2021

Research EK predictor

Jordi van Belzen, Dylan Arts

6/14/2021

Introduction

We are 2 students currently in the AI specialisation semester. For the open programme we wanted to do a project together, since we did our proftaak in the start semester we wanted to do something again.

A common interest we have is football. So with the EURO 2020, in 2021, we thought of the idea to see what we can do with machine learning and predicting games and see if we can create a model that is able to predict the probability of a win, loss or draw in a game. With the time in the open programme we want to explore this and see if we can come up with an application with a model in the backend that is able to make predictions on the football games.

If we have a well working model, we want to elaborate on this project even after this semester. We have both an entrepreneur mindset and we would like to come up with an

Contents

Research approach 3
DOT Framework 3
Research questions 4
Domain understanding 7
Machine learning predictions in football 7
Data storage solution 8
Societal impact 9
Sport betting 9
Individual player analysis 9
Our approach 10
Approach 1: Scoring points to a game 10
Approach 2: deducting points when the goal difference is higher as 2 10
Approach 3: Adding points when the goal difference is higher than 2 11
Approach 4: adding a coefficient based on previous seasons 11
Prototyping ..... **Error! Bookmark not defined.**
Sources 14

Research approach

DOT Framework

Research in ICT aims at creating an ICT product that fits the needs of the client, in this case this product will be a machine learning model that supports the client. To help bring structure to the research we will use the DOT framework. The DOT framework has multiple research domains with research strategies that can be used to support the decisions made in a project.

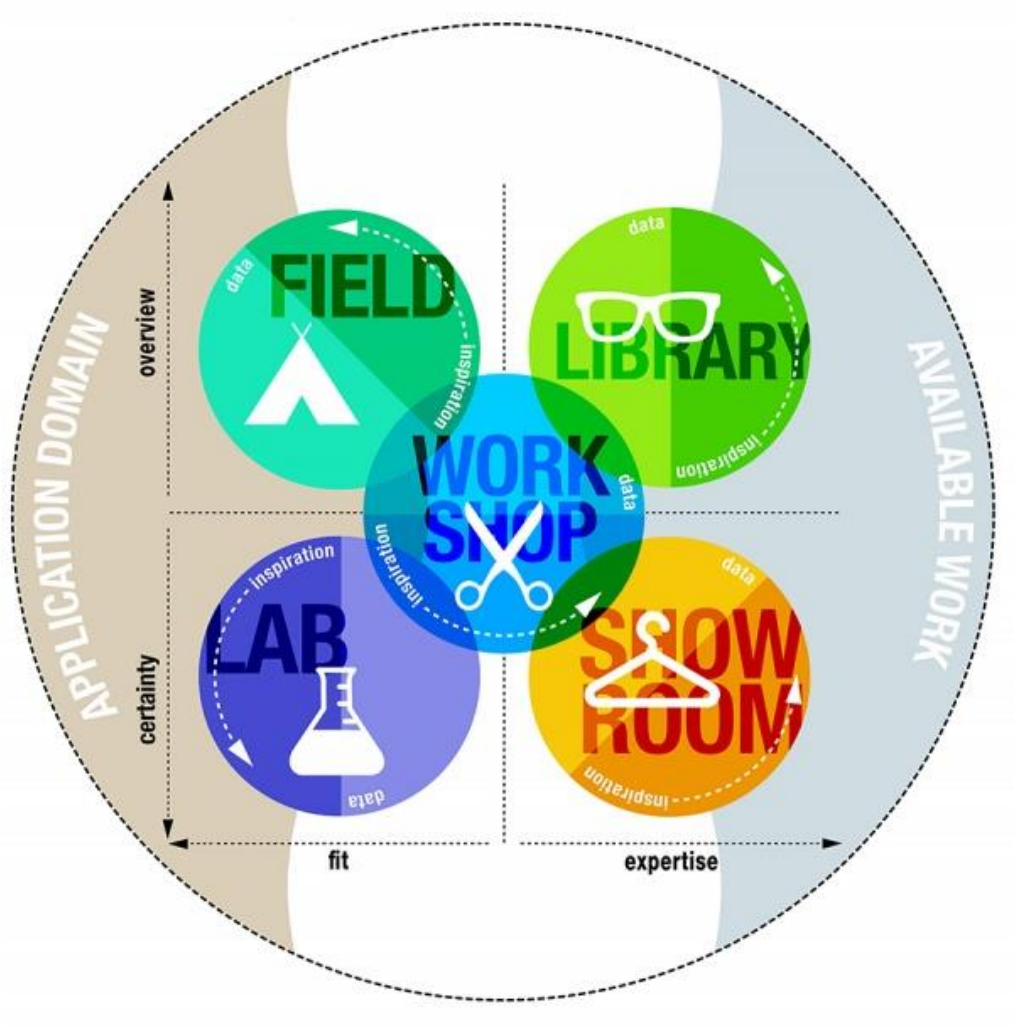

Figure 1: DOT framework domains

The DOT Framework is a tool that will be used in the research for this project. An overview will be made regarding the research questions and what domains will be used to answer these questions.

Research questions

The research questions are the questions set out to answer by the project. They are important in researching. Good research questions seek to improve knowledge on a topic and should be narrow and specific.

Main research question:

- Which machine learning model performs the best on predicting a football game?

Sub questions:

- Which ethical issues arise when making predictions on football games?
- Which storage solution for the data meets our requirements?
- Where can we find reliable and useful data for our application?
- How can we determine the quality of a team in a fair way?
- Which API or what API's are reliable and provides all the info needed in the deployment phase?

These research questions will be matched with the research strategies from the DOT Framework that will be used to find an answer.

Research question	DOT Framework research strategy
Which machine learning model performs the best on predicting a football game?	Library: Literature study, we will research online resources to see if there are already similar project out there and what machine learning models are being used in these if the exist. Field: Domain modelling, by looking into the domain we get a better understanding of the subject and what we would need to pay attention to. Lab: Data analytics, we will gather data and will be doing EDA and prepare it for modelling. Computer simulation, after preparing the model the data will be run through different models to predict outcomes of the game, this will be done by simulations of the models.
Which ethical issues arise when making predictions on football games?	workshop: Brainstorming, we will do a brainstorm session to come up with potential ethical issues we see with this product.
Where can we find reliable and useful data for our application?	Library: Literature, we will investigate multiple online sources that have football data and see if there are any sources that have the data we want and need for our model
How should the data be stored?	Library: Literature study, we will research potential solutions for storing our datasets online, we will be looking at cloud solutions and local solutions. Workshop: Brainstorming, after the research we will do a brainstorm session based on our results to
How can we determine the quality of a team in a fair way?	Library: Literature study, by investigating sources that researched this subject we can look for a good approach to tackle this problem. Lab: Data analytics, data analysis will be done over the gathered data and see if the results match up in a way we seem fair

Which API or what API's are reliable and provides all the info needed in the deployment phase?	Library

Domain understanding

Having a basic understanding of the domain you are working in will yield better results in the product. By knowing what you are trying to achieve and what the actual impact on the domain might be or the benefits of the project provider will help building a better understanding for the project. Therefore, looking into the domain is a first step into the project.

Machine learning predictions in football

To see if there are already existing projects out there, we set out on the internet to research machine learning models that predict football games. We found multiple articles that wrote about this subject. Football is the biggest sport on the planet, so that people tried to apply machine learning to predict games is not that surprising.

The one site that really stood out was https://kickoff.ai/. This site is made by 2 PhD students in machine learning. They made a website that shows their prediction of the outcome of the game. It also shows previous games so you can track if the prediction is correct. In the about section of the page there is a short description of how they tackled the problem of how to give a rating to a team.

"We model team strength dynamically
To predict future matches, a model needs to use data from the past. But how "far" in the past does it need to take data? Ideally, teams would have the same squad all the time and play frequently against each other. In practice, however, selected players change regularly and teams play only a few matches every year. To overcome this issue, our model allows the strength of a team to change over time. This enables us to take advantage of the many matches played over almost a century, while considering that recent confrontations should be more important to predict upcoming matches."

"We use Bayesian inference
This is a fancy way of saying that we are able to understand how confident we are about a particular prediction. As an example, take Argentina against Iceland. It is likely that Argentina will win—in fact, almost no one (except for Icelanders) would claim that Iceland has better chances. But just how much more likely is an Argentina win than an Icelandic win? 60%? 90% Perhaps 99%? On the one hand, Argentina is clearly the better team on paper, but on the other hand Iceland has never taken part in a World Cup final tournament and might perform over its usual level. Bayesian inference takes this (and much more) into account."

The approach they took is interesting and worth looking into for the project as well, since from what we can tell their model seems to be performing well overall. Football is an unpredictable game and anything can happen, but they seem to have a model that overall seems to be performing strong.

We found another approach by (kempa, M, 2020). The approach of this person is like the way we have been thought to tackle a machine learning problem. The first step is to webscrape data which will be used. Then some exploratory data visualisations were made to see if patterns could be spotted. The next thing he mentioned is the method he used to score teams, a number of the most recent games is taken and investigated. If a team wins they get + 3 points on their score, if they draw + 1 points and if the lose 0 points. With this the team gets a recent score which should give an impression of their current performance. After doing this a couple more steps were taken to prepare the data. Then the provisioning part is done.

Data storage solution

A suitable solution for storing the data needs to be setup. To do this we will investigate a couple solutions that are out there and could fit the need of our project.

Local

Storing the data local is not an option in our project. Since we are working remote from each other this would mean all the data would need to be send over every time a change happens in our data. This ending us up with no oversight on the data and a lot of files having to go back and forth.

OneDrive

What we don't have with storing locally we can manage with OneDrive. We can make a shared drive that we use for storing our data. This way we do not need to keep sending over data files back and forth, but we could just upload them to a shared drive and both have access to it.

Google Drive

Google drive is very similar to what OneDrive would do. We can store our datasets in drive and both have access to them. The issue however is that if we end up storing private data, do we want this to be on Google since their way of intruding in privacy over the years has shown that they don't hesitate to do this. So with storing a lot of data on drive, does Google handle this in a way we seem fit if private data is also stored.

GitHub

GitHub is a collaborative tool for programmers that allows for easy sharing and collaboration on code. It also allows for documents and other filetypes to be stored. Therefore, this can also be used for the project to store data.

We did decide to use GitHub as our way of sharing and collaborating for this project due to the ease of use and the power it gives us. We can easily go back to old revisions if something is not working and the sharing of files and code (jupyter notebooks) is great. This means we have all that we need in the same place making the oversight on the project a lot easier.

Societal impact

This is an important aspect within developing technology, which has been on the backburner for a long period where it was an afterthought. This led to a multiple scandals and misuse of technology. One of these recent examples in the Netherlands is de toeslagenaffaire. An algorithm was used to classify frauds. This led to a lot of people wrongfully classified as frauds which had devastating results for those people, they ended up divorcing their partners over this, accumulating huge debt and in the worst case there are cases where suicides are a direct result of this. The algorithm was also heavily biased and would classify people with a foreign background as more likely to fraud. To avoid having to attend these matters afterwards when the damage is done, by calculating these factors in development we can build in measures to prevent this from happening or at least decrease the likeliness of this happening.

Sport betting

If a game prediction tool in normal game circumstances works well, it might be abused to aid in betting on sport games. As humans we can determine the odds for some games well and make an overthought estimation on who will win, which is the input for the bet. The advantages of a system that can predict this is first the vast number of games it would be able to predict. Second it can make a better prediction on games that are a lot closer that in nature are close games.

As developers there is not much that can be done about this. Since there is no way to check if someone who used the application is also into sport betting or that the user is just curious about the results of a game.

Individual player analysis

Player analysis is already done on a big scale, since there are a lot of statistics available about players to give an indication of the performance of a player over a game. With the publicity of the professional player this is fair. This creates a timeline of their performance and this can be used by teams and scouts

However, this could also be used to predict performance of youngsters and there it becomes more of an ethical dilemma. If youngsters play for a professional football organisation already this might be part of the programme where they are monitored closely on their performance, but if the youngsters are not yet part of a professional football programme and are playing at local clubs, do we want to analyse their performance and make predictions over this. By doing this you sort of take the amateur players and analyse and predict them as professionals and them being underage as well.

This could be used to see if the performance of these youngster in the future would be good enough to advance to better teams or maybe even become professional players this could hinder the chances of players and gathering so much data on youngsters is in our opinion something that should be avoided. The privacy of these young players should be safeguarded and a system that makes predictions on player performance based on recent performance should not be available to scouts in younger leagues.

Our approach

After some brainstorming and the online research we have done we came to the following approach to give a team a score that we can use in machine learning. We came up with 3 variants that we want to test and see what the results are when used in modelling. We like the approach of giving teams a score based on the last couple of games played. So we took this approach and based to approached of this that try to take in some more variables.

Approach 1: Scoring points to a game

In this approach all is fair and no exemptions will be made. The downside with this approach is that we know some teams will perform lower by nature and since matchmaking for national teams is done by arrangements, a good team can arrange to play against less qualitative teams if they wanted to, so the scoring in our model would be in favour to them, while they only played lower tier teams. Same goes for the European championship and the world cup. Groups are made by draft and seed. If a team gets good rng in the draft they avoid better/equal teams and have an easier group stage. Again helping their score in our model.

Win	3 points
Draw	1 point
Lose	0 points

In the scenario where we look at 5 games, a team's max theoretical score is 15. Which would rank their performance as a team as good.

Approach 2: deducting points when the goal difference is higher as 2

In this approach we will deduct points from the winning team for the number of goals scored. The reason we put behind this is based on the idea that we can not keep track of all the variables of teams. If we look at a club like Real Madrid and compare this to Sparta for example, there is world of difference between the 2 clubs. When these teams get matched up against each other the chances of Real Madrid beating Sparta is high and the number of goals that Real Madrid would score in this game is likely to be high as well.

So we thought of this variant where if a the difference of goals scored in a game is higher then 2 we start deducting points of the score from the winning team. We deem it fair to have this 2-goal difference before taking away points. Usually, when good teams play against each other the difference we see is 2 goals. With this knowledge and with looking for a way to also keep the differences between teams in mind in a fair way we came up with the following scoring system.

Win	3 points * (1 - ((goal difference - 2) / 10))
Draw	1 point
Lose	0 points

Example, a team winning with a goal difference of 4 will get the following number of points:

3 * (1-((4-2)/10)) = 2.4 points

We think this approach is one way we can deal unfairness we can not solve with the base approach of the scoring system.

Approach 3: Adding points when the goal difference is higher than 2

Where we tried to address the issue if weaker and better teams with approach 2, there is another downside to it. What if a known good team wins with more then 2 goals from a team that is also known to be good. The team would be penalized in the system for winning with great numbers hindering their performance in our system. So we also want to investigate what happens if instead of deducting points, we score the dominant winning team more points. Boosting their performance.

Win	3 points * (1 + ((goal difference - 2) / 10))
Draw	1 point
Lose	0 points

Example, a team winning with a goal difference of 4 will get the following number of points:

3 * (1 + ((4 - 2) / 10)) = 3.4 points

This variant is more fair against well performing teams and will score them more generous boosting their odds in our predictions to win a game.

Approach 4: adding a coefficient based on previous seasons

Due to time constraints, we will not address this approach, but we have thought about it already. We want to take the standing of a previous season into account which would create a coefficient. In theory with this we can adjust the point scoring system in the best fair way. If the team that just got promoted to the top league and wins against the champion of last year, they will get way more points from it, being the odds of them winning the game being way lower then the team that won the title last year and reversed the team that won the title is highly likely to win from a team that just got promoted.

By making a coefficient for the league based on the result of last year, this is likely the fairest way to adjust the points, but also takes more time to implement. This would be our backup approach of the first 3 won't yield results we can use to predict games on.

Potential models

From the article Machine learning algorithms for football predictions by Kempa, M, it is also mentioned which models are being used and how they performed. In the article the classification algorithms KNN and random forests seem to be the best performing outscoring the closest regression with quite a big margin.

We will use the sklearn library and look at available regression and classification models to see what results these algorithms yield us, with a slightly different approach to try and make the scoring more even.

The algorithms we will use for classification are:

- KNN
- Random forests
- SVM

From the research we did, in the article it was mentioned KNN and random forests were the best performing models, so we want to use these as well. We also want to look further and explore more models if this fits in the limited time we have for the open programme.

The regression algorithms we will use are:

- Logistic regression
- Random forest regression

API choice

For our deployment we wanted to use a football API which we could call upon for gathering data, gathering matches and all these kinds of things that we would like to show in the application.

We investigated multiple API's, but we can to the conclusion that none of them were what we needed, or we needed to pay money for them otherwise we would be limited in the amount of requests we could do on a daily basis.

We discussed this issue internally and we decided that we would build our own artificial API. So, Dylan being the software student took this into his hands and set out to build the API that gathered the teams, stats, results and eventually it will also show the players.

Sources

Peng, N. Y. (2020, 12 mei). *What I've Learnt Predicting Soccer Matches with Machine Learning | Towards Data Science*. Medium. https://towardsdatascience.com/what-ive-learnt-predicting-soccer-matches-with-machine-learning-b3f8b445149d

Kickoff.ai: predicting football matches. (z.d.). Kickoff.Ai. Geraadpleegd op 16 juni 2021, van https://kickoff.ai/

Kempa, M. (2020, 24 september). *Machine Learning Algorithms for Football Predictions*. Medium. https://towardsdatascience.com/machine-learning-algorithms-for-football-prediction-using-statistics-from-brazilian-championship-51b7d4ea0bc8

Supervised learning — scikit-learn 0.24.2 documentation. (z.d.). SKlearn Classification. Geraadpleegd op 19 juni 2021, van https://scikit-learn.org/stable/supervised_learning.html#supervised-learning