Jackson Yu, Nicola Lambo, Naji El-Khouri

MAIS 202: DELIVERABLE 2 REPORT

1. Our goal is to make a model that could determine what players a given club should sign. Given that goals are the deciding factor in football matches, we decided to start simple and attempt to predict goals scored from other statistics, as the primary need for all clubs is to score goals.

2. We settled on the following two datasets:

   Premier League 2021 - 2022 player statistics:
   https://www.kaggle.com/datasets/omkargowda/football-players-stats-premier-league-20212022

   This dataset has 691 samples, and many relevant labels that describe all sorts of statistics for each of the 691 players. The labels are:
   - Player : Player's name
   - Team : Played club in 2021-2020
   - Nation : Player's nation
   - Pos : Position
   - Age : Player's age
   - MP : Matches played
   - Starts : Matches started
   - Min : Minutes played
   - 90s : Minutes played divided by 90
   - Gls : Goals scored or allowed
   - Ast : Assists
   - G-PK : Non Penalty Goals
   - PK : Penalty Kicks made
   - PKatt : Penalty Kicks attended
   - CrdY : Yellow Cards
   - CrdR : Red Cards
   - Gls : Goals scored per 90 mins
   - Ast : Assits per 90 mins
   - G+A : Goals and Assists per 90 mins
   - G-PK : Goals minus Penalty Kicks made per 90 mins
   - G+A-PK: Goals plus Assists minus Penalty Kicks made per 90 mins
   - xG : Expected Golas
   - npxG : Non-Penalty Expected Goals
   - xA : Expected Assits
   - npxG+xA : Non-Penalty Expected Goals plus Expected Assists
   - xG : Expected Golas per 90 mins
   - npxG : Non-Penalty Expected Goals made per 90 mins
   - xA : Expected Assits made per 90 mins

- npxG+xA : Non-Penalty Expected Goals plus Expected Assists made per 90 mins

—---

European Leagues 2021 - 2022 player statistics:
https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats
This dataset contains 2921 samples and an incredible amount of labels which represent different statistics for each player. We will not be using every label, since there are far too many. The labels on the dataset are:

- Rk : Rank
- Player : Player's name
- Nation : Player's nation
- Pos : Position
- Squad : Squad's name
- Comp : League that squat occupies
- Age : Player's age
- Born : Year of birth
- MP : Matches played
- Starts : Matches started
- Min : Minutes played
- 90s : Minutes played divided by 90
- Goals : Goals scored or allowed
- Shots : Shots total (Does not include penalty kicks)
- SoT : Shots on target (Does not include penalty kicks)
- SoT% : Shots on target percentage (Does not include penalty kicks)
- G/Sh : Goals per shot
- G/SoT : Goals per shot on target (Does not include penalty kicks)
- ShoDist : Average distance, in yards, from goal of all shots taken (Does not include penalty kicks)
- ShoFK : Shots from free kicks
- ShoPK : Penalty kicks made
- PKatt : Penalty kicks attempted
- PasTotCmp : Passes completed
- PasTotAtt : Passes attempted
- PasTotCmp% : Pass completion percentage
- PasTotDist : Total distance, in yards, that completed passes have traveled in any direction
- PasTotPrgDist : Total distance, in yards, that completed passes have traveled towards the opponent's goal
- PasShoCmp : Passes completed (Passes between 5 and 15 yards)
- PasShoAtt : Passes attempted (Passes between 5 and 15 yards)
- PasShoCmp% : Pass completion percentage (Passes between 5 and 15 yards)
- PasMedCmp : Passes completed (Passes between 15 and 30 yards)
- PasMedAtt : Passes attempted (Passes between 15 and 30 yards)

- PasMedCmp% : Pass completion percentage (Passes between 15 and 30 yards)
- PasLonCmp : Passes completed (Passes longer than 30 yards)
- PasLonAtt : Passes attempted (Passes longer than 30 yards)
- PasLonCmp% : Pass completion percentage (Passes longer than 30 yards)
- Assists : Assists
- PasAss : Passes that directly lead to a shot (assisted shots)
- Pas3rd : Completed passes that enter the 1/3 of the pitch closest to the goal
- PPA : Completed passes into the 18-yard box
- CrsPA : Completed crosses into the 18-yard box
- PasProg : Completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area
- PasAtt : Passes attempted
- PasLive : Live-ball passes
- PasDead : Dead-ball passes
- PasFK : Passes attempted from free kicks
- TB : Completed pass sent between back defenders into open space
- PasPress : Passes made while under pressure from opponent
- Sw : Passes that travel more than 40 yards of the width of the pitch
- PasCrs : Crosses
- CK : Corner kicks
- CkIn : Inswinging corner kicks
- CkOut : Outswinging corner kicks
- CkStr : Straight corner kicks
- PasGround : Ground passes
- PasLow : Passes that leave the ground, but stay below shoulder-level
- PasHigh : Passes that are above shoulder-level at the peak height
- PaswLeft : Passes attempted using left foot
- PaswRight : Passes attempted using right foot
- PaswHead : Passes attempted using head
- TI : Throw-Ins taken
- PaswOther : Passes attempted using body parts other than the player's head or feet
- PasCmp : Passes completed
- PasOff : Offsides
- PasOut : Out of bounds
- PasInt : Intercepted
- PasBlocks : Blocked by the opponent who was standing it the path
- SCA : Shot-creating actions
- ScaPassLive : Completed live-ball passes that lead to a shot attempt
- ScaPassDead : Completed dead-ball passes that lead to a shot attempt
- ScaDrib : Successful dribbles that lead to a shot attempt
- ScaSh : Shots that lead to another shot attempt
- ScaFld : Fouls drawn that lead to a shot attempt
- ScaDef : Defensive actions that lead to a shot attempt
- GCA : Goal-creating actions
- GcaPassLive : Completed live-ball passes that lead to a goal

- GcaPassDead : Completed dead-ball passes that lead to a goal
- GcaDrib : Successful dribbles that lead to a goal
- GcaSh : Shots that lead to another goal-scoring shot
- GcaFld : Fouls drawn that lead to a goal
- GcaDef : Defensive actions that lead to a goal
- Tkl : Number of players tackled
- TklWon : Tackles in which the tackler's team won possession of the ball
- TklDef3rd : Tackles in defensive 1/3
- TklMid3rd : Tackles in middle 1/3
- TklAtt3rd : Tackles in attacking 1/3
- TklDri : Number of dribblers tackled
- TklDriAtt : Number of times dribbled past plus number of tackles
- TklDri% : Percentage of dribblers tackled
- TklDriPast : Number of times dribbled past by an opposing player
- Press : Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball
- PresSucc : Number of times the squad gained possession withing five seconds of applying pressure
- Press% : Percentage of time the squad gained possession withing five seconds of applying pressure
- PresDef3rd : Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball, in the defensive 1/3
- PresMid3rd : Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball, in the middle 1/3
- PresAtt3rd : Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball, in the attacking 1/3
- Blocks : Number of times blocking the ball by standing in its path
- BlkSh : Number of times blocking a shot by standing in its path
- BlkShSv : Number of times blocking a shot that was on target, by standing in its path
- BlkPass : Number of times blocking a pass by standing in its path
- Int : Interceptions
- Tkl+Int : Number of players tackled plus number of interceptions
- Clr : Clearances
- Err : Mistakes leading to an opponent's shot
- Touches : Number of times a player touched the ball. Note: Receiving a pass, then dribbling, then sending a pass counts as one touch
- TouDefPen : Touches in defensive penalty area
- TouDef3rd : Touches in defensive 1/3
- TouMid3rd : Touches in middle 1/3
- TouAtt3rd : Touches in attacking 1/3
- TouAttPen : Touches in attacking penalty area
- TouLive : Live-ball touches. Does not include corner kicks, free kicks, throw-ins, kick-offs, goal kicks or penalty kicks.
- DriSucc : Dribbles completed successfully
- DriAtt : Dribbles attempted
- DriSucc% : Percentage of dribbles completed successfully

- DriPast : Number of players dribbled past
- DriMegs : Number of times a player dribbled the ball through an opposing player's legs
- Carries : Number of times the player controlled the ball with their feet
- CarTotDist : Total distance, in yards, a player moved the ball while controlling it with their feet, in any direction
- CarPrgDist : Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal
- CarProg : Carries that move the ball towards the opponent's goal at least 5 yards, or any carry into the penalty area
- Car3rd : Carries that enter the 1/3 of the pitch closest to the goal
- CPA : Carries into the 18-yard box
- CarMis : Number of times a player failed when attempting to gain control of a ball
- CarDis : Number of times a player loses control of the ball after being tackled by an opposing player
- RecTarg : Number of times a player was the target of an attempted pass
- Rec : Number of times a player successfully received a pass
- Rec% : Percentage of time a player successfully received a pass
- RecProg : Completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area
- CrdY : Yellow cards
- CrdR : Red cards
- 2CrdY : Second yellow card
- Fls : Fouls committed
- Fld : Fouls drawn
- Off : Offsides
- Crs : Crosses
- TklW : Tackles in which the tackler's team won possession of the ball
- PKwon : Penalty kicks won
- PKcon : Penalty kicks conceded
- OG : Own goals
- Recov : Number of loose balls recovered
- AerWon : Aerials won
- AerLost : Aerials lost
- AerWon% : Percentage of aerials won

On this dataset, some simple processing was done, in particular, we converted all of the non-number features such as Nation, Position, etc, into numbers such that the model could use these features effectively. On the second dataset, there was a feature that was leading to inconsistent results and that we were very likely not to use anyway, so we decided to remove it entirely from the dataset, and was GcaDrib.

3. For now, we used a simple linear regression model, because we believed goals could feasibly be predicted with a linear regression given certain features. We later find out this is not so much the case.

   (a) We used pandas for data processing, matplotlib for graphing, numpy to perform matrix operations to build the linear regression model.

   (b) We based our splits on a relatively arbitrary basis, based on which statistics were relevant to the main figure we are trying to predict (Expected Goals), since we noticed an insignificant amount of change while choosing different splits. Those would become the training variables. The training/testing range would vary from 70%/30% to 30%/70%. No hyperparameters were used in the making of this model.

   (c) We tested our models using mean absolute error, mean squared error and R squared. We noticed that the models would overfit when given too many features, and so we decided to graph the R-squared values for each feature individually to see which ones had the most impact on their own, and future models will take this into major consideration.

   (d) We think the major challenge is more about processing the dataset that we want for testing, we sometimes need to assign numerical values for categorical data and we need to clean the dataset as well so it doesn’t have any empty cells and NaN’s. So for a simple model like this, it’s more about just getting the right dataset, and we just need to put that into our model.

4. This submodel that we have built will serve as a template for the other components for a larger predictive model that will provide the inquiring club with the optimal player for its team. Evaluating it will set a relatively flexible standard for the remaining components we are planning to put into place.

   For the first model, we use 3 features: Min' for Minutes Played, 'Ast' for Assists, and 'PK' for Penalty Kicks Scored and label being goals scored. And here is the metrics we get:

   ```
   MAE: 1.4125611625386008
   MSE: 4.713309285201418
   RMSE: 2.1710157266131027
   R-squared: 0.5063216352179305
   ```

   So on average, the model predicts 1.41 goals away from the real goal value. This is a very good result considering we only used three features. Additionally, the r-squared value is about 0.50. This tells us that these three features are somewhat related, but not entirely related to our label, which is the number of goals that we are trying to predict.

As for the second model, we used an exhaustive database to estimate the number of goals scored by a player during the current season. The metric that we used to measure an individual statistic’s effect on the dependent variable (goals scored in this context) was R squared errors. The closer a stat’s RSE is to 1, the more of an effect it has. The closer to -1 it is, the less it fits within the context of what it is we want to predict. The corresponding plot can be found using the GitHub link to the team’s repository (we couldn’t fit it on the doc): football-player-best-transfer-selector/graph1.png at main · nianlape/football-player-best-transfer-selector (github.com)

5. We are very likely to change our model drastically once we have a firm grasp on neural networks, since we were advised to use them for our idea by our TPM. For now, the linear regression model performs relatively well in the task of predicting goals scored when given the right amount of relevant features. However we have indications that linear regression may be far too simple to accurately predict goals given some features, as we are running into things such as negative R-squared values which have led us to conclude that other models must be explored to fully predict goals consistently.