# Applying Data Mining to Scouting

Markus Wehr
Nazih Kalo
Stephen Stark
Tam Nguyen
WooJong Choi


# AGENDA


| | |
|---|---|
| **01** | Business Use Case |
| **02** | Analytics in the Marketplace |
| **03** | Data Overview, EDA, Engineering |
| **04** | Client Pipeline |
| **05** | Model Engineering |
| **06** | Anomaly Detection |
| **07** | Bid Prediction |
| **08** | Next Steps |

# Business Use Case

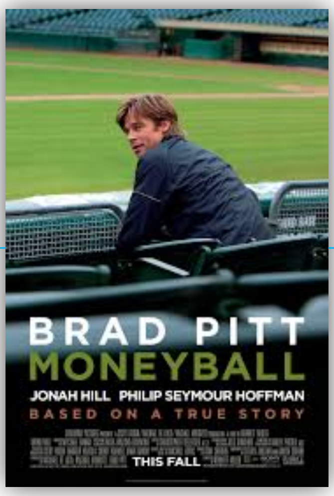
# Business Use Case

## Scouting meets Advanced Analytics

| Scouting yesterday | | Scouting today |
|---|---|---|
| • Observation<br>• Rudimentary data<br>• Intuition | VS | • Advanced analytics<br>  • availability of massive data<br>  • ability to process & capture insight |

### The hit movie: Moneyball (2011)

- The potential of unconventional sabermetrics in sport
- Scouting in soccer is a global challenge.




- Poor performing clubs face relegation which has a immediate impact on the club's bottom line
- Important for small $$$ teams to use analytics to compete with larger clubs

## Business Objectives

We are positioning ourselves as a scouting agency that:

- uses the **FIFA 2018 dataset** and
- apply various **data mining methods** to:

1. Enhance the **discovery of talents**
2. Help soccer clubs better understand the **dynamics (features)** that come into play when determining the **value** of a player

## Key Assumptions

- Our dataset reflects information up to Summer 2018.
- Market values are not biased and reflect the true intrinsic value of the player. We understand that may not be the case, but for the purpose of our models, we assume that it is.
- All feature scores, which are developed by an independent third party, are accurate and reflective of the true player style. These features are reflective of historical performance

# Hart Zwingelberg – Manager, Business Intelligence, Chicago Fire

“This (referring to soccer analytics) wasn’t a thing even five years ago,"...“To see (teams) starting to switch to a more analytically based and project-oriented front office, it’s really great. And it’s only going to explode from here.”

**Highlights of our meeting with Hart:**

- Chicago Fire uses advanced analytics for internal team assessment
- Due to the global nature of the game, the Chicago Fire prefers to outsource its scouting function to 3rd party resources (who include advanced analytics in their arsenal of player assessment)
- Focus on defining success metrics by position that fit within their overall team strategy/style
- Hart sees the potential for advanced analytics in sport and is interested in coordinating a project with the MScA program in the future

# Data Overview, EDA, Engineering

# Data - Overview

## Dataset

- **CSV**
- **18207(R) x 89(C)**

## Features

### Profile

| | | |
|---|---|---|
| • ID | • Club | • Joined |
| • Name | • Club Logo | • Loaned From |
| • Age | • Preferred Foot | • Contract Valid Until |
| • Height | • Weak Foot | • Int. Reputation |
| • Weight | • Body Type | • Photo |
| • Nationality | • Real Face | |
| • Flag | • Jersey Number | |

### Position Related

| | | | |
|---|---|---|---|
| • Position | • LS | • LAM | • LWB |
| | | • CAM | |
| | • ST | • RAM | • RWB |
| | • RS | • LM | • LB |
| | • LW | • LCM | • LCB |
| | • LF | • CM | |
| | | • RCM | • CB |
| | • CF | • RM | • RCB |
| | • RF | • LDM | |
| | | • CDM | |
| | • RW | • RDM | • RB |

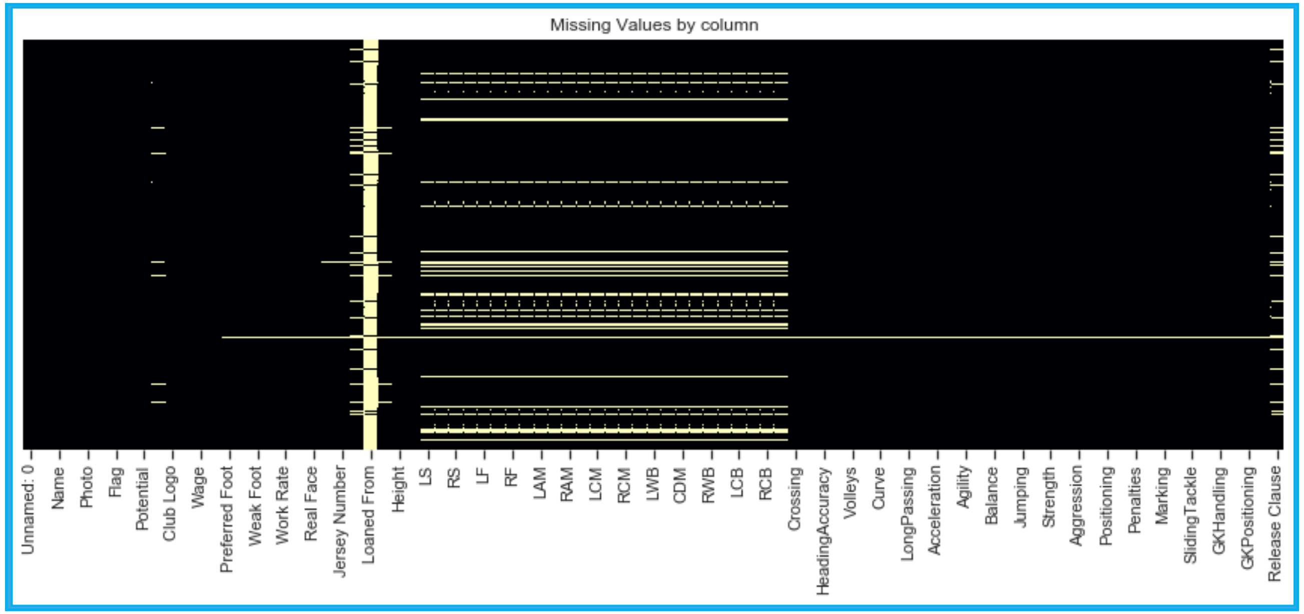
### Attributes/Skills

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| • Overall | • Crossing | • Dribbling | • Acceleration | • ShotPower | • Aggression | • Marking | • GKDiving |
| • Potential | • Finishing | • Curve | • SprintSpeed | • Jumping | • Interceptions | • StandingTackle | • GKHandling |
| • Special | • HeadingAccuracy | • FKAccuracy | • Agility | • Stamina | • Positioning | • SlidingTackle | • GKKicking |
| • Skill Moves | • ShortPassing | • LongPassing | • Reactions | • Strength | • Vision | | • GKPositioning |
| • Work Rate | • Volleys | • BallControl | • Balance | • LongShots | • Penalties | | • GKReflexes |
| | | | | | • Composure | | |

### $$$

| | | |
|---|---|---|
| • Value | • Wage | • Release Clause |

## MISSING VALUES




FIFA
For the Good of the Game

# Data Processing & Feature Engineering

## Original Data

| Feature | Data Type | Missing Values |
|---|---|---|
| **ID** | Categorical | - |
| **Name** | Text | - |
| **Age** | Numerical | - |
| **Height** | Text | 48 |
| **Weight** | Text | 48 |
| **Nationality** | Categorical | - |
| **Flag** | Categorical | - |
| **Club** | Categorical | 241 |
| **Club Logo** | Text | |
| **Preferred Foot** | Categorical | 48 |
| **Weak Foot** | Numerical | 48 |
| **Body Type** | Categorical | 48 |
| **Real Face** | Categorical | 48 |
| **Jersey Number** | Categorical | 60 |
| **Joined** | Date | 1553 |
| **Loaned From** | Categorical | 16943 |
| **Contract Valid Until** | Date | 289 |

## After Data Processing & Feature Engineering

| Processing/Feature Engineering | Imputation / Drop | Data Type |
|---|---|---|
| Dropped | - | - |
| Dropped | - | - |
| - | - | Numerical |
| Converted inches to centimeters | 48 missing rows dropped | Numerical |
| Removed the text "lbs" and converted to integer | 48 missing rows dropped | Numerical |
| Dropped and new column “Continent” created to assign continent instead | 0 | Dummy |
| Dropped | - | - |
| Dropped and new column “Club Reputation” created by taking the mean of 'International Reputation' for players for each club | Filled in missing values with "No_club" | Numerical |
| Dropped | - | - |
| Converted to Binary: 0 = left, 1 = right | 48 missing rows dropped | Categorical |
| No change | 48 missing rows dropped | Numerical |
| Removed one-off body types and replaced them with either "lean", "stocky" and "normal" based on domain knowledge | 48 missing rows dropped | Numerical |
| Converted to Binary: 0 = No, 1 = Yes | 48 missing rows dropped | Categorical |
| No change | 48 missing rows dropped. 12 remaining missing values were filled in using the mode Jersey Number of the player's position | Categorical |
| Converted to int: 2019/1/1 - Joined Date | Filled in missing values with 0 | Numerical |
| Converted to Binary: 0 = Not on loan, 1 = On loan | Missing value means the players is not on loan. These missing values are assigned 0 | Categorical |
| Converted to int: years of contract left from 2018 | Filled in missing values with 0 (expired) | Numerical |

# Data Processing & Feature Engineering

## Original Data

| Feature | Data Type | Missing Values |
|---|---|---|
| **Position** | Categorical | 60 |
| **LS** | Text | 2085 |
| | Text | 2085 |
| | Text | 2085 |
| **24 columns** | Text | 2085 |
| | Text | 2085 |
| | Text | 2085 |
| | Text | 2085 |
| RB | Text | 2085 |

## After Data Processing & Feature Engineering

| Processing/Feature Engineering | Imputation / Drop | Data Type |
|---|---|---|
| Position_Group column created that assigns one of the following to the player: Forward, Midfielder, Defender, GoalKeeper, Other (no position) | Players assigned Other originally did not have a position, but later imputed based on the players' max ability from Attacking, Defending, GoalKeeping | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |
| "+ int" removed and a new column created to capture just the int. Column converted to integer. | 2025 missing values are Goalkeepers, who do not have a value for this column | Dummy |

# Data Processing & Feature Engineering

## Original Data

| Feature | Data Type | Missing Values |
|---|---|---|
| **Overall** | Numerical | - |
| **Potential** | Numerical | - |
| **Special** | Numerical | - |
| **Skill Moves** | Numerical | 48 |
| **Work Rate** | Categorical | 48 |
| *** Attributes x 34** | Numerical | 48 |
| **Value** | Text | |
| **Wage** | Text | |
| **Release Clause** | Text | 1564 |

## After Data Processing & Feature Engineering

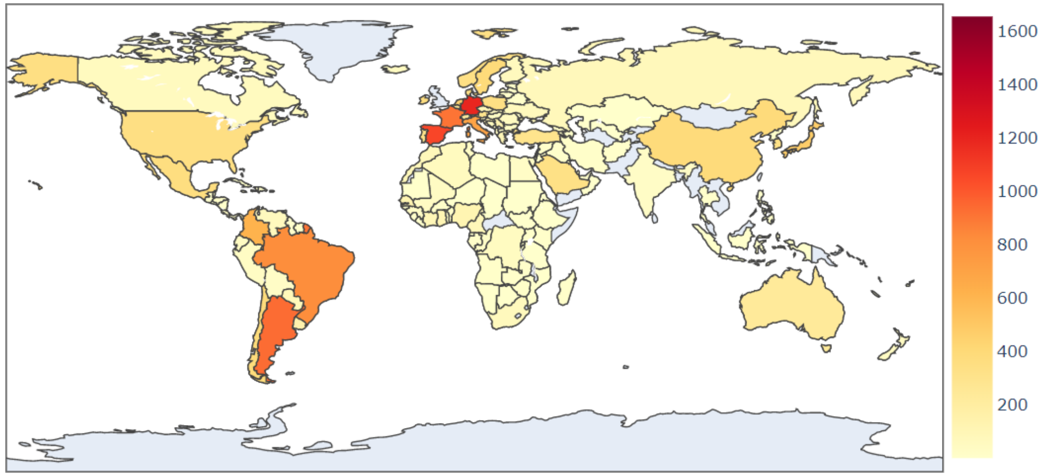
| Processing/Feature Engineering | Imputation / Drop | Data Type |
|---|---|---|
| - | - | Numerical |
| - | - | Numerical |
| - | - | Numerical |
| - | 48 missing rows dropped | Numerical |
| Dropped and created new columns "Attack_WR" and "Defense_WR" | 48 missing rows dropped | Numerical |
| 7 New columns created “Attack”, “Skill”, “Movement”, “Power”, Mentality“, “Defending”, “GoalKeeping” and assigned with means of attributes that belong to the group | 48 missing rows dropped | Numerical |
| Removed currency signs and converted to integer. | | Numerical |
| Removed currency signs and converted to integer. | | Numerical |
| Removed currency signs and converted to integer | Missing values filled in with 0 | Numerical |

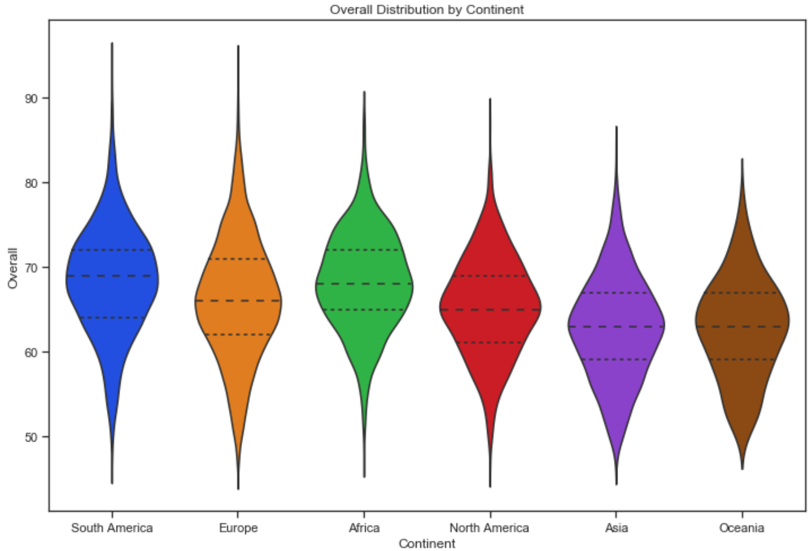
# Summary:

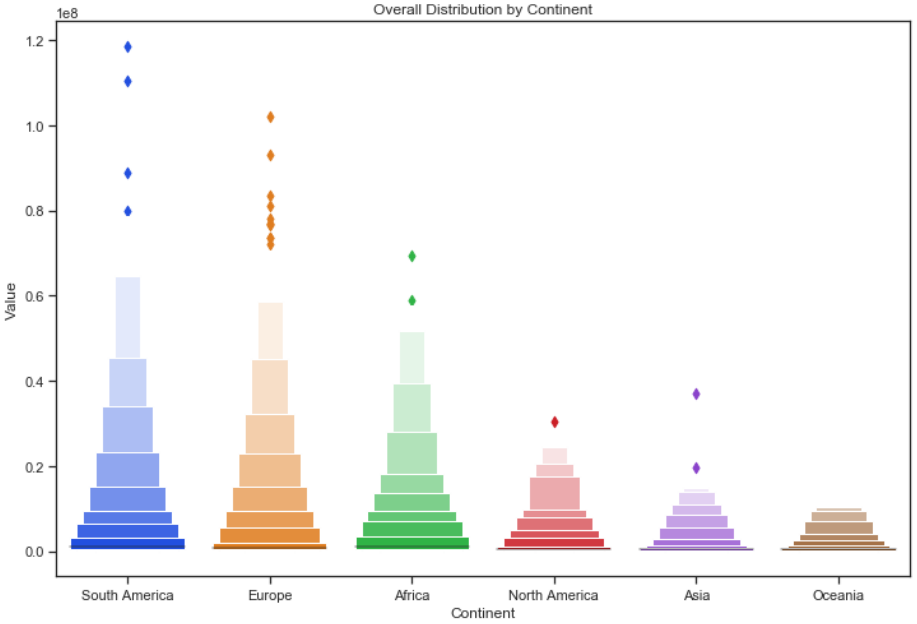
18159 rows x 125 columns

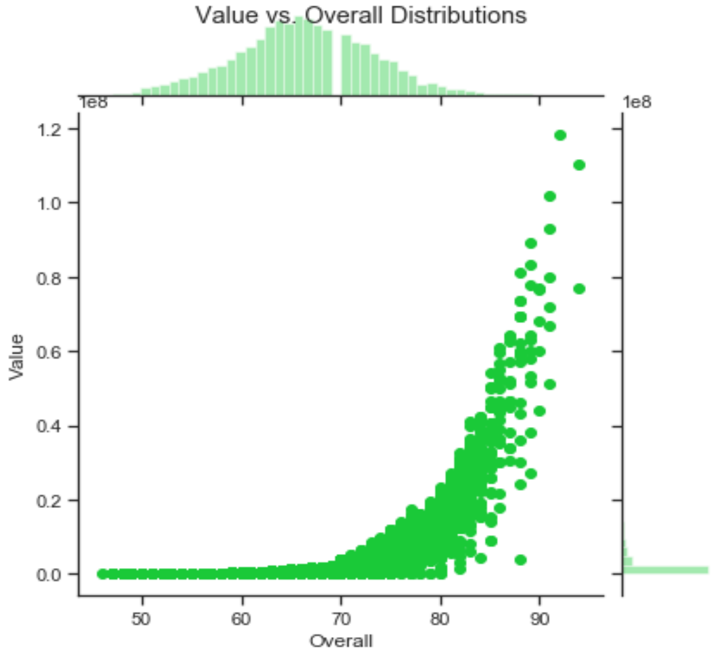
For the Good of the Game

# EDA – Visualization (1/3)



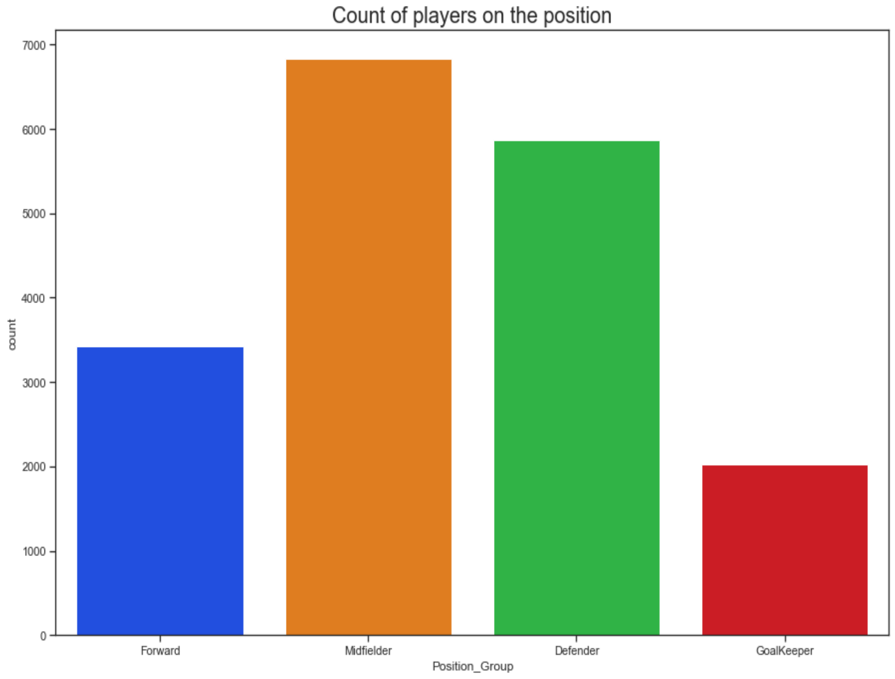



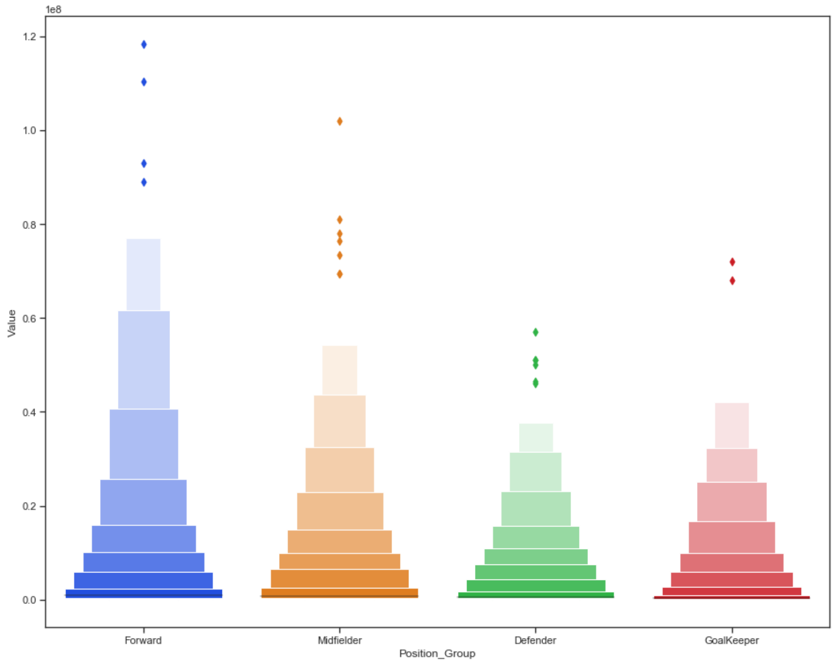



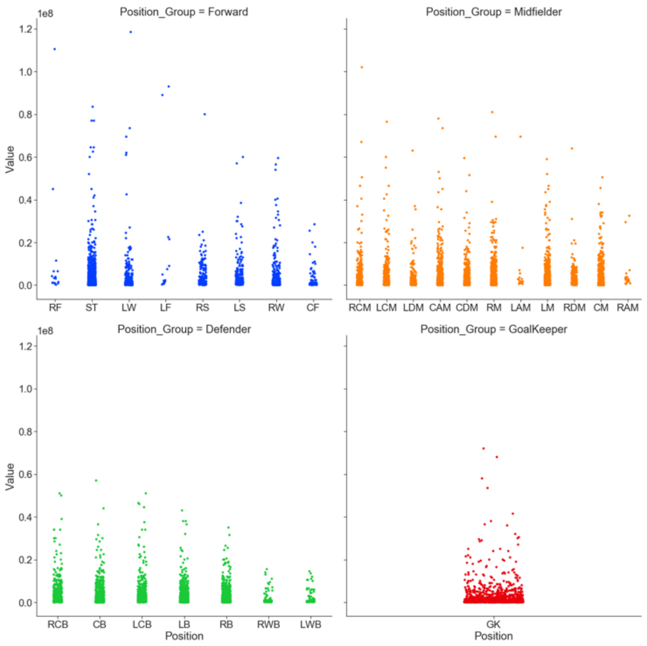



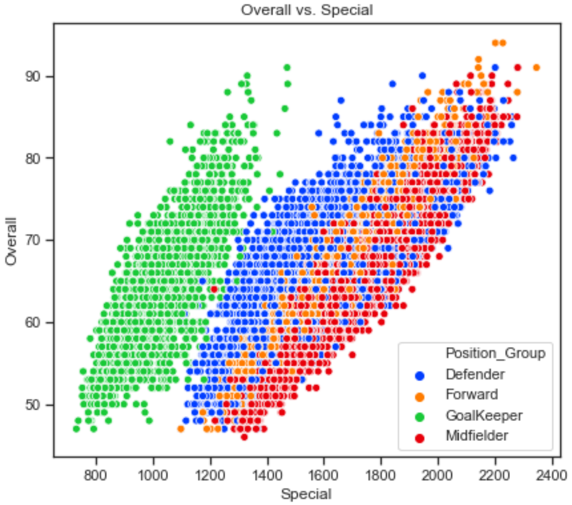

For the Good of the Game

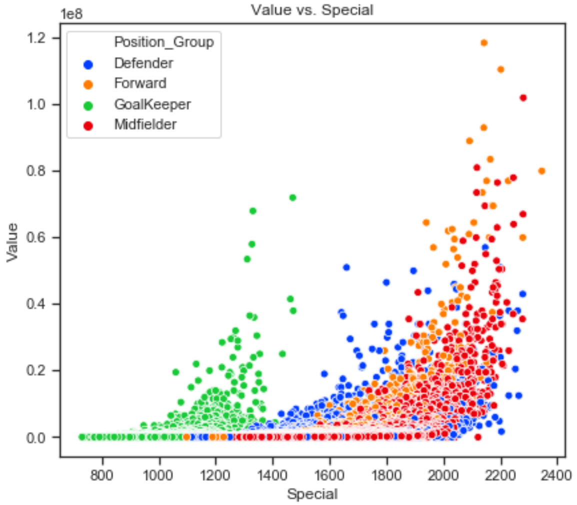
# EDA – Visualization (2/3)



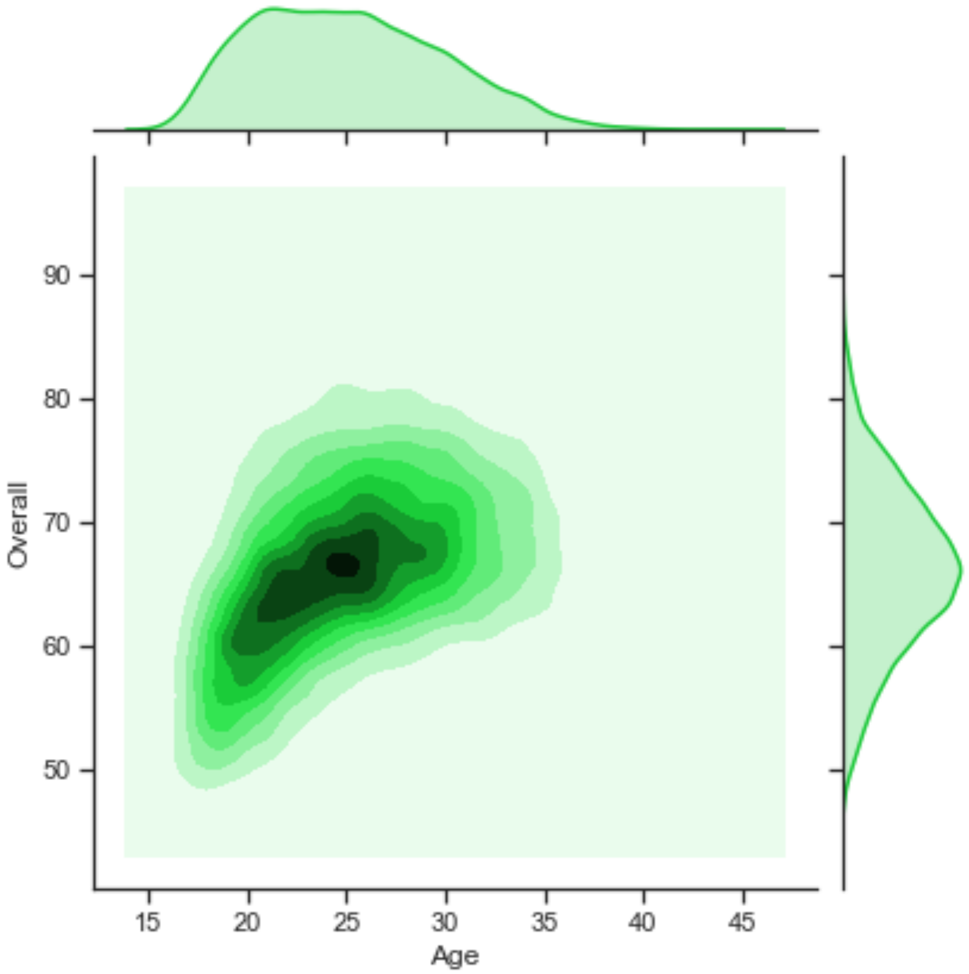



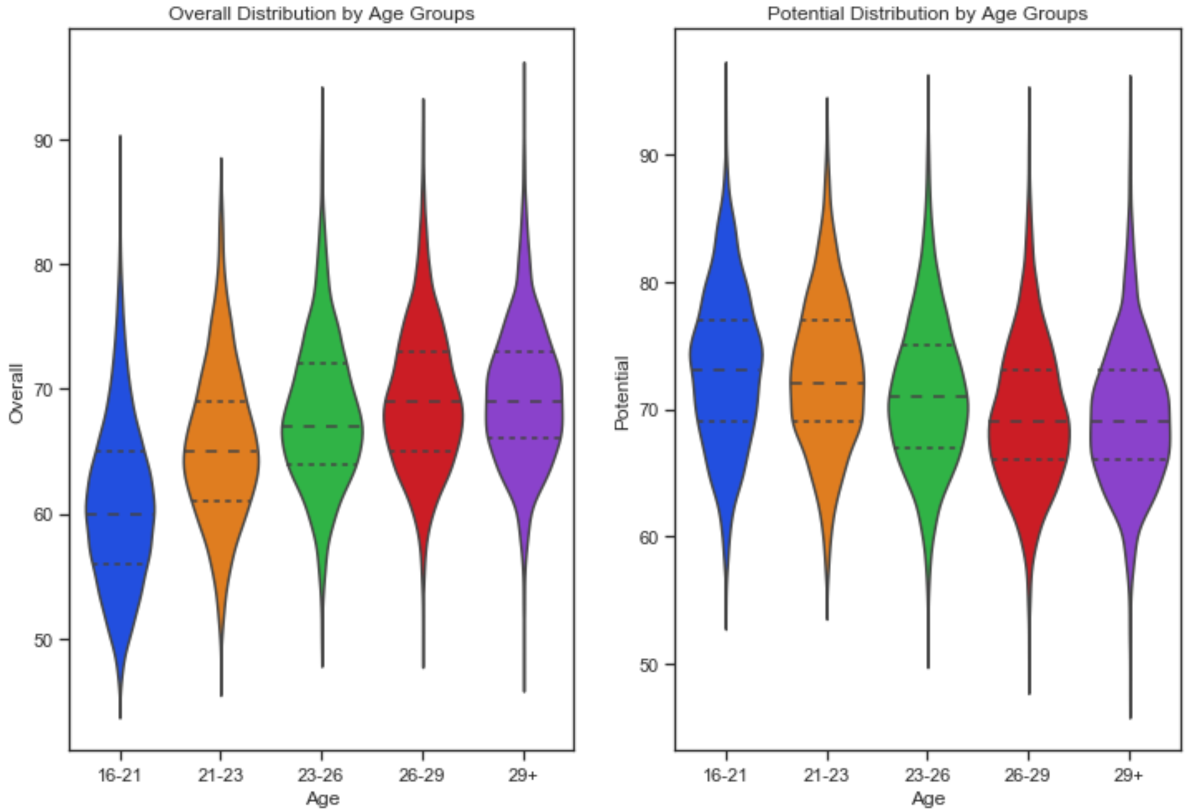




FIFA
For the Good of the Game

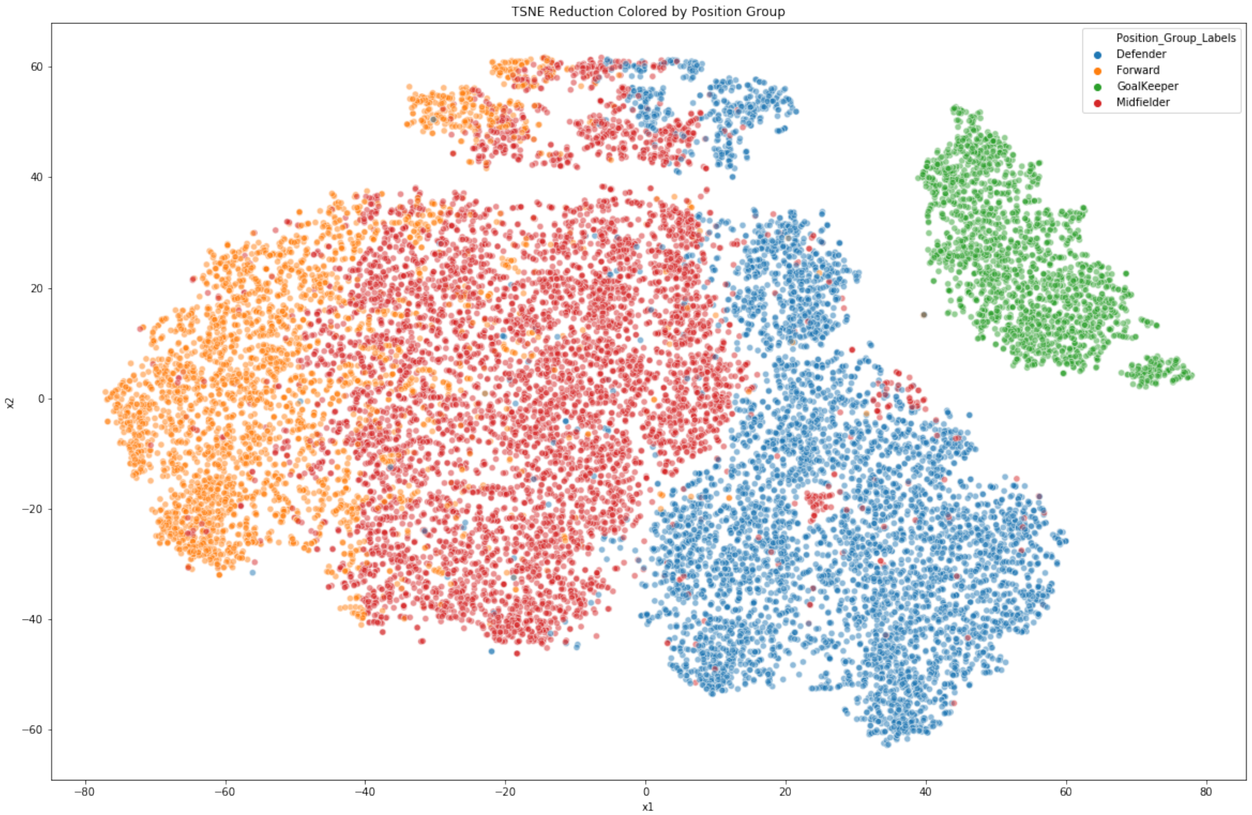
# EDA – Visualization (3/3)

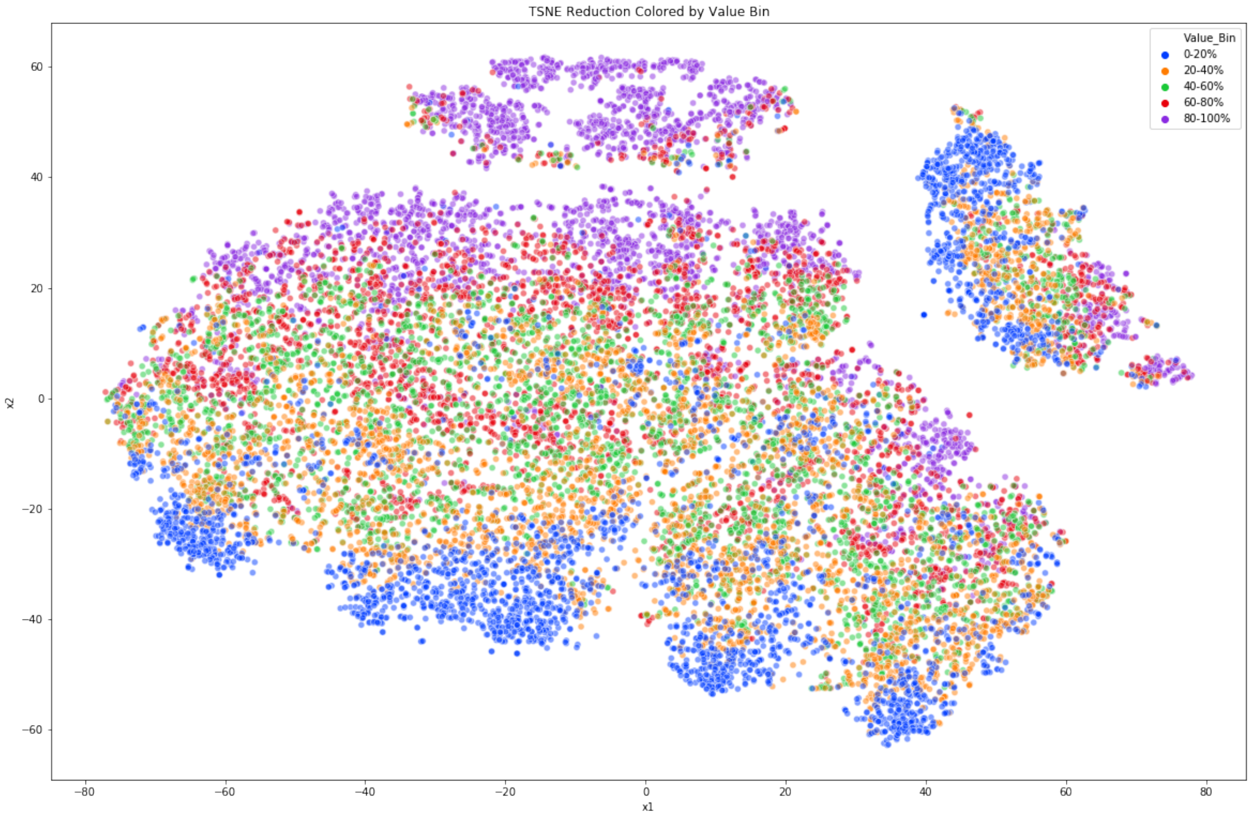



FIFA
For the Good of the Game

# TSNE




TSNE reduction shows clustering of position groups…

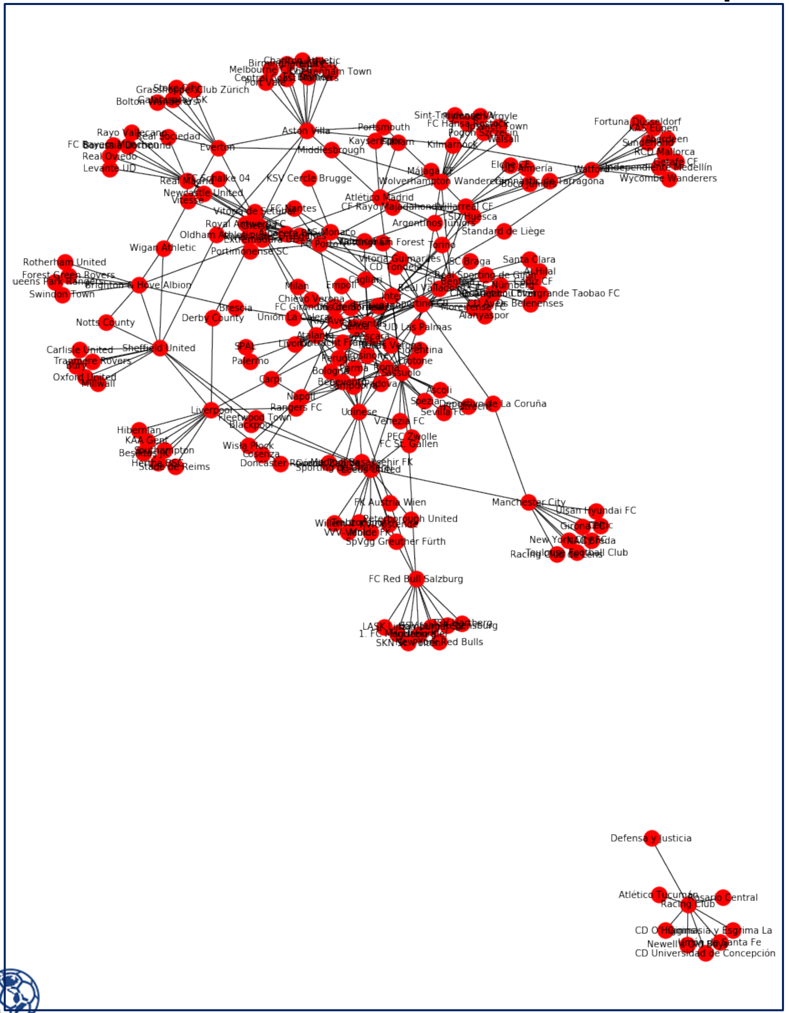
and within these position group clusters there is additional clustering of players by valuation ($) level.

For the Good of the Game

# Analysis on players on loan: Graph

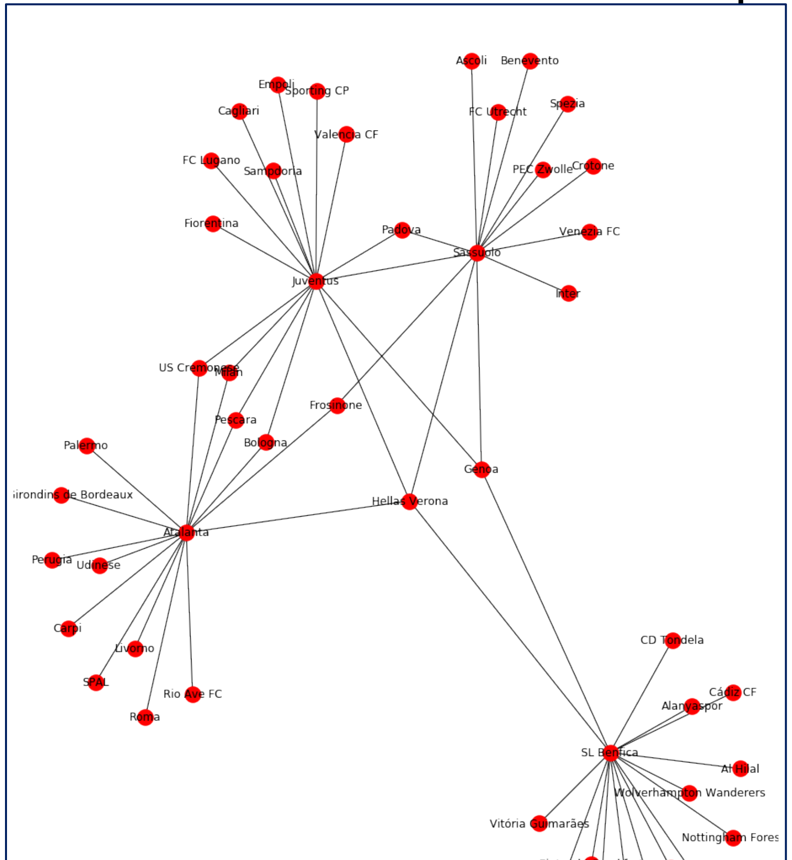
## Players on loan: 1265

Clubs that loaned out **10 or more** players



| | |
|---|---|
| England | 10 |
| Italy | 8 |
| Portugal | 3 |
| Spain | 3 |
| France | 1 |
| Austria | 1 |

Clubs that loaned out **15 or more** players




**Degree centrality**

| | |
|---|---|
| SL Benfica | 0.326531 |
| Juventus | 0.306122 |
| Atalanta | 0.306122 |
| Sassuolo | 0.265306 |

| | |
|---|---|
| Italy | 3 |
| Portugal | 1 |

**Why?** Italy doesn't have B teams, so they send young players out on loan to give them playing time.

(B teams allowed in Italy from 2019 so this pattern may change)

For the Good of the Game

# Client Pipeline Process

**01**

**Create Restricted Set of Recommended Players**

- K-Nearest Neighbors

**02**

**Isolate Outlier Players**

Anomaly Detection:
- SVM-One Class
- Local Outlier factor
- Isolation Forest
- DBSCAN

**03**

**Predict Bid Price for Isolate Outlier Players**

- Linear Regression
- Decision Tree
- Random Forest
- XGBoost
- SVR



For the Good of the Game

# Model Engineering

# Client Pipeline Process

Give me 300 hundred players similar to **“M. Salah”**


## 01
**Create Restricted Set of Recommended Players**

- K-Nearest Neighbors

## 02
**Isolate Outlier Players**

Anomaly Detection:
- SVM-One Class
- Local Outlier factor
- Isolation Forest
- DBSCAN

## 03
**Predict Bid Price for Isolate Outlier Players**

- Linear Regression
- Decision Tree
- Random Forest
- XGBoost
- SVR



FIFA
For the Good of the Game

# Filtering Functions

## Option #1:

### filter_players(position, ovr_min = 0, ovr_max= 100)

Accepts a position name and overall range and returns a filtered list & dataframe of the players that meet those criteria

**Step 1:** Enter the position looking for:

```
CM
```

**Step 2:** What is the min overall?:

```
74
```

**Step 3:** What is the max overall?:

```
86
```

## Output

```
Here are the filtered players based on your criteriea:

['Thiago',
 'S. Milinković-Savić',
 'Jorginho',
 'I. Gündoğan',
 'N. Keïta',
 'C. Tolisso',
 'A. Rabiot',
 'L. Goretzka',
 'J. Draxler',
 'Cesc Fàbregas',
 'M. Dembélé',
 'Rodri',
```

Here are the filtered players' features based on your criteriea:

| | Age | Overall | Potential | Special | Preferred Foot | International Reputation | Weak Foot | Skill Moves | Real Face | Height | Weight | LS | ST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **67** | 27 | 86 | 86 | 2190 | 1 | 3.0 | 3.0 | 5.0 | 1 | 175 | 154 | 75 | 75 |
| **78** | 23 | 85 | 90 | 2206 | 1 | 2.0 | 4.0 | 4.0 | 1 | 190 | 168 | 81 | 81 |
| **121** | 26 | 84 | 87 | 2136 | 1 | 2.0 | 3.0 | 3.0 | 0 | 180 | 148 | 70 | 70 |
| **136** | 27 | 84 | 84 | 2138 | 1 | 3.0 | 4.0 | 4.0 | 1 | 180 | 176 | 75 | 75 |
| **161** | 23 | 83 | 88 | 2082 | 1 | 2.0 | 4.0 | 4.0 | 1 | 173 | 141 | 73 | 73 |
| **162** | 23 | 83 | 88 | 2207 | 1 | 2.0 | 3.0 | 3.0 | 1 | 180 | 179 | 78 | 78 |
| **168** | 23 | 83 | 87 | 2184 | 0 | 2.0 | 3.0 | 3.0 | 1 | 193 | 176 | 77 | 77 |
| **169** | 23 | 83 | 88 | 2203 | 1 | 3.0 | 4.0 | 3.0 | 1 | 188 | 174 | 77 | 77 |
| **184** | 24 | 83 | 86 | 2112 | 1 | 3.0 | 5.0 | 4.0 | 1 | 188 | 170 | 79 | 79 |

## Option #2:

### recommended_k_players_df(player, k_players = 100)

Accepts a player's name and number of players to recommend and returns a dataframe of the recommended players and a list of their names. The recommendations are limited to players from the same position group.

**Step 1:** Enter the player you are looking for:

```
M. Salah
```

**Step 2:** Enter the number of similar players you are looking for:

```
300
```

## Output

```
Here are 300 players similar to M. Salah:

0                  L. Messi
1         Cristiano Ronaldo
2                 Neymar Jr
4              K. De Bruyne
5                 E. Hazard
6                 L. Modrić
7                 L. Suárez
10          R. Lewandowski
11                 T. Kroos
13              David Silva
15                P. Dybala
16                  H. Kane
17             A. Griezmann
20          Sergio Busquets
21                E. Cavani
```

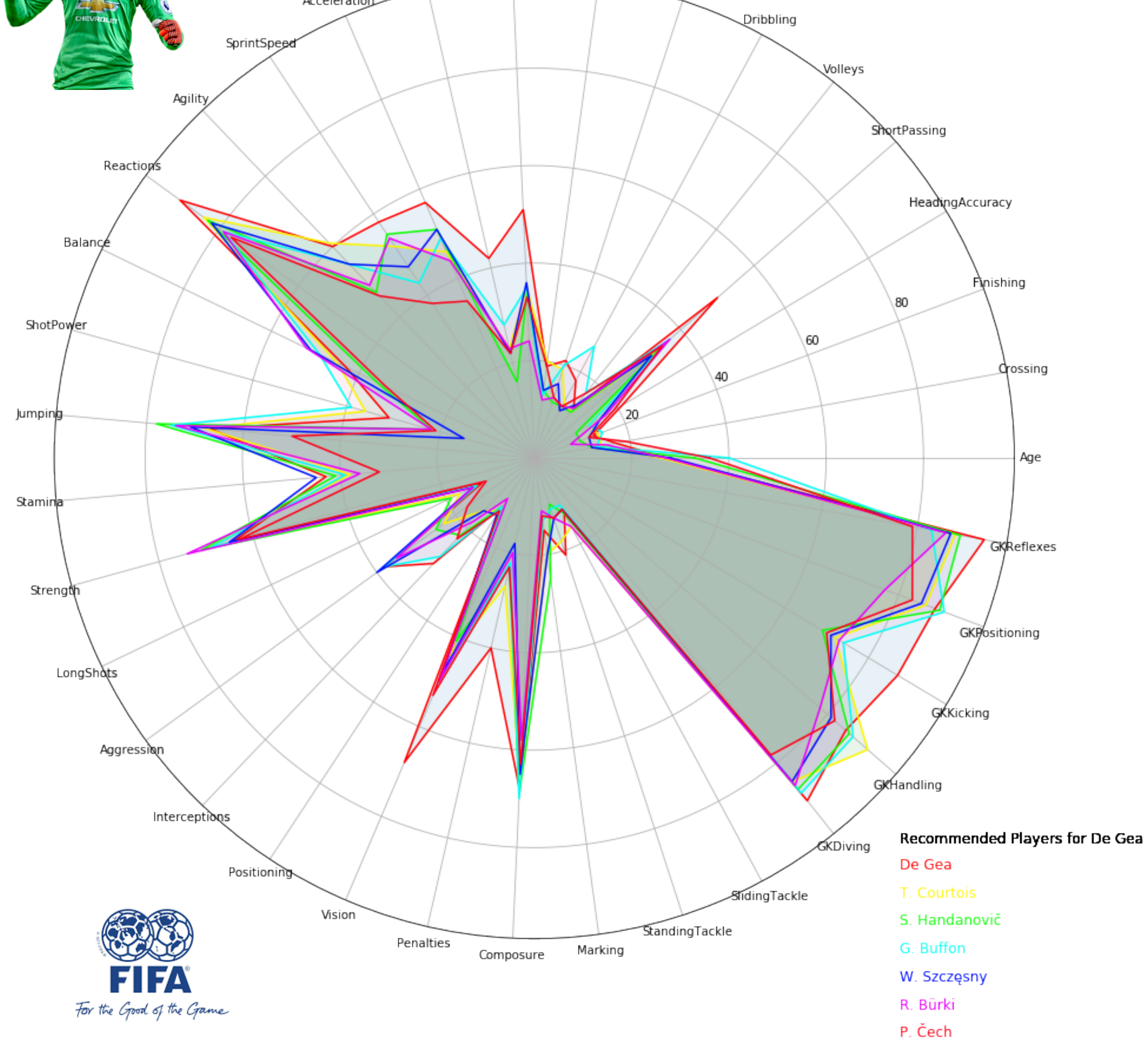
Here are the players' features

| | Age | Overall | Potential | Special | Preferred Foot | International Reputation | Weak Foot | Skill Moves | Real Face | Height | Weight | LS | ST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **G. Bale** | 28 | 88 | 88 | 2279 | 0 | 4.0 | 3.0 | 4.0 | 1 | 185 | 181 | 86 | 86 |
| **A. Griezmann** | 27 | 89 | 90 | 2246 | 0 | 4.0 | 3.0 | 4.0 | 1 | 175 | 161 | 86 | 86 |
| **M. Reus** | 29 | 86 | 86 | 2172 | 1 | 4.0 | 4.0 | 4.0 | 1 | 180 | 157 | 82 | 82 |
| **R. Lewandowski** | 29 | 90 | 90 | 2152 | 1 | 4.0 | 4.0 | 4.0 | 1 | 183 | 176 | 87 | 87 |
| **A. Sánchez** | 29 | 85 | 85 | 2172 | 1 | 4.0 | 3.0 | 4.0 | 1 | 170 | 163 | 81 | 81 |
| **P. Pogba** | 25 | 87 | 91 | 2247 | 1 | 4.0 | 4.0 | 5.0 | 1 | 193 | 185 | 81 | 81 |
| **I. Perišić** | 29 | 85 | 85 | 2199 | 1 | 3.0 | 5.0 | 4.0 | 1 | 185 | 176 | 82 | 82 |
| **E. Cavani** | 31 | 89 | 89 | 2161 | 1 | 4.0 | 4.0 | 3.0 | 1 | 185 | 170 | 85 | 85 |

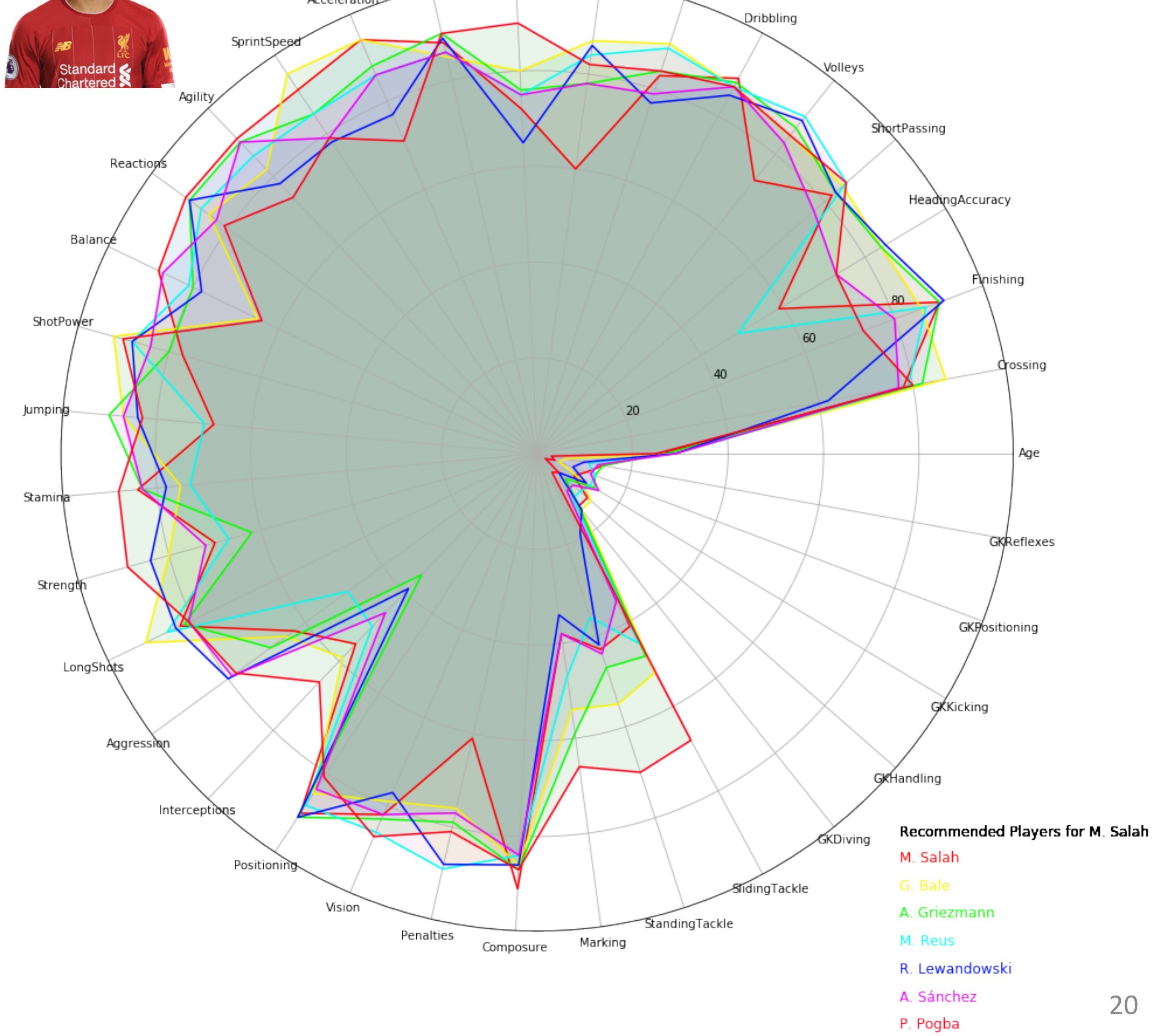
FIFA
For the Good of the Game

# Analyzing Recommendation Feature Similarities

## Recommendation Works for Goalkeepers...




## ...and for Forwards



# Anomaly Detection

# Client Pipeline Process

01

Create Restricted Set of Recommended Players

- K-Nearest Neighbors

02

Isolate Outlier Players

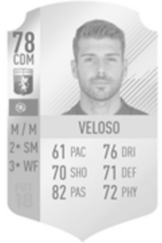
Anomaly Detection:
- SVM-One Class
- Local Outlier factor
- Isolation Forest
- DBSCAN

03

Predict Bid Price for Isolate Outlier Players

- Linear Regression
- Decision Tree
- Random Forest
- XGBoost
- SVR





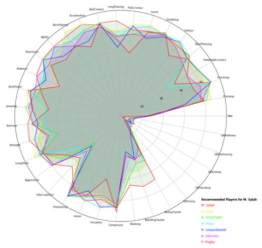


FIFA
For the Good of the Game

# Anomaly Detection Process

Give me 300 hundred players similar to **“M. Salah”**

**Recommendation Model**



**Anomaly Detection Model**
“Value”, “Wage”, “Release clause”

**300 players**

**Anomaly Detection**
- Value
- Wage
- Release Clause

**Normal players**

**Abnormal players**

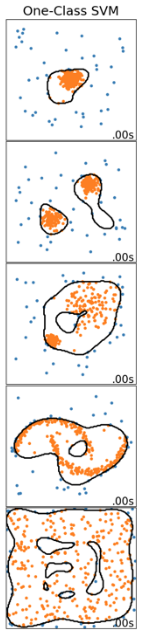
**< Mean value of 300 players**

**Under Valued Players**

# Anomaly Detection Methods

## How they work ?

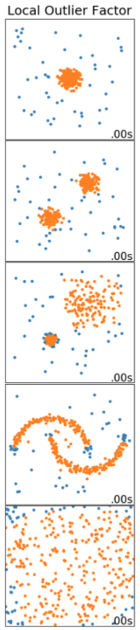
## Parametric: **OneClass SVM**




1. Provide normal training data
2. Algorithm creates a representational model of this data (boundary).
3. If newly encountered data is too different it is labeled as out-of-class.

**Suitable with novelty detection**

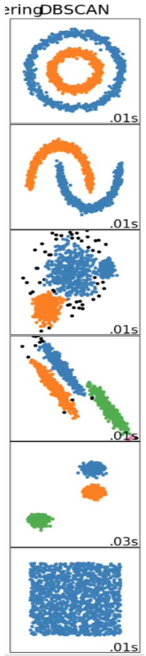
## Density Based: **LOF**




1. Pick a k value (# of neighbors)
2. Calculate k-distance as distance to kth neighbor
3. Smooth k-distance to get reachability distance = max[k-d & d(a,b)]
4. The local reachability density: lrd(a) = 1/(sum(reach-dist(a,n))/k)
5. Compare lrd of 'a' to its k-neighbors and get k-ratio
6. If k-ratio >1 : outlier

**Interpret k ratio depends on business knowledge and experience**

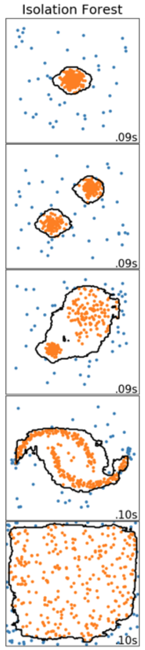
## Density Based: **DBSCAN**



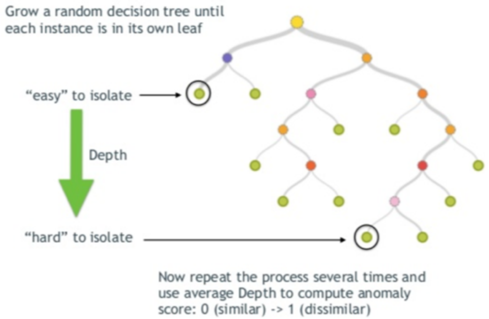

1. Define eps and min.samples
2. Core point if a minimum number of points are within a given distance
3. A point is reachable if there is a path consisting of core points from start to end
4. Any point that is not reachable is considered an outlier

**Depends on how we choose eps and min_samples**

## Ensemble: **Isolation Forest**




1. Build forest of decision trees
2. For each tree, select a random feature and a random split point.
3. Outliers should be identified closer to the roots of the trees on average >> score
4. S = 1: anomaly, S<0.5 normal
5. If all scores close to 0.5, then no clear anomalies.




---

**OneClass SVM**

Pros:
- Scales well to high dimensional data

Cons:
- Difficult to understand and interpret the final model
- Difficult to tune hyperparameters gamma & nu
- One-class SVM approach does not control over the false alarm rate (class imbalance)

**LOF**

Pros:
- Effective when the distribution of values in the feature space can not be assumed.
- Intuitive and easily interpretable

Cons:
- No specific rule of thumb to detect outlier based on k- ratio.
- Need to find appropriate distance metric
- Struggles with high dimensionality data

**DBSCAN**

Pros:
- Great at handling outliers within dataset
- Separates clusters of high/low density

Cons:
- Struggles with high dimensionality data
- Struggles with clusters of similar density

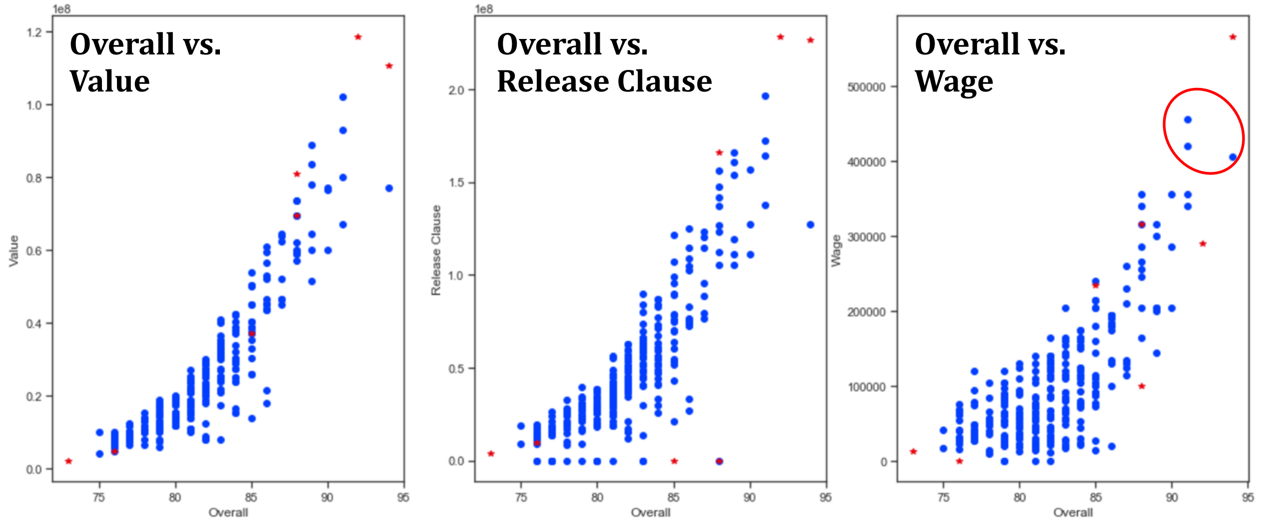
**Isolation Forest**

Pros:
- Can handle high dimensional data
- Low linear time-complexity and a small memory-requirement
- Does not employ distance/density and only considers isolation

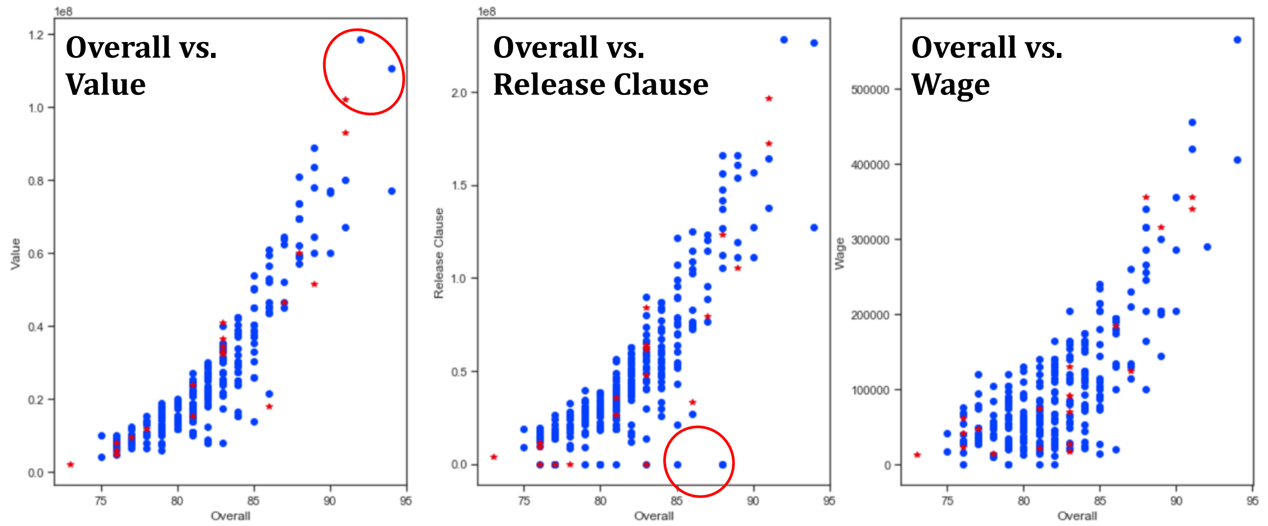
Cons:
- Not ideal if we have a model or good understanding of outliers (I.e. if there is training data)

# Anomaly Detection - Compare across methods

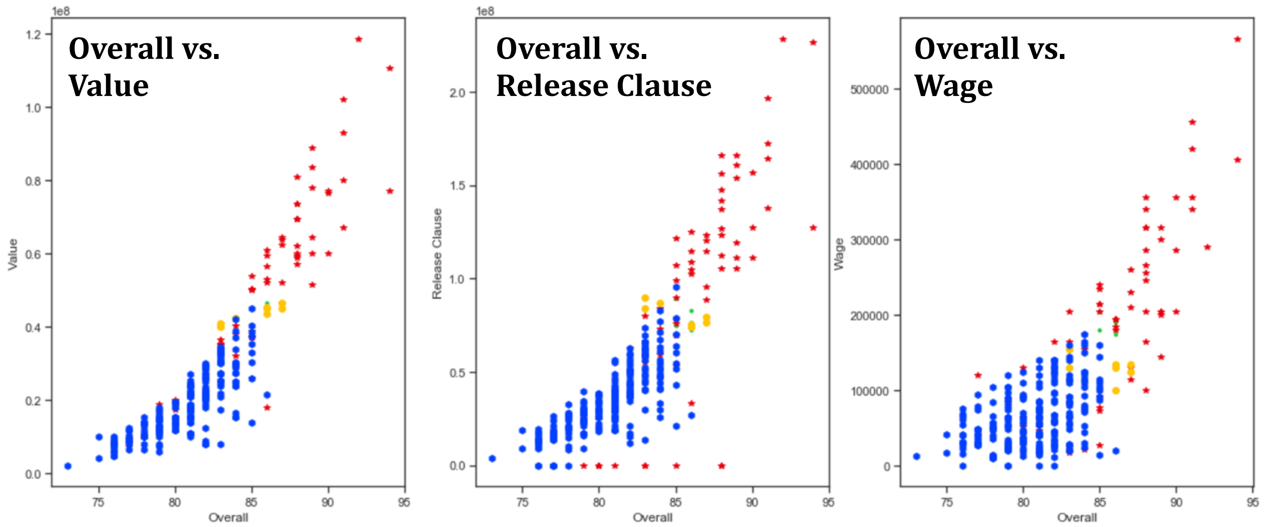
## SVM ONE CLASS



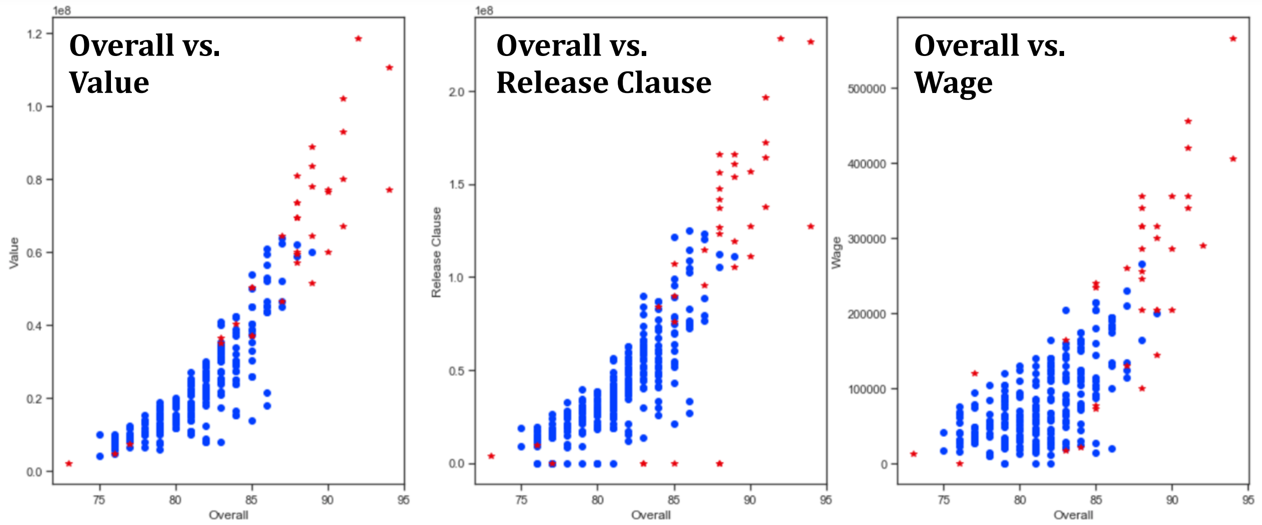

## LOCAL OUTLIER FACTOR




## DBSCAN




## ISOLATION FOREST



# Bid Prediction

# Client Pipeline Process

## 01
**Create Restricted Set of Recommended Players**

- K-Nearest Neighbors

## 02
**Isolate Outlier Players**

Anomaly Detection:
- SVM-One Class
- Local Outlier factor
- Isolation Forest
- DBSCAN

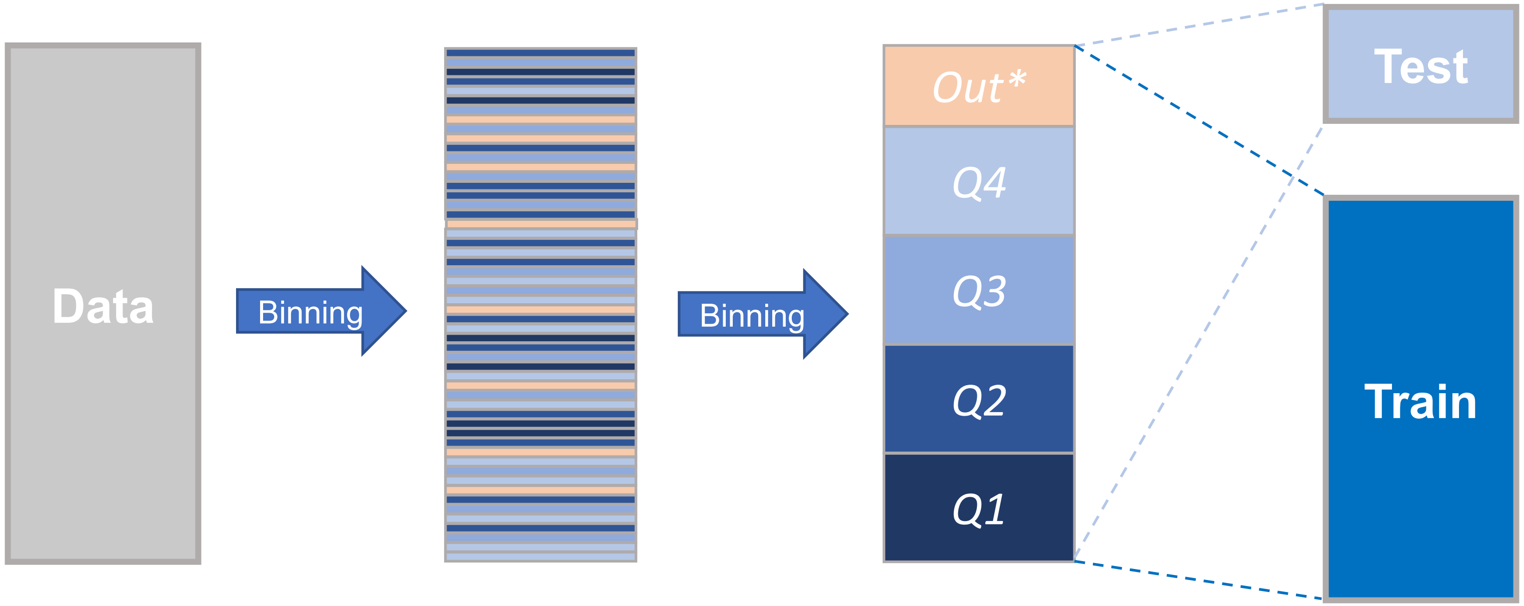
## 03
**Predict Bid Price for Player**

- Linear Regression
- Decision Tree
- Random Forest
- XGBoost
- SVR


FIFA
For the Good of the Game

# Bid Prediction – Data Pre-Processing

## 1 Stratified train and test sampling

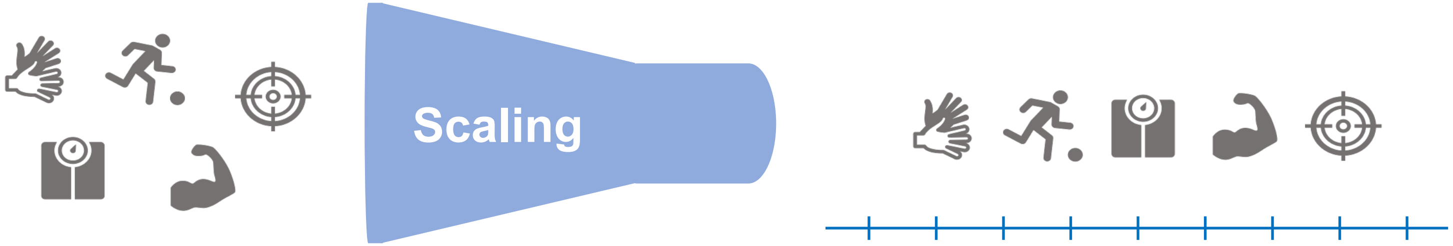



**Goal:**
Have same distribution of values in training and test set

- Stratified sampling of training and test set based on player value
- Outliers account for ~13% and build their own group
- Remaining data are binned based on quartiles

* $x > Q75_{value} + (Q75_{value} - Q25_{value}) * 1.5 = Outlier$

## 2 Scaling




**Goal:**
Normalizing range of independent features

- Scaling all numerical features that are not categorical
- After scaling, each feature has *mean = 0* and *standard deviation = 1*

FIFA
For the Good of the Game

# Bid Prediction – Model Consideration

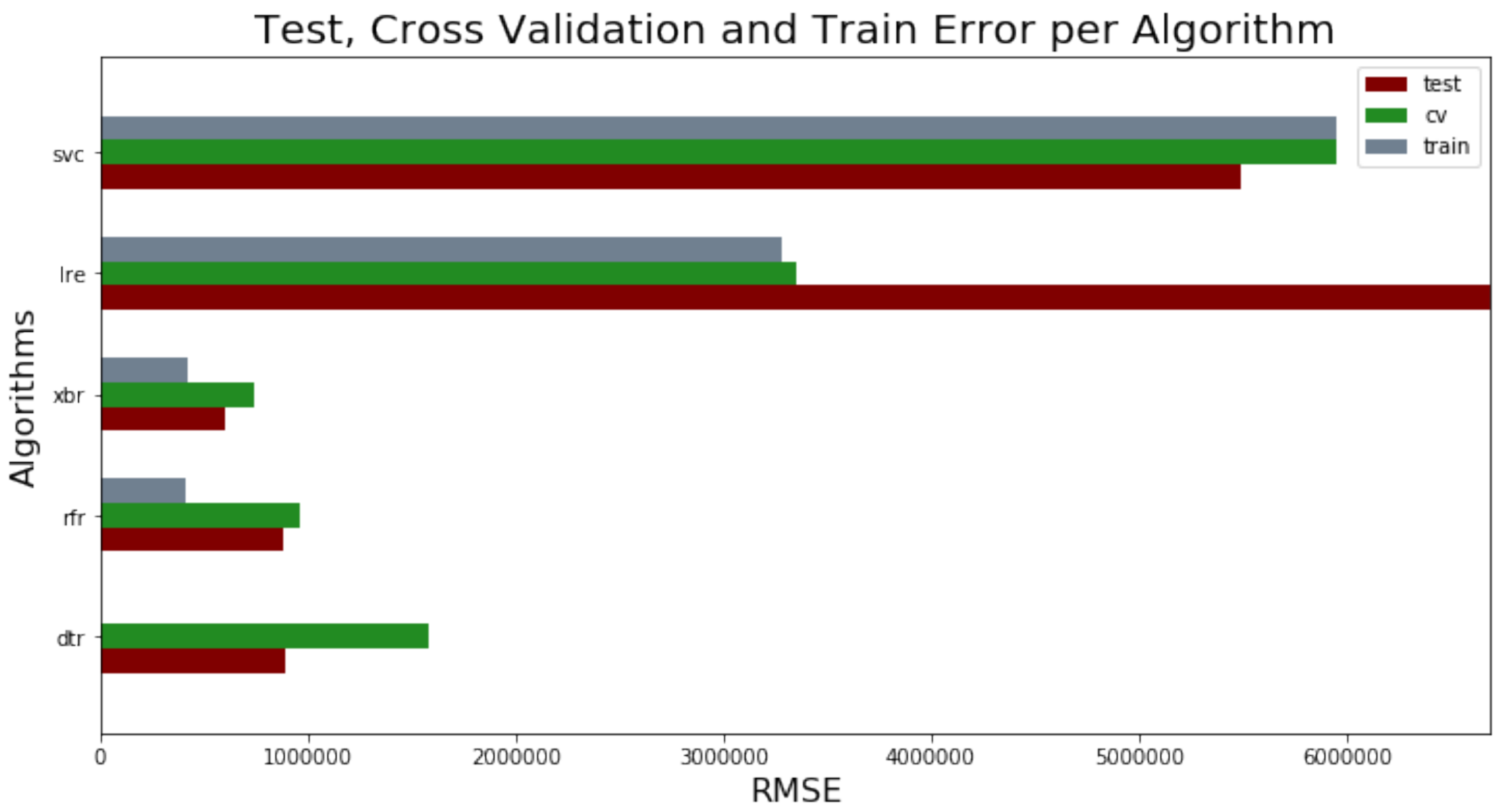
Linear

Non-linear

| Model Type | Strengths | Weaknesses |
|---|---|---|
| **Linear Regression** | • Simple<br>• Easy to understand relationships (Interpretable coefficients)<br>• Inference focused | • Poor performance with non-linear data relationships between dependent and independent variables<br>• Not naturally flexible enough to capture more complex patterns, and adding the right interaction terms & polynomials difficult. |
| **Support Vector Regression** | • Can handle non-linear relationships without changing the explanatory variables through "kernel trick"<br>• Effective in the higher dimension | • Difficult to tune hyperparameters<br>• Difficulty specifying the 'right' kernel function |
| **Decision Tree** | • Capable of understanding non-linear relationships<br>• Handles collinearity efficiently.<br>• No assumptions on distribution of data | • Greedy algorithm<br>• Prone to overfit when complexity not controlled |
| **Random Forest** | • Same as DT +<br>• More resistant to over-fitting<br>• RF is much easier to tune than GBM.<br>• Biased in favor of categorical variables with attributes with more levels | • Computationally expensive<br>• Not a well descriptive model over the prediction. |
| **Gradient Boosting** | • Same as DT +<br>• Learns sequentially<br>• Deals with unbalanced datasets better than RF | • Prone to overfit to noisy data<br>• Slower than RF because trees are built sequentially<br>• Harder to tune than RF |

FIFA
For the Good of the Game

# Bid Prediction – Baseline Model Results




- *Using RMSE as evaluation metric**
- Support Vector Regression most stable model
- Linear Regression with extremely high test error
- Decision Tree with virtually no training value
- Random Forest shows some variance, but has a relatively low bias overall
- *XGBoost with the best result, weighing variance and bias*

*

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

FIFA
For the Good of the Game

# Bid Prediction – Feature Selection

1

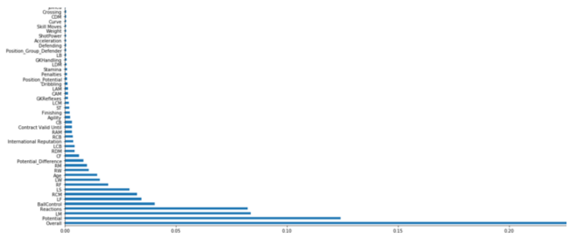
| | Specs | Score | Expl_percent |
|---|---|---|---|
| 5 | International Reputation | 10926.667128 | 1.020489e+01 |
| 1 | Overall | 9247.932429 | 8.637050e+00 |
| 75 | Club_Reputation | 8635.202466 | 8.064794e+00 |
| 2 | Potential | 7144.411856 | 6.672479e+00 |
| 54 | Reactions | 5942.582653 | 5.550038e+00 |
| 66 | Composure | 3632.566134 | 3.392613e+00 |
| 8 | Real Face | 3565.014584 | 3.329523e+00 |
| 3 | Special | 2416.453120 | 2.256832e+00 |
| 64 | Vision | 2126.810904 | 1.986322e+00 |
| 81 | Mentality | 1937.306594 | 1.809336e+00 |
| 44 | ShortPassing | 1773.834349 | 1.656662e+00 |

**F-Value:**

- Start with constant model $M_0$
- Try all models $M_1$ consisting of just one feature and pick the best according to the F statistic
- Try all models $M_2$ consisting of $M_1$ plus one other feature and pick the best

$$\frac{F_{score}}{\sum_{n=1}^{N} F_{score,n}} > 0.01$$

2

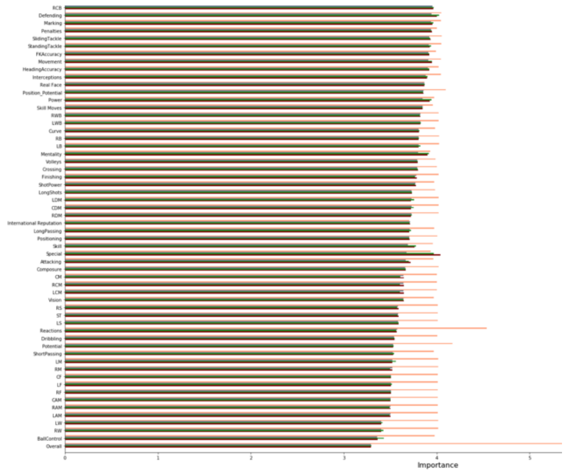


**Tree Regressor**

- Based on Extra Tree Regressor (Decision Tree with random splits)
- Total reduction of the criterion brought by that feature (Gini importance)
- Rank by total reduction

***Top ten features***

3



**RMSE-based:**

- Try all models $M_1$ consisting of just one feature and calculate the RMSE for each of the baseline models
- Rank by lowest RMSE

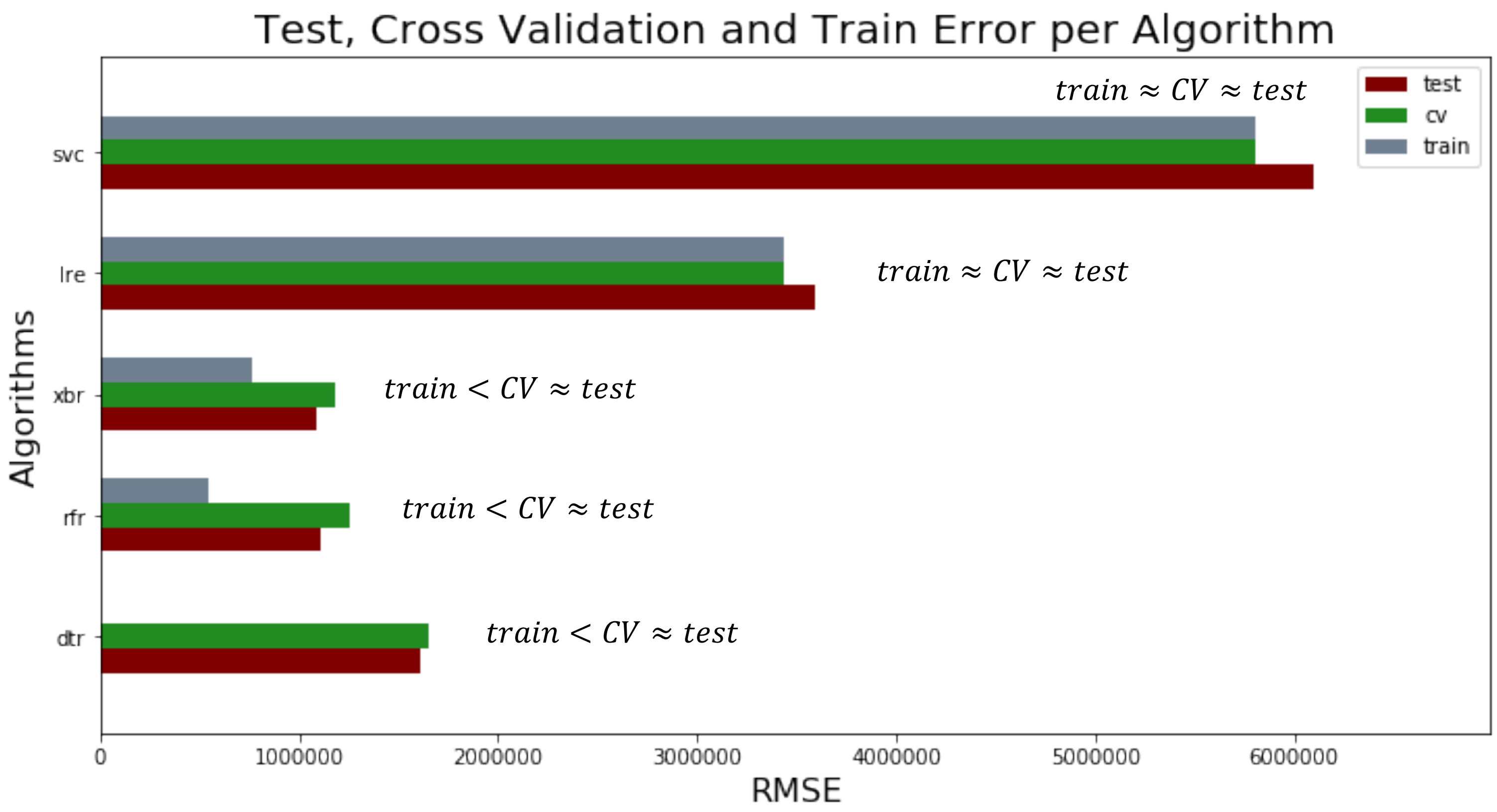
***Top ten features***

**Final feature selection**

FIFA

For the Good of the Game

# Bid Prediction – Prediction on Reduced Features




- Errors became more stable for most of the models, as compared to baseline model
- Especially Linear Regression improved significantly
- Bias similar to baseline models, therefore, we did not loose much information by reducing number of features

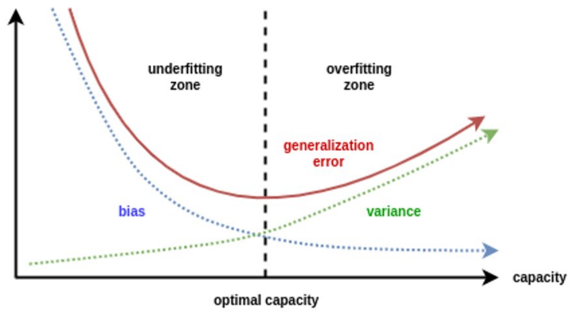
- XGBoost, Random Forest and Decision Tree show signs of overfitting
- Parameter tuning needed

FIFA
For the Good of the Game

# Bid Prediction – Parameter Tuning

## 1 Setting the goal

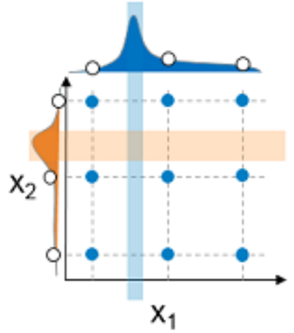
- *Problem:* Setting the optimal parameters for each model to find the sweet spot between variance and bias
- Decrease complexity for XGBoost, Random Forest and Decision Tree
- If possible, decrease bias without significantly increasing variance for all models




## 2 GridSearch

- GridSearch is an exhaustive method to find optimal hyperparameters

| Model | # of parameters | # of fits |
|---|---|---|
| Decision Tree | 4 | 8,000 |
| Random Forest | 4 | 243 |
| XGBoost | 5 | 324 |
| Support Vector Reg. | 2 | 60 |



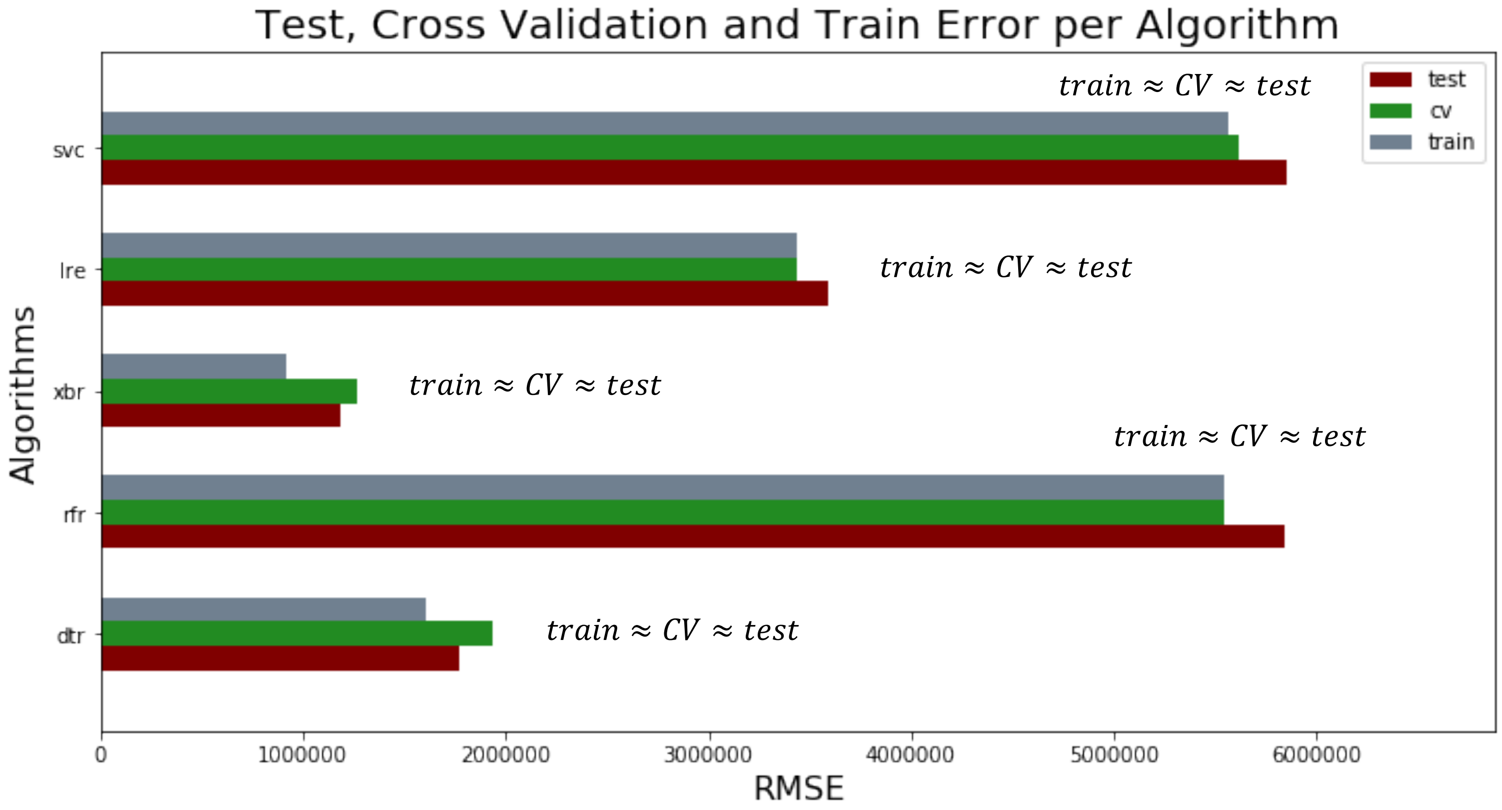
Standard Grid Search

## 3 Manual adjustments

- GridSearch is optimizing MSE, but not considering variance-bias tradeoff
- To balance variance and bias, manually adjustment is needed (Trial and Error process)

FIFA
For the Good of the Game

# Bid Prediction – Final Evaluation




- In terms of variance, all models are more or less stable
- XGBoost and Decision Tree show somewhat more variance than other models

- Lowest RMSE by far for XGBoost and maybe Decision Tree
- Even though XGBoost show a little more variance, we accept this in turn for a lower bias

**Using XGBoost as our model for final bid prediction**

For the Good of the Game

# Results

# Dashboard

Please choose between the 2 options below:

## Option 1: PLAYER

Select a player ▼


## Option 2: POSITION AND SCORE

Select a position ▼

CM
CDM
RW
…

Min Overall score ▼

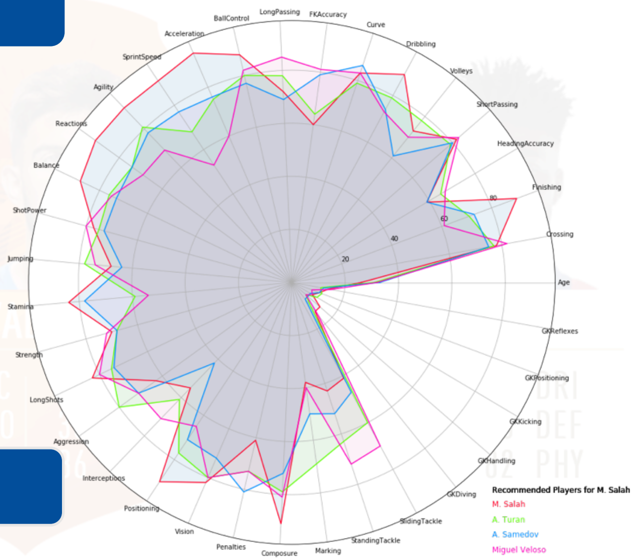
76

Max Overall score ▼

88

## Here is your first bid players suggestion




Suggested value:

| | | |
|---|---|---|
| 3,612,241.2 | 894,736.06 | 303,069.7 |

# Next Steps

# Next Steps

- Expand dataset to include historical data
- Incorporate intra-match statistics, including geospatial data as well as personal health data such as heart-rate monitoring
- Develop analytics to assess coaching style and style of play
- Maintain communication with the Chicago Fire for future potential projects

FIFA
For the Good of the Game

# Thank you!