PREDICTIVE MODELING OF SOCCER GOALS

Group #8:
Brandon Sigman
Zhexu Li
Shawn Oh
Yue Wu
Chia-Wei Hsieh

Nov. 11, 2021

PROJECT OBJECTIVE

- Analyze how different actions affects the probability of scoring a goal, and assess the strengths and weaknesses of player and teams' shot selections.

- Create machine learning model to predict various metrics of success in soccer, such as the probability of a given shot resulting in a goal

- Quantify the quality of shot opportunities in a game and establish a foundation for using a data science approach to modeling the game

DATASET

Wyscout Events Dataset:

dataset link: https://figshare.com/collections/Soccer_match_event_dataset/4415000/2
paper: https://www.nature.com/articles/s41597-019-0247-7

- Posted 6/2019, updated 1/2020
- A large dataset containing all notable actions recorded in a season of professional European soccer games, such as passes, fouls, shots, etc, and their corresponding metadata.
- Contains over 3 million actions recorded in 1941 matches played by 3603 players in 142 teams.
- Clean and detailed dataset with over 40k shots

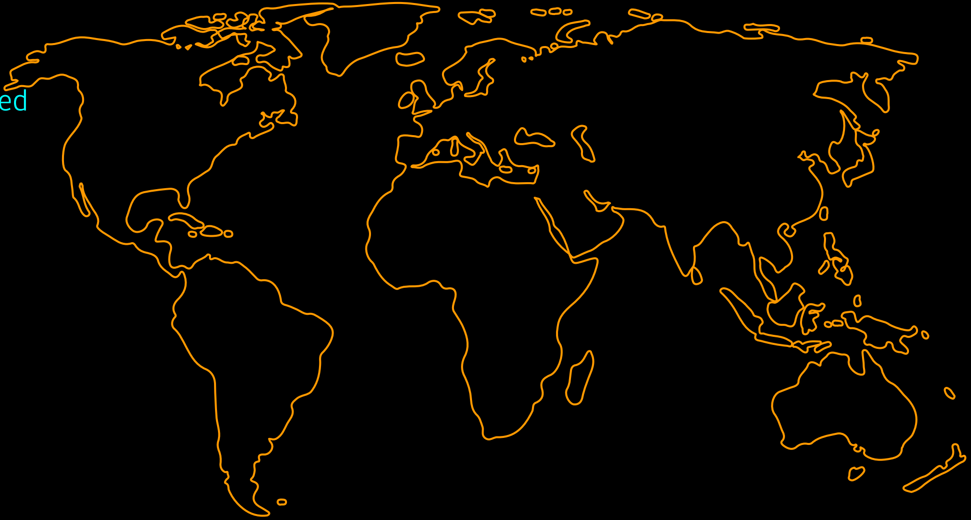
Data Analysis

- **Goal vs. Actions**

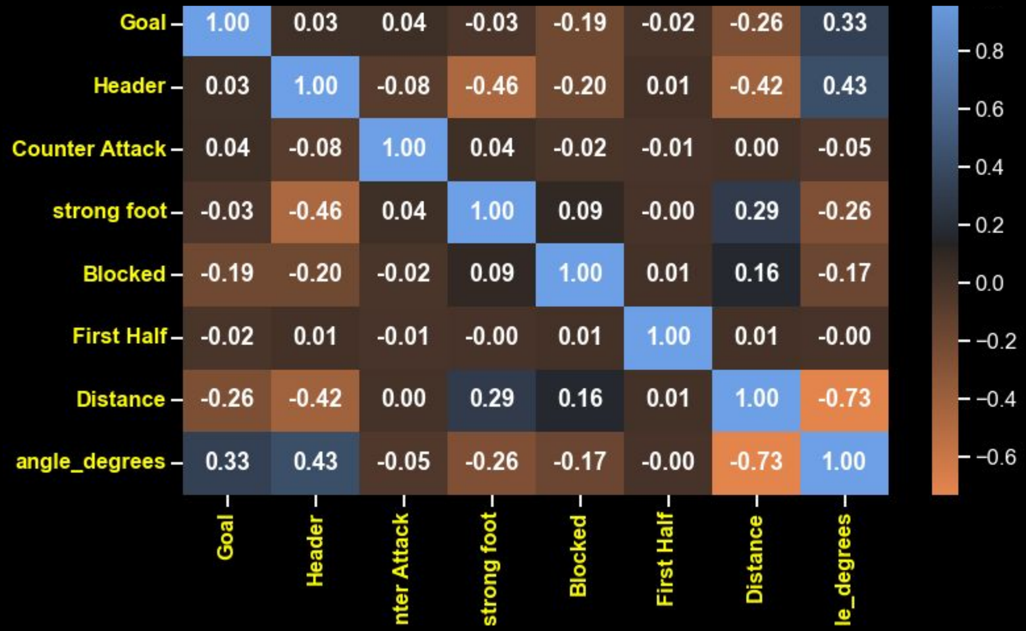
  Analyze how different shot conditions related to the probability of scoring.

- **Teams and Players**

  Analyze team and players performance.



Correlation Matrix



	Goal	Header	nter Attack	strong foot	Blocked	First Half	Distance	le_degrees
Goal	1.00	0.03	0.04	-0.03	-0.19	-0.02	-0.26	0.33
Header	0.03	1.00	-0.08	-0.46	-0.20	0.01	-0.42	0.43
Counter Attack	0.04	-0.08	1.00	0.04	-0.02	-0.01	0.00	-0.05
strong foot	-0.03	-0.46	0.04	1.00	0.09	-0.00	0.29	-0.26
Blocked	-0.19	-0.20	-0.02	0.09	1.00	0.01	0.16	-0.17
First Half	-0.02	0.01	-0.01	-0.00	0.01	1.00	0.01	-0.00
Distance	-0.26	-0.42	0.00	0.29	0.16	0.01	1.00	-0.73
angle_degrees	0.33	0.43	-0.05	-0.26	-0.17	-0.00	-0.73	1.00

Correlation between Goal and Distance: -0.26

Correlation between Goal and Angle: 0.33

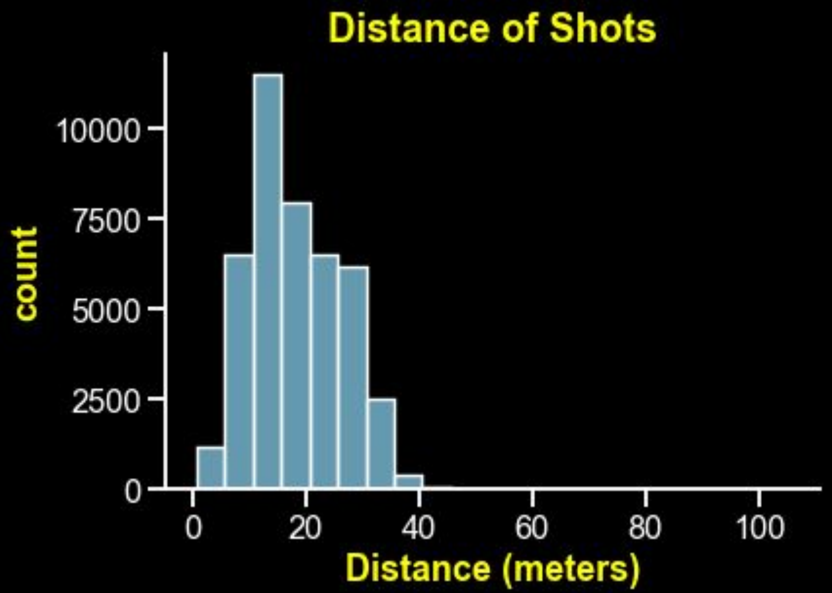
Correlation between Goal and Blocked: -0.19

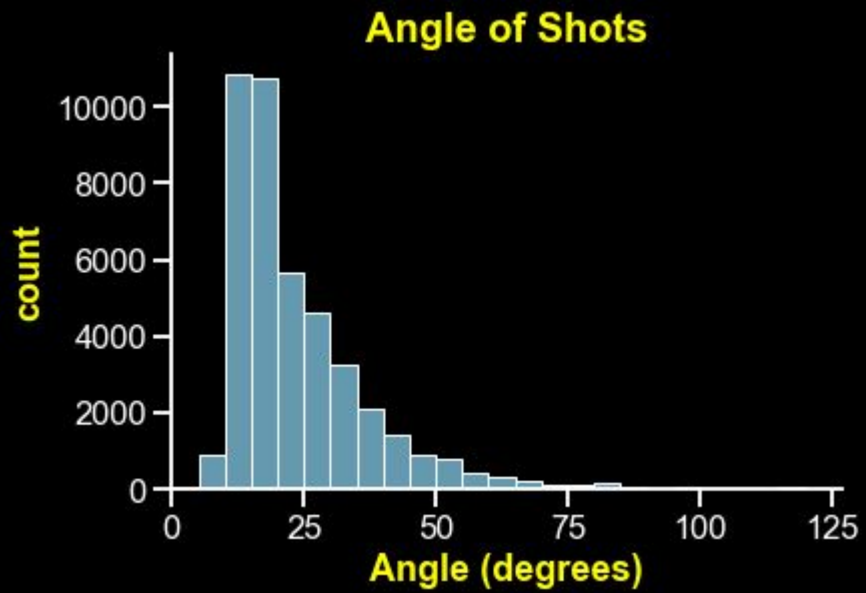
Other features don't have mentionable correlation with Goal.

Shot close and with high angle while not being blocked to have better chance of getting goals.

CORRELATION

Distance and Angle

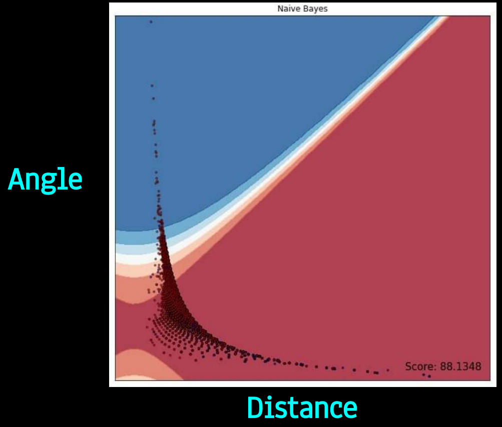




Over 95% shots occurred within 32 meters from the goal. Median distance is 16.8 meters.
Over 95% shots had angle between 5 to 52 degrees. Median angle is 20 degrees.

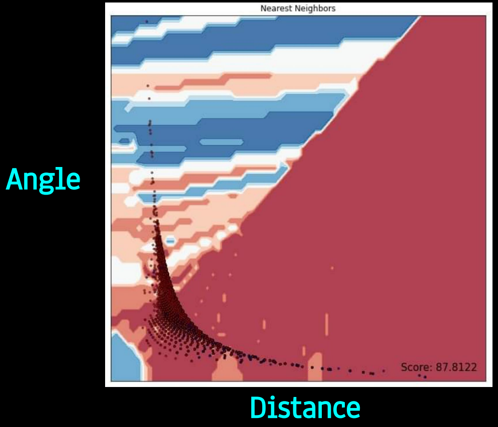
Analyze Angle and Distance with ML Classifier

- goal = 0 (bad)
- goal = 1 (good)



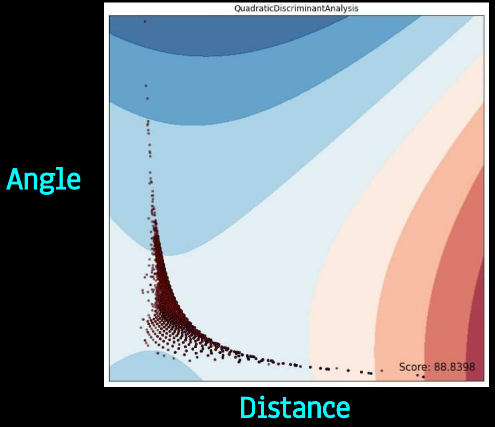
- **Gaussian Naive Bayes**: Compute the posterior probability for different distances and angles to predict the probability of goal. If P(g = 1) is larger than P(g = 0), guess 1.

  Model: GaussianNB()



- **K-Nearest Neighbors**: Find closest K points. The prediction is considered as the category that has most amount of points.

  Model: KNeighborsClassifier(k=5)

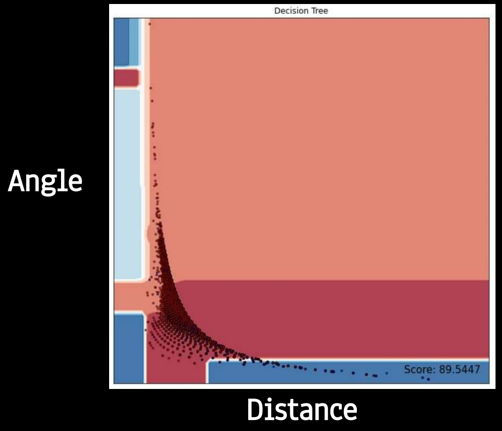


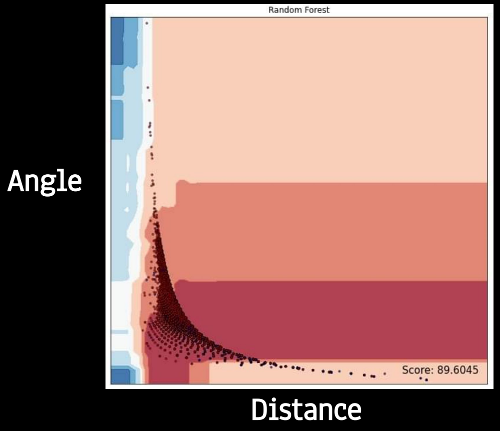
- **Quadratic Discriminant Analysis**: Another linear discriminant analysis algorithm but additionally calculates the covariance of two variables (here is angle and distance) to get the relationship between the variables.

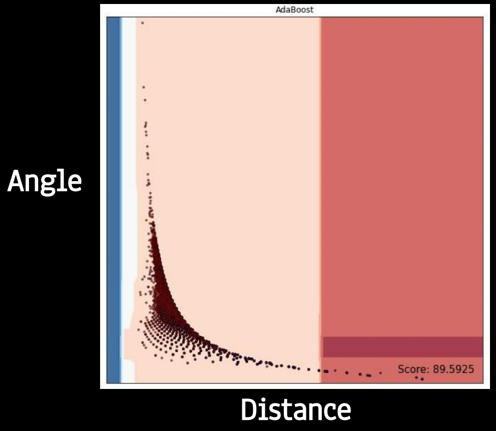
  Model: QuadraticDiscriminantAnalysis()

Analyze Angle and Distance with ML Classifier

- goal = 0 (bad)
- goal = 1 (good)







- **Decision Tree**:
  Learn to offer a series of questions through the features of training data, and then predict the category.

  Model:
  DecisionTreeClassifier(max_depth=5)

- **Random Forest**:
  Randomly allocate training data to build different Decision Trees, and take the majority decision as the prediction.

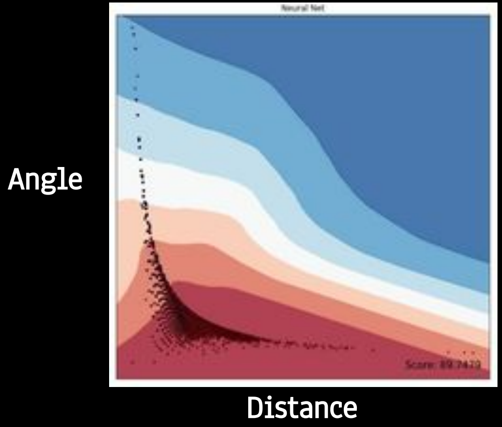
  Model:
  RandomForestClassifier(max_depth=5, n_estimators=10)

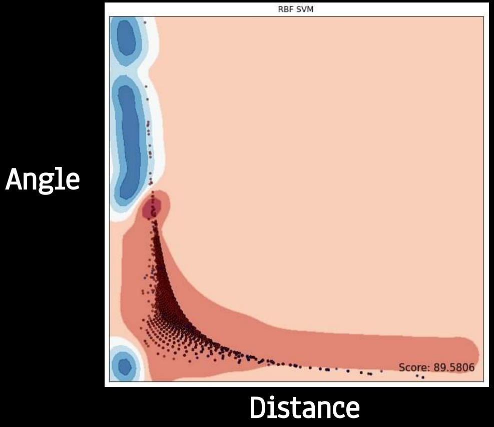
- **AdaBoost**:
  Initialize the weights distribution of the training data. Then update the weights when training Decision Tree. And adopted the updated weights of data to train the next Decision Tree.

  Model:
  AdaBoostClassifier()

Analyze Angle and Distance with ML Classifier

- goal = 0 (bad)
- goal = 1 (good)





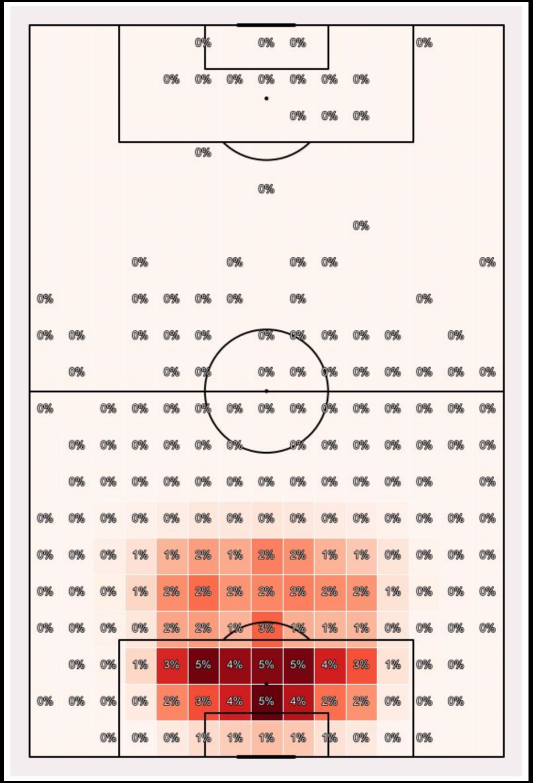
- **MLPClassifier**:
  Use neural network with default 100 hidden layers to train a model as a non-linear classifier.

  Model:
  MLPClassifier(alpha=0.01, max_iter=1000)

- **RBF SVM**:
  Project all data to higher dimensions so it can be easier to find linear hyperplanes for classifying.

  Model:
  SVC(gamma=2, C=1)

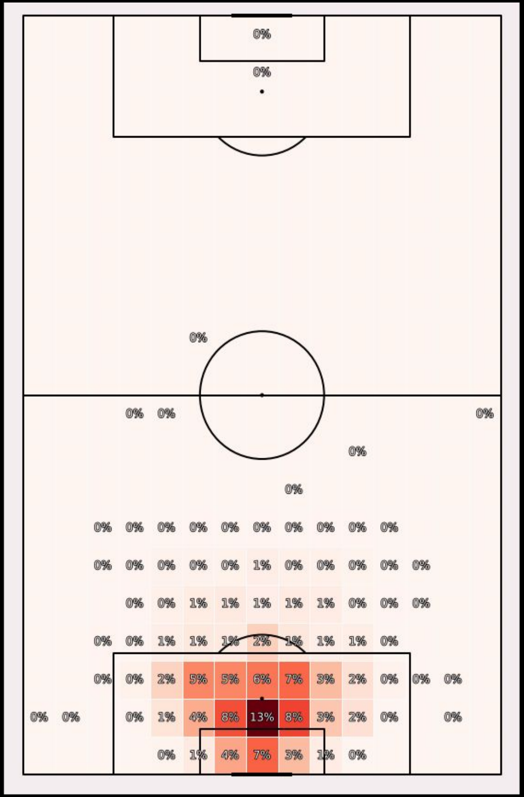
- In general, Distance ↑
  => The probability of bad shots↑
- From the KNN model, MLPClassifier and Gaussian Naive Bayes, we can observe the effect from different angles. But the skewness of the angle doesn't play a main role in prediction.
- All scores are above 87%
  => Prove that our original assumption of angle and distance as the main influence of goal is correct



Most shots were distributed inside the full back region.

Shots Distribution

Goals Distribution



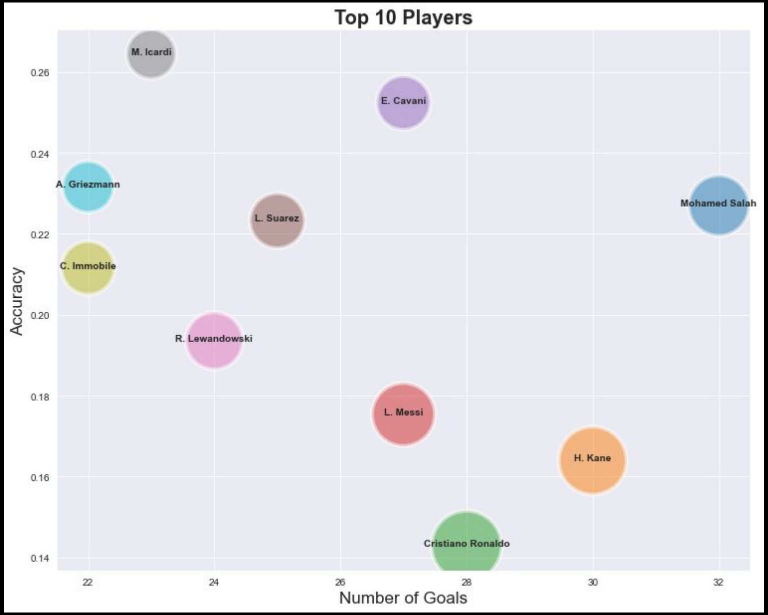
Most goals were distributed inside the penalty box.

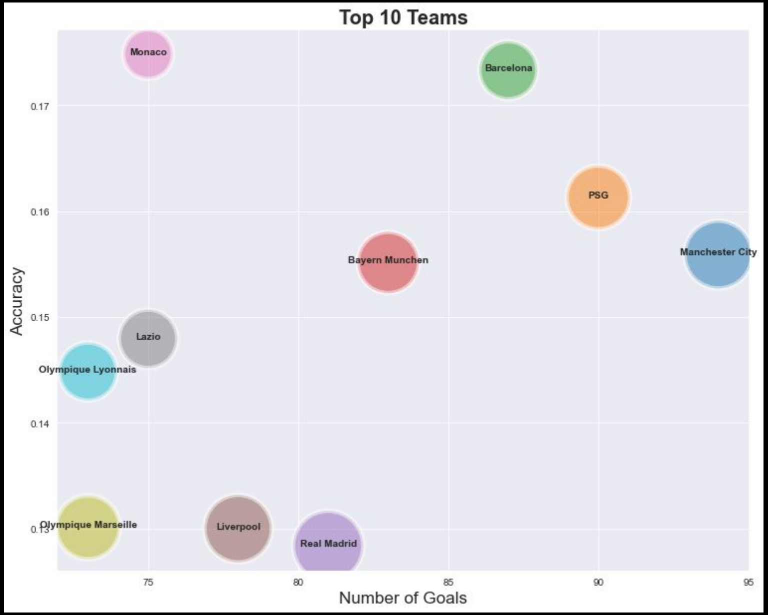
Distribution

Players and Teams

Mohamed Salah obtained the highest number of goals, and his shots were relatively accurate.

Similarly, Manchester city has the highest number of goals while maintained a relatively high accuracy.





Modeling Expected Goals

Prediction Models

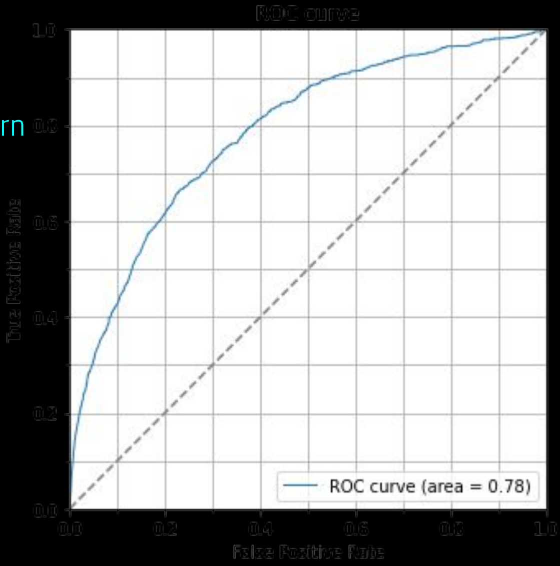
Python Module: soccer_xg

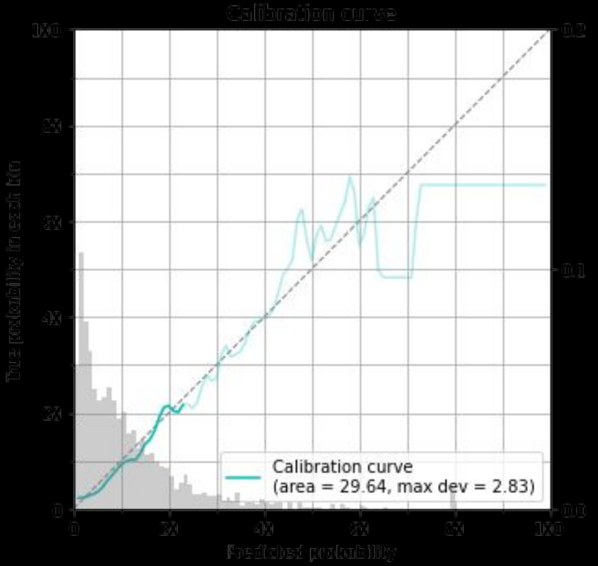
Default Scikit-learn model
basically a wrapper around three separate Scikit-learn pipelines

- open play shots
- free kicks
- penalties

Training Set: ESP, ITA, FRA and GER
Testing Set: ENG for validate and test

Parameter
Max Deviation: 31.46

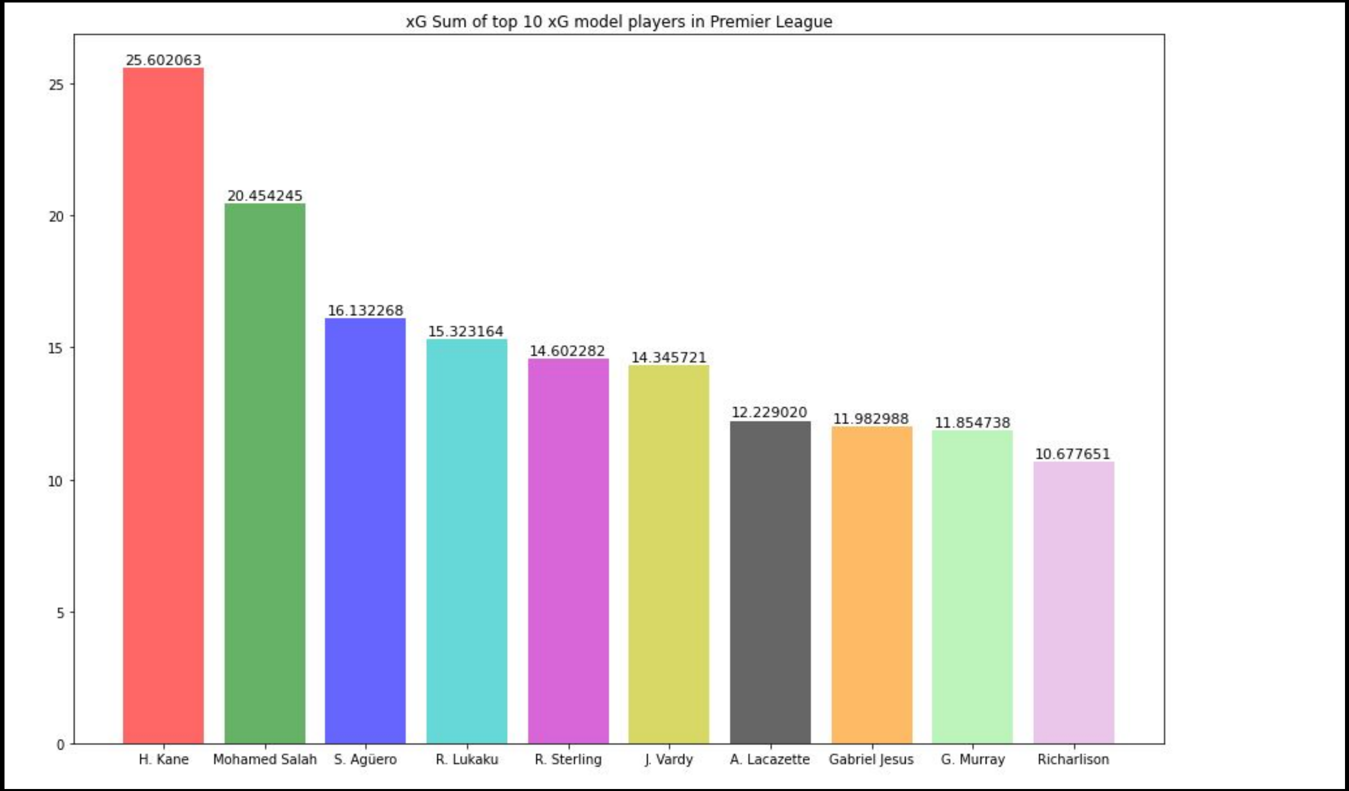




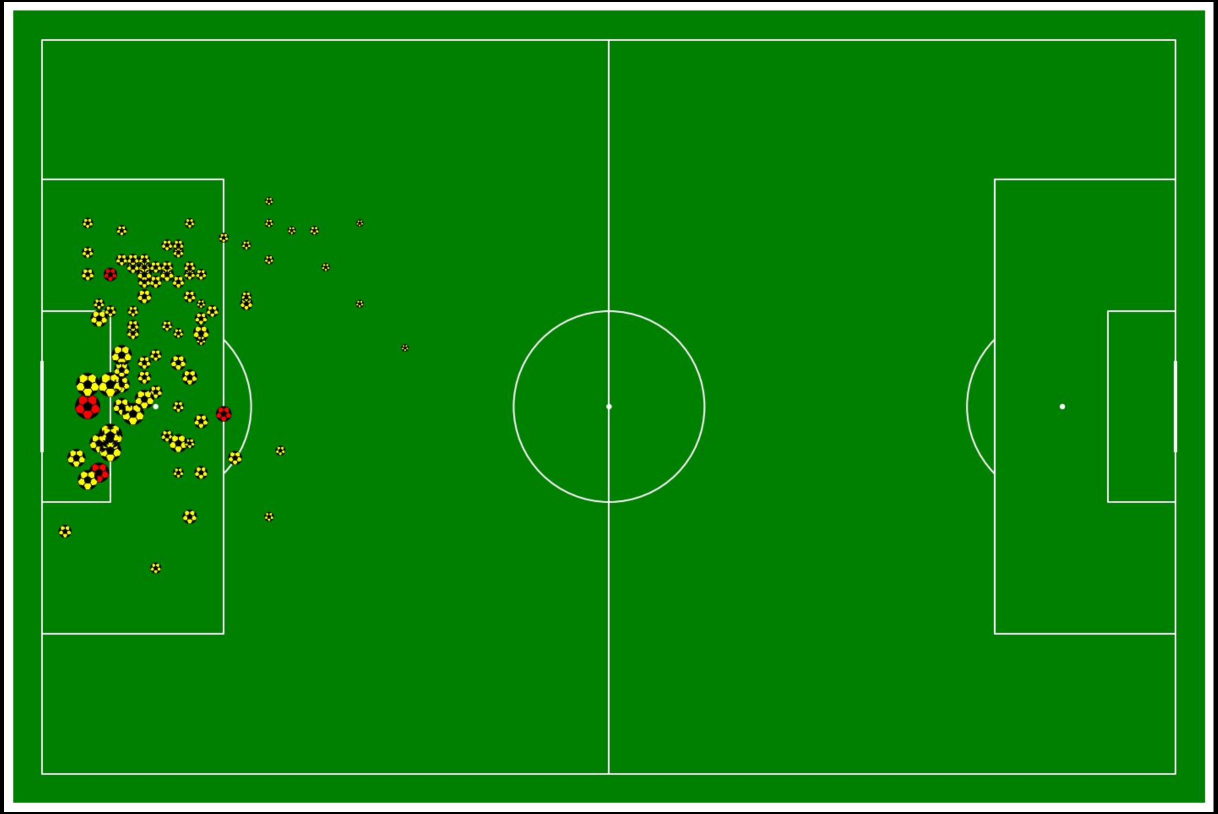
Logistic Regression vs Neural Network

Performance Metric	Logistic Regression	Neural Network
Balanced Error Rate	.28	.287
Precision	.238	.242
Recall	.708	.679
F1	.356	.356
Total Accuracy	.728	.739

Top 10 xG players in Premier League



Richarlison Xg map(inefficient)

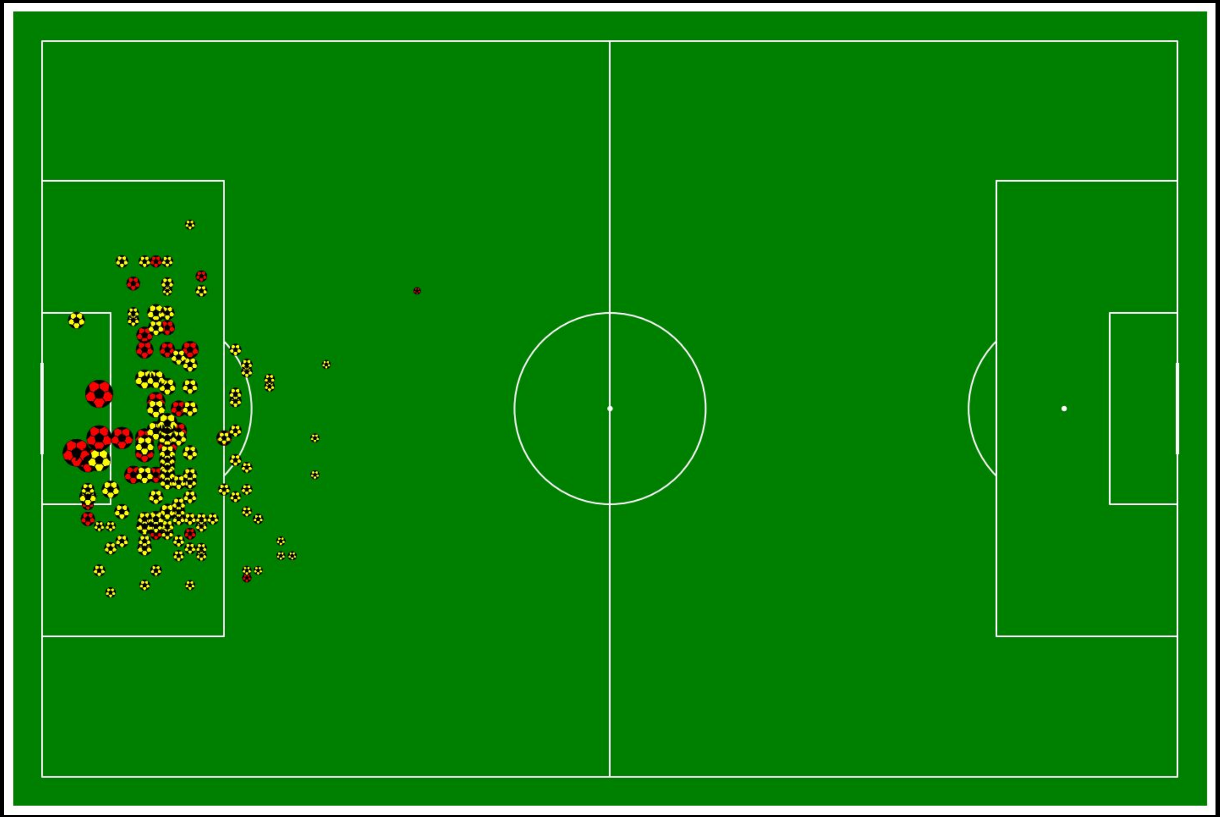


Bad Shot selection

.06 goals per shot

.116 xg per shot

Mo Salah Xg map (Efficient Shooter)



10% higher xg per shot than richarlison

.126 xg/shot

.227 goals/ shot

METHODOLOGY

Data Cleaning and Processing

- The data cleaning and feature extraction process were conducted using numpy and pandas.

Data Analysis

- Analysis was conducted using numpy and pandas.

Visualization

- Visualizations were created using matplotlib and seaborn

Models

- Logistic Regression and Neural Network trained with and without addressing class imbalance
- Decision Tree, Random Forest, AdaBoost, MLP, SVM, Gaussian, QDA and KNN trained on two variables - angles and distance to predict it as a bad or good goal
- Technical and theoretical details are included in the report notebook

ADDITIONAL RESOURCES

- https://figshare.com/collections/Soccer_match_event_dataset/4415000/2
- https://slidesgo.com/theme/soccer-player-portfolio#search-Sport&position-6&results-89

Thank you!