Football Match Prediction and Betting using Data Science and Machine Learning

Sahil Bharodiya

12011008

Chaganti Venkat Rami Reddy

12011041

People love to wager on the outcome of matches, so the betting industry is booming. Cricket, football, horseback riding, and wrestling are just a few available sports. Football, which is a major sport in Europe and North America, brings in a lot of money for the betting sector. People wager on the outcome of the game as well as on the outcome of the play event. Each football league produces a large amount of data. Several strategies and methodologies are used to forecast the outcome of a football match. Many of these strategies are based on past performance and goal totals.

English Premier League is arguably the most competitive, so it is the most natural sport to study. There has never been an analytic definition of football because the sport is in a constant state of change, with certain characteristics becoming less significant over time. Here, machine learning could prove to be invaluable.

Welcome!

Let’s Start, The purpose of this project is to explore the application of data science, statistical and machine learning methods in an effort to make predictions regarding the outcomes of matches such as win, lose, draw, goal differentials, and total goals by using historical data pertaining to the teams that participated in the match. The majority of the project consists of constructing a framework that, when given raw match data, generates appropriate features by drawing on historical statistics regarding match outcomes. Following that, we will measure the accuracy of the model utilizing a variety of different techniques, all of which are based on machine learning.

Begin with an outline or agenda to give a big picture view

01 **Football and Data Science**

02 **Motivation**

03 **Aim**

04 **Objective**

05 **Background**

06 **Ethical Aspects**

Used Tools

- Python
- Numpy
- Pandas
- Matplotlib & Seaborn
- BeautifulSoup
- Scikit-Learn

Data

Data is extracted from https://www.besoccer.com/competition using **BeautifulSoup** module of python.

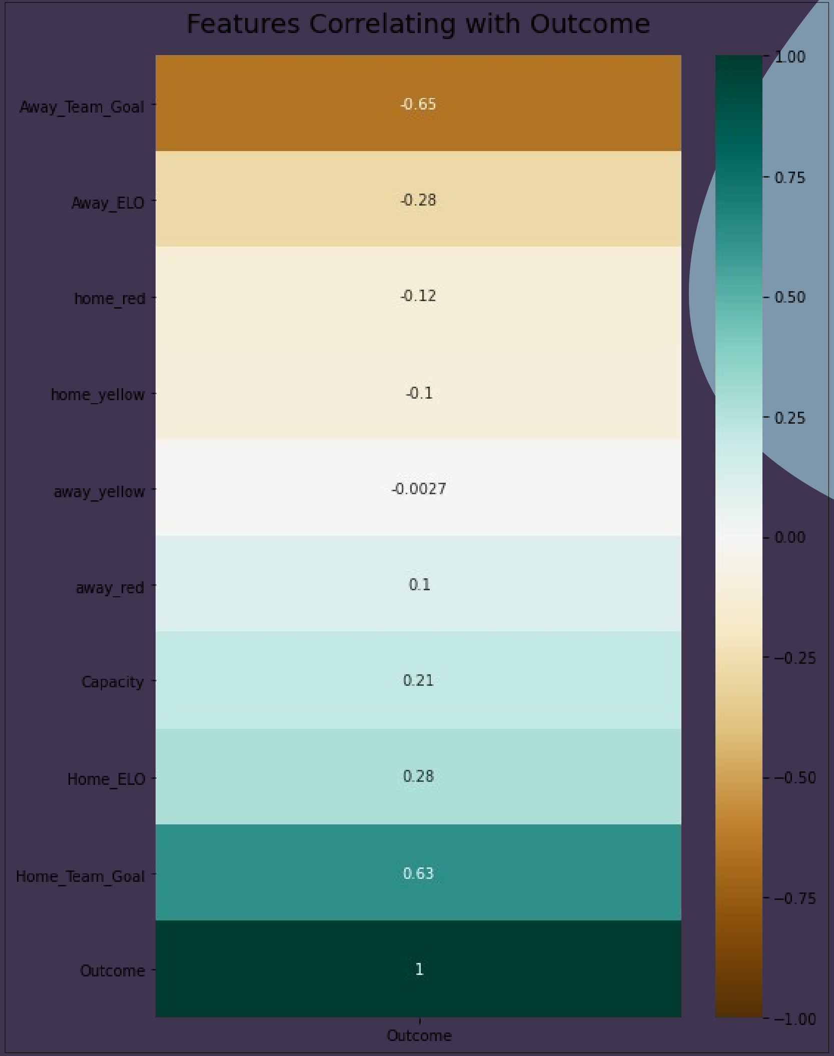
- Home_Team = Name of Home Team
- Away_Team = Name of Away Team
- Result = Outcome of Match
- Link = URL of the match
- Season = Year
- Round = Week of the Season
- League = Name of the league
- Home_ELO = ELO rating of Home Team
- Away_ELO = ELO rating of Away Team
- Date = Date of Match played
- Home_Yellow = No. of Yellow card given to the Home Team
- Home_Red = No. of Red card given to the Home Team
- Away_Yellow = No. of Yellow card given to the Away Team
- Away_Red = No. of Red card given to the Away Team

Data Insights

- We have a total of 12164 match records from season * 1990 to 2021.
- There were 52 teams played in this duration.
- Total of 32110 Goals have been scored by Teams in this duration.
- Total of 18245 Goals were scored by Home Team while 13865 Goals were scored by Visiting Team
- Average 2.63 Goals registered per Match.
- 3182 Matches have been drawn.

*currently we have data from 2012 to 2021

Feature Co-relation



New Features Creation

Since the high co-relation features are in very less amounts, we have to create new features.

The new features that we will create will have a high co-relation value.

New features will be derived from the existing features.

New Features

- **Last_5_Home_Team_avgGoal** : Average Goal of Home Team from last 5 match
- **Last_5_Away_Team_avgGoal** : Average Goal of Away Team from last 5 match
- **Last_5_Home_Team_Home_avgGoal** : Average Goal of the Home Team when they played at Home in the Last 5 match
- **Last_5_Away_Team_Away_avgGoal** : Average Goal of Away Team when they played Away in the Last 5 match
- **Last_5_Home_Team_All_Streak** : Home Team's Last 5 Result(sum).
- **Last_5_Away_Team_All_Streak** : Away Team's Last 5 Result(sum).
- **Last_5_Home_Team_Home_Streak** : Home Team's Last 5 match results when played at home
- **Last_5_Away_Team_Away_Streak** : Away Team's Last 5 match results when played at away

Training and Testing

Used Model

- K Nearest Neighbors
- Decision Tree
- Naive Bayes
- Support Vector Machine
- Random Forest Regressor
- AdaBoost
- Logistic Regression

Optimization

Since we have multiple models with multiple accuracies, to pickup the best from that will use GridSearchCV algorithms.

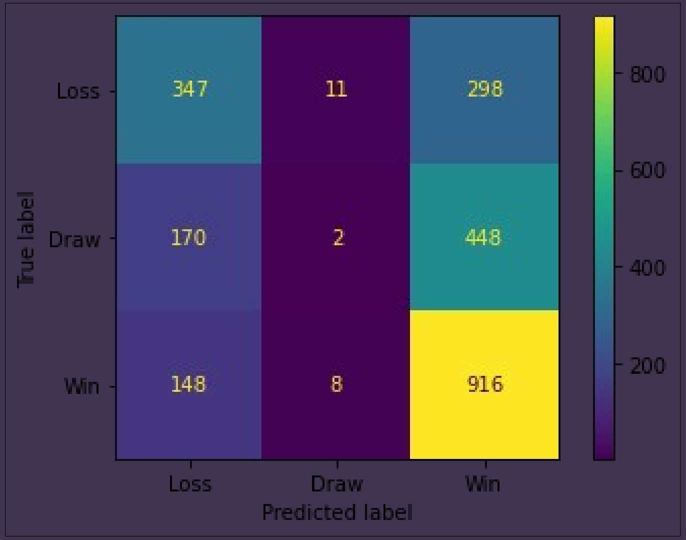
In the GridSearchCV algorithm, we provide a parameter grid of arguments in model like max_depth, alpha, n_estimators, learning_rate, etc.

GridSearchCV will try all the possible combinations and give us the best model.

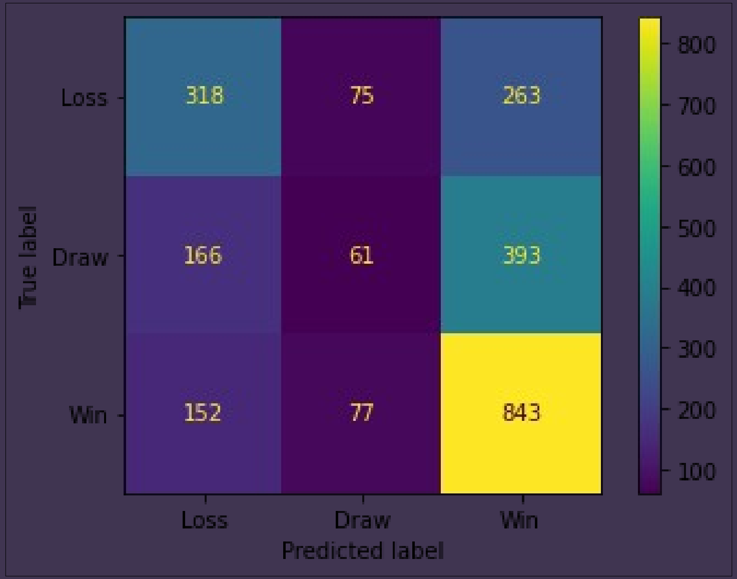
Accuracy Comparison

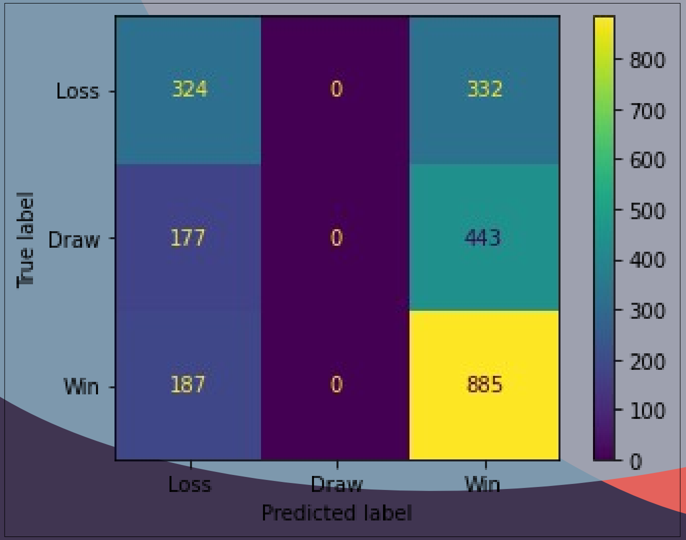
Model	Accuracy	F1 Score
GaussianNB	0.746744	0.748221
KNeighbours	0.535456	0.515414
SVC	0.604920	0.512565

Confusion Matrix



K
Neighbors





SVM

GaussianNB

Kelly Betting

The Kelly criterion is a mathematical formula developed by John L. Kelly Jr. while working at AT&T's Bell Laboratories. The formula is related to the growth of capital over a long period of time. It is used to determine how much money should be invested in a particular asset to achieve the greatest possible increase in one's wealth over time. It decides how much money should be wagered on each round. In order to achieve an optimal level of expected value, it makes use of the fundamental concepts underlying probability theory.

In Kelly Betting we have the formula for Odds.

We also calculate the Betting amount in each section using Kelly Betting.

Using Odds we will calculate profit in each section.

Deployment

We have created three models,

1. Goal Difference Prediction
2. Outcome Prediction
3. Total Goal Prediction

Will create a Flask server in localhost with HTML and CSS, then deploy it on the Google Cloud Platform.

Let’s Look at What we have built