Business understanding

Our interest in football, especially in Premier League, goes back to childhood. Markus has played football from age of 3 and Del has always been really interested in the sport. Thus, we consider us as experts and occasionally bet on the games, but often losing some money, it has got us thinking, if there is any way to predict (correctly) by using machine learning. Since football has always been considered as unpredictable, because there are so many factors, but last year Markus got top 20 (out of 10000 contestants) in the world in a Premier League score predicting competition. We have a goal to predict this year's league table (as it would stand at the time of presentation). Also, our goal is to see what teams/games are more highly correlated to have a certain result, for example in Markus's predicting career, he has noted, that Southampton hosting Liverpool has often ended in draw, but Liverpool is obviously much more stronger. In addition, we would like to predict the various elements of the game, which you can bet on - for example, how many corners one team had or both teams to score etc.

The success criteria is a bit hard to define, since sport is always hard to predict, but we would like to predict this year's league table and confirm if we have got results right or league positions of our favorite teams right (Chelsea and Arsenal, respectively). In addition, we would like to identify some strong correlations and results between certain teams. If it is possible, we would also like to see how correlated are yellow/red cards to the outcome. Because if a team gets a red card, it is in a massive disadvantage.

Our resources consists of our brains, all the data from the last 5 years of Premier League, two computers and some algorithms, which we would test out.

The data is verified and official, so we would not have to check it and since it is official, it is legal and available for everyone.

In this project we have many risks, since football is considered unpredictable and we are scared that the data is not correlated at all (hopefully not!) and we would have nothing notable.

The project does not cost anything but our hours from life, which are priceless, but if we can find some correlations, we could easily make some money by betting on the games or compete in predicting competitions, which usually have some kind of buy-in.

We will use the models, which we have learned in this course and find out the best one, by having training data and test data split. Then we will check the accuracy of this season scores/outcomes and finally, select the best one to present. We could visualize the data by having both league tables (real-life and predicted one) side-by-side. And also by pointing out the most correlated scores and teams.

Success criterias would be at least 60% of correct outcome, at least 3% of correct scores and anything interesting, which we don't even know about right now.

Data understanding

We already have the data gathered, but it should be made up of all the statistics and info of the game, for example the teams playing, the full-time score, half-time score, statistics (corners, fouls, yellow cards, shots, shots on target etc). The data exists and is available, but there are so many versions of data, which we could use - for example, our dataset has various sites' coefficients, which have been confirmed one hour before the game, instead of having all the lineups etc, because they ultimately show the same thing. Obviously, this limits our predicting quality, but since the data is very popular and available, we could use it as easily. We would use the five documents of five Premier League seasons.
Our dataset has the following information about the game -
Div - League division
Date - Match date (dd/mm/yy)
HomeTeam - Home team
AwayTeam - Away team
FTHG - Full time home team goals
FTAG - Full time away team goals
FTR - Full time result
HTGH - Half time home team goals
HTAG - Half time away team goals
HTR - half time result
Referee - Match referee
HS - Home team shots
AS - Away team shots
HST - Home team shots on target
AST - Away team shots on target
HF - Home team fouls committed
AF - Away team fouls committed
HC - Home team corners
AC - Away team corners
HY - Home team yellow cards
AY - Away team yellow cards
HR - Home team red cards
AR - Away team red cards
B365H - Bet365 home win odds
B365D - Bet365 draw odds
B365A - Bet365 away win odds
BWH - Bet&Win home win odds
BWD - Bet&Win draw odds
BWA - Bet&Win away win odds
IWH - Interwetten home win odds
IWD - Interwetten draw odds
IWA - Interwetten away win odds
PSH - Pinnacle home win odds

PSD - Pinnacle draw odds
PSA - Pinnacle away win odds
WHH - William Hill home win odds
WHD - WIlliam Hill draw odds
WHA - William Hill away win odds
VCH - VC Bet home win odds
VCD - VC Bet draw odds
VCA - VC Bet away win odds
Bb1x2 - Number of BetBrain bookmakers used to calculate match odds averages and maximums
BbMxH - BetBrain maximum home win odds
BbAvH - BetBrain average home win odds
BbMxD - BetBrain maximum draw odds
BbAvD - BetBrain average draw odds
BbMxA - BetBrain maximum away win odds
BbAvA - BetBrain average away win odds
BbOU - Number of BetBrain bookmakers used to calculate over/under 2.5 total goals averages and maximums
BbMx>2.5 - BetBrain maximum over 2.5 goals
BbAv>2.5 - BetBrain average over 2.5 goals
BbMx < 2.5 - BetBrain maximum under 2.5 goals
BbAv < 2.5 - BetBrain average under 2.5 goals
BbAH - Number of BetBrain bookmakers used to calculate Asian handicap averages and maximums
BbAHh - BetBrain size of handicap (home team)
BbMxAHH - BetBrain maximum Asian handicap home team odds
BbAvAHH - BetBrain average Asian handicap home team odds
BbMxvAHA - BetBrain maximum Asian handicap away team odds
BbAvAHA - BetBrain average Asian handicap away team odds
Data is definitely going to need some cleaning up for different predictions, for example if we want to predict home team goals of a certain match, we probably don’t need the odds of over/under 2.5 goals.

Project plan

1. To agree on how to split the data into training and test data - 0.5 hours
   (Since there are 5 seasons, we can't take them all as the same and would need to apply some kind of coefficients for the first seasons, because within five years there have been a lot of changes)
2. Start training with various models, like Random forest, KNN etc, which we have learned in this course - at least 10 hours
   (Since we don't know which model will be the best nor if the data responds accurately, we would test for at least 10 hours and more if needed, because we don't want to predict the score, but also identify how much various statistical features of the match change the outcome, if possible)
3. Test the data with test data - 2 hours
   (We need to test the models we have trained on test data, to see if it is possible, and if it is possible, how accurately it is predicting. The biggest question is that, is it possible to predict more accurately than random guessing)
4. Test the data with real-life results from current season - 3 hours
   (We would need to check if the results/outcomes from current season actually match or are as accurate, because it would not help us, if we could predict past results)
5. Visualize the data with current league table and compare them - 5 hours
   (We think that this is the best visualization, because it shows the differences and how much it differentiates)
6. Start testing the correlations between teams, cards, shots - 10 hours
   (Since our biggest criteria is predicting the outcomes, we would have to do that first, but later we could start noticing patterns or correlations, which could happen between teams, but we can't see them ourselves)
7. Results
   (Make results and conclusions of how did we do within these three weeks)

Methods:
Use everything we have learned, but also everything we can find ourselves from Google or Kaggle.
Tools are probably course materials, python libraries, everything Google offers us.