Sports analytics

Number of goals scored prediction for European Football Leagues

Why do we need to predict the number of goals scored?

Why do we need to predict the number of goals scored?

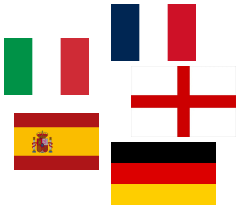
- Focus on the model interpretation, not on the model results.
- How specific parameters/values change the number of goals scored.
- Find the parameters that can be predicted accurately.

10 top leagues

300+ clubs

10 seasons

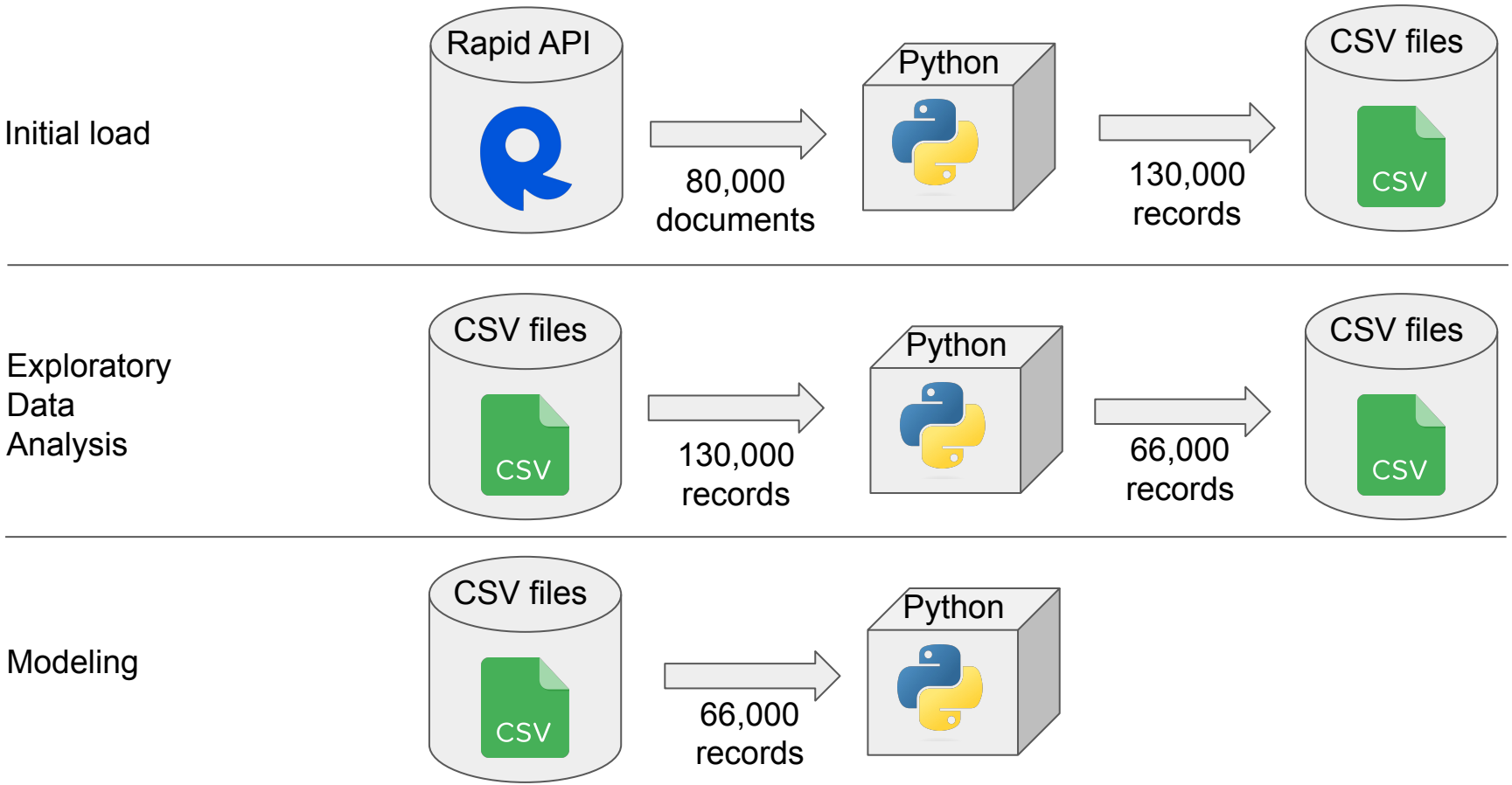
20,000+ players



170+ nationalities

$20 budget

Architecture

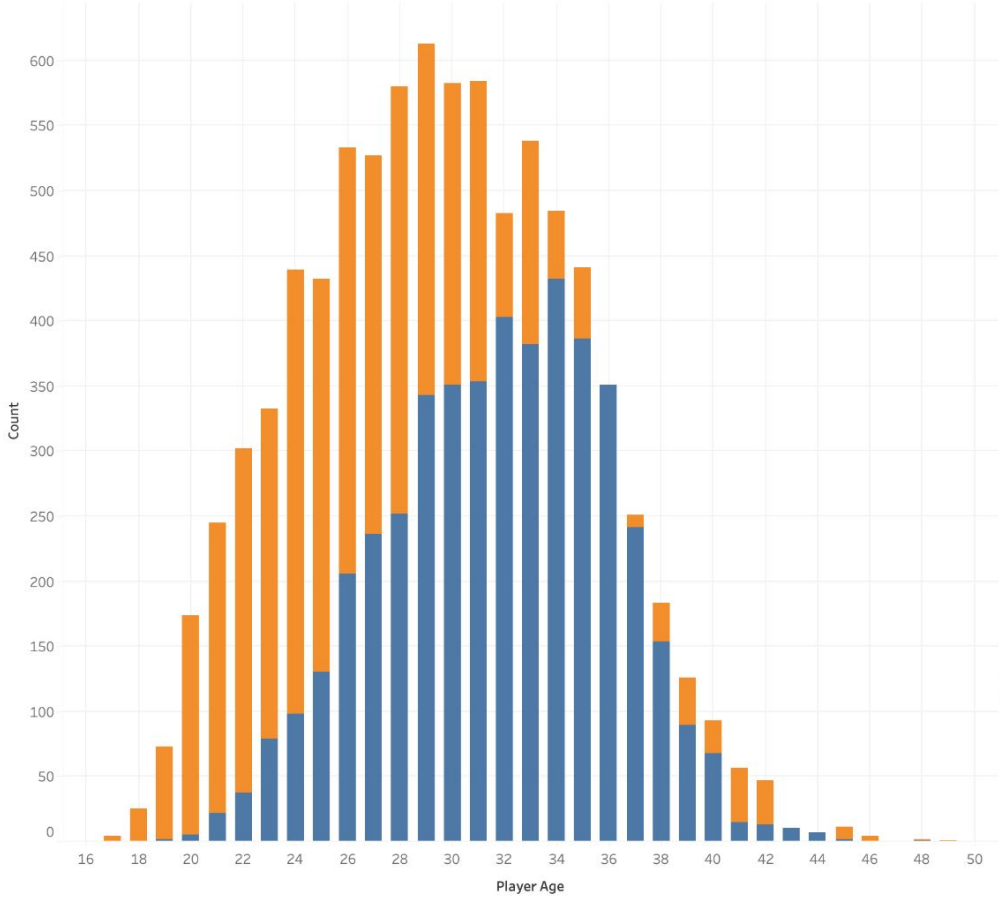


Results

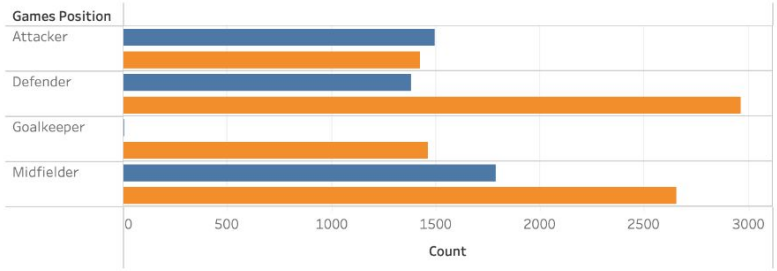
	Linear Regression	SGD Regression	KNN	Decision Tree	SVR
R2	0.65	0.65	0.68	0.76	0.76
MAE	1.02	1.03	0.79	0.68	0.84
MdAE	0.65	0.68	0.27	0.21	0.5
RMSE	1.72	1.72	1.65	1.44	1.42
Learning time, mins	0.02	0.01	0	0.01	4.17
Prediction time, mins	0	0	1.26	0	0.11

Correct/incorrect predictions

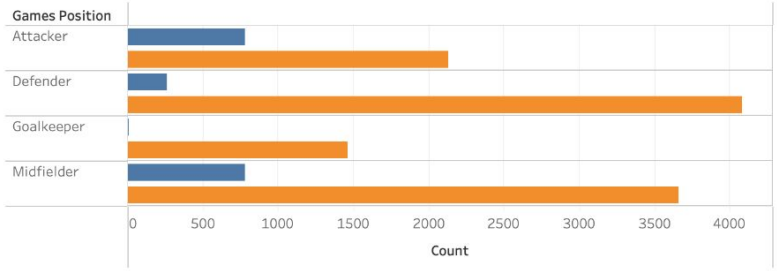
Player Age



Position



Position (including Mean Absolute Error)



Is prediction correct
No
Yes

Linear Regression (numerical features)

Worked well	Didn't work well
● To score **1 goal** a player needs to take **3 shots** ● To score **1 goal** a player needs to make **13 assists**	● To score **1 goal** a player needs to be **200 years** older ● To score **1 goal** the venue capacity should be **400,000** lower

Linear Regression (categorical features)

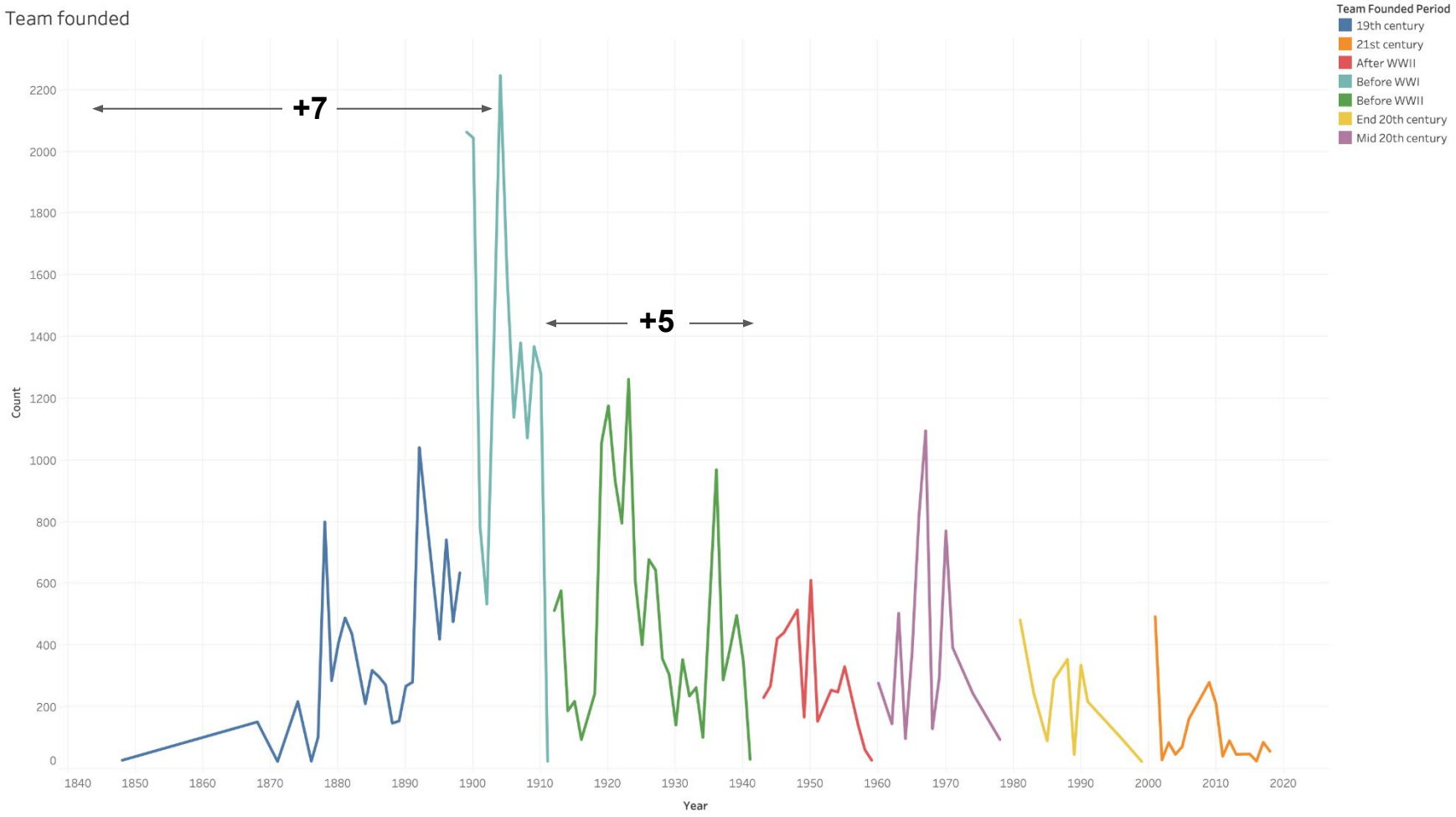
	Top 3		
	1st	**2nd**	**3rd**
Nationality	Colombia (0.8)	Tunisia (0.7)	Uruguay (**0.66**) Argentina (**0.66**)
League	Ukraine (0.15)	Spain (**0.1**)	Turkey (0.02)
Team	Barcelona (**0.66**)	Bayern Munich (0.62)	Club Brugge KV (0.58)
Stadium	Barcelona (**0.66**)	Bayern Munich (0.62)	Borussia Dortmund (0.46)

Linear Regression (categorical features)

	Top 3		
	1st	**2nd**	**3rd**
Nationality	Colombia (0.8)	Tunisia (0.7)	Uruguay (**0.66**) Argentina (**0.66**)
League	Ukraine (0.15)	Spain (**0.1**)	Turkey (0.02)
Team	Barcelona (**0.66**)	Bayern Munich (0.62)	Club Brugge KV (0.58)
Stadium	Barcelona (**0.66**)	Bayern Munich (0.62)	Borussia Dortmund (0.46)

Decision Tree

Condition	Result
Plays less than **50%** of games	Scores **<5** goals
No injuries	Scores **+5** goals
Starting lineup	Scores **+14** goals
Weight is between **65** than **83** kg	Scores **+8** goals
Younger than **33**	Scores **+8** goals
Plays in Manchester United / PSG / Ajax / Bayern Munich	Scores **+9** / **+8** / **+5** / **+5** goals
Nationality is Slovenia / Côte d'Ivoire	Scores **+10** / **+21** goals
Team was founded before **1904**	Scores **+7** goals
Team was founded between the world wars	Scores **+5** goals

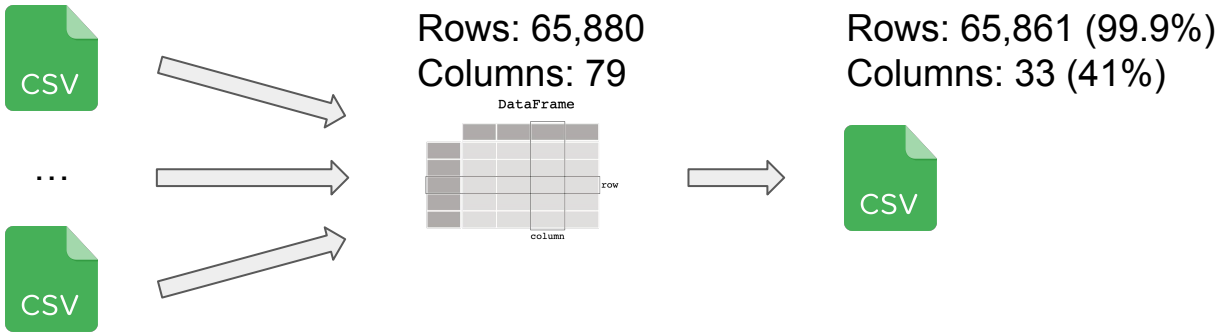

Team founded
Team Founded Period
19th century
21st century
After WWII
Before WWI
Before WWII
End 20th century
Mid 20th century
+7
+5
Count
Year
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020

APPENDIX.

Features and target

Features	Target
Player: age, height, weight, nationality, position	Number of the goals scored for a season
Player statistics: games played, shots, assists, passes, tackles, dribles, fouls	
Club: name, foundation date	
Stadium: name, capacity, surface	
League: name	

Cleaning and transformation



Cleaning and transformation

Drop columns:

- All values are the same
- A lot of distinct values (categorical)
- With very weak correlation
- Correlation was not statistically significant
- Multicollinear features
- Not relevant for ML (pictures, names, IDs, duplicated info)

Drop rows:

- Key values are null
- Duplicates
- Negative numeric numbers

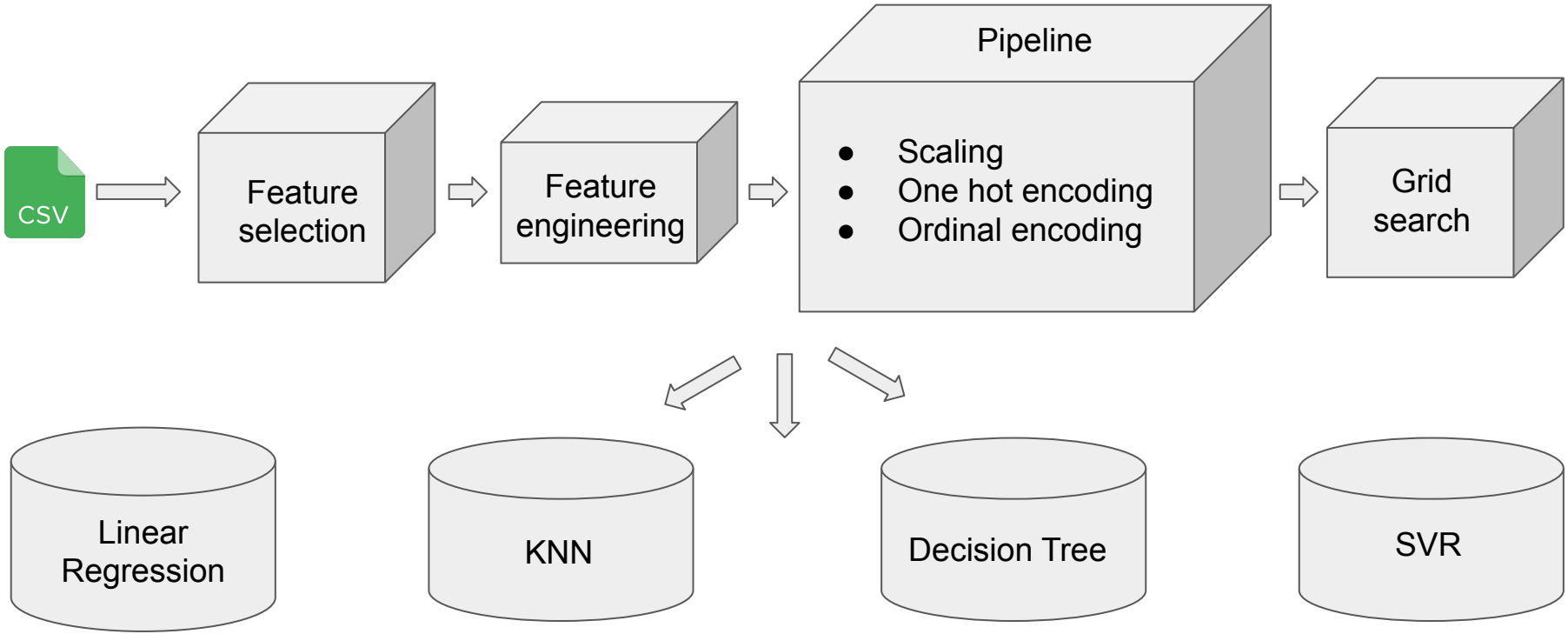
Remove values:

- Special characters
- Measures (cm, kg)

Update values:

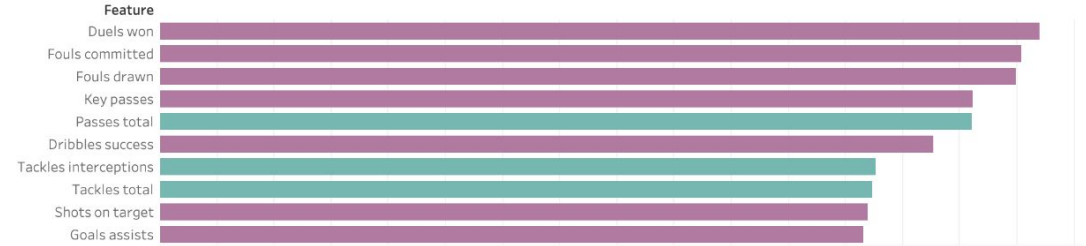
- Fill nulls with 0 for numeric
- Fill nulls with text for categorical

Modeling

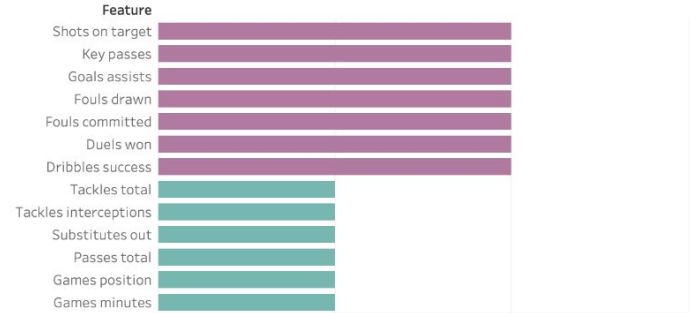

CSV
Feature selection
Feature engineering
Pipeline
Scaling
One hot encoding
Ordinal encoding
Grid search
Linear Regression
KNN
Decision Tree
SVR

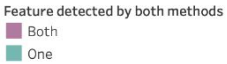
Feature importance

PCA



KBest and PCA





KBest