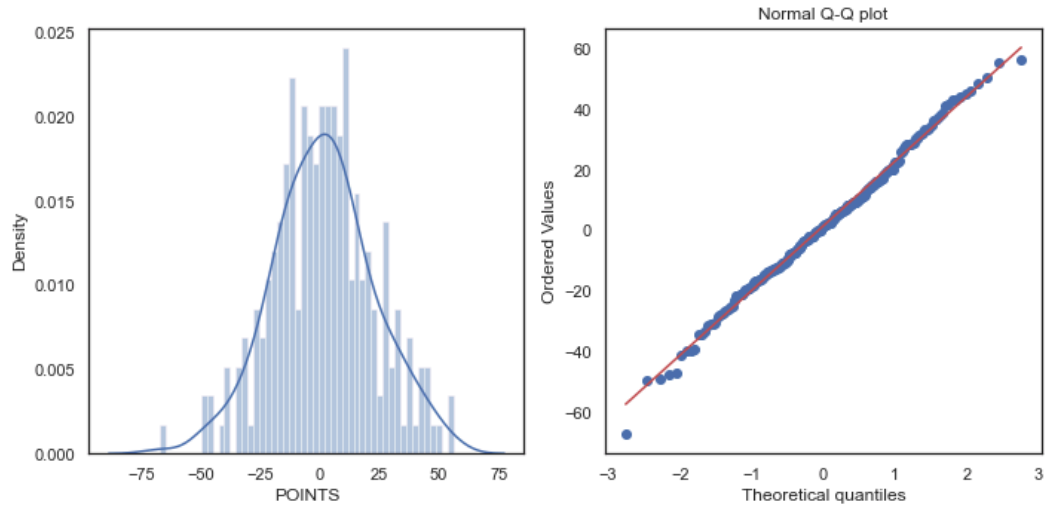




LASSO REGULARIZATION - FE MODEL

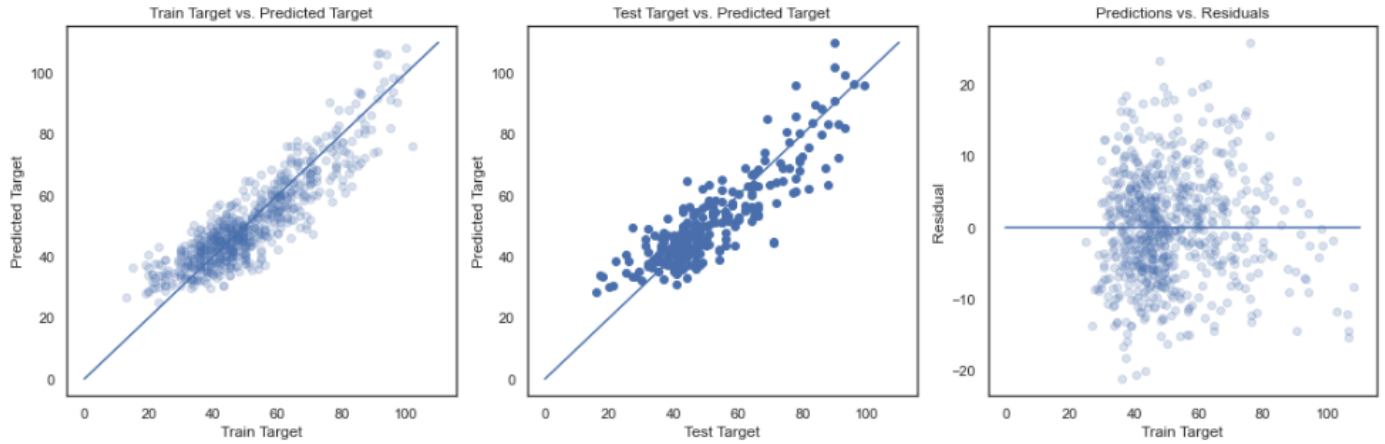
```
Mean Absolute Error (MAE) (train/val): 5.904
Mean Absolute Error (MAE) test         : 6.478
```

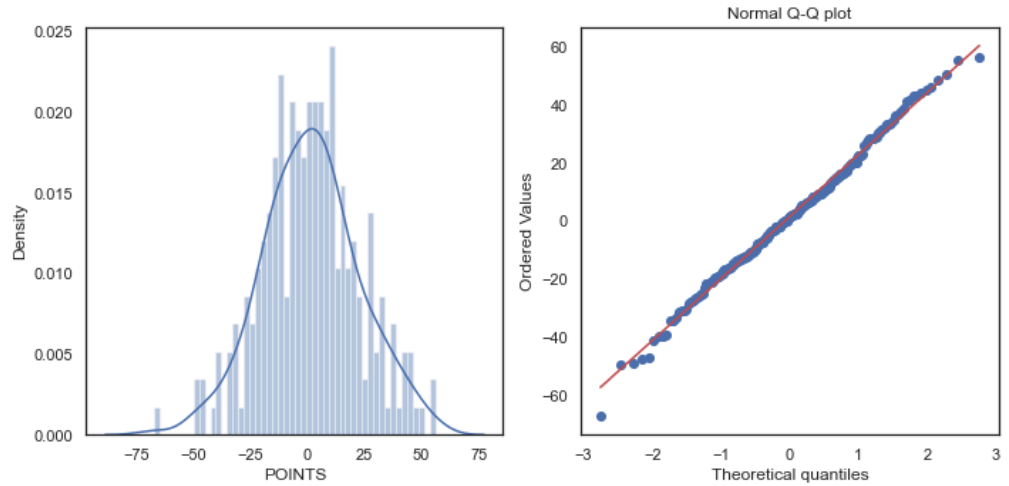




RIDGE REGULARIZATION - INITIAL MODEL

```
Mean Absolute Error (MAE) (train/val): 5.971
Mean Absolute Error (MAE) test         : 6.597
```





RIDGE REGULARIZATION - FE MODEL

```
Mean Absolute Error (MAE) (train/val): 5.904
Mean Absolute Error (MAE) test         : 6.473
```