FOOTBALL MATCHES PREDICTIONS USING DIFFERENT APPROACHES

Author: Menti Tommaso – VR504908

Tutor: Matteo Denitto

Professor: Marco Cristani

Univr Master's Degree Artificial intelligence

Introduction

In this project we aim to predict football matches of the Premier League season 2008/2009. In order to do this we tried different approaches with models: **TiDE**, **Arima**, **Linear Regression** and **Transformer**.
We focus our predictions on: the points obtained in each of the last 8 games and the cumulative points of each team. We used these predictions to compare the real final ranking of the season with our predictions for every approach.

What is the best dataset?

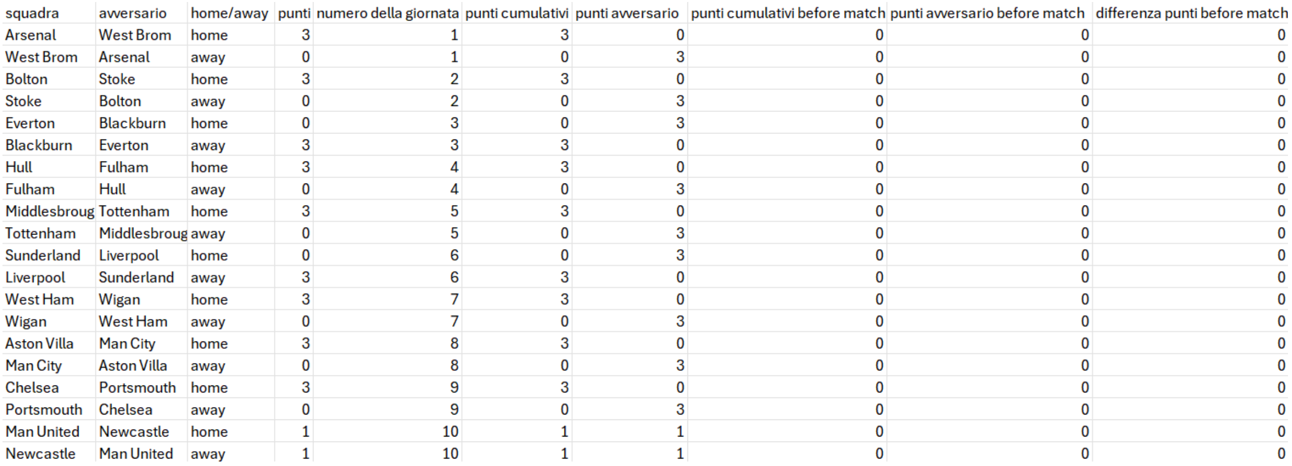
I found the datasets of every season starting from 2008/2009 to 2019/2020. We discussed about how much data use for our predictions and what would be the strenghts and weaknesses about using all the datasets or only one. We realized that different seasons vary significantly from one other. For example in the 2015/2016 season, Leicester City, that was always placed in the middle of the final ranking, won the premier league, and in the following seasons it performed worse than the previous years.
Keeping track of the all seasons can create some problems due to the impredictability of the sport.
We decided to work on only one season, because in general it is more stable and the teams tends to maintain the same trend throughout the season.
We decided to work on the 2008/2009 season.

Data preprocessing

Data processing in one crucial step in our prediction process. First of all we separated each match into two rows, every row shows the statistics of the match from the point of view of each team.
From the original dataset we derived different features:

> squadra, avversario, home/away, numero della giornata, punti cumulativi, punti avversario, punti cumulativi before match, punti avversario before match, differenza punti before match

The first match for all the teams looks like this:

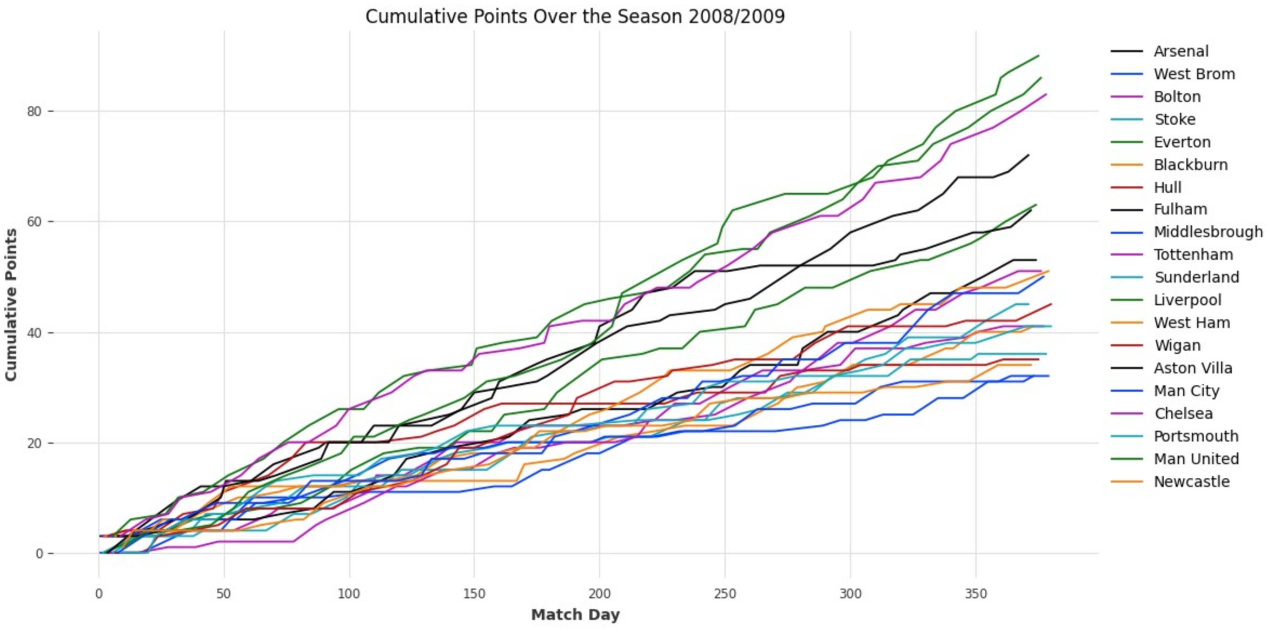


squadra	avversario	home/away	punti	numero della giornata	punti cumulativi	punti avversario	punti cumulativi before match	punti avversario before match	differenza punti before match
Arsenal	West Brom	home	3	1	3	0	0	0	0
West Brom	Arsenal	away	0	1	0	3	0	0	0
Bolton	Stoke	home	3	2	3	0	0	0	0
Stoke	Bolton	away	0	2	0	3	0	0	0
Everton	Blackburn	home	0	3	0	3	0	0	0
Blackburn	Everton	away	3	3	3	0	0	0	0
Hull	Fulham	home	3	4	3	0	0	0	0
Fulham	Hull	away	0	4	0	3	0	0	0
Middlesbroug	Tottenham	home	3	5	3	0	0	0	0
Tottenham	Middlesbroug	away	0	5	0	3	0	0	0
Sunderland	Liverpool	home	0	6	0	3	0	0	0
Liverpool	Sunderland	away	3	6	3	0	0	0	0
West Ham	Wigan	home	3	7	3	0	0	0	0
Wigan	West Ham	away	0	7	0	3	0	0	0
Aston Villa	Man City	home	3	8	3	0	0	0	0
Man City	Aston Villa	away	0	8	0	3	0	0	0
Chelsea	Portsmouth	home	3	9	3	0	0	0	0
Portsmouth	Chelsea	away	0	9	0	3	0	0	0
Man United	Newcastle	home	1	10	1	1	0	0	0
Newcastle	Man United	away	1	10	1	1	0	0	0

Prediction methods

We used different forecasting methods to see the strengths and weaknesses of each one. TiDE, Linear Regression and Transformer supports multivariate forecasting, so we made a prediction of the point obtained on each of the last 8 matches and the cumulative points for the whole season for every team.
Arima doesn't support multivariate forecasting so we predicted only the cumulative points.
For Linear Regression, TiDE and Transformer we used as covariates the enemy team and if it is a home or away match.

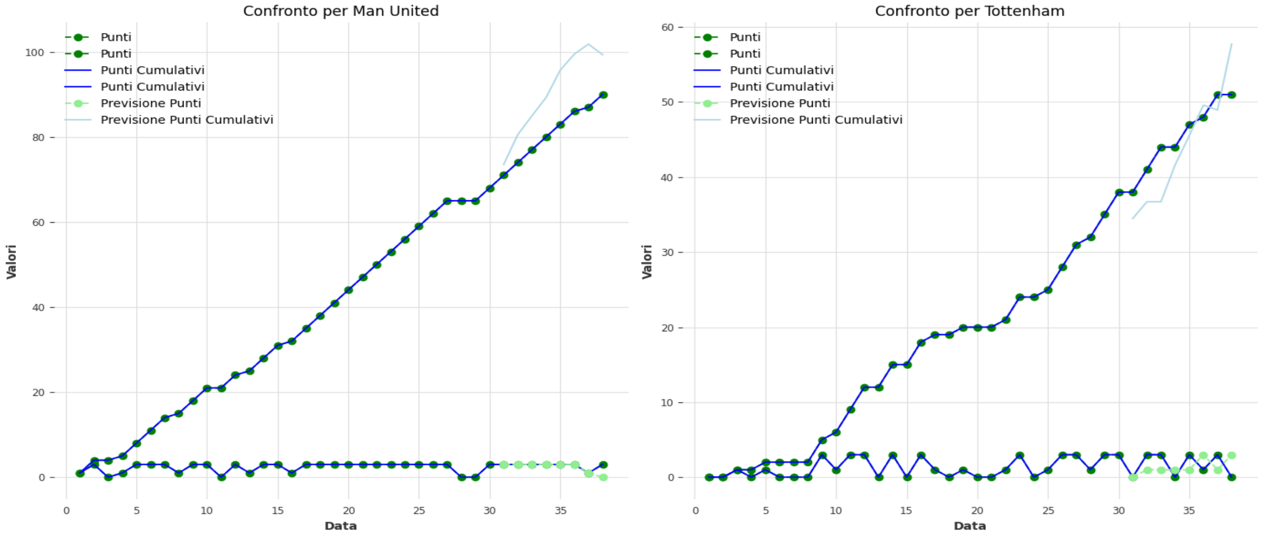
Cumulative points over the entire season for every team:



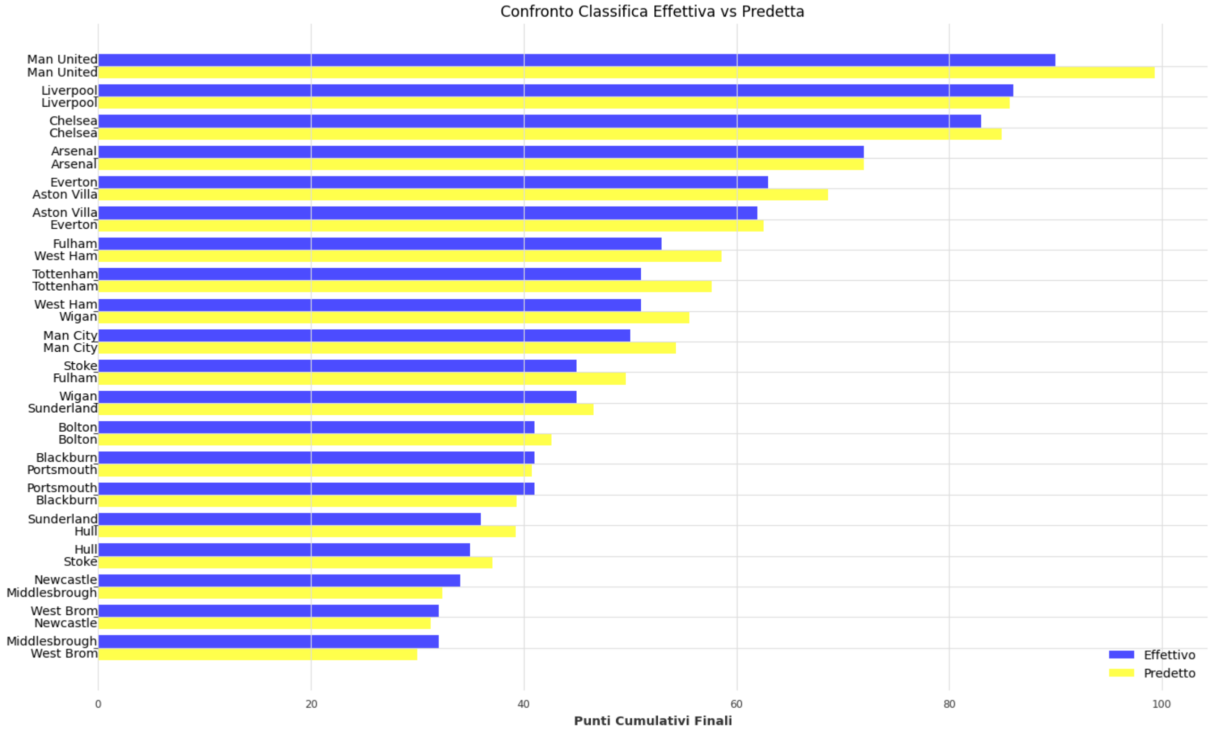
TiDE

TiDE is a linear model, it is made by multi-layer perceptron (MLP) based on encoder-decoder model. It was meant to be used for long-term timeseries forecasting but we tried a different approach and it performed quite well. Its main strenght is the simplicity and the possibility to handle non-linear dependencies.
The task of the Encoder is to map the past and the covariates of a time-series to a dense representation of the feature.
The Decoder maps the encoded hiddend representations intro future predictions.
Residual Block is the basic layer of the architecture.

- Plot of the predictions for Manchester United and Tottenham:



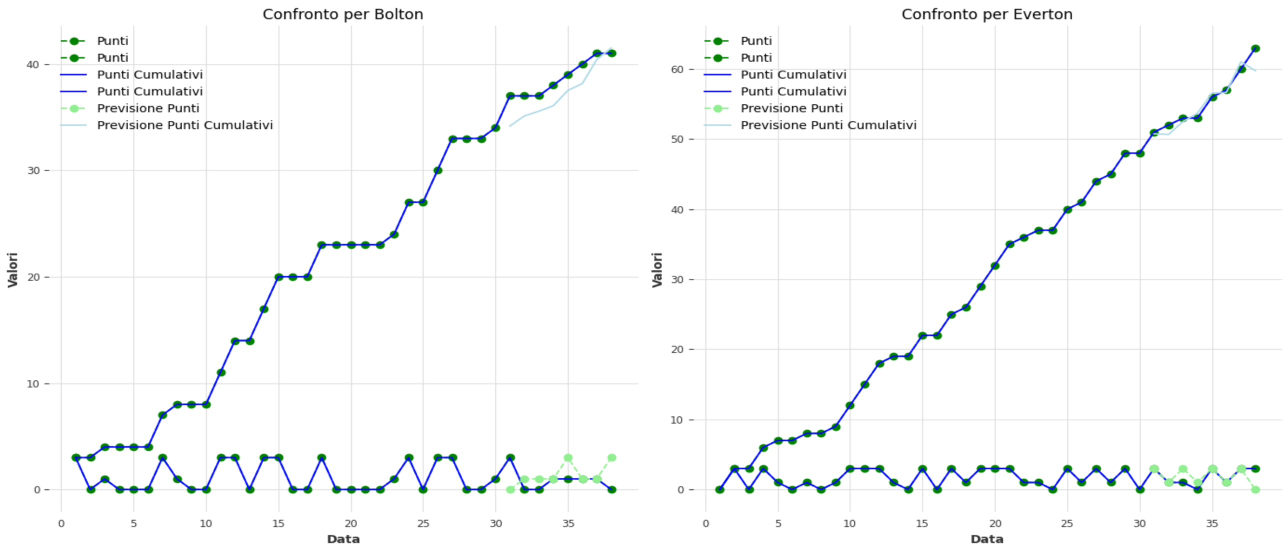
- Comparison between true final rank and predicted final rank:



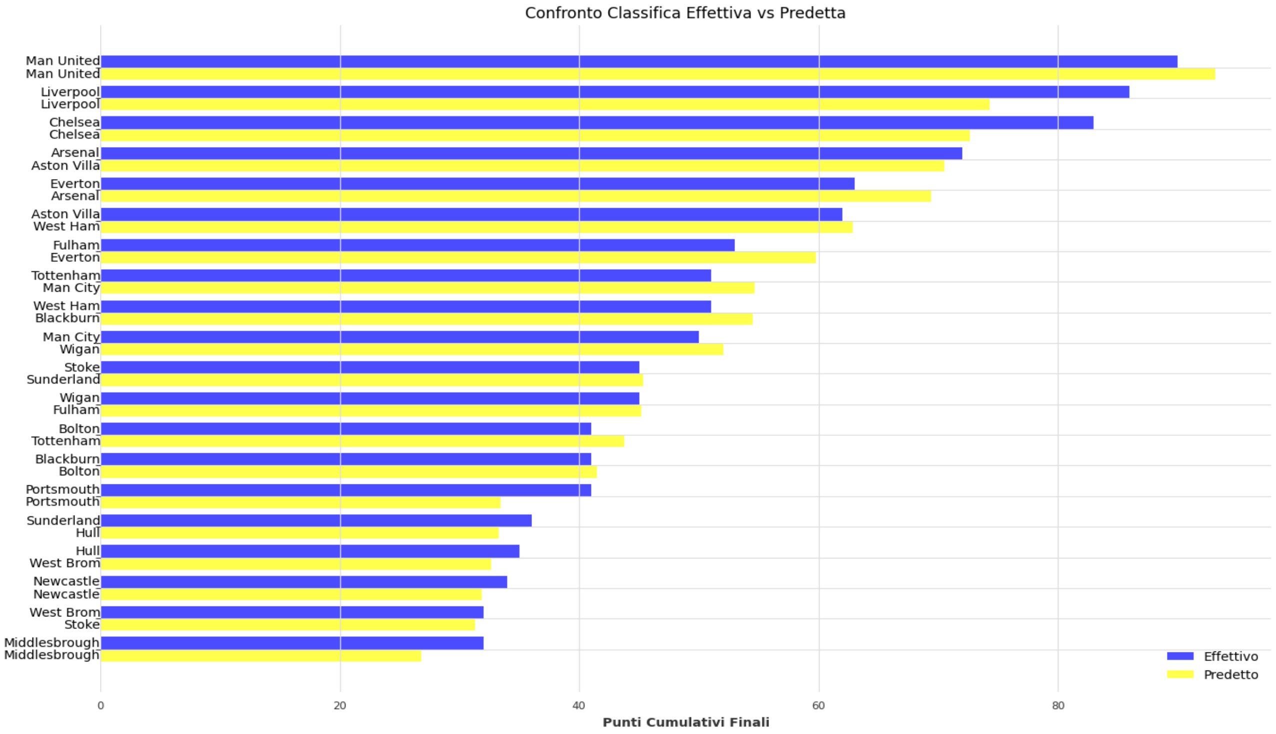
Linear Regression

Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. It measures the association between them. The goal is to find the best fitting line that minimizes the sum of squared differences between the observed and predicted values.
It can be applied when the relation between independent and dependent variables is linear, data is homoskedastic and residuals are independent and normally distributed. Linear regression analysis is the most widely used of all statistical techniques.

- Plot of the predictions for Bolton and Everton:



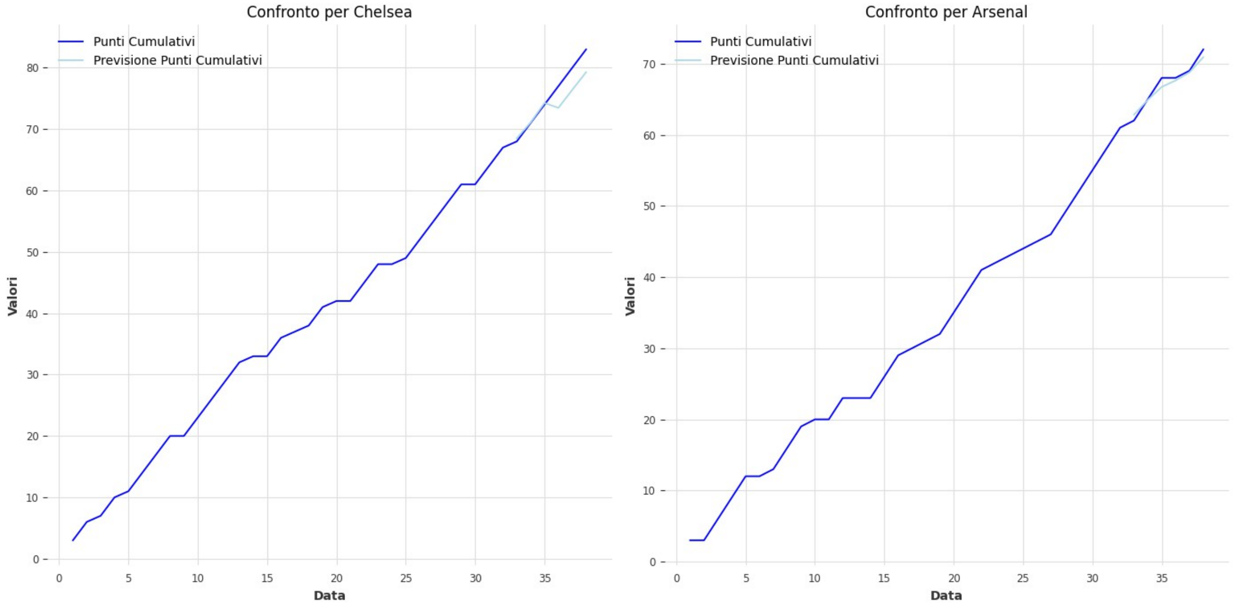
- Comparison between true final rank and predicted final rank:



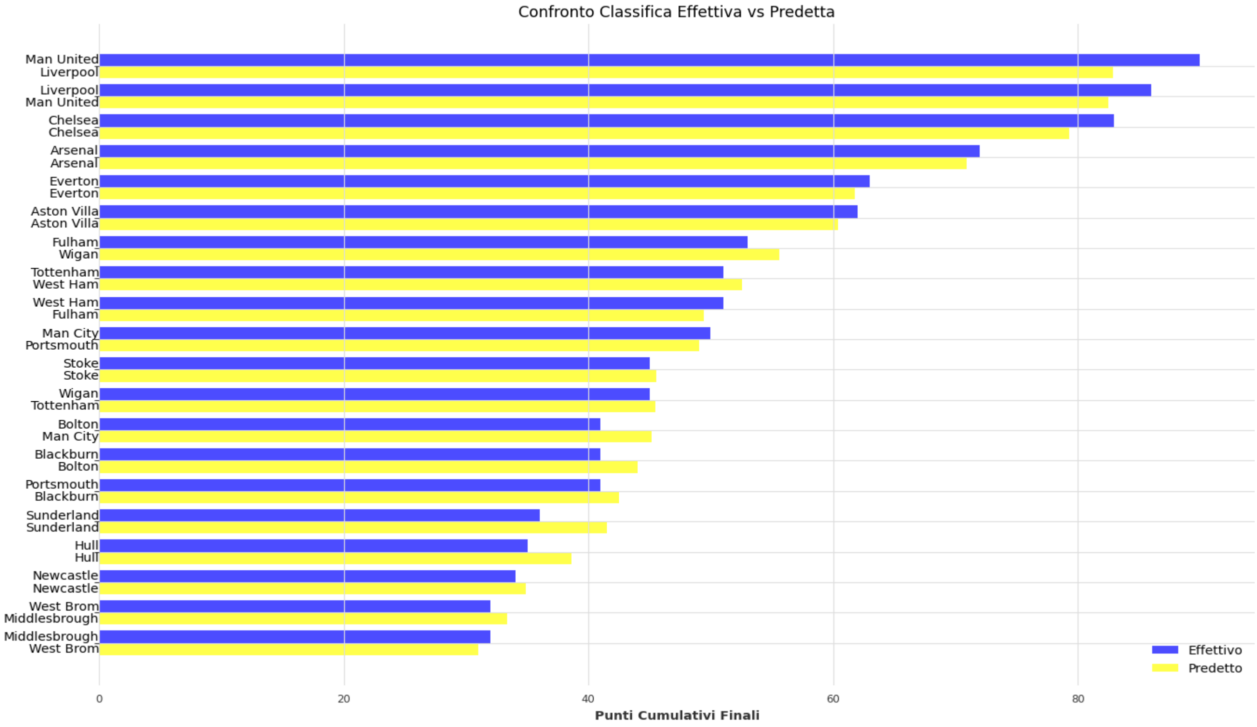
ARIMA

ARIMA is a statistical model used for analyzing and forecasting timeseries data. It combines three different component: autoregression, differencing and moving average.
Autoregression captures the relationship between an obstervation and the past.
Differencing removes non-stationarity, the time-series data becomes stable in terms of mean and variance over the time.
Moving average models the error of time-series as linear combination of past error terms.

- Plot of the predictions for Chelsea and Arsenal:



- Comparison between true final rank and predicted final rank:



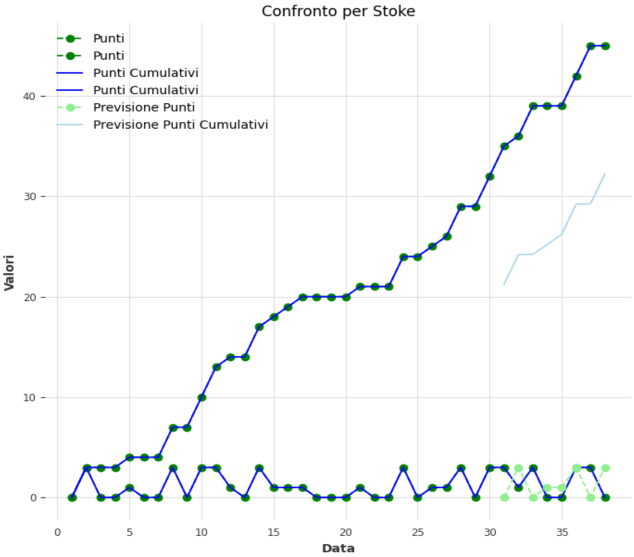
Transformer

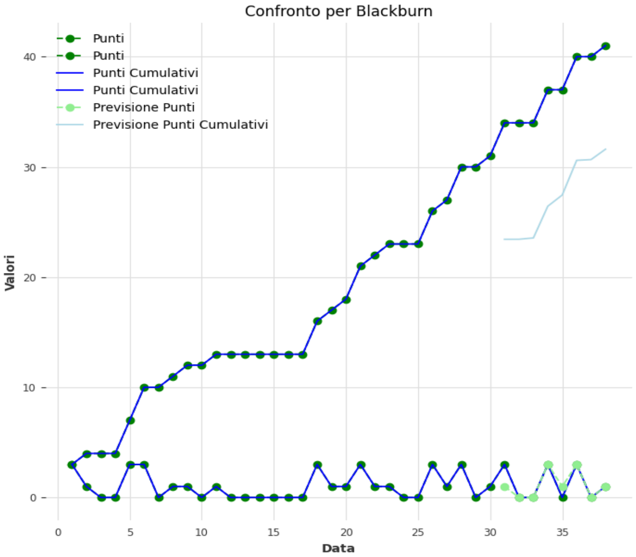
The Transformer model is a deep learning architecture primarily designed for handling sequential data. The architecture is made by two main components:

- Self-Attention mechanism: it weight the importance of data in a sequence, capturing complex dependencies.
- Encoder-Decoder: the encoder process the input sequence into a contect vector, the decoder generates the output squence.

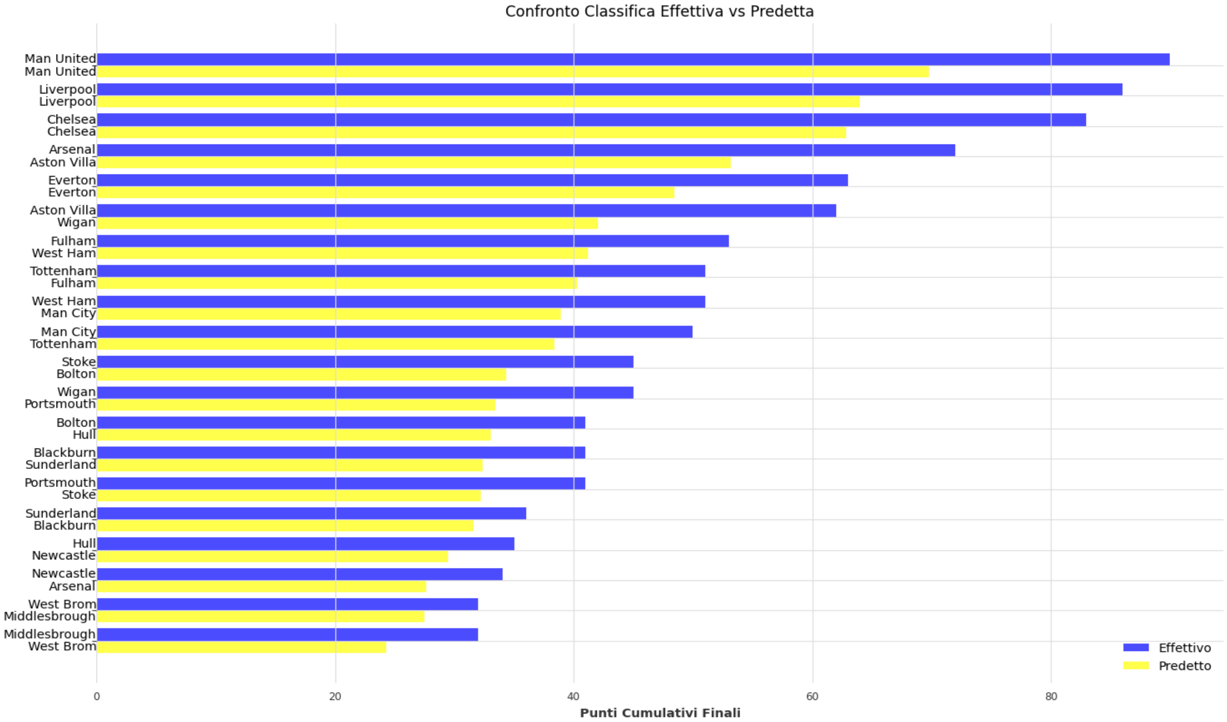
It also uses Feed Forward neural network to introduce non linearity.

- Plot of the predictions for Stoke and Blackburn:





- Comparison between true final rank and predicted final rank:



Comparison

	TiDE	Linear Regression	ARIMA	Transformer
Mean Squared Error	4,18	4,43	3,28	11,55
Accuracy	33,75%	38,75%	X	43,12%

Considerations

In this project we compared different models for the predictions of football matches during the Premier League season 2008/2009. These models are TiDE, Linear Regression, ARIMA and Transformer. Let's see the strengths and the weaknesses of each one.
TiDE and **Linear Regression** has similar MSE. It is a good result, in fact, considering the final rank based on the prediction, these models switch some teams but the ranking is more or less the same to the true one.
Trasformer has the highest MSE, but as we can see in the graph, it is able to understand the trends, it is only shifted downward. On the other hand it has the highest accuracy on match predictions.
ARIMA has the lowest MSE and it is the most precise in the prediction of cumulative points, it does not support multivariate forecasting so we do not consider single match prediction accuracy.

Future works

Several improvements can be made to increase the precision of predictions. The primary and most important is incorporating a significantly larger dataset. Including a wider range of variables such as team composition, individual player statistics, injury reports, weather conditions, yellow and red cards, and match statistics (for example number of corners and shots) can all impact game outcomes and should be considered. Additionally, considering data from multiple seasons could make the models able to recognize patterns in games with higher goal-scoring frequencies or higher number of cards such as derbies. Another possible extension would be to predict not just match outcomes, but also specific statistics like the number of goals scored adn assists. This project can be integrated with advanced machine learning techniques and more complex model's architecture in order to reach higher accuracy in prediction task.

Conclusions

This project shows how hard is to predict events in football.
Football matches are unpredictable due to numerous variables, for example the team form, player injuries, or random variables like the mood.
The variability of team performance across different season makes the predictions even harder, that's why we focused on only one season.
There is no model that is clearly better than the others.
If we are focusing on cumulative points ARIMA is the model with the best performance.
Transformer performs better on the single match prediction, but the accuracy isn't high enough to consider it good.