FIFA 21 DATA ANALYSIS

TABLE OF CONTENTS

01

INTRODUCTION

PROBLEM STATEMENT

- Managers often find it **difficult to estimate** accurately a player’s wage.

- This could be due to factors such as incorrectly evaluating a certain attribute of a player

- Leading to over evaluation and thus giving a wrong wage.

GOAL

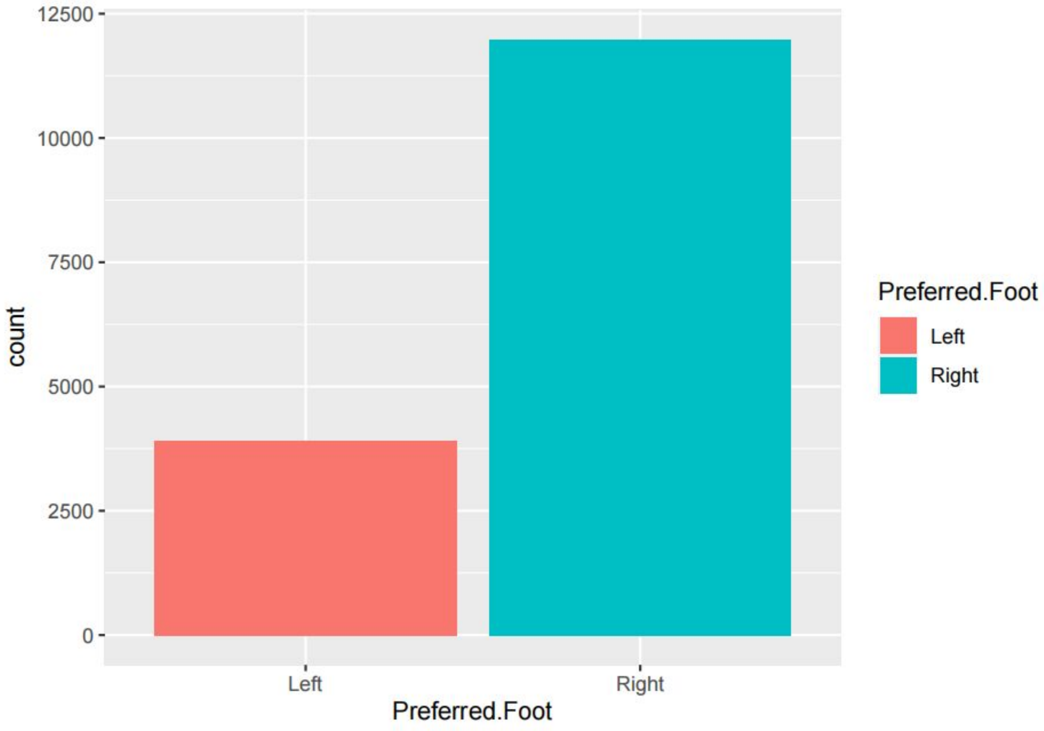
- To build a prediction model which can help to estimate the appropriate wage for a player based on their attributes and skill levels.

- Helping the managers create an ideal team composition within their budget constraints.

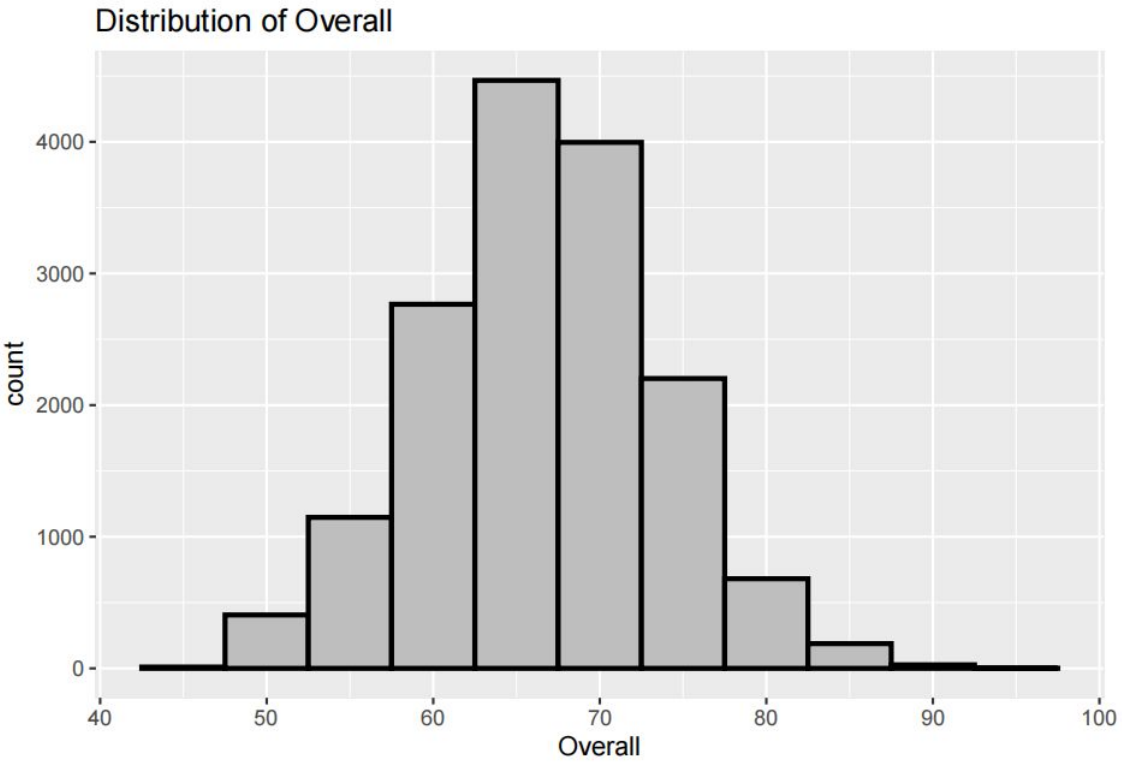
02

EXPLORATORY DATA ANALYSIS

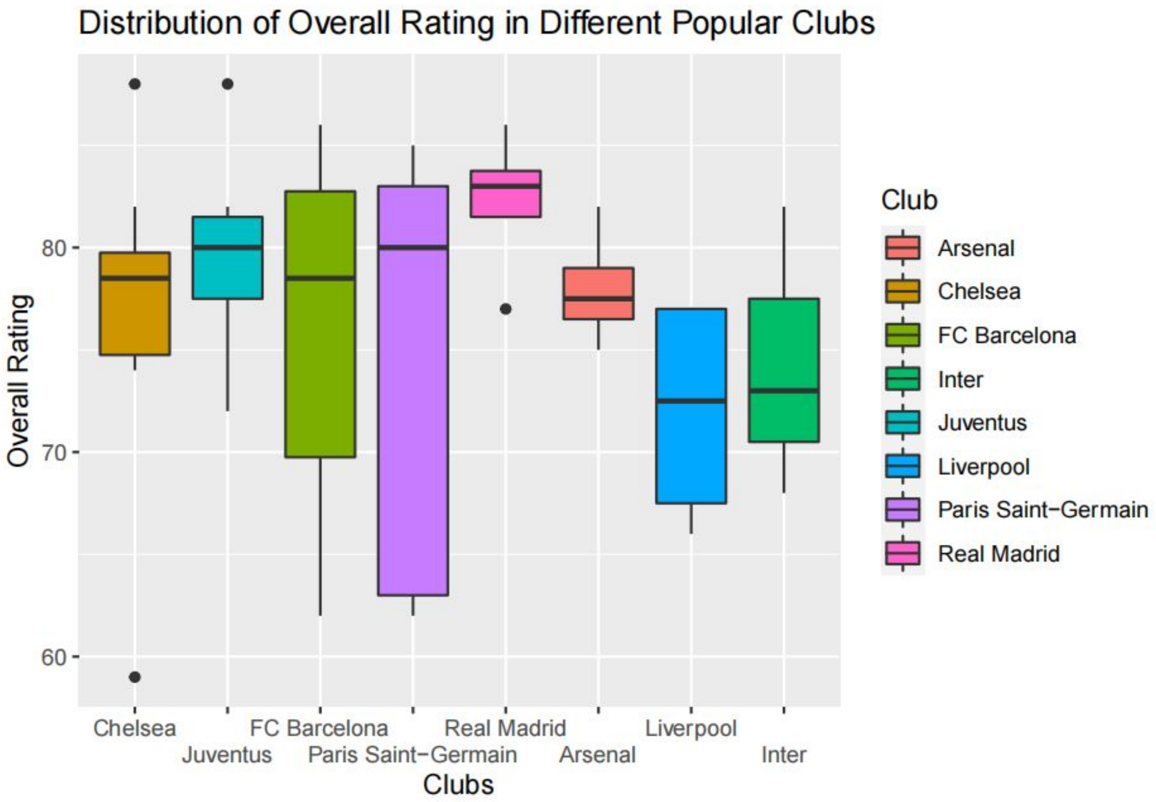
PREFERRED FOOT



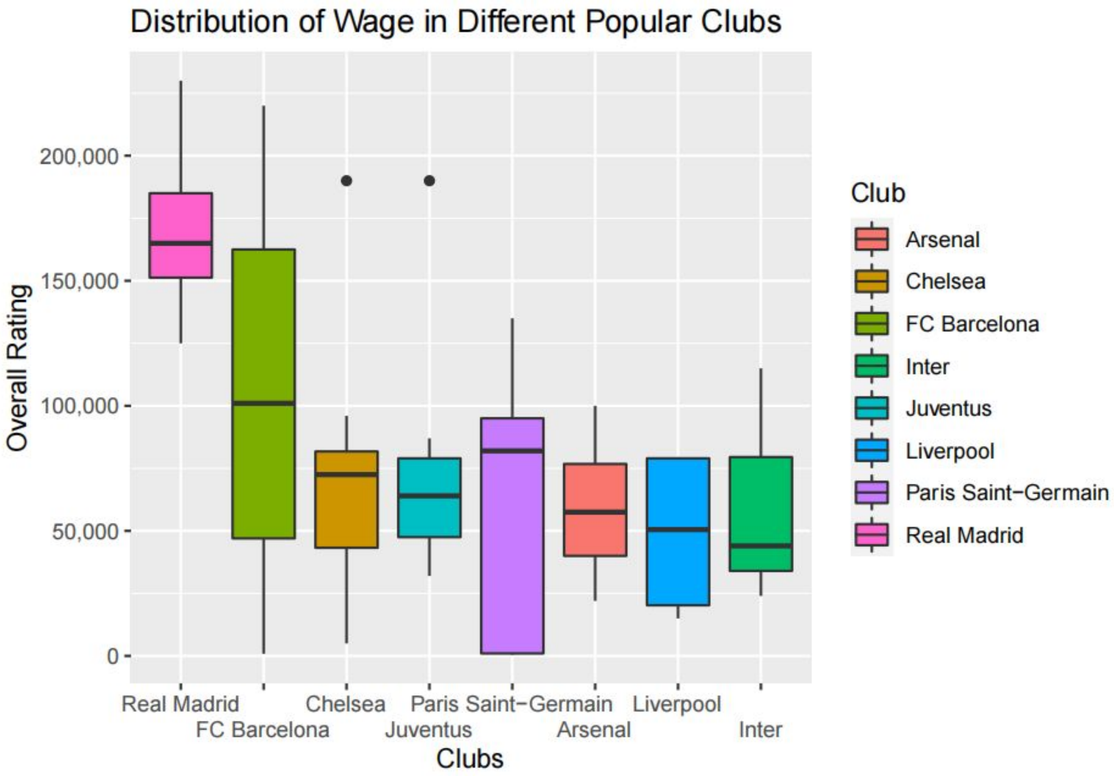
DISTRIBUTION OF OVERALL RATINGS



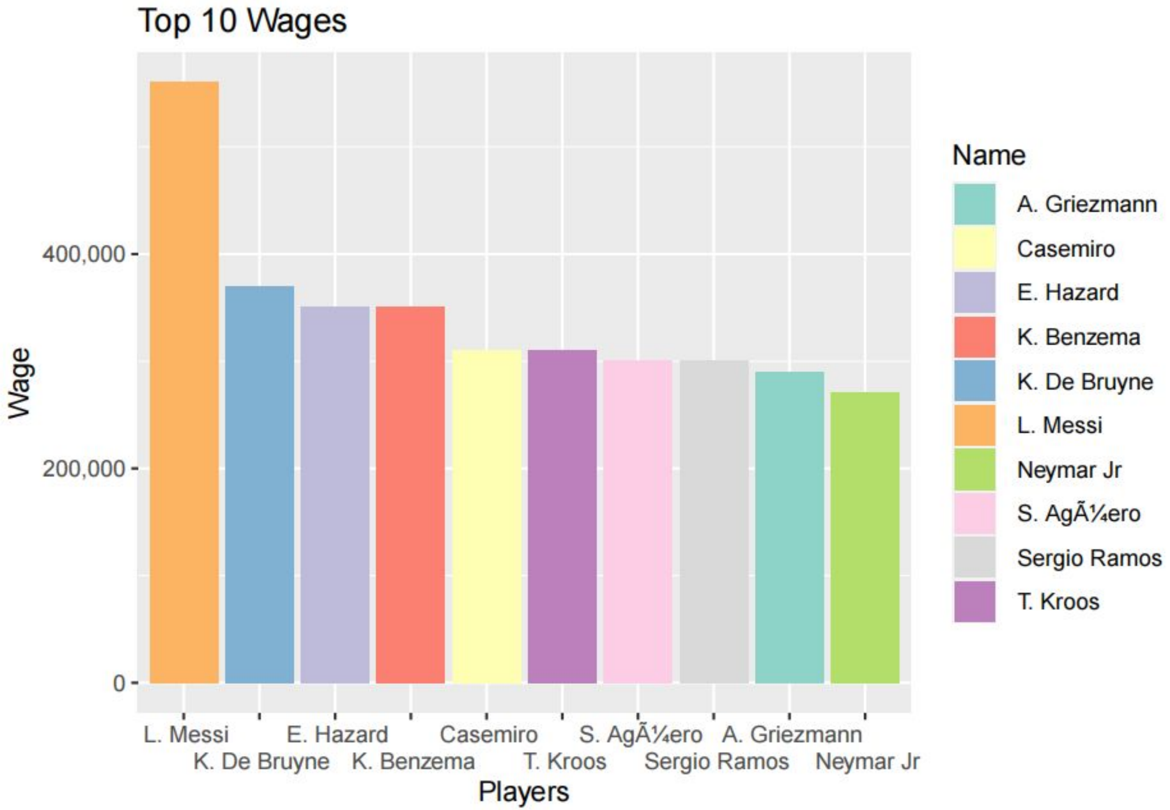
DISTRIBUTION OF OVERALL RATINGS IN POPULAR SOCCER TEAMS



DISTRIBUTION OF WAGE IN POPULAR SOCCER TEAMS

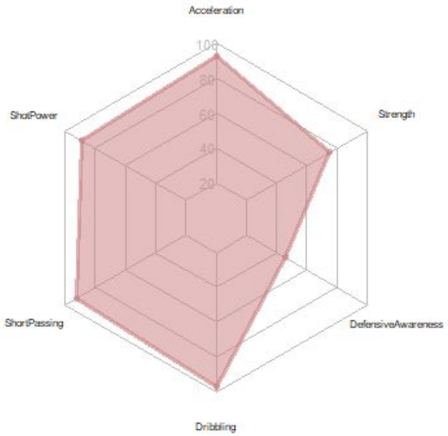


TOP 10 HIGHEST WAGES



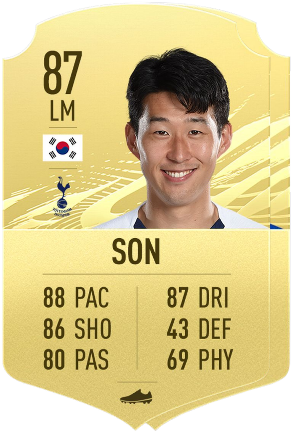
EDA: Comparison of Selected Players' Statistics

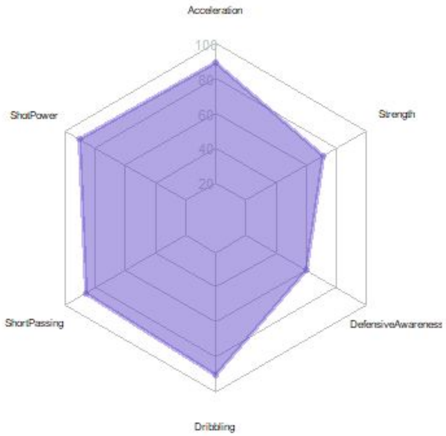
- Lionel Messi
- Wage: €560,000



EDA: Comparison of Selected Players' Statistics

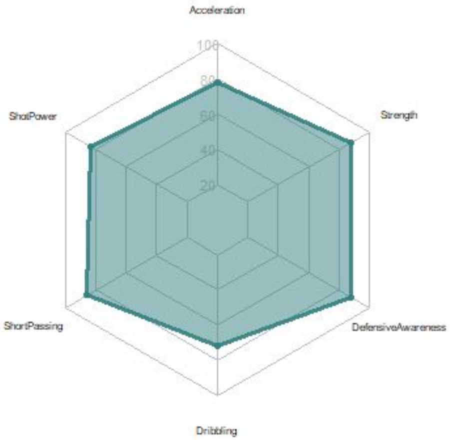
- Son Heung Min
- Wage: €165,000





EDA: Comparison of Selected Players' Statistics

- Sergio Ramos
- Wage: €300,000



03

MODEL SELECTION

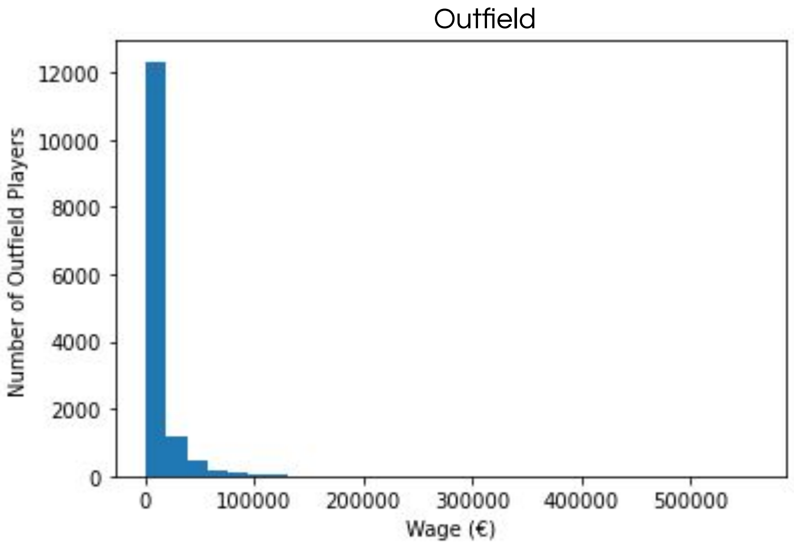
Dataset Used

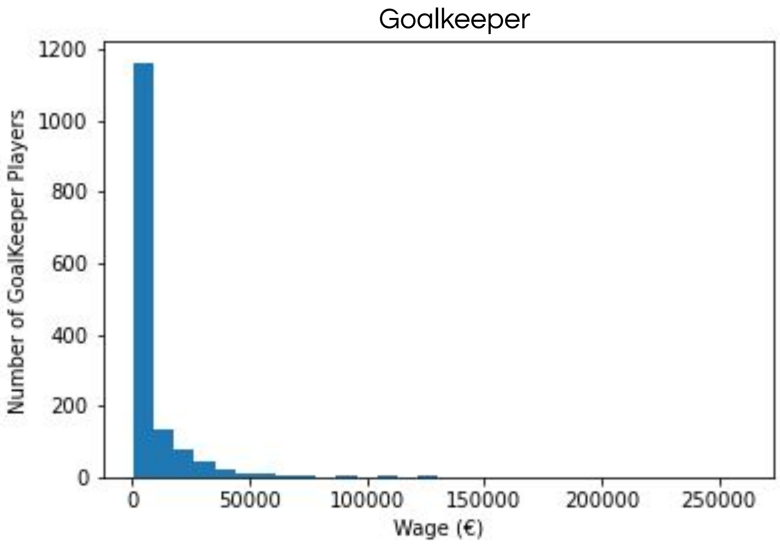
FIFA21 Dataset
Was split into 2
separate datasets

Outfielders
(OF)

Goalkeepers
(GK)

Heavy Tailed Wage





- Observed that observe that it is Right Skewed in both Datasets
- **Log** transformed **Wage** target for prediction which seems to improve the distribution slightly towards normal

Model Selection and Comparison

Selected Metric for Comparison:

Root Mean Square Log Error (**RMSLE**)
(which is *RMSE* but in log scale
since y variable (Wage) was log transformed)

Models with the **lowest RMSLE** will be chosen
as the best performing model respectively

Algorithms considered

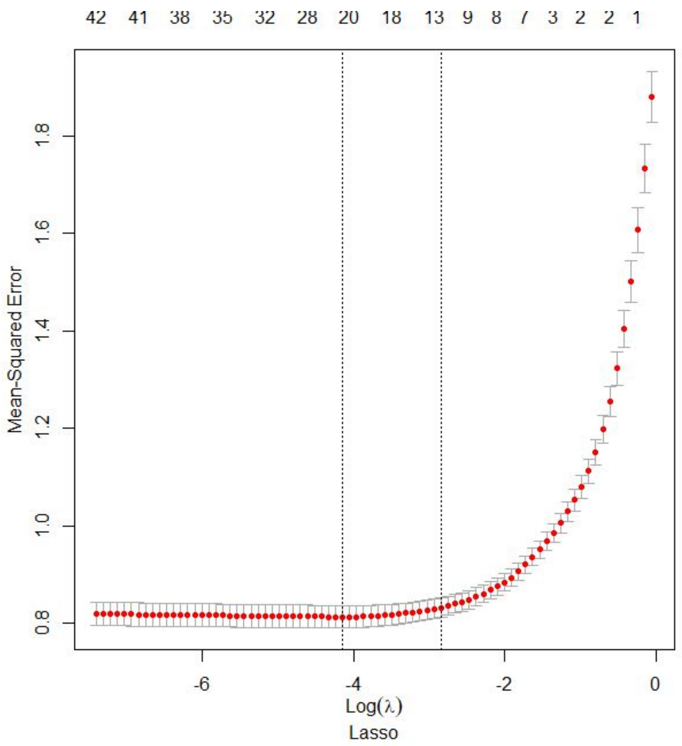
1. **Multiple Linear Regression** (Lasso, Scad)
2. **Random Forests**
3. **Support Vector Regression**

LASSO

- Lasso is able to yield simpler and more interpretable models involving only a subset of variables.
- We select the optimal tuning parameter for Lasso by performing 10-fold cross-validation (CV) and choosing the largest whose CV error is within 1 standard error of the minimum.
- This will allow us to yield a more parsimonious model compared to the minimum CV method.

LASSO

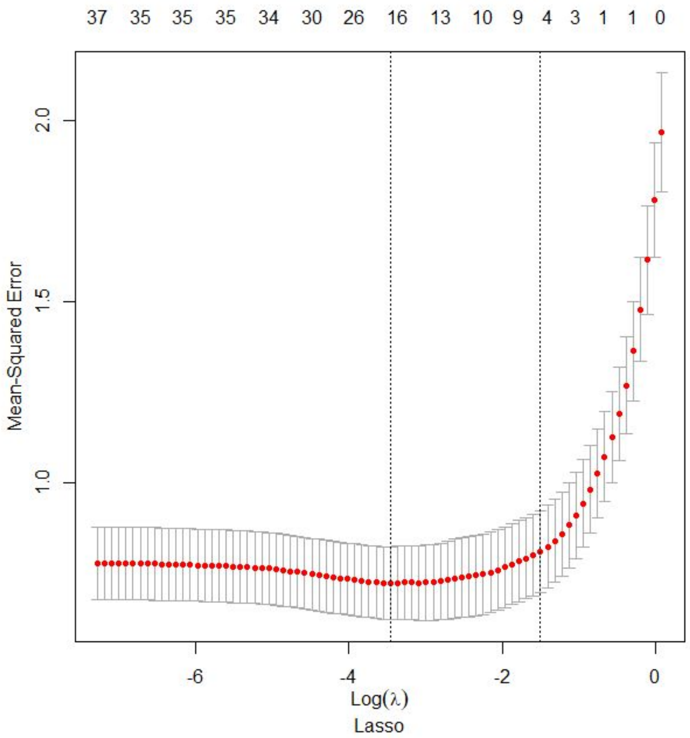
Outfield



- From the CV plot above, for outfield players, the 1 s.e. rule selects a more parsimonious model of size 13 as compared to the minimum CV error which selects a model of size 21

LASSO

Goalkeepers



- From the CV plot above, for goalkeepers, the 1 s.e. Rule selects a more parsimonious model of size 5 as compared to the minimum CV error which selects a model of size 18

LASSO

Outfield

	Test Error
RMSLE	0.907
Median Abs Error	€1709.44

Top 3 variables for outfield:

	Feature	Importance
1	International.Reputation	0.216
2	Reactions	0.0436
3	Potential	0.0383

Goalkeepers

	Test Error
RMSLE	0.885
Median Abs Error	€1320.56

Top 3 variables for goalkeepers:

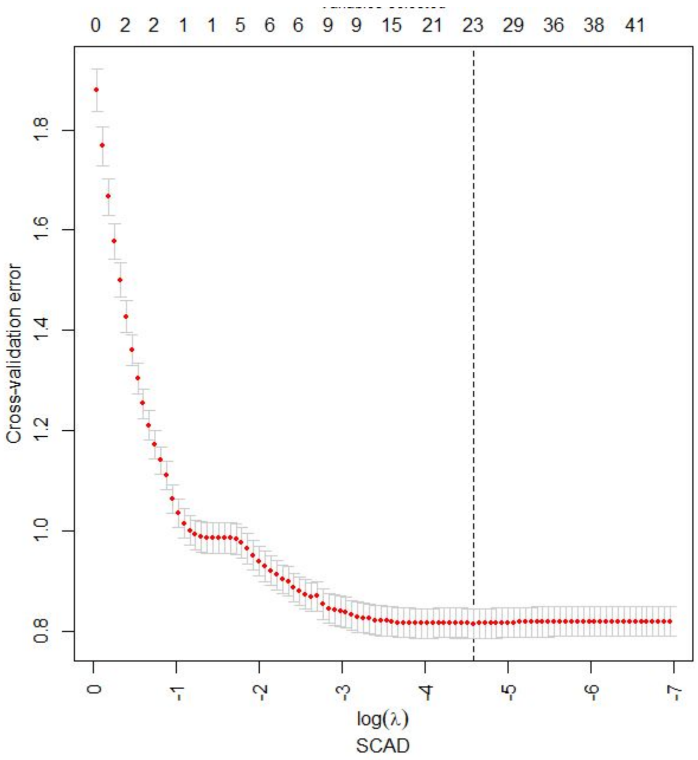
	Feature	Importance
1	GKReflexes	0.0776
2	Composure	0.0118
3	Reactions	0.0104

SCAD

- As Lasso shrinkage causes estimates of non-zero coefficients to be biased towards zero, we try SCAD next to reduce the bias from Lasso.
- Similar to Lasso, CV was used to estimate the penalty parameters.

SCAD

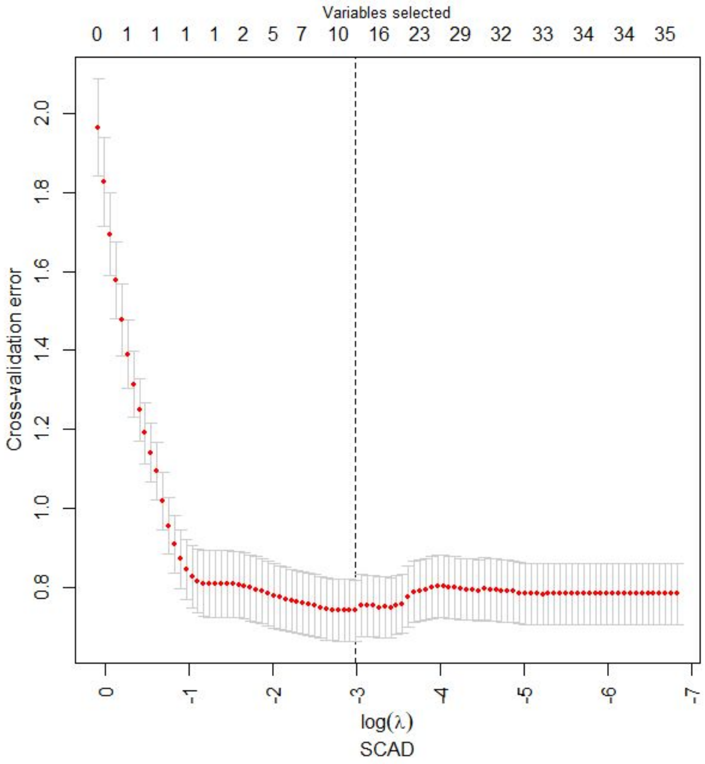
Outfield



- For outfield player, SCAD selects 9 variables

SCAD

Goalkeeper



- For goalkeepers, SCAD selects only 1 variable

SCAD

Outfield

	Test Error
RMSLE	0.908
Median Abs Error	€1661.59

Goalkeepers

	Test Error
RMSLE	0.898
Median Abs Error	€1158.26

Top 3 variables for outfield:

	Feature	Importance
1	Age	0.0768
2	Potential	0.0750
3	International.Reputation	0.0598

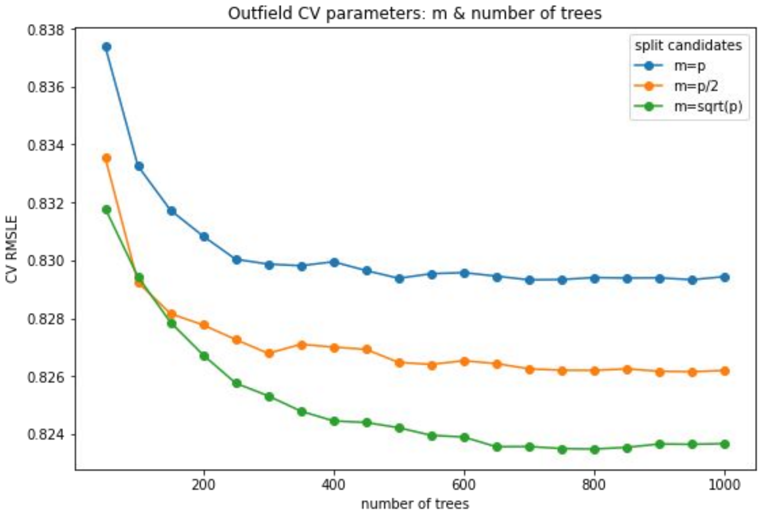
Top 3 variables for goalkeepers:

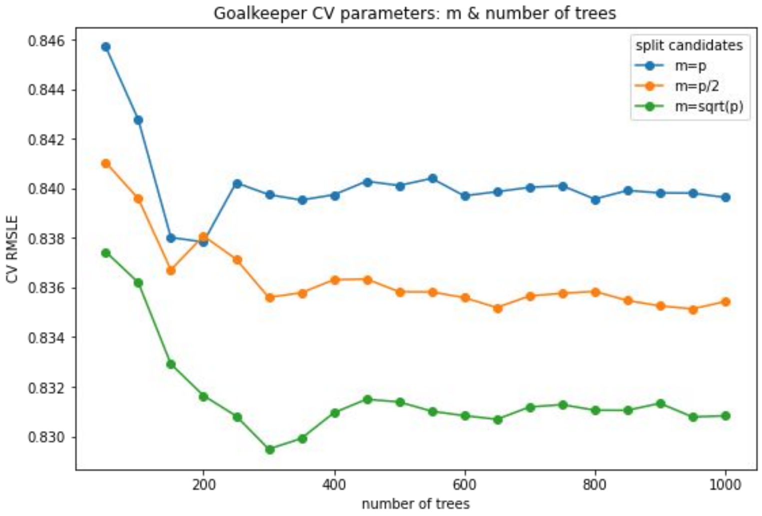
	Feature	Importance
1	GKReflexes	0.0776

Random Forest Regression

To **reduce high variance** of a single regression tree, bootstrap samples are used for to build multiple trees taking average

To **reduce highly correlated trees,** only random subset `m` of all predictors are chosen as split candidate at each tree split.





Random Forest Regression

Outfield

	Test Error
RMSLE	0.822
Median Abs Error	€1372.67

Top 3 variables for outfield:

	Feature	Importance
1	Reactions	0.224216
2	Composure	0.106332
3	ShortPassing	0.083784

Goalkeepers

	Test Error
RMSLE	0.83
Median Abs Error	€1071.03

Top 3 variables for goalkeepers:

	Feature	Importance
1	GKReflexes	0.229475
2	GKHandling	0.156626
3	Reactions	0.112376

Support Vector Regression (Supervised Learning)

Acknowledge presence of non-linearity in the data
and then provides a proficient prediction model.

Hyperparameters	What are they?
Hyperplane	Decision boundaries used to predict the continuous output.
Kernel	Mathematical functions to transform data into the required form in the higher dimensional space Linear, polynomial, radial basis function (RBF)
Boundary Lines	Epsilon (ε)

Parameters to tune

Kernel	Parameters to consider/tune
Linear	C (regularization parameter) - Strictly positive, Ɛ
Polynomial	C, Gamma, degree of polynomial, Ɛ
Radial basis function (RBF)	C, gamma, Ɛ

Choosing the best kernel for GK

Kernel	RMSLE
Linear	**1.780**
Polynomial	**1.655**
RBF	**0.845**

Choosing the best kernel for OF

Kernel	RMSLE
Linear	**1.781**
Polynomial	**1.773**
RBF	$\boxed{\textbf{1.372}}$

Support Vector Regression

Outfield

	Test Error
RMSLE	1.372
Median Abs Error	€3051

Avg log Wage: 8.28

Median Wage: €3000

Goalkeepers

	Test Error
RMSLE	0.845
Median Abs Error	€1280

Avg log Wage: 7.99

Median Wage: € 3000

Test Evaluation: Outfield

Summary

Model	RMSLE
LASSO	0.907
SCAD	0.908
Random Forest (RF)	0.822
SVM	0.871

Best Model: Random Forest

RMSLE: 0.822

	Feature	Importance
1	Reactions	0.224216
2	Composure	0.106332
3	ShortPassing	0.083784

Test Evaluation: Goalkeeper

Summary

Model	RMSLE
LASSO	0.885
SCAD	0.898
Random Forest (RF)	0.830
SVM	0.845

Best Model: Random Forest

RMSLE: 0.83

	Feature	Importance
1	GKReflexes	0.229
2	GKHandling	0.157
3	Reactions	0.112

04

RECOMMENDATIONS

Recommendations

Using the top 3 feature importances from goalkeeper and outfield, managers may consider the following metrics to justify a player's wage:

<table>
<tr><th>Feature</th><th>Estimated by</th><th>Player Type</th></tr>
<tr><td>Composure</td><td>Player Turnover Rate</td><td>Outfield</td></tr>
<tr><td>Short Passing</td><td>Player miss-pass Rate</td><td>Outfield</td></tr>
<tr><td>Reactions</td><td rowspan="2">Reaction Tests</td><td rowspan="2">Outfield and Goalkeeper</td></tr>
<tr><td>GK Reflexes</td></tr>
<tr><td>GKHandling</td><td>Percentage of successful catches out of attempted catches</td><td>Goalkeeper</td></tr>
</table>

Thank You!