**Machine Learning Engineer Nanodegree**

**Capstone Report**

**“Win/Lose prediction in Soccer games based on Team Line-up”**

***Maged Makramalla, 2019***

Contents

[Definition 3](#_Toc774958)

[Project Overview 3](#_Toc774959)

[Problem Statement 3](#_Toc774960)

[Metrics 3](#_Toc774961)

[Analysis 5](#_Toc774962)

[Data Exploration 5](#_Toc774963)

[Exploratory Visualization 5](#_Toc774964)

[Algorithms and Techniques 5](#_Toc774965)

[Benchmark 6](#_Toc774966)

[Methodology 7](#_Toc774967)

[Data Processing 7](#_Toc774968)

[Implementation 7](#_Toc774969)

[Refinement 7](#_Toc774970)

[Results 9](#_Toc774971)

[Model Evaluation and Validation 9](#_Toc774972)

[Justification 9](#_Toc774973)

[Conclusion 10](#_Toc774974)

[Free-Form Visualization 10](#_Toc774975)

[Reflection 10](#_Toc774976)

[Improvement 10](#_Toc774977)

# Definition

## Project Overview

Sports Betting Industry rely heavily on developing algorithms that can help predict win/lose percentage based on roster selection and formation, rather than team general expected results.

Machine learning is heavily being used for establishing betting odds and win-lose percentages. [This paper](https://eecs.qmul.ac.uk/~norman/papers/Spurs_final.pdf) actually introduces a similar approach to the one suggested, while focusing on Tottenham Hotspur Football club data in the early 2000s.

[This paper](https://www.sciencedirect.com/science/article/pii/S2210832717301485) also provides a solid starting point for establishing a proper framework using Neural Networks.

There is also a couple of interesting blog posts and discussion boards such as [Link1](https://www.wagerbop.com/machine-learning-in-sports-betting/), [Link2](https://medium.com/fansunite/machine-learning-in-bookmaking-9311bbf8bd3f) and [Link3](https://medium.com/@jxbailey23/applying-data-science-to-sports-betting-1856ac0b2cab), that provide relevant and up-to-date information of the different approaches used in the current betting industry.

However, as a fan, one could easily predict an outcome of a specific game based on team selection and injuries and general profiles of certain players.

It is therefore important to add a team’s lineup into the calculation or at least show that one could predict an outcome with a high level of certainty based on the chosen lineup alone before adding the historical and ranking data to establish a more accurate estimation overall.

In this section, look to provide a high-level overview of the project in layman’s terms. Questions to ask yourself when writing this section:

*Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?*

*Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?*

## Problem Statement

Win/Lose Percentage is usually based on recent team performance and historical encounters between teams. The objective of this project is to develop an algorithm that tis more focused on team selection and lineup in matches.

In brief this algorithm will first identify the players of each team that has usually the greatest impact on the result, and accordingly will predict the W/L outcome based on the team lineup.

Accordingly, when providing the team’s lineup of player as an input, one should receive a win/lose classification as an output.

In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:

Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?

Have you thoroughly discussed how you will attempt to solve the problem?

Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?

## Metrics

As most sports fans already know, predicting the outcome of any game with a high certainty is almost an impossible task. There will always be surprises and upsets in all types of sports and competitors.

However, to develop a model that can provide a decent amount of precision, the data set will be divided into training and testing set as usual. The accuracy score of the precited outcome of both the testing and training set will be calculated.

The aim is to try a couple of different classification models and to obtain an accuracy of over 70% for both the training and testing prediction to avoid over and underfitting.

In this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:

Are the metrics you’ve chosen to measure the performance of your models clearly discussed and defined?

Have you provided reasonable justification for the metrics chosen based on the problem and solution?

# Analysis

## Data Exploration

This project will use the European [Soccer Database](https://www.kaggle.com/hugomathien/soccer) available on Kaggle.

This data set should have all the information required for the project.

It consists of 7 different data sets:

1. Country: 11 x 2 (rows x column)
2. League: 11 x 3 (rows x column)
3. Match: 26k x 115 (rows x column)
4. Player: 11.1k x 7 (rows x column)
5. Player\_Attribute: 184k x 42 (rows x column)
6. Team: 299 x 5 (rows x column)
7. Team\_Attribute: 1458 x 25 (rows x column)

The data includes:

* Data gathered from 11 different European countries in the seasons between 2008-2016
* Player’s and Team’s attributes sourced from EA Sports data
* Team line up and formations of each game
* Betting odds from 10 different providers
* Detailed match statistics:
  + Stage
  + Goals
  + Players
  + Possession

The target value will rely heavily on the Match dataset since it includes most of the information needed, along with the Team and Player dataset to support the prediction.

In order to use this dataset, a couple of conversions and additions haver to be performed first.

First of all the sqlite file has to be converted to a csv file to make it easier to inspect and to use using the common libraries such as pandas.

It is also worth mentioning that there are no win/lose variable indicated in any of the datasets, so this crucial variable will have to be obtained using the provided data such as home and away goals.

Additionally, it will be a bit challenging to differentiate between the home and away teams along with all of their stats, which will be more elaborated in the following sections.

Lastly, there are also some cells with missing data, which will be properly dealt with based on the application and necessary calculations.

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?

If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?

If a dataset is not present for this problem, has discussion been made about the input space or input data for your problem?

Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

## Exploratory Visualization

Since visualizing the dataset might not be that beneficiary, this section will focus on how each dataset will be used more thoroughly.

3 data sets will mainly be used:

* Player
* Team
* Match

Player:

The player dataset contains 7 columns as per the below snippet:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| id | player\_api\_id | player\_name | player\_fifa\_api\_id | birthday | height | weight |
| 1 | 505942 | Aaron Appindangoye | 218353 | 2/29/1992 0:00 | 182.88 | 187 |
| 2 | 155782 | Aaron Cresswell | 189615 | 12/15/1989 0:00 | 170.18 | 146 |
| 3 | 162549 | Aaron Doran | 186170 | 5/13/1991 0:00 | 170.18 | 163 |
| 4 | 30572 | Aaron Galindo | 140161 | 5/8/1982 0:00 | 182.88 | 198 |
| 5 | 23780 | Aaron Hughes | 17725 | 11/8/1979 0:00 | 182.88 | 154 |

In this case on the player\_api\_id and the player\_name could be used to effectively determine the most valuable players in any given team

Team:

The team data set consist of only 5 columns as per the below snippet.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | team\_api\_id | team\_fifa\_api\_id | team\_long\_name | team\_short\_name |
| 1 | 9987 | 673 | KRC Genk | GEN |
| 2 | 9993 | 675 | Beerschot AC | BAC |
| 3 | 10000 | 15005 | SV Zulte-Waregem | ZUL |
| 4 | 9994 | 2007 | Sporting Lokeren | LOK |
| 5 | 9984 | 1750 | KSV Cercle Brugge | CEB |

In this case the team\_api\_id and the team\_long\_name (or team\_short\_name) will be used to choose the team that should be analyzed.

Match:

This dataset contains most of information required for the analysis. It contains a total of 115 columns, so a snippet won’t be that helpful for visualizing purposes.

Instead the main feature that will be highlighted using the backward design or analysis approach, by exploring the desired inputs and outputs of the algorithm

Algorithm Input:

The input of the algorithm should be a list of 11 players that are participating in the game. Although this information is already available, one must consider if the team in question is considered the home or away team in the game. More technical details will be available in the Data Processing section.

The features used are therefore: home\_team\_api, away\_team\_api, home\_player\_1 – home\_player\_11, away\_player\_1 – away\_player\_11; in order to create: team\_player\_1 – team\_playe\_r11

Algorithm Output:

The output of this algorithm should be a Win / Lose classification indicated as ‘1’ for win and ‘0’ for lose.

Since this feature is not available in the data set, this result will be obtained using the goal difference, i.e. goals scored – goals conceded. In this algorithm, a draw will also be indicated, but will count as a lost game. More technical details will be available in the Data Processing section.

The features used are therefore: home\_team\_api, away\_team\_api, home\_team\_goal , home\_team\_goal; in order to create team\_win, team\_draw, team\_lose (and goal difference for further regression analysis)

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

*Have you visualized a relevant characteristic or feature about the dataset or input data?*

*Is the visualization thoroughly analyzed and discussed?*

*If a plot is provided, are the axes, title, and datum clearly defined?*

## Algorithms and Techniques

In order to obtain the most effective results, several algorithms will be used throughout this process.

First of all, of all numpy and pandas will be used to process the data accordingly.

Since this is considered a classification algorithm, several classification algorithms will be used in order to obtain different results, such as Decision Trees, Random Forest and Support Vector Machines.

The accuracy score metrics will be used to calculate the efficiency of each algorithm based on both the testing and training data.

Additionally, an MLP Classifier will also be used just for reference purposes.

Finally, a histogram will be used to compare all algorithms performance.

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

*Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*

*Are the techniques to be used thoroughly discussed and justified?*

*Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

## Benchmark

The project will use the available data of the betting odds available in the same data set to use as a benchmark to measure the efficiency of the developed algorithm.

Moreover, I believe a simple linear regression model dictating a 50/50 Win-Lose ratio could serve as a great baseline for this algorithm as a sanity check for the developed algorithms

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

Has some result or value been provided that acts as a benchmark for measuring performance?

Is it clear how this result or value was obtained (whether by data or by hypothesis)?

# Methodology

## Data Processing

Before applying the machine learning algorithms the dataset has to be properly prepared.

This process is all captures in the python file called *team\_games.py*.

This script starts by first selecting the team id, in this case the Spanish Side ‘Barcelona’ was chosen.

First the script captures all the matches that the team has played, either as a home or an away game.

Then the features indices of the players will change by substituting home\_player\_n to team\_player\_n and away\_player\_n to opp\_player\_n (n being the player number)

Five new columns are created:

* Team\_goal\_scored
* Team\_goal\_conceded
* Team\_win
* Team\_loss
* Team\_draw

And finally, a loop will run through all the rows that does the following:

* Choosing the correct player for team and opposition based on team id
* Calculating the goal difference
* Assign the correct win – draw – loss values based on goal difference and team id

It is also worth mentioning that some of the missing data within the dataset will be properly dealt with using the next machine learning scripts, since it is usually handled based on its influence and application.

As a last step a new file labeled ‘team\_games’ is created that includes all the necessary information for the algorithm such as: team\_win, team\_draw, team\_lose, team\_player\_1 – team\_player\_11.

The newly created file will now be used to all further calculations and predictions.

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?

Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?

If no preprocessing is needed, has it been made clear why?

## Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

*Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*

*Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*

*Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

## Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

Has an initial solution been found and clearly reported?

Is the process of improvement clearly documented, such as what techniques were used?

Are intermediate and final solutions clearly reported as the process is improved?

# Results

## Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

*Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*

*Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*

*Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*

*Can results found from the model be trusted?*

## Justification

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

Are the final results found stronger than the benchmark result reported earlier?

Have you thoroughly analyzed and discussed the final solution?

Is the final solution significant enough to have solved the problem?

# Conclusion

## Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

Have you visualized a relevant or important quality about the problem, dataset, input data, or results?

Is the visualization thoroughly analyzed and discussed?

If a plot is provided, are the axes, title, and datum clearly defined?

## Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

Have you thoroughly summarized the entire process you used for this project?

Were there any interesting aspects of the project?

Were there any difficult aspects of the project?

Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

## Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

Are there further improvements that could be made on the algorithms or techniques you used in this project?

Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?

If you used your final solution as the new benchmark, do you think an even better solution exists?

**Before submitting, ask yourself. . .**

* Does the project report you’ve written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your analysis, methods, and results?
* Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?
* Is the code that implements your solution easily readable and properly commented?
* Does the code execute without error and produce results similar to those reported?

**Domain Background:**

**Data Sets and Inputs:**

**Solution Statement:**

Line-ups and team selections play a huge role in predicting the outcome of any anticipated game. Missing players in key positions in the formation may alter any expected results. Thus, determining these key players in each position that greatly influence the final results, would help get a better and more accurate results.

This Algorithm will be able to first predict the most valuable players in each team and thus a better prediction of a game’s outcome based on the lineup of both team.

**Benchmark Model:**

The project will use the available data of the betting odds available in the same data set to use as a benchmark to measure the efficiency of the developed algorithm.

Moreover, I believe a simple linear regression model dictating a 50/50 Win-Lose ratio could serve as a great baseline for this algorithm as a sanity check for the developed algorithms

**Evaluation Metrics:**

The Results will be represented in Win/Draw/Lose Percentage similar to the percentages usually provided by Sports betting companies

Since this is considered a Regression prediction algorithm “R Square” will most probably be used to evaluate and measure the model’s overall performance.

**Project Design**

The solution will be developed in several steps.

First the key players in each position will be identified based on their contribution to the game. As a starting point, the consider will only consider games won/drawn/lost as the main factor for getting this job done.

This information will then used to first establish the Win/Lose value already included in the dataset enhance the percentages provided by the betting agencies to provide better insights and proper predictions.

Since the data is already labeled, the algorithm will use the usual regression rather classification, the usual algorithms we learned during the course.

In a later phase, the algorithm could get more into details by rating the players based on their direct contribution, such as goal scored, assists, goal conceded etc.