Supervised Learning - FIFA 23

DTSA 5509
Final Project

Project Agenda

1. Project Introduction
2. Data Cleaning
3. Exploratory Data Analysis
4. Regression Modeling
5. Conclusion

1. Project Introduction

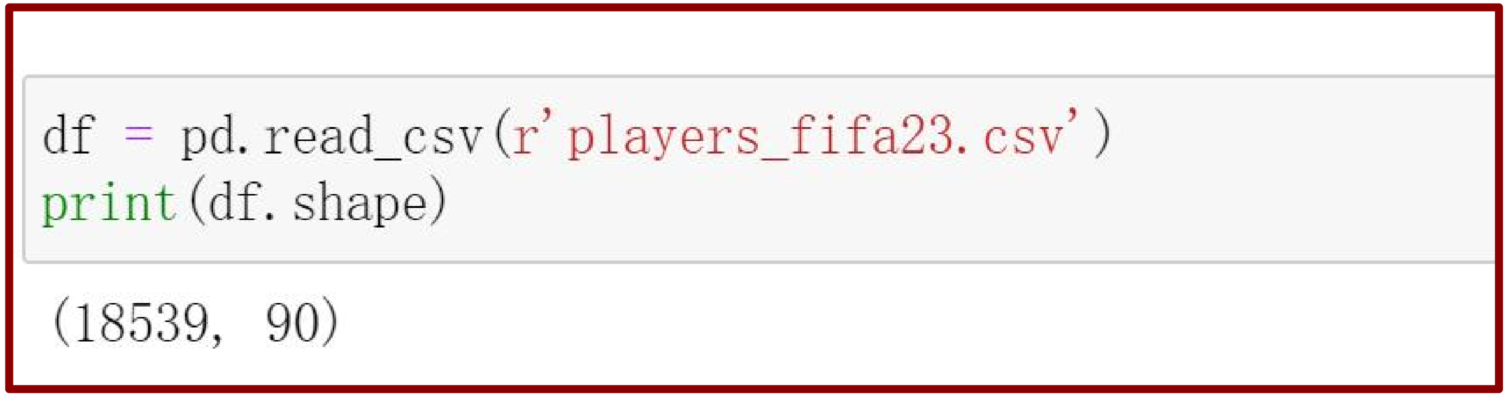
Dataset

- □ FIFA23 1st Edition Player Database
- □ From Kaggle Website
- □ Players' Information and Attributes

Goal

Predict Players' Market Value
in the Game by
Supervised Learning Methods

2. Data Cleaning



```
df = pd.read_csv(r'players_fifa23.csv')
print(df.shape)
```

```
(18539, 90)
```

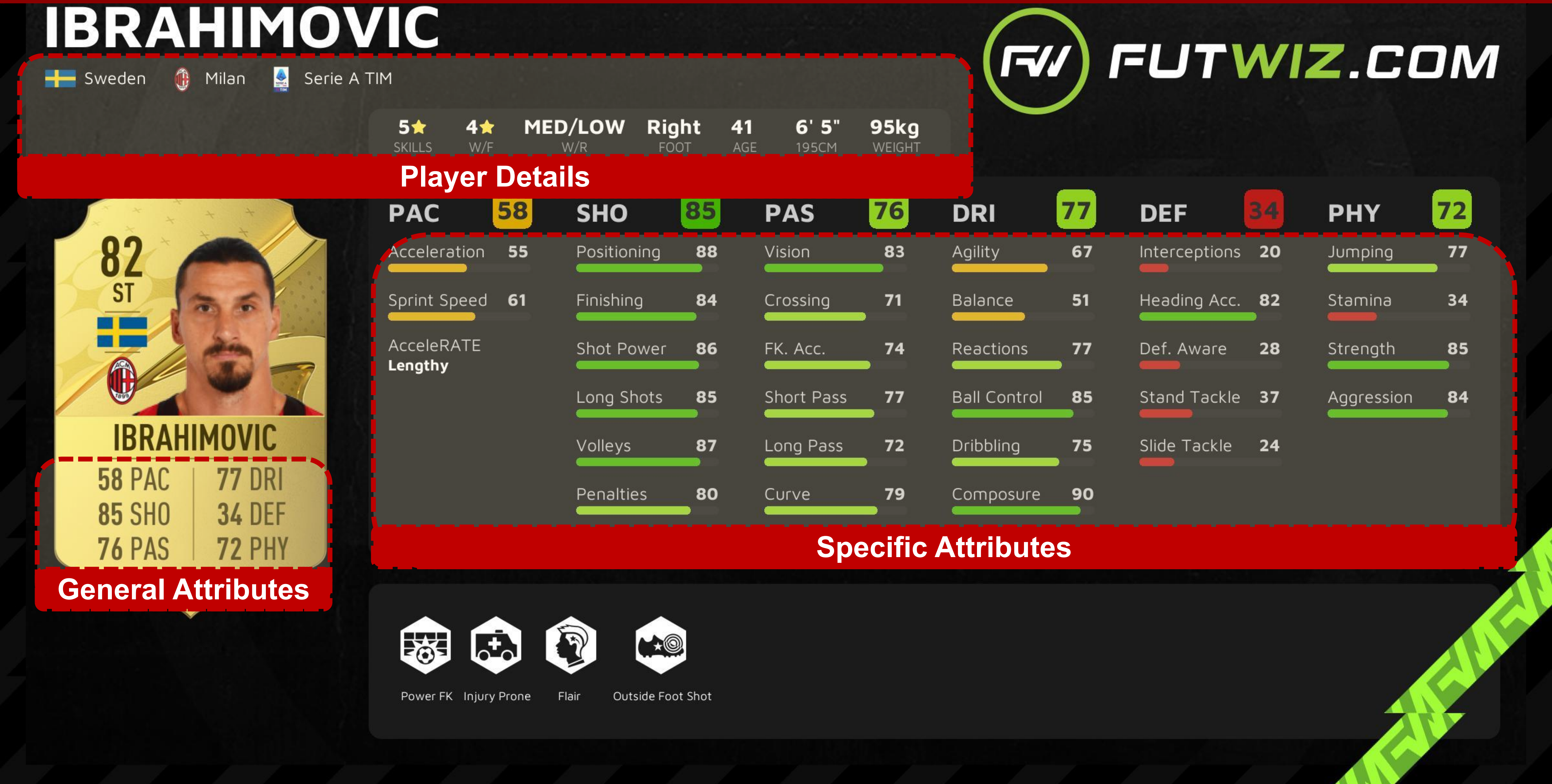
Players

➢ **18539**

Attributes

➢ **90**

2.1 Feature Description



IBRAHIMOVIC

FUTWIZ.COM

Sweden Milan Serie A TIM

5★	4★	MED/LOW	Right	41	6' 5"	95kg
SKILLS	W/F	W/R	FOOT	AGE	195CM	WEIGHT

Player Details

82
ST
IBRAHIMOVIC

58 PAC	77 DRI
85 SHO	34 DEF
76 PAS	72 PHY

General Attributes

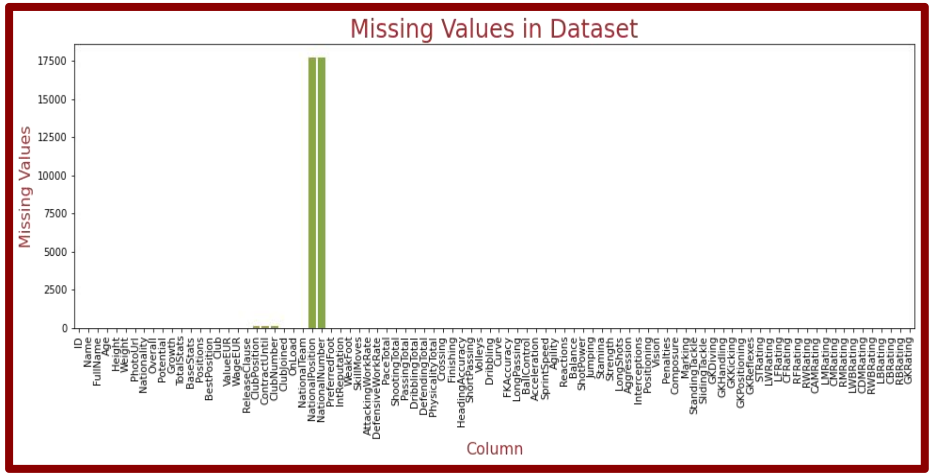
PAC	58	SHO	85	PAS	76	DRI	77	DEF	34	PHY	72
Acceleration	55	Positioning	88	Vision	83	Agility	67	Interceptions	20	Jumping	77
Sprint Speed	61	Finishing	84	Crossing	71	Balance	51	Heading Acc.	82	Stamina	34
AcceleRATE **Lengthy**		Shot Power	86	FK. Acc.	74	Reactions	77	Def. Aware	28	Strength	85
		Long Shots	85	Short Pass	77	Ball Control	85	Stand Tackle	37	Aggression	84
		Volleys	87	Long Pass	72	Dribbling	75	Slide Tackle	24		
		Penalties	80	Curve	79	Composure	90				

Specific Attributes

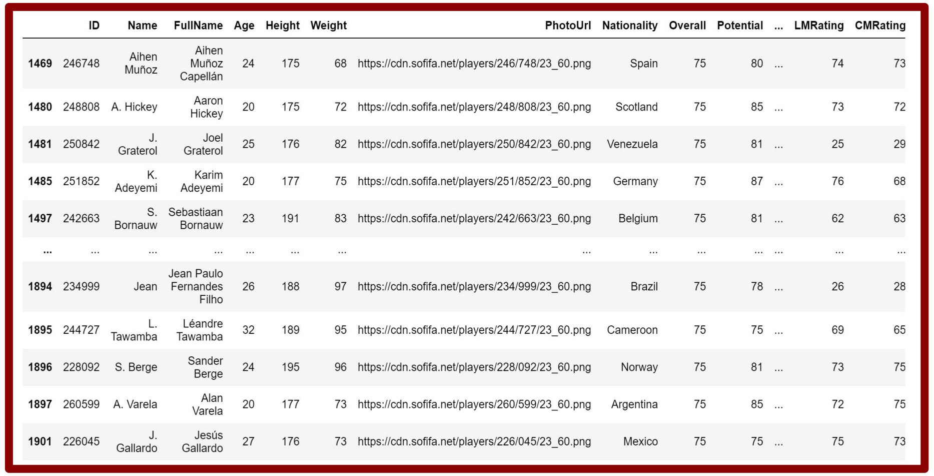
Power FK Injury Prone Flair Outside Foot Shot

2.2 Cleaning Steps

Refill Missing Values

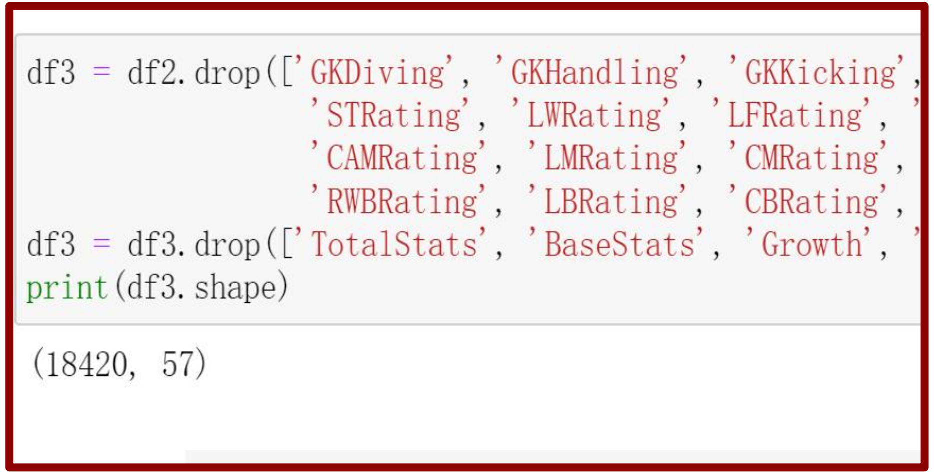


Delete Duplicates



	ID	Name	FullName	Age	Height	Weight	PhotoUrl	Nationality	Overall	Potential	...	LMRating	CMRating
1469	246748	Aihen Muñoz	Aihen Muñoz Capellán	24	175	68	https://cdn.sofifa.net/players/246/748/23_60.png	Spain	75	80	...	74	73
1480	248808	A. Hickey	Aaron Hickey	20	175	72	https://cdn.sofifa.net/players/248/808/23_60.png	Scotland	75	85	...	73	72
1481	250842	J. Graterol	Joel Graterol	25	176	82	https://cdn.sofifa.net/players/250/842/23_60.png	Venezuela	75	81	...	25	29
1485	251852	K. Adeyemi	Karim Adeyemi	20	177	75	https://cdn.sofifa.net/players/251/852/23_60.png	Germany	75	87	...	76	68
1497	242663	S. Bornauw	Sebastiaan Bornauw	23	191	83	https://cdn.sofifa.net/players/242/663/23_60.png	Belgium	75	81	...	62	63
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1894	234999	Jean	Jean Paulo Fernandes Filho	26	188	97	https://cdn.sofifa.net/players/234/999/23_60.png	Brazil	75	78	...	26	28
1895	244727	L. Tawamba	Léandre Tawamba	32	189	95	https://cdn.sofifa.net/players/244/727/23_60.png	Cameroon	75	75	...	69	65
1896	228092	S. Berge	Sander Berge	24	195	96	https://cdn.sofifa.net/players/228/092/23_60.png	Norway	75	81	...	73	75
1897	260599	A. Varela	Alan Varela	20	177	73	https://cdn.sofifa.net/players/260/599/23_60.png	Argentina	75	85	...	72	75
1901	226045	J. Gallardo	Jesús Gallardo	27	176	73	https://cdn.sofifa.net/players/226/045/23_60.png	Mexico	75	75	...	75	73

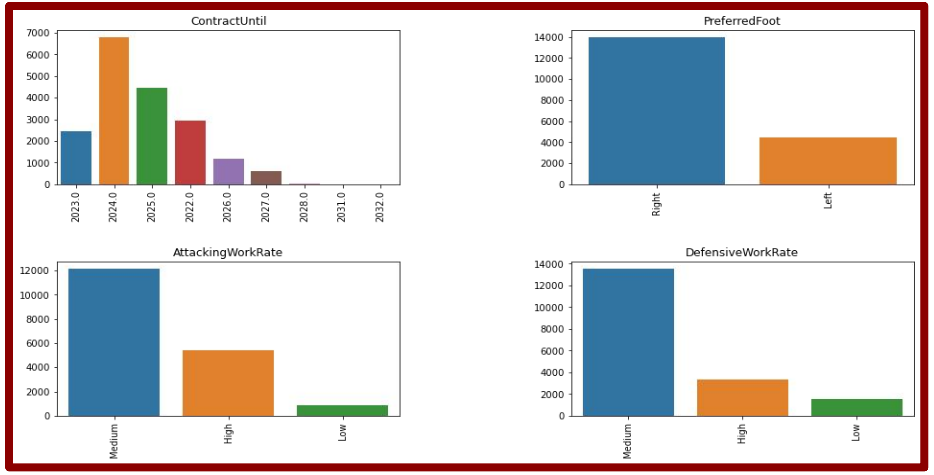
Remove Unrelevant



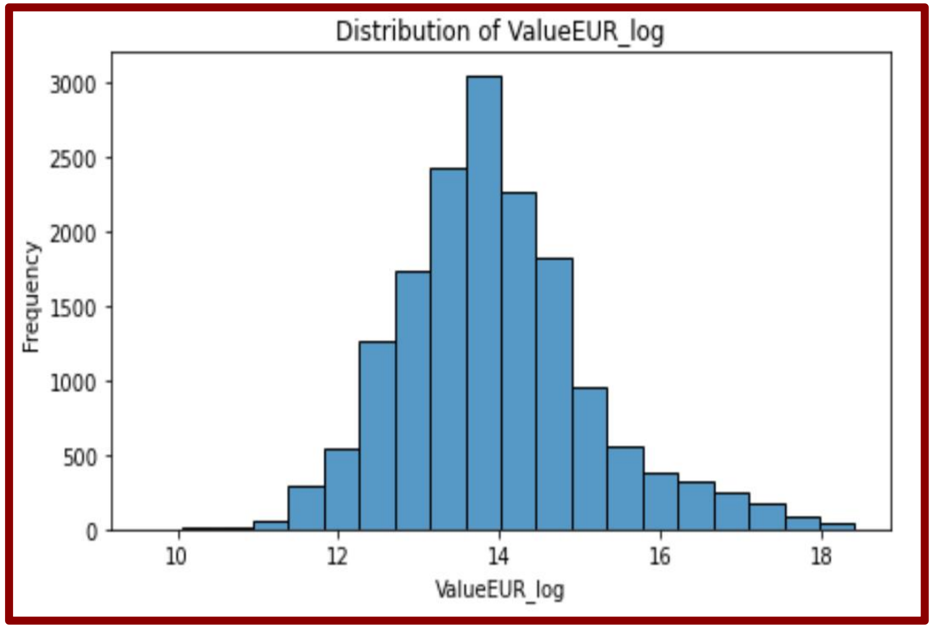
```
df3 = df2.drop(['GKDiving', 'GKHandling', 'GKKicking',
               'STRating', 'LWRating', 'LFRating', '
               'CAMRating', 'LMRating', 'CMRating',
               'RWBRating', 'LBRating', 'CBRating',
df3 = df3.drop(['TotalStats', 'BaseStats', 'Growth', '
print(df3.shape)
```

```
(18420, 57)
```

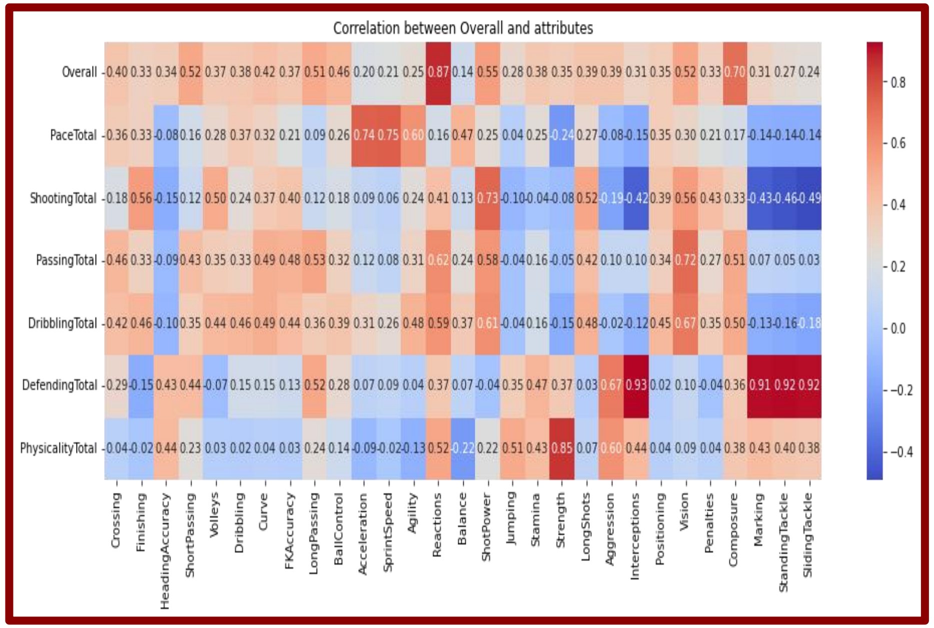
Create Dummy Variable



3. Exploratory Data Analysis



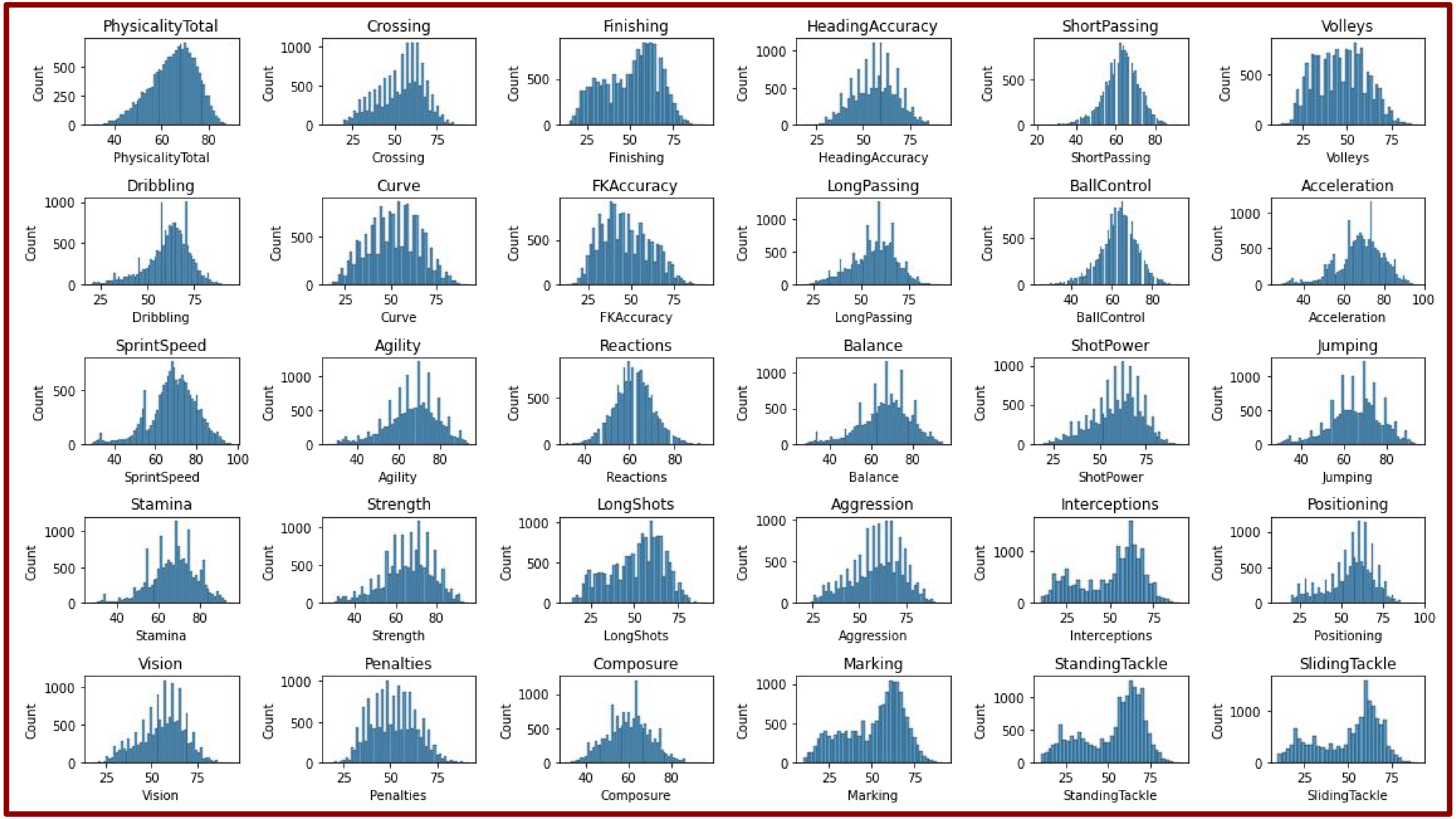
Distribution



	Crossing	Finishing	HeadingAccuracy	ShortPassing	Volleys	Dribbling	Curve	FKAccuracy	LongPassing	BallControl	Acceleration	SprintSpeed	Agility	Reactions	Balance	ShotPower	Jumping	Stamina	Strength	LongShots	Aggression	Interceptions	Positioning	Vision	Penalties	Composure	Marking	StandingTackle	SlidingTackle
Overall	0.40	0.33	0.34	0.52	0.37	0.38	0.42	0.37	0.51	0.46	0.20	0.21	0.25	0.87	0.14	0.55	0.28	0.38	0.35	0.39	0.39	0.31	0.35	0.52	0.33	0.70	0.31	0.27	0.24
PaceTotal	0.36	0.33	-0.08	0.16	0.28	0.37	0.32	0.21	0.09	0.26	0.74	0.75	0.60	0.16	0.47	0.25	0.04	0.25	-0.24	0.27	-0.08	-0.15	0.35	0.30	0.21	0.17	-0.14	-0.14	-0.14
ShootingTotal	0.18	0.56	-0.15	0.12	0.50	0.24	0.37	0.40	0.12	0.18	0.09	0.06	0.24	0.41	0.13	0.73	-0.10	-0.04	-0.08	0.52	-0.19	-0.42	0.39	0.56	0.43	0.33	-0.43	-0.46	-0.49
PassingTotal	0.46	0.33	-0.09	0.43	0.35	0.33	0.49	0.48	0.53	0.32	0.12	0.08	0.31	0.62	0.24	0.58	-0.04	0.16	-0.05	0.42	0.10	0.10	0.34	0.72	0.27	0.51	0.07	0.05	0.03
DribblingTotal	0.42	0.46	-0.10	0.35	0.44	0.46	0.49	0.44	0.36	0.39	0.31	0.26	0.48	0.59	0.37	0.61	-0.04	0.16	-0.15	0.48	-0.02	-0.12	0.45	0.67	0.35	0.50	-0.13	-0.16	-0.18
DefendingTotal	0.29	-0.15	0.43	0.44	-0.07	0.15	0.15	0.13	0.52	0.28	0.07	0.09	0.04	0.37	0.07	-0.04	0.35	0.47	0.37	0.03	0.67	0.93	0.02	0.10	-0.04	0.36	0.91	0.92	0.92
PhysicalityTotal	0.04	-0.02	0.44	0.23	0.03	0.02	0.04	0.03	0.24	0.14	-0.09	-0.02	-0.13	0.52	-0.22	0.22	0.51	0.43	0.85	0.07	0.60	0.44	0.04	0.09	0.04	0.38	0.43	0.40	0.38

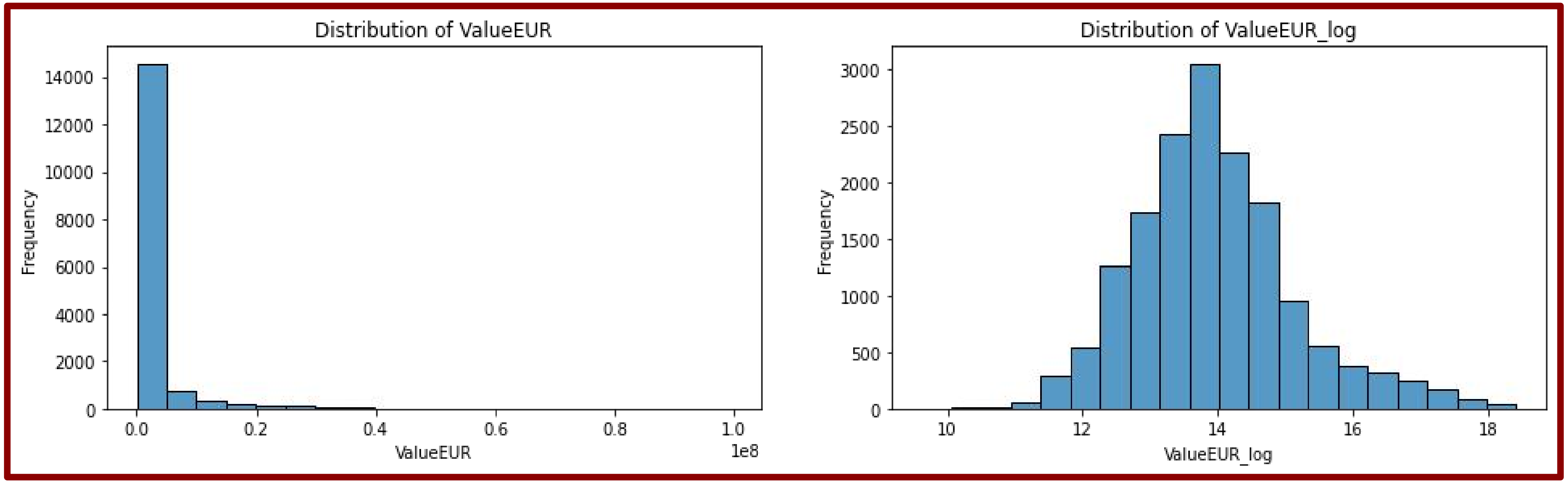
Correlation

3.1 Checking Distribution



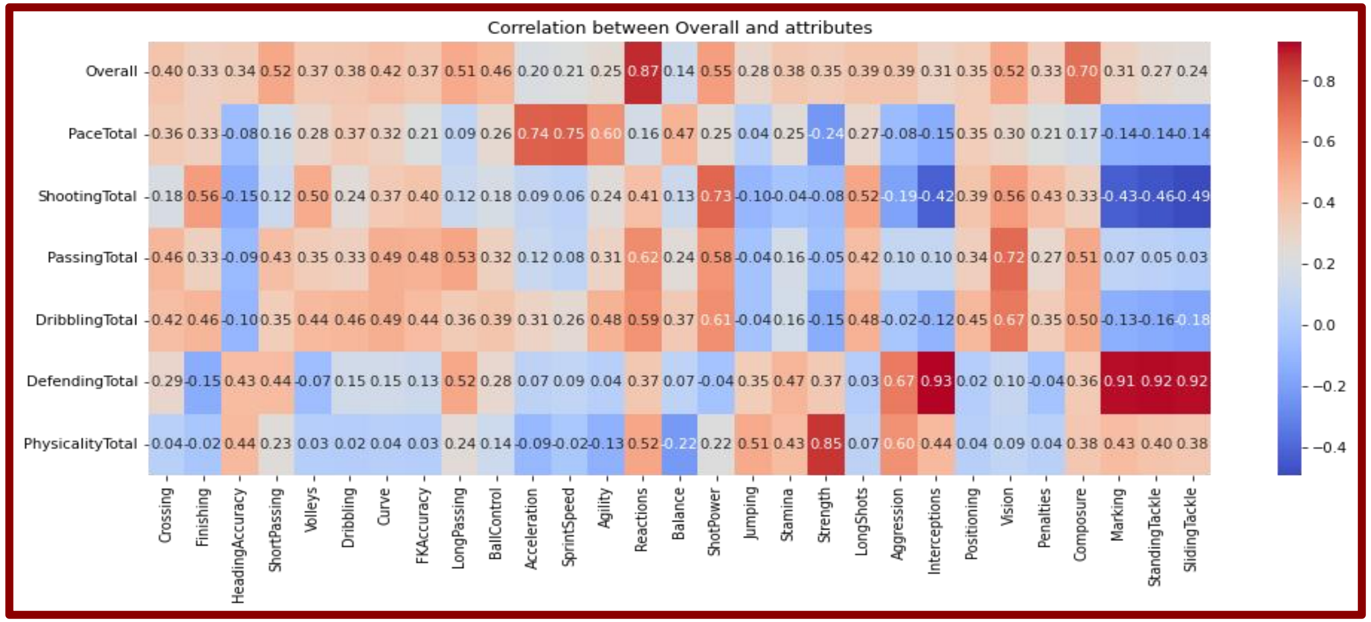
3.1 Checking Distribution

- The predicted values are heavily skewed to the left
- Most players show very low Market Values
- Close to normal distribution after log transformation



3.2 Checking Correlation

- **General Rating is caculated by Specific Values**
- **We will remove the category ratings to avoid multicollinearity.**



Correlation between Overall and attributes

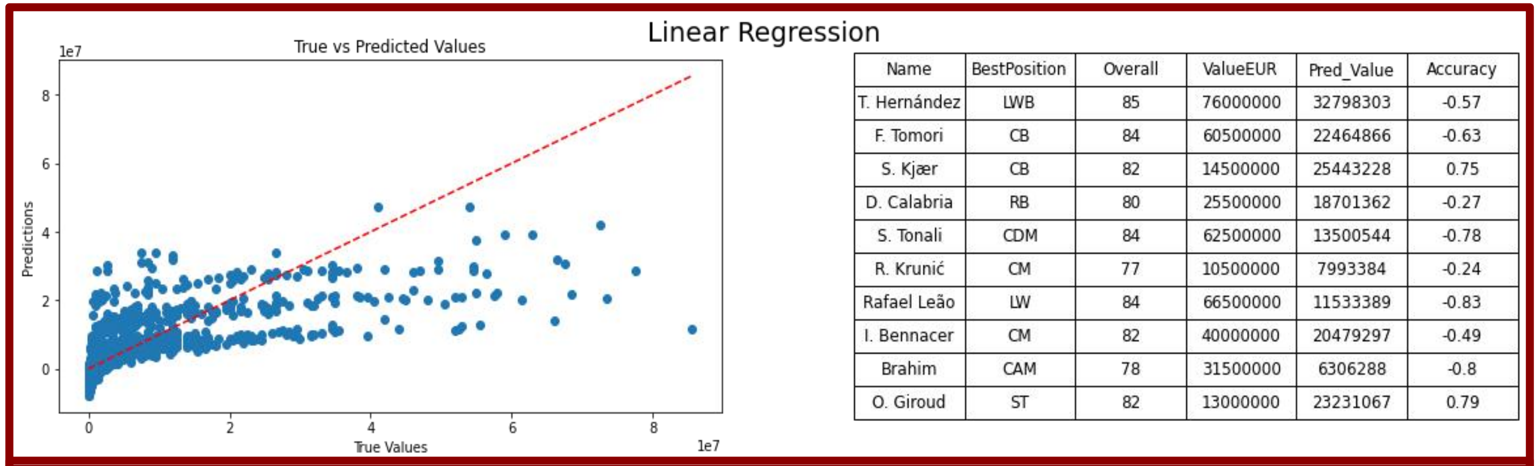
	Crossing	Finishing	HeadingAccuracy	ShortPassing	Volleys	Dribbling	Curve	FKAccuracy	LongPassing	BallControl	Acceleration	SprintSpeed	Agility	Reactions	Balance	ShotPower	Jumping	Stamina	Strength	LongShots	Aggression	Interceptions	Positioning	Vision	Penalties	Composure	Marking	StandingTackle	SlidingTackle
Overall	0.40	0.33	0.34	0.52	0.37	0.38	0.42	0.37	0.51	0.46	0.20	0.21	0.25	0.87	0.14	0.55	0.28	0.38	0.35	0.39	0.39	0.31	0.35	0.52	0.33	0.70	0.31	0.27	0.24
PaceTotal	0.36	0.33	-0.08	0.16	0.28	0.37	0.32	0.21	0.09	0.26	0.74	0.75	0.60	0.16	0.47	0.25	0.04	0.25	-0.24	0.27	-0.08	-0.15	0.35	0.30	0.21	0.17	-0.14	-0.14	-0.14
ShootingTotal	0.18	0.56	-0.15	0.12	0.50	0.24	0.37	0.40	0.12	0.18	0.09	0.06	0.24	0.41	0.13	0.73	-0.10	-0.04	-0.08	0.52	-0.19	-0.42	0.39	0.56	0.43	0.33	-0.43	-0.46	-0.49
PassingTotal	0.46	0.33	-0.09	0.43	0.35	0.33	0.49	0.48	0.53	0.32	0.12	0.08	0.31	0.62	0.24	0.58	-0.04	0.16	-0.05	0.42	0.10	0.10	0.34	0.72	0.27	0.51	0.07	0.05	0.03
DribblingTotal	0.42	0.46	-0.10	0.35	0.44	0.46	0.49	0.44	0.36	0.39	0.31	0.26	0.48	0.59	0.37	0.61	-0.04	0.16	-0.15	0.48	-0.02	-0.12	0.45	0.67	0.35	0.50	-0.13	-0.16	-0.18
DefendingTotal	0.29	-0.15	0.43	0.44	-0.07	0.15	0.15	0.13	0.52	0.28	0.07	0.09	0.04	0.37	0.07	-0.04	0.35	0.47	0.37	0.03	0.67	0.93	0.02	0.10	-0.04	0.36	0.91	0.92	0.92
PhysicalityTotal	0.04	-0.02	0.44	0.23	0.03	0.02	0.04	0.03	0.24	0.14	-0.09	-0.02	-0.13	0.52	-0.22	0.22	0.51	0.43	0.85	0.07	0.60	0.44	0.04	0.09	0.04	0.38	0.43	0.40	0.38

4. Regression Modeling

Linear Models	Linear	RIDGE	LASSO	Multi-Nominal
non-parametric Models	Decision Tree	KNN	SVM	
Ensemble Models	Random Forest	Gradient Boosting	Adaboost	

4.1 Linear Regression

- Although log transformation improves the performance
- Milan players still have an error rate around 40%.



Linear Regression

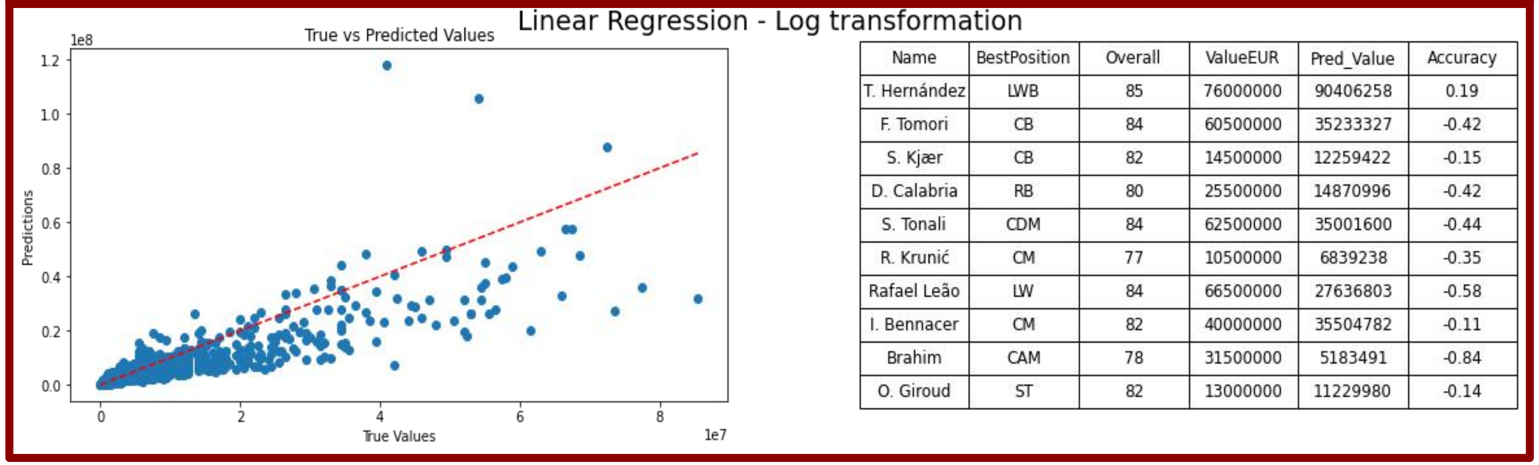
Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	32798303	-0.57
F. Tomori	CB	84	60500000	22464866	-0.63
S. Kjær	CB	82	14500000	25443228	0.75
D. Calabria	RB	80	25500000	18701362	-0.27
S. Tonali	CDM	84	62500000	13500544	-0.78
R. Krunić	CM	77	10500000	7993384	-0.24
Rafael Leão	LW	84	66500000	11533389	-0.83
I. Bennacer	CM	82	40000000	20479297	-0.49
Brahim	CAM	78	31500000	6306288	-0.8
O. Giroud	ST	82	13000000	23231067	0.79

R2 Score: 0.5013

MAE: 2430090

MSE: 21612327480836

Runing: 0.0554 s



Linear Regression - Log transformation

Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	90406258	0.19
F. Tomori	CB	84	60500000	35233327	-0.42
S. Kjær	CB	82	14500000	12259422	-0.15
D. Calabria	RB	80	25500000	14870996	-0.42
S. Tonali	CDM	84	62500000	35001600	-0.44
R. Krunić	CM	77	10500000	6839238	-0.35
Rafael Leão	LW	84	66500000	27636803	-0.58
I. Bennacer	CM	82	40000000	35504782	-0.11
Brahim	CAM	78	31500000	5183491	-0.84
O. Giroud	ST	82	13000000	11229980	-0.14

R2 Score: 0.7453

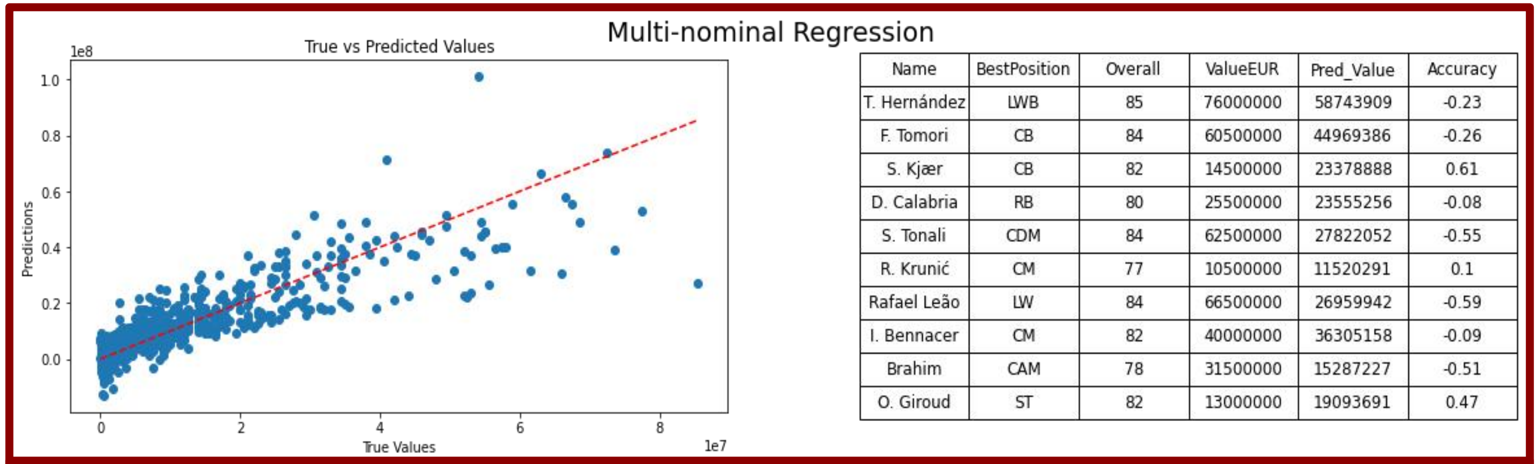
MAE: 988093

MSE: 11036989963568

Runing: 0.0419 s

4.2 Multi-nominal Regression

- This model fits the market prices of players quite well.
- The error of Milan's players is mostly within 20%.


Multi-nominal Regression

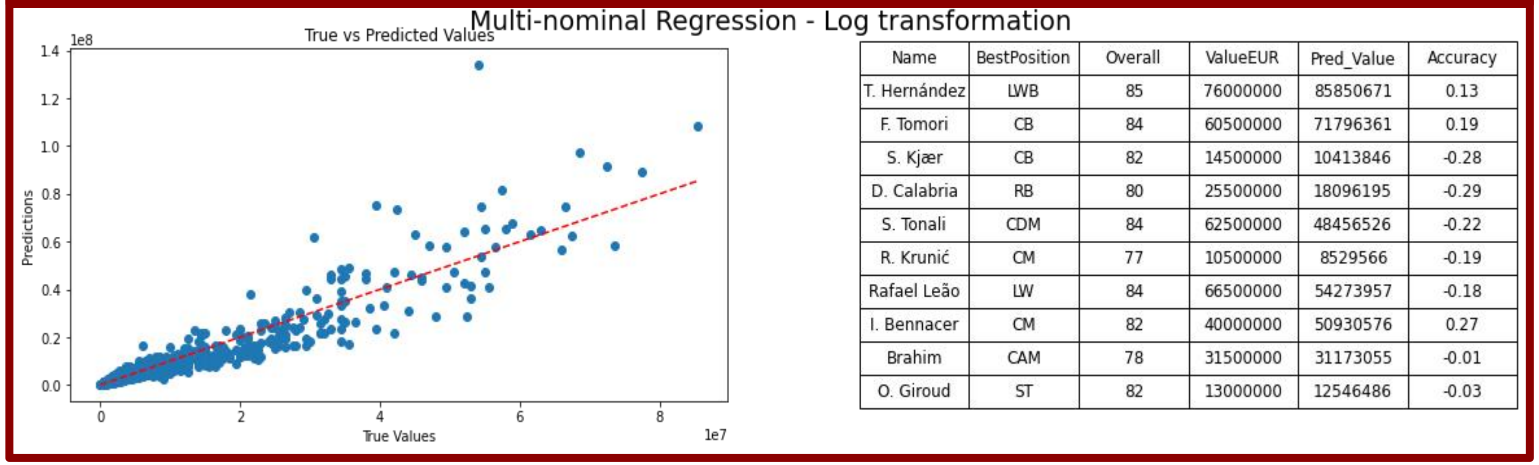
Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	58743909	-0.23
F. Tomori	CB	84	60500000	44969386	-0.26
S. Kjær	CB	82	14500000	23378888	0.61
D. Calabria	RB	80	25500000	23555256	-0.08
S. Tonali	CDM	84	62500000	27822052	-0.55
R. Krunić	CM	77	10500000	11520291	0.1
Rafael Leão	LW	84	66500000	26959942	-0.59
I. Bennacer	CM	82	40000000	36305158	-0.09
Brahim	CAM	78	31500000	15287227	-0.51
O. Giroud	ST	82	13000000	19093691	0.47

R2 Score: 0.7785

MAE: 1663694

MSE: 9601522135004

Running: 2.1948 s


Multi-nominal Regression - Log transformation

Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	85850671	0.13
F. Tomori	CB	84	60500000	71796361	0.19
S. Kjær	CB	82	14500000	10413846	-0.28
D. Calabria	RB	80	25500000	18096195	-0.29
S. Tonali	CDM	84	62500000	48456526	-0.22
R. Krunić	CM	77	10500000	8529566	-0.19
Rafael Leão	LW	84	66500000	54273957	-0.18
I. Bennacer	CM	82	40000000	50930576	0.27
Brahim	CAM	78	31500000	31173055	-0.01
O. Giroud	ST	82	13000000	12546486	-0.03

R2 Score: 0.8763

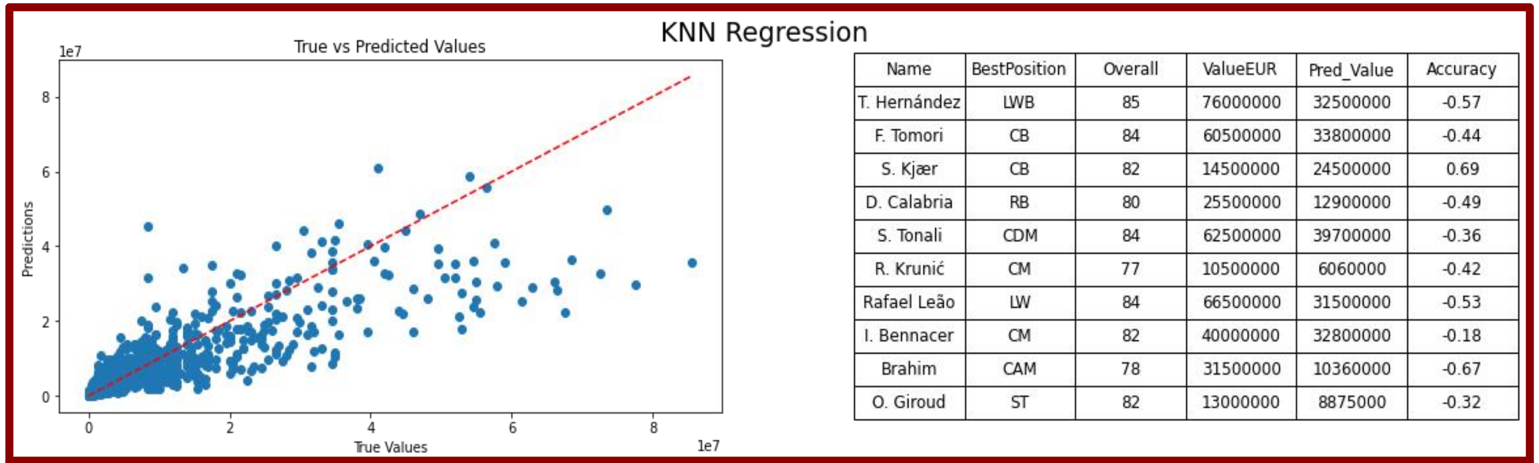
MAE: 629980

MSE: 5359653235589

Running: 2.5147 s

4.3 KNN Regression

- Log transformation make the result worse.
- Most of the predictions for Milan's players have a significant error.


KNN Regression

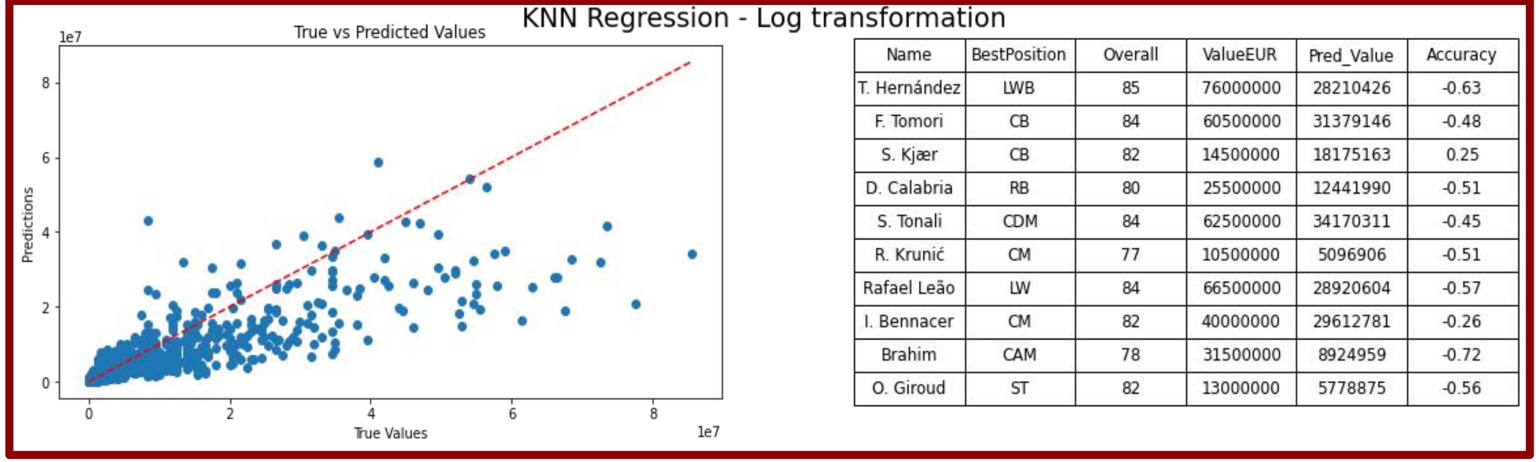
Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	32500000	-0.57
F. Tomori	CB	84	60500000	33800000	-0.44
S. Kjær	CB	82	14500000	24500000	0.69
D. Calabria	RB	80	25500000	12900000	-0.49
S. Tonali	CDM	84	62500000	39700000	-0.36
R. Krunić	CM	77	10500000	6060000	-0.42
Rafael Leão	LW	84	66500000	31500000	-0.53
I. Bennacer	CM	82	40000000	32800000	-0.18
Brahim	CAM	78	31500000	10360000	-0.67
O. Giroud	ST	82	13000000	8875000	-0.32

R2 Score: 0.7428

MAE: 1094924

MSE: 11145900528328

Running: 3.535 s


KNN Regression - Log transformation

Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	28210426	-0.63
F. Tomori	CB	84	60500000	31379146	-0.48
S. Kjær	CB	82	14500000	18175163	0.25
D. Calabria	RB	80	25500000	12441990	-0.51
S. Tonali	CDM	84	62500000	34170311	-0.45
R. Krunić	CM	77	10500000	5096906	-0.51
Rafael Leão	LW	84	66500000	28920604	-0.57
I. Bennacer	CM	82	40000000	29612781	-0.26
Brahim	CAM	78	31500000	8924959	-0.72
O. Giroud	ST	82	13000000	5778875	-0.56

R2 Score: 0.6932

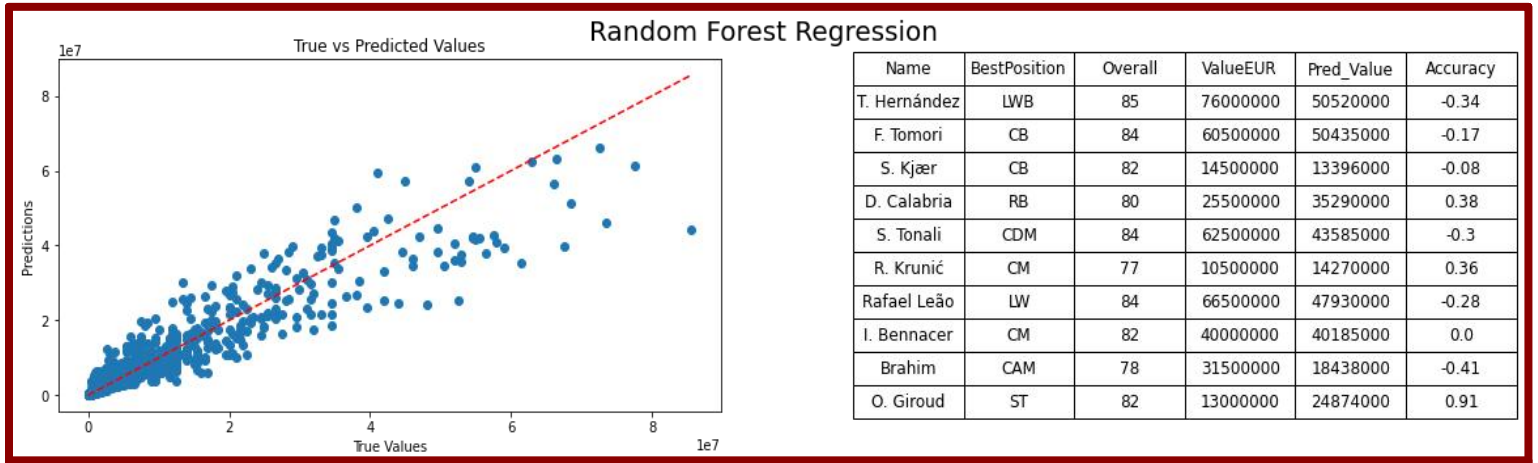
MAE: 1122900

MSE: 13296664187068

Running: 3.164 s

4.4 Random Forest

- One of the best model based on the score and MAE
- But the predictions for high-value players are inaccurate.



Random Forest Regression

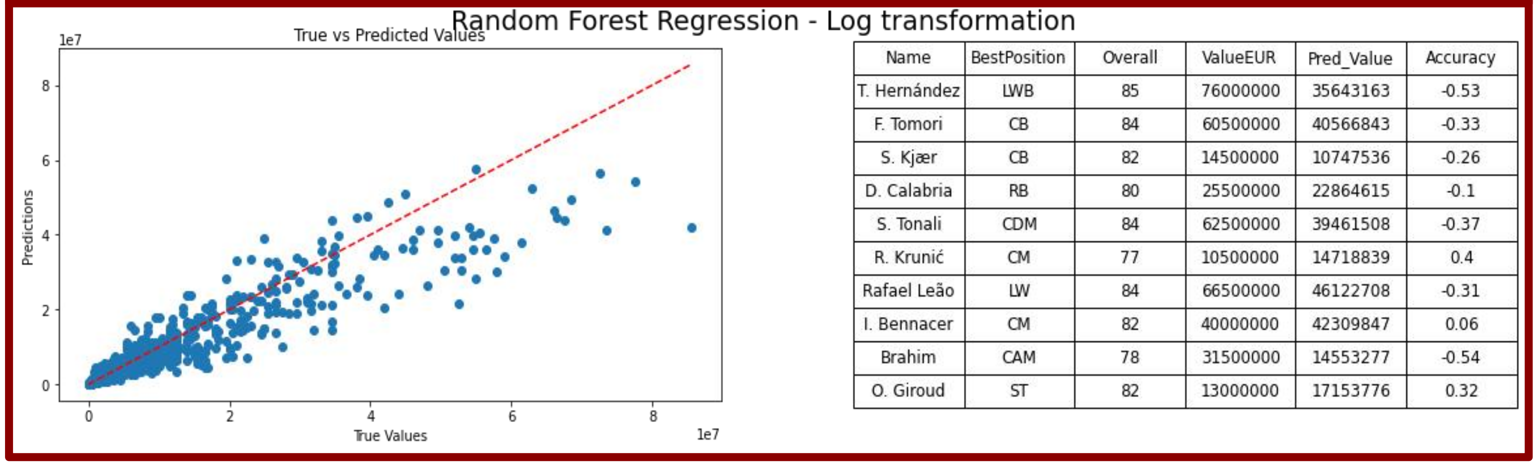
Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	50520000	-0.34
F. Tomori	CB	84	60500000	50435000	-0.17
S. Kjær	CB	82	14500000	13396000	-0.08
D. Calabria	RB	80	25500000	35290000	0.38
S. Tonali	CDM	84	62500000	43585000	-0.3
R. Krunić	CM	77	10500000	14270000	0.36
Rafael Leão	LW	84	66500000	47930000	-0.28
I. Bennacer	CM	82	40000000	40185000	0.0
Brahim	CAM	78	31500000	18438000	-0.41
O. Giroud	ST	82	13000000	24874000	0.91

R2 Score: 0.8933

MAE: 697402

MSE: 4623413076250

Running: 57.7969 s



Random Forest Regression - Log transformation

Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	35643163	-0.53
F. Tomori	CB	84	60500000	40566843	-0.33
S. Kjær	CB	82	14500000	10747536	-0.26
D. Calabria	RB	80	25500000	22864615	-0.1
S. Tonali	CDM	84	62500000	39461508	-0.37
R. Krunić	CM	77	10500000	14718839	0.4
Rafael Leão	LW	84	66500000	46122708	-0.31
I. Bennacer	CM	82	40000000	42309847	0.06
Brahim	CAM	78	31500000	14553277	-0.54
O. Giroud	ST	82	13000000	17153776	0.32

R2 Score: 0.8769

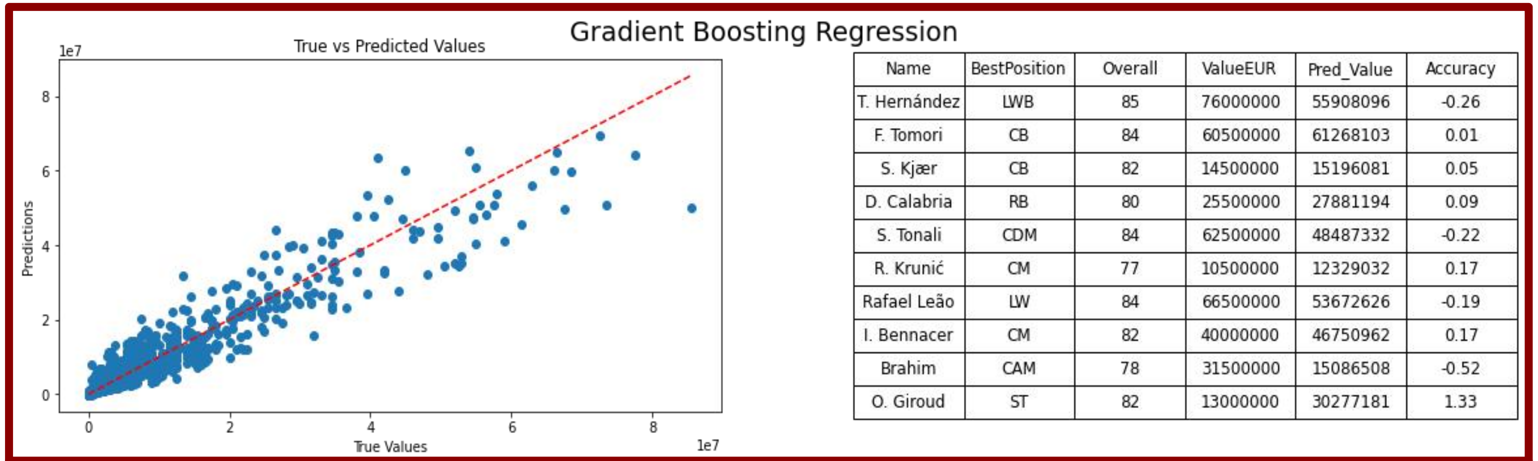
MAE: 679591

MSE: 5334894705692

Running: 52.8876 s

4.5. Gradient Boosting Regression

- One of the best model based on the score and MAE
- The value of most of Milan's players has been predicted accurately.



Gradient Boosting Regression

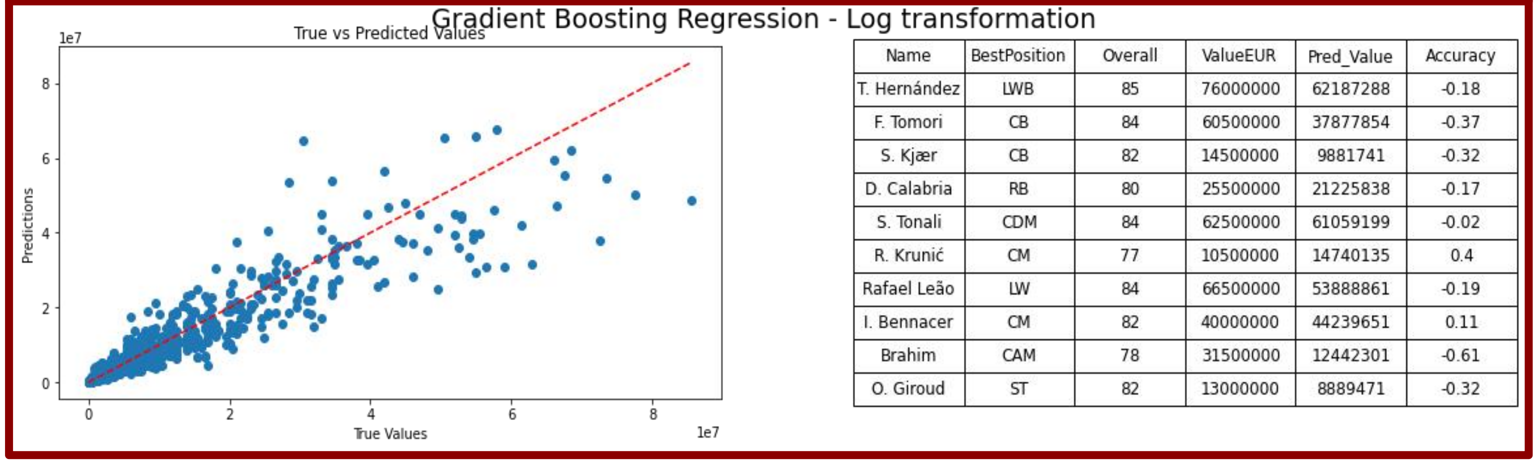
Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	55908096	-0.26
F. Tomori	CB	84	60500000	61268103	0.01
S. Kjær	CB	82	14500000	15196081	0.05
D. Calabria	RB	80	25500000	27881194	0.09
S. Tonali	CDM	84	62500000	48487332	-0.22
R. Krunić	CM	77	10500000	12329032	0.17
Rafael Leão	LW	84	66500000	53672626	-0.19
I. Bennacer	CM	82	40000000	46750962	0.17
Brahim	CAM	78	31500000	15086508	-0.52
O. Giroud	ST	82	13000000	30277181	1.33

R2 Score: 0.9147

MAE: 741671

MSE: 3694760751632

Running: 12.8189 s



Gradient Boosting Regression - Log transformation

Name	BestPosition	Overall	ValueEUR	Pred_Value	Accuracy
T. Hernández	LWB	85	76000000	62187288	-0.18
F. Tomori	CB	84	60500000	37877854	-0.37
S. Kjær	CB	82	14500000	9881741	-0.32
D. Calabria	RB	80	25500000	21225838	-0.17
S. Tonali	CDM	84	62500000	61059199	-0.02
R. Krunić	CM	77	10500000	14740135	0.4
Rafael Leão	LW	84	66500000	53888861	-0.19
I. Bennacer	CM	82	40000000	44239651	0.11
Brahim	CAM	78	31500000	12442301	-0.61
O. Giroud	ST	82	13000000	8889471	-0.32

R2 Score: 0.8883

MAE: 650905

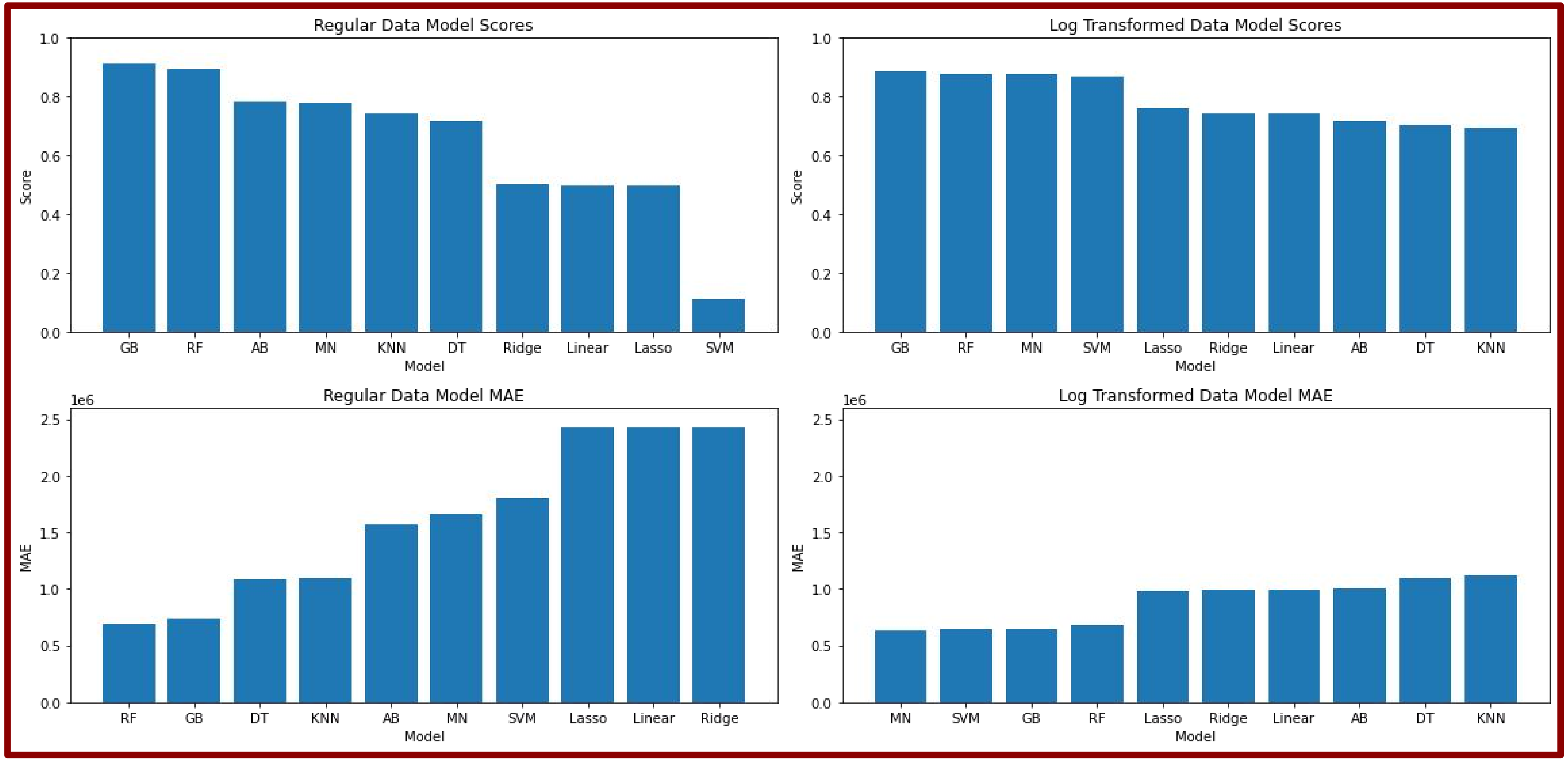
MSE: 4839335412423

Running: 14.1918 s

4.6Model Conclusion

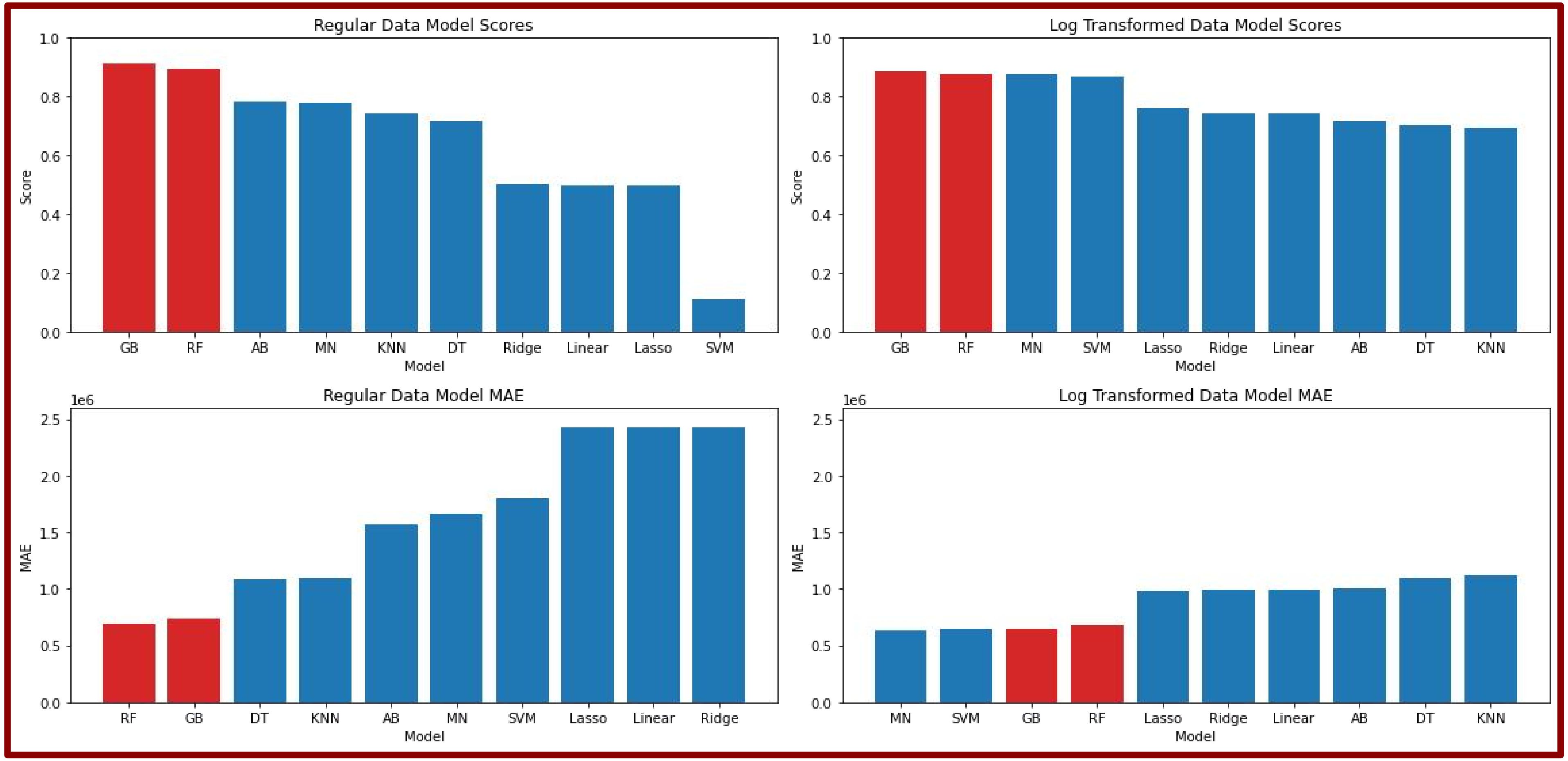
Achieving a score of 90% and relatively accurate predictions.

- **Present the performance of 10 models on the FIFA23 dataset**
- **Sorted from best to worst according to their Score and MAE.**



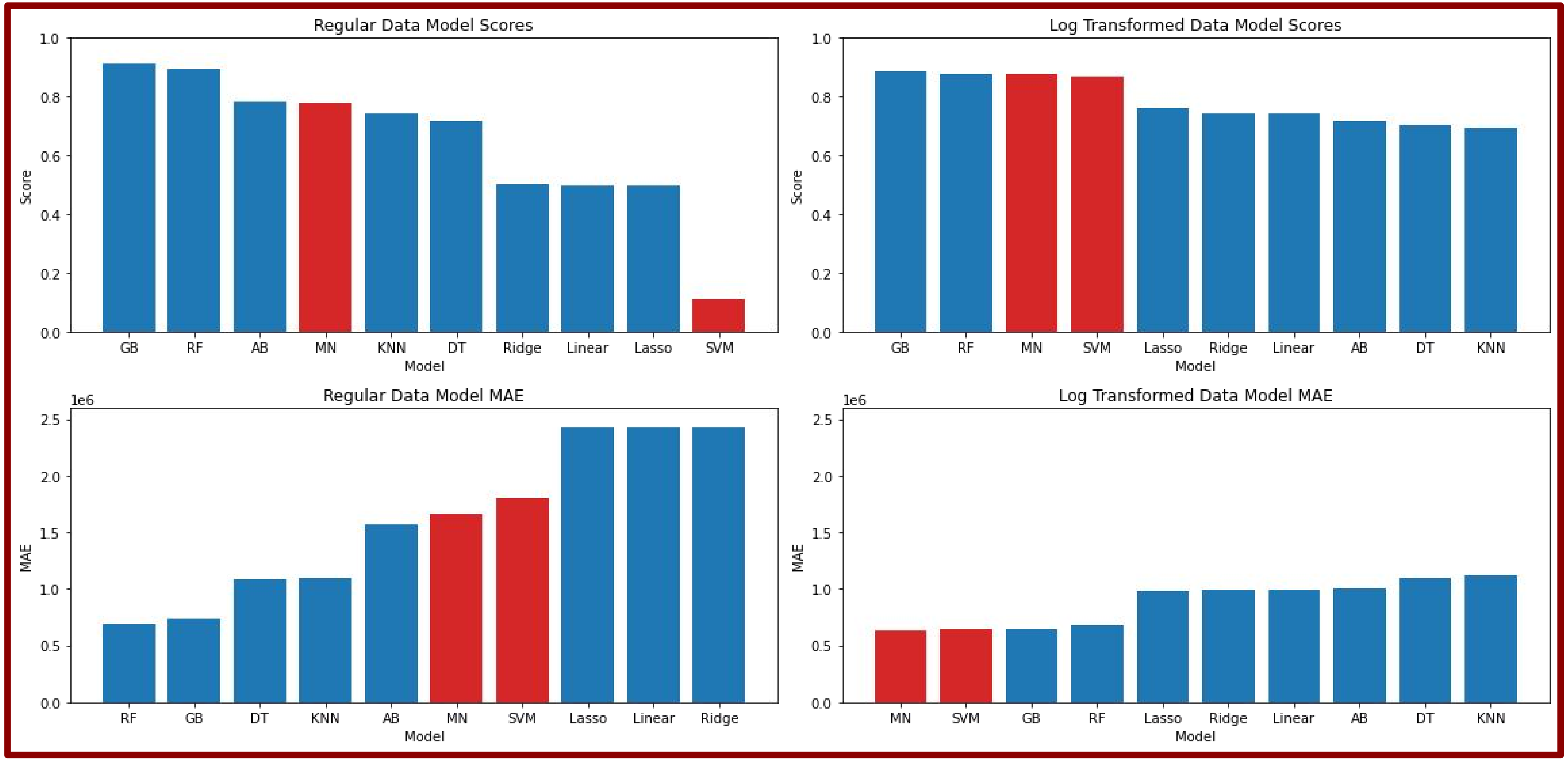
Model Conclusion 1

Gradient Boosting and Random Forest regression models achieve scores of around 90% and low MAE values on both the original and log-transformed targets.



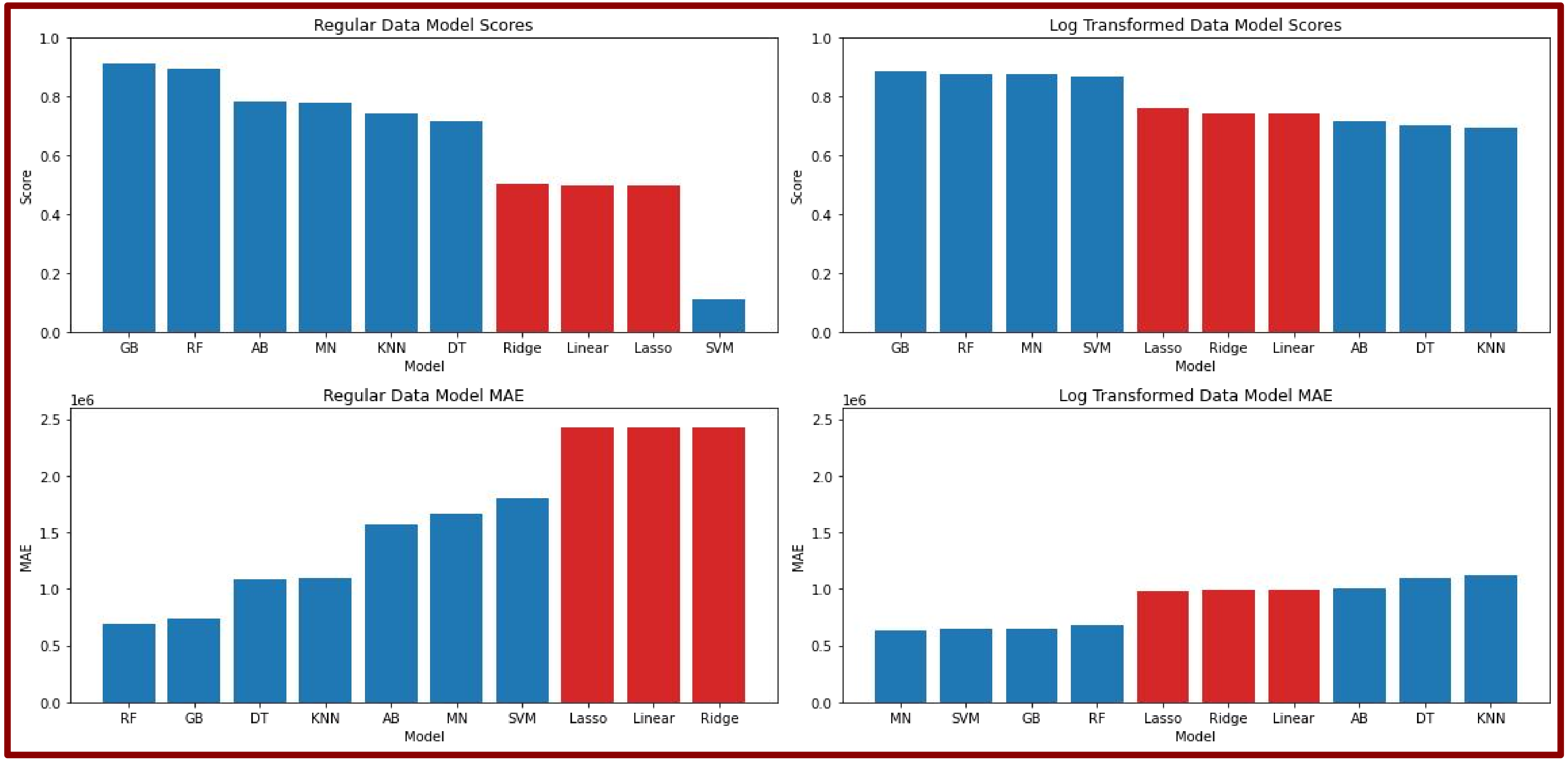
Model Conclusion 2

The Multi-nominal and SVM regression models have the lowest MAE on log transformation, but they perform poorly on the original dataset.



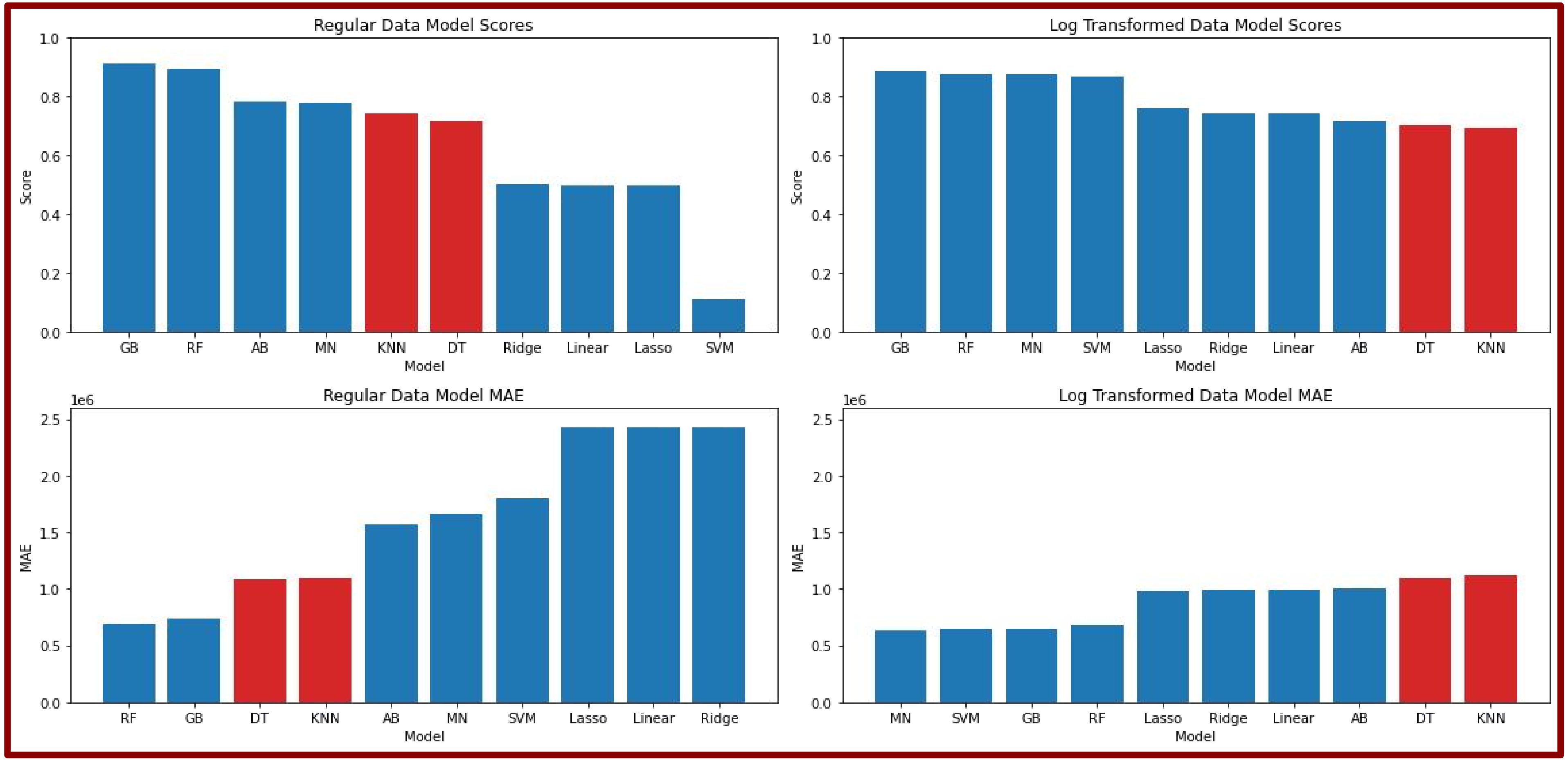
Model Conclusion 3

Although Log transformation improves some performance, Linear regression model fits poorly on this dataset, and neither Ridge nor Lasso provide a significant improvement.



Model Conclusion 4

Decision Tree and KNN regression perform better on the original dataset than on the log-transformed dataset, which means that log transformation still loses some of the original data information.



Regular Data Model Scores

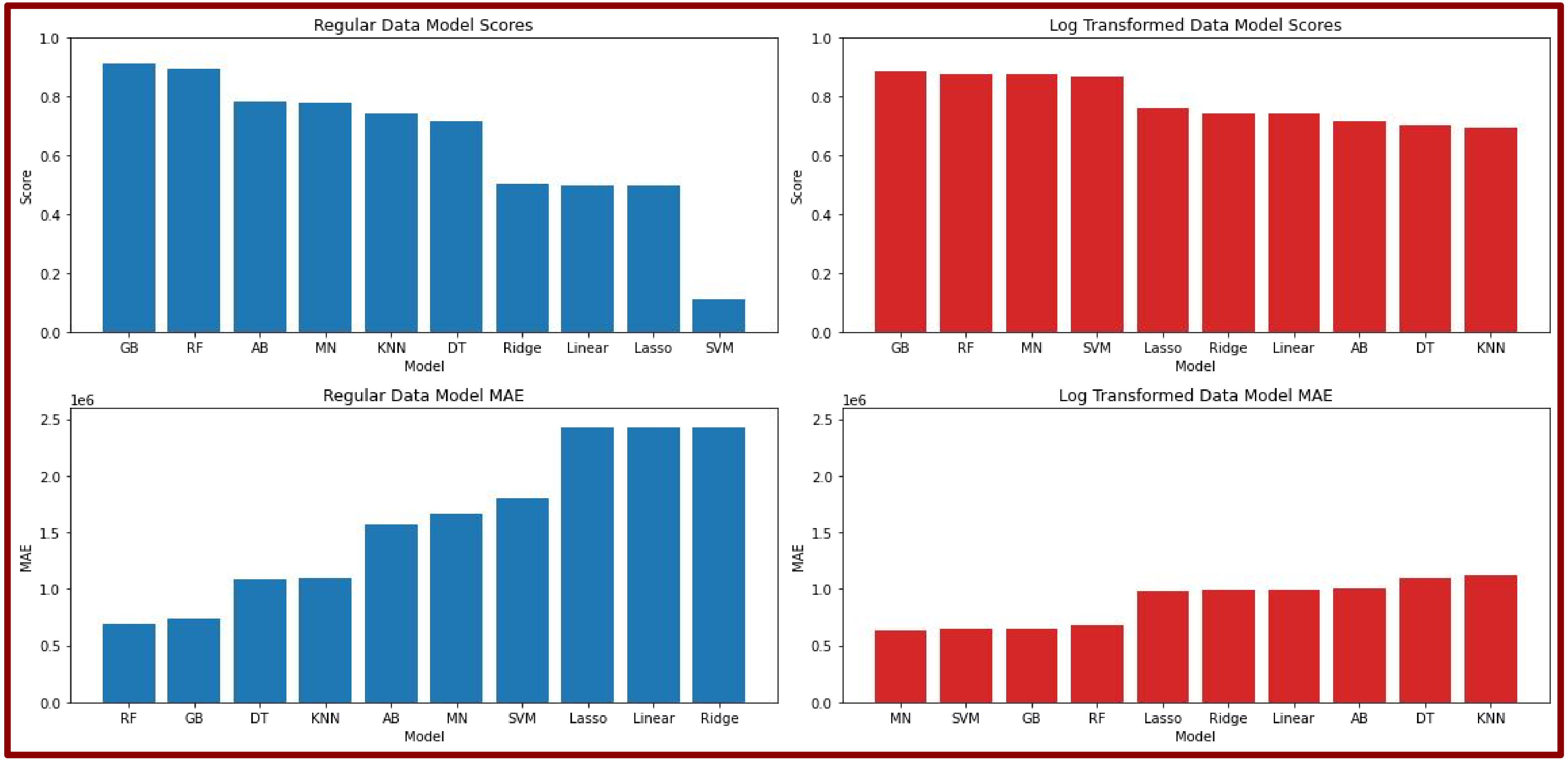
Log Transformed Data Model Scores

Regular Data Model MAE

Log Transformed Data Model MAE

Model Conclusion 5

Overall, the MAE of the models after log transformation is much lower, indicating that log transformation should be considered for variables with severe skewness when performing regression prediction.



5. Final Conclusion

No Free Lunch Theorem	No perfect algorithm, only continuous experimentation to find the best method.
Ensemble Methods	Multiple weak models combined together often result in a stronger model.
Feature Engineering	Right feature leads to the successful result.

Supervised Learning - FIFA 23

Thank you for watching!