**IST 718 PROJECT REPORT**

**OMG – Oh My Goal!**An Analytical Overview of the Beautiful Game

Logo

Description automatically generated

**Project by:  
  
Team #21**

Shruti Varma | Shaunak Edakhe

Emmanuel Victor Kamya

Table of Contents

1. Abstract3
2. Data Collection/Cleaning3
3. Data Exploration Insights6
4. Methodology9
5. Model Prediction10

**5.1 Prediction #1 (Value of a shot)**10

5.1.1 Inference12

**5.2 Prediction #2 (Value of shot: Cristiano Ronaldo and Lionel Messi)**15

5.2.1 Inference17

**5.3 Prediction #3 (Predict time to score a goal based: Cristiano Ronaldo and Lionel Messi)**20

5.3.1 Inference21

6. Conclusion22

7. Roadblocks23

1. **Abstract**

With its origins being traced back to mid-19th century England, football has been one of the slowest adopters of modern analytics techniques. Big data is now being used to influence better decision-making. Organizations are utilizing end-to-end data pipelines consisting of data engineering, data analysis, and machine learning to maximize their chances of winning on-the-field as well as off it.

The Football Events dataset obtained from Kaggle contains more than 900,000 events from 9,074 football games across the top 5 leagues in Europe, based on text commentary (link to the dataset: <https://www.kaggle.com/datasets/secareanualin/football-events>) with reference from a project on the same dataset (link to project: <https://www.databricks.com/blog/2018/07/09/analyze-games-from-european-soccer-leagues-with-apache-spark-and-databricks.html>.) The dataset contains 951,121 rows and 41 columns. The top 5 leagues include the English Premier League, the German Bundesliga, the La Liga in Spain, Serie A in Italy, and Ligue 1 in France. The data spans from the 2011/2012 season to the 2016/2017 season, making up data worth 6 seasons of football.

3 predictions will be made on this data set including an Expected Goals model for the English Premier League, using on-field factors to predict the value of a shot for Cristiano Ronaldo and Lionel Messi, and Recommending a probable time to score a goal given an opponent.

Here are some interesting insights from the data exploration stage of the analysis to get started:

* About 16 players scored more than 75 goals in 6 seasons including Cristiano Ronaldo and Lionel Messi
* Own goals always found to cross the goal line at the centre of the goal.
* During 6 seasons, Messi didn’t receive any red cards while Ronaldo received 2.

1. **Data Collection/Cleaning**

The main dataset consists of individual game-related events ordered chronologically and includes significant data such as:

* id\_odsp - unique identifier of game
* time - minute of the game
* event\_type - primary event
* event\_team - the team that produced the event
* player - name of the player involved in the main event
* shot\_place - placement of the shot, 13 possible placement locations
* shot\_outcome - 4 possible outcomes
* location - location on the pitch where the event happened, 19 possible locations
* is\_goal - binary variable if the shot resulted in a goal (own goals included)

Below are the steps for data cleaning and engineering:

1. To define the schema by defining the column name, column data type
2. Deal with missing values by filling them either with NA for string type columns and 99 for numerical columns
3. Create a dictionary for that will be used for label encoding for each categorical variable**.** The categorical variables will be mapped using a user defined function named **mapVals**
4. Map the categorical columns using mapVals and join the eventsDf dataframe to the gameinfDf data frame

When mapping game related events, below are the categories for each variable:

|  |  |
| --- | --- |
| Event Name | Event |
| evtTypeMap | 0: Announcement  1: Attempt  2: Corner  3: Foul  4: Yellow Card  5: Second yellow card  6: Red card  7: Substitution  8: Free kick won  9: Offside  10: Hand ball  11: Penalty conceded  99: NA |
| evtTyp2Map | 12: Key Pass  13: Failed through ball  14: Sending off  15: Own goal  99: NA |
| sideMap | 1: Home  2: Away |
| shotPlaceMap | 1: Bit too high  2: Blocked  3: Bottom left corner  4: Bottom right corner  5: Centre of the goal'  6: High and wide  7: Hits the bar  8: Misses to the left  9: Misses to the right  10: Too high  11: Top centre of the goal  12: Top left corner  13: Top right corner  99: NA |
| shotOutcomeMap | 1: On target  2: Off target  3: Blocked  4: Hit the bar  99: NA |
| locationMap | 1: Attacking half  2: Defensive half  3: Centre of the box  4: Left wing  5: Right wing  6: Difficult angle and long range  7: Difficult angle on the left  8: Difficult angle on the right  9: Left side of the box  10: Left side of the six yard box  11: Right side of the box  12: Right side of the six yard box  13: Very close range  14: Penalty spot  15: Outside the box  16: Long range  17: More than 35 yards  18: More than 40 yards  19: Not recorded  99: NA |
| bodyPartMap | 1: Right foot  2: Left foot  3: Head  99: NA |
| assistMethodMap | 0: None  1: Pass  2: Cross  3: Headed pass  4: Through ball  99: NA |
| situationMap | 1: Open play  2: Set piece  3: Corner  4: Free kick  99: NA |
| countryCodeMap | germany: DEU  france: FRA  england: GBR  spain: ESP  italy: ITA |

1. **Data Exploration Insights Using Standard Statistical Techniques**

**Teams that scored more than 300 goals**

Barcelona and Real Madrid (both from Spain) have the highest percentage of goals scored, with 8.94% and 8.75% respectively.

**Chart, pie chart

Description automatically generated**

**Top scoring players, filtered on players who scored more than 75 goals.**

About 16 players scored more than 75 goals in 6 seasons including Cristiano Ronaldo and Lionel Messi being the players that scored the most goals.

Chart, bar chart

Description automatically generated

**Shot placement location plot**

The scatter plot shows the goals scored at different time frames in the game. Based on the selection of these data points, we observe the influence it has on each categorical variable like Region, Pass, Shot location. It can be observed by the blue plots that there are many goals scored at the center of the goal.

**Chart, box and whisker chart

Description automatically generated**

**Top player shot placement and goal timing plots**

We use two plots below to visualize and understand the scoring patterns for 4 top strikers, namely:

* Edinson Cavani
* Lionel Messi
* Thomas Muller
* Pierre Emerick-Aubameyang

The Sankey plot below shows the top 4 goal scorers mapping the foot used to shoot and the location of the goal where the ball entered, including bottom right corner, top left corner, top right, and bottom left as well as the center of the goal. Goals scored on different time frames can help us see how they influence each player and what body part they use to score the goal and the location for the shot used to score the goal.

Graphical user interface

Description automatically generated

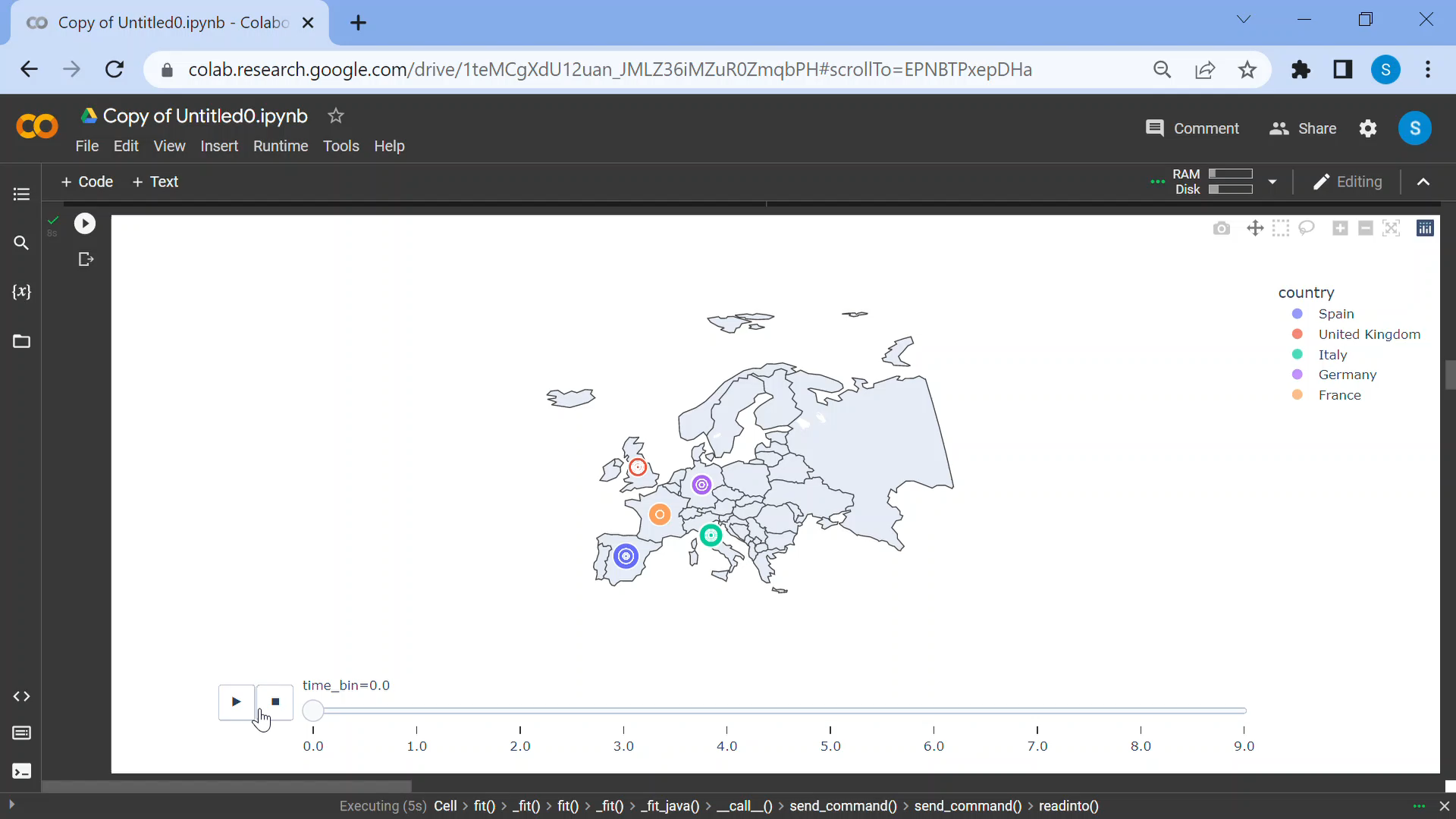
The graph below shows the goals scored by the Top 4 players based on Home and Away metrics - the sub plots are categorized based on the location of the shot on the goal at different times of the game. The plots show Messi dominating the goals shot on the bottom left at various game timings

Graphical user interface, application

Description automatically generated

**Visualizations showing how each region scores at different time frames of the game**

An animation containing the map of Europe below shows how scores for each region changed at every point in time during the 6 seasons, for the countries Spain, United Kingdom, Italy, Germany, and France. The bigger the size of the circle the more goals scored. Spain is seen to have the most goals for each time bin.

****

A supporting graph that visualizes the goal shot placement and timing bins when the goal was scored, grouped by a country basis.

**Text

Description automatically generated with medium confidence**

1. **Methodology**

By its very nature, football is a low scoring game and so, it would be difficult to build an accurate model that predicts whether a team/player will score using on-field factors. However, since many chances are created during a match that can easily lead to a goal, we will be predicting what we term as an **expected goal.** To predict an expected goal, we use the on-field events including:

1. type of event
2. team
3. shot placement location
4. assist method
5. situation and
6. country code

We are trying to predict whether the on-field factors above will predict whether a goal will be scored or not. This is therefore a classification problem.

Below are the machine learning algorithms implemented to answer various questions throughout this project:

1. Random Forest Classifier
2. Gradient Boosting Classifier
3. Multilayer Perceptron Classifier
4. Euclidean Distance

For each of the machine learning algorithms, a grid search was performed to find the optimized hyperparameters that will generate accurate predictions.

1. **Model Prediction**

**5.1 Prediction #1: Value of a shot**

Predicting **the value of a shot** for the English Premier League across the 6 seasons using features below:

1. type of event
2. team
3. shot placement location
4. assist method
5. situation and
6. country code

Within the machine learning pipeline, String indexing, one hot encoding, and vector assemblers were implemented on the above columns.

The data was split into **training validation** and **testing** with **75%, 12.5% and 12.5%** respectively. The data was trained, validated, and tested Gradient Boosting, Random Forest, and Multilayer Perceptron. For the three machine learning algorithms, the predictions were evaluated.

For each of the machine learning algorithms, a grid search was performed to find the optimized hyperparameters that will generate accurate predictions.

**Gradient Boost grid search:**

Text

Description automatically generated

**Random Forest grid search:**

Text

Description automatically generated

**Multilayer Perceptron grid search:**

Text

Description automatically generated

The 3 machine learning algorithms, Gradient Boosting, Random Forest, and Multilayer perceptron were optimized using the hyperparameters below:

|  |  |
| --- | --- |
| **Machine Learning Algorithm** | **Hyperparameters** |
| Gradient Boosting | 1. Max Depth = 5 2. Max Iterations = 20 |
| Random Forest | 1. Max Depth = 5 2. Number of trees = 7 |
| Multilayer Perceptron | 1. Step size = 0.2 2. Max Iterations = 200 3. Training size = 0.5 4. Number of neurons = 10 |

**5.1.1 Inference: Value of a shot**

The multilayer perceptron is observed to have the highest accuracy.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Gradient Boosting | 73.35% |
| Random Forest | 72.97% |
| Multilayer Perceptron | 97.72% |

**5.1.1.1 Evaluation of the Gradient Boost model**

**ROC curve**

The ROC curve below is very close to the ideal vector (1,1). This indicates that the GBT is performing as expected and making predictions especially when it comes to distinguishing between a goal or no goal.

**Shape, square

Description automatically generated**

**Precision and Recall Curve**

An excellent model has AUC near to the 1 which means it has a good measure of separability. The Area Under the PR Curve however relatively smaller with an area under curve of 0.59. This means that the GBT model performed separability between goal and no goal at an accuracy of 73.35%.

**Chart, line chart

Description automatically generated**

**5.1.1.2 Evaluation of the Random Forest model:**

**ROC curve**

The ROC curve below is very close to the ideal vector (1,1). This indicates that the Random Forest is performing as expected and making predictions especially when it comes to distinguishing between a goal or no goal.

Shape, square

Description automatically generated

**Random Forest Precision and Recall Curve**

An excellent model has AUC near to the 1 which means it has a good measure of separability. The Area Under the PR Curve however relatively smaller with an area under curve of 0.66. This means that the Random Forest model performed separability between goal and no goal at an accuracy of 72.97%.

Chart, line chart

Description automatically generated

* + - 1. **Evaluation of the Multilayer Perceptron model:**
* **Accuracy**: An accuracy of **97.72%** shows how well the model performed as a whole. This means that 97.72% of predictions of whether it was a goal or not a goal were predicted correctly.
* **Precision**: A precision of **98.46%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **Recall**: A precision of **99%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **F1**: An F1 score of **97%** shows that the model has a high rate of false positives and false negatives.

**5.2: Prediction #2 Value of a shot (Cristiano Ronaldo and Lionel Messi)**

Predicting **the value of a shot** for the **Cristiano Ronaldo** and **Lionel Messi** across the 6 seasons using features below:

1. type of event
2. team
3. shot placement location
4. assist method
5. situation

To perform prediction, separate data set for each player had to be extracted. These dataframes are **ronaldo\_df** and **messi\_df.** Each data set was trained, validated, and tested using logistic regression and multilayer perceptron.

In order to the fit the data into the model, the following feature engineering was performed: string indexers, one hot encoders and vector assemblers.

Each data was split into **training validation** and **testing** with **75%, 12.5% and 12.5%** respectively. The data was trained, validated, and tested using logistic regression and multilayer perceptron. For the 2 machine learning algorithms, the predictions were evaluated.

For each of the machine learning algorithms, a grid search was performed to find the optimized hyperparameters that will generate accurate predictions:

**Logistic Regression**

Text

Description automatically generated

**Multilayer Perceptron**

Text

Description automatically generated

Hyperparameters for machine learning algorithms implemented for both data sets:

|  |  |
| --- | --- |
| **Machine Learning Algorithm** | **Hyperparameters** |
| Logistic Regression | 1. RegParam = default (0.0) 2. Maximum Iteration = 100 3. Elastic net Param = default (0.0) |
| Multilayer Perceptron | 1. Step size = 0.2 2. Max Iterations = 200 3. Training size = 0.5 4. Number of neurons = 10 |

**5.2.1 Inference: Value of a shot (Cristiano Ronaldo and Lionel Messi)**

Logistic regression is observed to have a higher accuracy than the Multilayer Perceptron for both data sets, for both players with 96.64% and 90.58% for Cristiano Ronaldo and Lionel Messi respectively. Regarding Cristiano Ronaldo, the goals scored (27) are higher than the goals predicted (26). This is because of the unknown series of events that might not have been captured in our data. Nevertheless, we achieved the highest possible accuracy.

**Cristiano Ronaldo**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Goals Predicted** | **Goals scored** |
| Logistic Regression | 94.39% | 26 | 27 |
| Multilayer Perceptron | 92.5% | 27 | 27 |

* + - 1. **Logistic Regression**

**Evaluation of the model:**

* **Accuracy**: An accuracy of **94.39%** shows how well the model performed as a whole. This means that **94.39%** of predictions of whether Lionel Messi scored a goal or not were predicted correctly.
* **Precision**: A precision of **95.87%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **Recall**: A precision of **97.89%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **F1**: An F1 score of **94.16%** shows that the model has a high rate of false positives and false negatives.

**ROC curve for Cristiano Ronaldo**

The ROC curve below is very close to the ideal vector (1,1) this indicates that logistic regression model is performing as expected and making predictions especially when it comes to distinguishing whether Cristiano Ronaldo scored a goal or not.

Line chart

Description automatically generated with low confidence

**5.2.1.2 Evaluation of the Multilayer Perceptron model:**

* **Accuracy**: An accuracy of **92.5%** shows how well the model performed as a whole. This means that **92.5%** of predictions of whether Lionel Messi scored a goal or not were predicted correctly.
* **Precision**: A precision of **94.8%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **Recall**: A precision of **96.8%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **F1**: An F1 score of **92.2%** shows that the model has a high rate of false positives and false negatives.

**Lionel Messi**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Goals Predicted** | **Goals scored** |
| Logistic Regression | 91.9% | 25 | 25 |
| Multilayer Perceptron | 88.8% | 28 | 25 |

**5.2.1.3 Logistic Regression**

**Evaluation of the model:**

* **Accuracy**: An accuracy of **91.9%** shows how well the model performed as a whole. This means that **91.9%** of predictions of whether Lionel Messi scored a goal or not were predicted correctly.
* **Precision**: A precision of **95.45%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **Recall**: A precision of **95.45%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **F1**: An F1 score of **91.9%** shows that the model has a high rate of false positives and false negatives.

**ROC curve for Lionel Messi**

The ROC curve below is very close to the ideal vector (1,1) this indicates that logistic regression model is performing as expected and making predictions especially when it comes to distinguishing between a goal or no goal.

A picture containing line chart

Description automatically generated

**5.2.1.4 Evaluation of the Multilayer Perceptron model:**

* **Accuracy**: An accuracy of **88.8%** shows how well the model performed as a whole. This means that **88.8%** of predictions of whether Lionel Messi scored a goal or not were predicted correctly.
* **Precision**: A precision of **93.75%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **Recall**: A precision of **93.75%** means that the proportion of correctly predicted positive observations over total number of predicted positive observations is high.
* **F1**: An F1 score of **88.8%** shows that the model has a high rate of false positives and false negatives.

**Prediction #3: Predict time to score a goal against given opponent**

Recommending probable time to score a goal against a given opponent based on **commentary text**. **Principle Component Analysis (PCA)** will be implemented. Similarly, like the last prediction, comparison of probable time to score a goal will be compared between **Cristiano Ronaldo** and **Lionel Messi.** To fit the commentary text into the PCA algorithm, preprocessing of the commentary text must be performed using the following steps:

1. Regular expressions tokenizer
2. Stop word remover
3. Count vectorizer and
4. IDF
5. Standard Scaler

K=2 is the hyperparameter used for the PCA model.

The data was split into **training validation** and **testing** with **75%, 12.5% and 12.5%** respectively.

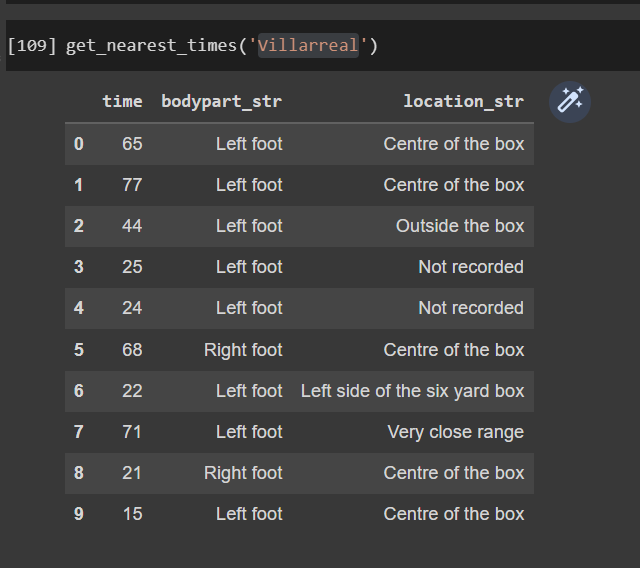
**5.3.1 Inference: Predict time to score a goal against given opponent.**

**Cristiano Ronaldo**

Here we used the Euclidean distance and Cristiano Ronaldo’s data set to find probable times to score against La Liga opposition Villareal

* 6 goals in the 1st half
* 4 goals in the 2nd half
* Heavy left foot usage

Majority of goals struck from the centre of the box as shown in the table below:



**Lionel Messi**

Using Euclidean distance and Lionel Messi’s data to find probable times to score against La Liga opposition Real Sociedad

* All goals scored in the 2nd half
* Equal mix of right and left feet

Majority of goals struck from the centre of the box as shown in the table below:

Graphical user interface

Description automatically generated

1. **Conclusion**

In conclusion, 3 goals or predictions were focused upon.   
The first goal (Prediction #1) predicted the value of a shot in the English Premier League.   
The next 2 goals (Prediction #2 and Prediction #3) were focused on two top goal scorers of all time in football, namely Cristiano Ronaldo and Lionel Messi. Seeing that football (soccer) is a low scoring game, predictions were made using expected goals.

* The Multilayer Perceptron model performed the best for Goal #1 with an accuracy of 97.72%.
* For Goal #2, when predicting the value of shots for 2 of the top scorers, Logistic Regression was the best performing model predicting the value of a shot with an accuracy of 94.39% and 91.9% for Cristiano Ronaldo and Lionel Messi respectively.
* Goal #3 predicted that based on historical commentary text, Cristiano Ronaldo is likely to have 6 expected goals in the first half and 4 expected goals in the second half, while Lionel Messi is likely to score all expected goals in the second half.

1. **Roadblocks**
2. Being based on text commentary, the data was mostly textual in nature and had to be mapped to values. Numeric data was comparatively much fewer in number.
3. As the data was massive (~950,000 rows & 41 columns), running GridSearch would take hours without reaching completion at times. Better computing power would make the process easier.
4. Data was somewhat unusual to the extent where minute events that are recorded in football games such as chances created, tackles missed were recorded, but simple data such as the full-time match scores, goals scored by a team in a particular match, etc., weren’t.