DATA WAREHOUSING AND DATA MINING REPORT

Module: CA4010

Names: Toba Toki – 15349476

Adam Craig – 14490662

Predicting - Who will win the Euro's?

Table of Contents

Declaration

We declare that this material, which We now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. We have read and understood the Assignment Regulations. We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.

1. Introduction

This report will provide a concise explanation of our approach to making future predictions based on a dataset. It will describe our dataset and how we constructed it and transformed it. It will also mention the process we used to form our predictions and create our prediction model. Finally we will provide some results of our predictions and analysis and a conclusion based on this, and the discoveries we made along the way as well as things learned.

What was our Idea?

We would preform analysis on the form of teams which have qualified for the UEFA European championship over the previous 2years before the tournament begins.

We aimed to see if we could identify any trends between certain statistics which map to progressing in the tournament or exiting in the group stages.

We took data such as a team's results(W/L/D) and the important statistics of the match such as the goals scored, goals conceded, clean sheets etc. Then we wanted to take this data, rid the data of any noise and transform it so that we would be able to get as much information that is relevant to our prediction. We would then take our transformed data and run it through the appropriate algorithm that would return a prediction for us.

2. Dataset

The dataset we used was created by us from scratch. We used various websites to acquire the information such as:

- Soccerway - https://uk.soccerway.com/ - Used to gather attributes / statistics for each team
- Wikipedia - https://en.wikipedia.org . – Used to declare qualified teams for each tournament in our training data and the class (Staged they eventually reached)

We eventually would end up a training dataset containing:

- 2 Years Statistics for 16 teams which qualified for the euro's that year.
- 4 tournaments (from Euro 2000 – Euro 2012) to train our model on. = 64 teams in total.

Each dataset we had were split into the years of when the Euros was happening. We had the statistics that we needed for the countries. At first, each country had just 9 parameters that we wanted to use to predict but as we went on we decided that we needed more information if we wanted our prediction to be better. We used Data Transformation techniques such as Aggregation, Generalisation and Attribute Construction to increase these 9 parameters for each team to 21 to give us much greater detail in helping to identify any trends.

The picture below is what our data looked like originally, before our **Data Workshop** where we used the **data transformation** techniques mentioned below.

TournamentYear	Team	Total Games Played	Total goals scored	Total goals conceded	Total Clean Sheets	Total Win	Total Draw	Total Loss	Qualified As	Round Reached
2012	Croatia	20	33	15	8	11	3	6	Runner Up	Group
2012	Czech Republic	20	29	20	9	9	6	5	Runner Up	QuarterFinal
2012	Denmark	18	28	19	7	10	5	3	Winner	Group
2012	England	17	30	13	9	11	2	4	Winner	QuarterFinal
2012	France	23	37	12	13	15	2	6	Winner	QuarterFinal
2012	Germany	22	57	29	7	14	3	5	Winner	Semi-Finals
2012	Greece	21	23	13	10	11	1	9	Winner	QuarterFinal
2012	Italy	20	27	13	10	11	5	4	Winner	Final
2012	Netherlands	21	54	19	11	14	3	4	Winner	Group
2012	Poland	25	32	23	9	11	5	9	Host	Group
2012	Portugal	18	41	20	8	9	4	5	Runner Up	Semi-Finals
2012	Republic of Ireland	24	41	15	14	12	4	7	Runner Up	Group
2012	Russia	21	25	10	13	11	3	7	Winner	Group
2012	Spain	22	54	23	9	16	4	2	Winner	Winner
2012	Sweden	23	44	17	10	13	5	5	Runner Up	Group
2012	Ukraine	22	36	36	3	9	8	5	Host	Group

Not having enough attributes in our dataset was troublesome to deal with and was something that was not practical for us to work with. Having done some analysis we identified that there was little correlation between most attributes and they were too broad. We couldn't tell very much from these attributes upon first analysis besides the obvious biased opinions being football fans ourselves.

After attaining the results of each team, we decided to take a closer look at teams with high stats. We analysed our data and we noticed that certain teams played against easier oppositions in the lead up to the Euros than other teams. In turn, these teams had more wins, scored more goals and conceded less. We decided to use **Attribute Construction** to solve this.

3. Data Preparation

The workshop we chose was Data Transformation & Integration. We chose this workshop because after first analysis we could not really tell much from our dataset which we created ourselves, with the attributes varying so much. We realised that no one or two attributes alone could determine an

approximate stage a team may reach in a tournament. We agreed that we need to further transform our dataset.

Data Transformation

Attribute Construction is when new attributes are constructed from given attributes and added in order to help improve the accuracy and understanding of structure in high-dimensional data.

We gathered information about the top 10 ranked teams in the world at the time for each of the years. If a team had played a team on this list, we constructed attributes to record this data in order to help weigh our data for our predictions. Doing this gave us more detailed information which we could later use in our prediction model.

TournamentYear	Team	Total Games Played	Games Played vs Top 10 ranked	Wins vs Top 10 ranked	Loss vs Top 10 ranked	Draw vs Top 10 ranked	Goals Scored vs Top 10 ranked	Goals Conceded vs Top 10 ranked	Clean Sheet vs Top 10 ranked	Total goals scored	Total goals conceded	Total Clean Sheets	Total Win	Total Draw	Total Loss	Total Win vs Normal	Total Lose vs Normal	TotalWin%	TotalLose%	GoalsPerGame%	Qualified As	Round Reached
2012	Croatia	20	0	0	0	0	0	0	0	33	15	8	11	3	6	11	6	0.55	0.3	1.65	runner-up	Group
2012	Czech Republic	20	3	0	3	0	3	8	0	29	20	9	9	6	5	9	2	0.45	0.25	1.45	runner-up	QuarterFinal
2012	Denmark	16	5	1	3	1	6	12	0	28	19	7	10	5	3	9	0	0.555555556	0.166667	1.555555556	Winner	Group
2012	England	17	2	1	1	0	3	3	1	30	13	9	11	2	4	10	3	0.647058824	0.235294	1.764705882	Winner	QuarterFinal
2012	France	23	4	3	0	1	5	2	2	37	12	13	15	2	6	12	6	0.652173913	0.26087	1.608695652	Winner	QuarterFinal
2012	Germany	22	4	3	0	1	9	6	1	57	29	7	14	3	5	11	5	0.636363636	0.227273	2.590909091	Winner	Semi-Finals
2012	Greece	21	2	1	0	1	2	0	2	23	13	10	11	1	9	10	9	0.523809524	0.428571	1.095238095	Winner	QuarterFinal
2012	Italy	20	4	1	1	1	3	3	0	27	13	10	11	5	4	10	3	0.55	0.2	1.35	Winner	Final

Having applied **Attribute Construction** on our dataset, we went from knowing little about our data to having a deeper understanding about how each team performed leading up to the competition, meaning if their high goals scored was down to playing easy opposition, or them actually being an attacking force against top ranked opposition.

To be able to summarise some our data, we had to apply **Data Aggregation**.

Data Aggregation is any process in which information is gathered and expressed in a summary form. This is also a form of data reduction. We summarised the total win, total loss and goals per game of each team in percentage form.

From this we able to apply **Data Generalisation** to our dataset. This is where low-level data is replaced by higher-level concepts using hierarchies. Based on the **data aggregated** results, we took each country's win percentage entering the tournament, generalised them in categories called "Elite, Moderate or Lower".

	A	B	C	S	T
1	TournamentYear	Team	Total Games Played	Total Win%	Level Based On Total Win%
2	2012	Croatia	20	0.55	Moderate
3	2012	Czech Republic	20	0.45	Lower
4	2012	Denmark	18	0.555555556	Moderate
5	2012	England	17	0.647058824	Elite
6	2012	France	23	0.652173913	Elite
7	2012	Germany	22	0.636363636	Elite
8	2012	Greece	21	0.523809524	Moderate
9	2012	Italy	20	0.55	Moderate
10	2012	Netherlands	21	0.666666667	Elite
11	2012	Poland	25	0.44	Lower
12	2012	Portugal	18	0.5	Lower
13	2012	Republic of Ireland	24	0.5	Lower
14	2012	Russia	21	0.523809524	Moderate
15	2012	Spain	22	0.727272727	Elite
16	2012	Sweden	23	0.565217391	Moderate
17	2012	Ukraine	22	0.409090909	Lower

Doing this gave us a clearer view of how good or bad a team's form was prior to them entering the competition. It helped us read the data as in the beginning the attributes were very spread but based on important attributes, we were able to generalise them into the three classes. Once again, lot of the classes we generalised the teams into were as expected prior from our own biased knowledge.

The workshop we chose involved a lot of work for us, as we had to gather further information for each of the 4 tournaments we chose as our training dataset. After our workshop we had a larger dataset of 21 attributes. Our workshop gave us an insight into which possible attributes together effected a team's performance in the tournament. We discovered this through various transformation techniques and correlation coefficients of different attributes.

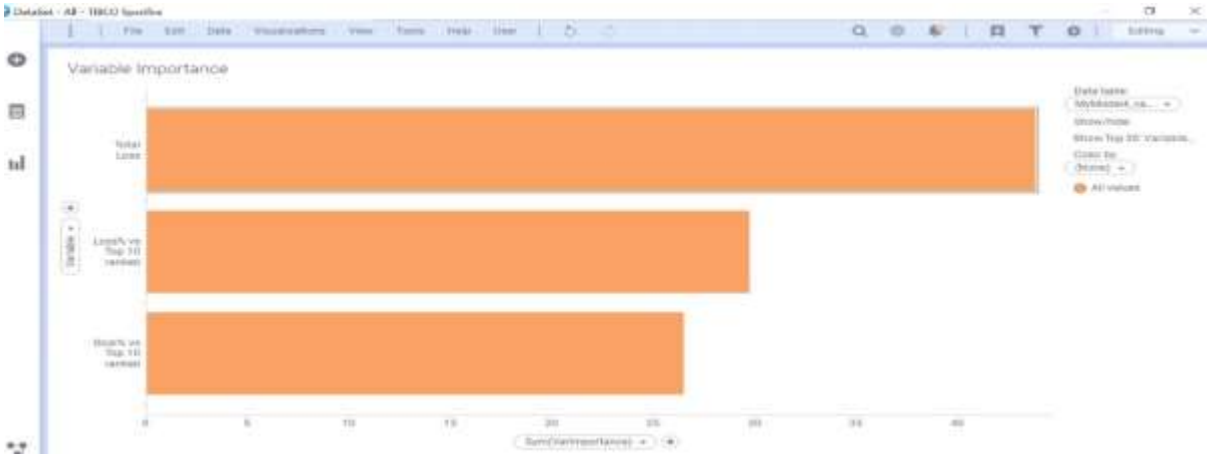
4. Algorithm Description & Implementation

From our workshop we had a better insight into which attributes were important, and which were not. We chose a Classification approach and felt this was very suitable to our problem.

We looked at several algorithms which may be suited to our problem such as Bayes Classifier, k-NN Classification and Decision Trees. With our data being numeric and so spread, we felt that a Decision Tree was the best approach to classify teams and through a set of rules. We chose a very powerful algorithm called Top-Down Induction of Decision Trees.

Attributes Selection

From our data preparation workshop, we discovered certain attributes with were deemed import or had an influence on the resulting class, however remaining attributes were still too spread with little correlation between them. We were finding it difficult as they were all numeric. We used a Machine Learning application **TIBCO Cloud Sportifre** which you can upload your dataset to, select an algorithm, and run it and view the predictions. We used this as it tells you which attributes are important or have a lot of influence on the resulting outcomes. We wanted to back up our own analysis before starting to build our decision tree to ensure our selections were not biased as we felt the start of the model was the most important.



We chose the below attributes from our dataset to use in our Decision Tree.

- Loss % vs Top 10 ranked team
- Goal % vs Top 10 ranked team
- Total No. of Losses
- Total Games Played
- Total Clean Sheets
- Total Goals Conceded
- Stage Reached (CLASS)

From our analysis these attributes had the most importance as they not only **showed a correlation to the stage a team may reach**, but from a information perspective they took into account the quality of a team both attacking and defensively, their performance against difficult opposition(which they will face in a tournament), and also the possible effect of fatigue(playing too many games).

Ranking Attributes

We looked the range of each set of attributes and **split these into subsets** based on the Stage reached in the tournament. For example, every team who reached the final had between 2-6 losses.

Stage	Loss % vs Top 10	Goal % vs Top 10	No. of Total Loss	Total Games Played	Total Clean Sheets	Total Goals Conceded
Final	0-.66	.33-2.5	2-6	16-22	8-13	13-23
Semi Final	0-.66	.33-2.5	2-11	17-25	5-11	9-29
Quarter Final	0-1	.33-3	1-9	11-25	5-18	8-28
Group	0-1	0-2.33	1-10	17-26	2-15	9-36

The **attributes selection strategy** we chose to implement was **Random**. There was nothing to be

gained by using the 'take first' or 'take last' methods as the numeric attributes varied and intersected so much we realised we would have to be as precise and careful as possible in creating branches and rules.

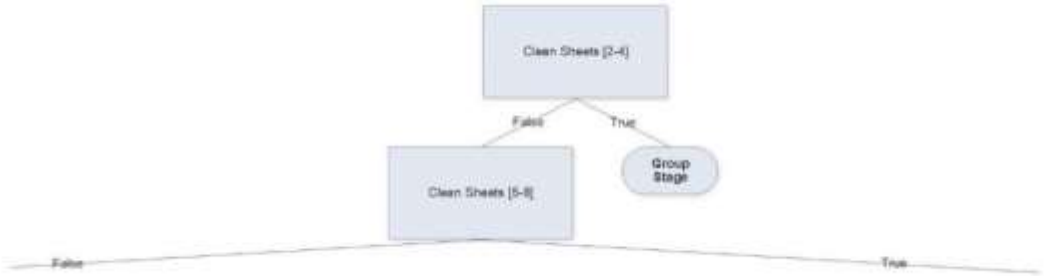
Information Gain

We implemented a **(TDIDT)Top-Down Induction of Decision Trees**. The goal was to eliminate or classify as large a set of teams at each rule as possible. To do this we needed to calculate the information gain that each rule would provide, that is, after putting a set of teams through a rule, how large a subset would this rule create. We were essentially calculation the Gain ratio each rule in the tree provided.

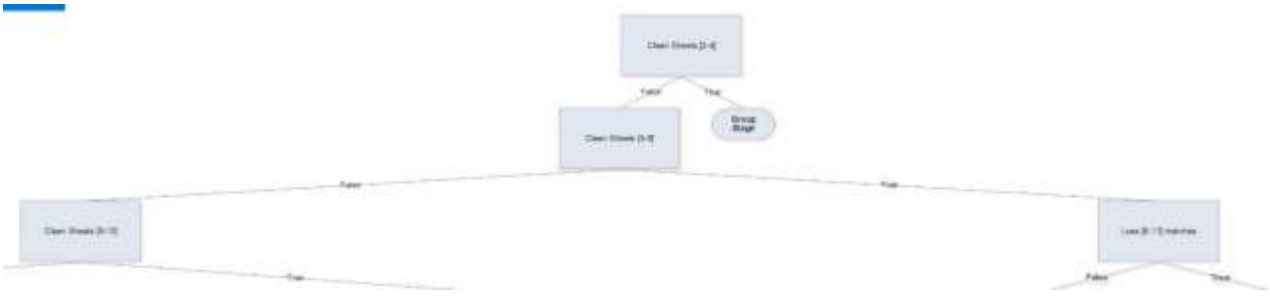
$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

The aim was the have the rules which had the largest ration first, as the further down the tree you, the more the sets of ranges intersect and more precise the rules become.

Our first rule with the highest information gain ratio which classified 20/64 teams was ***'between 2-4(inclusive) clean sheets '.***



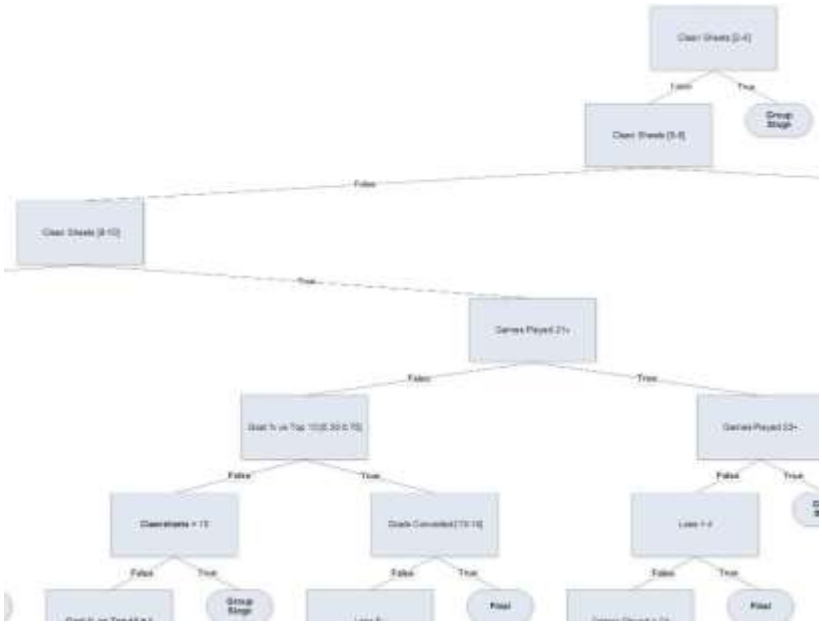
Our TDIDT continued to classify by choosing the subset with the largest information gain. We observed through this process the importance of defensive attributes as opposed to offensive (such as goals scored).



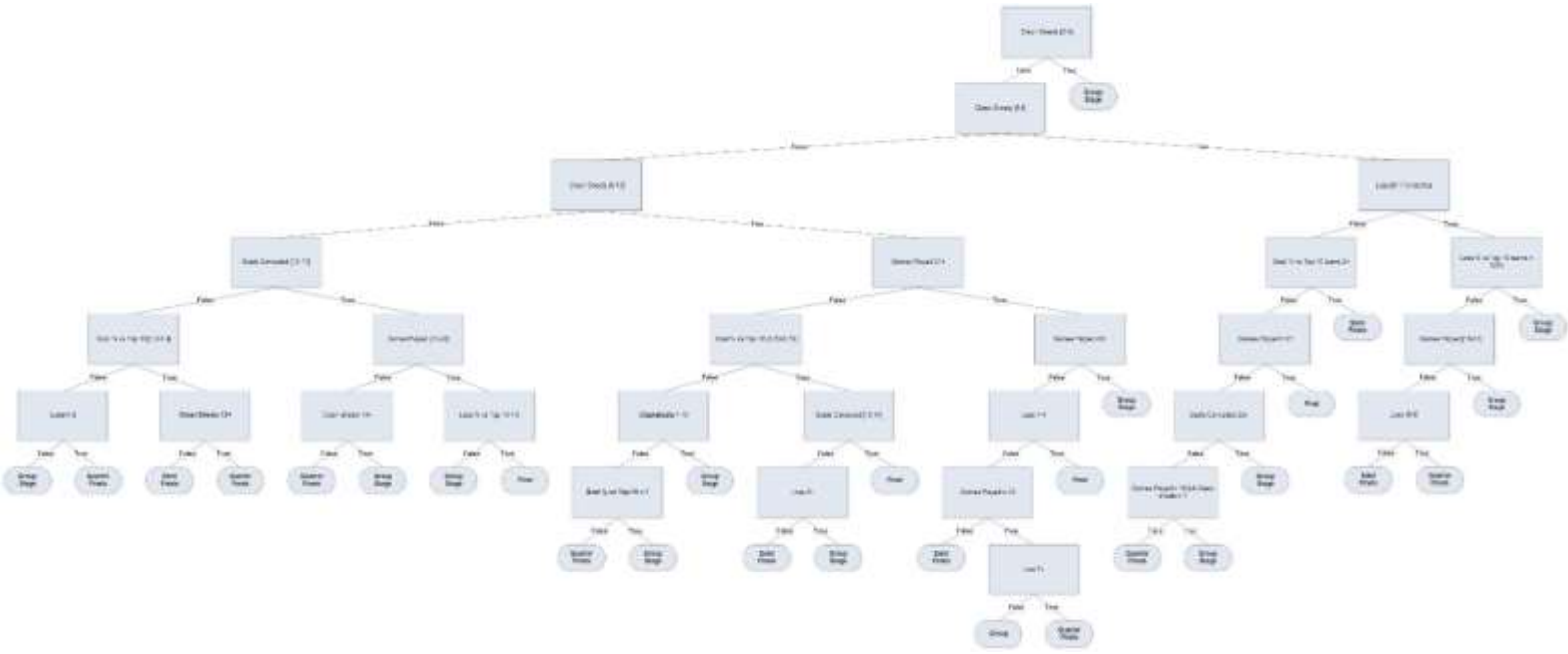
The further down the tree we went the more difficult it became to partition the subsets and classify the teams with 1 rule alone. We had to **create further smaller subsets** from our original ones as there was intersections between other attributes.

For example: if we had 3 teams that qualify for 1 rule or set with a high information gain ratio, due to other statistics the 3 teams may not have in common we cannot use this rule.

The below image illustrates the middle of the Decision Tree. We can see from this already how the sets are being partitioned smaller and rules becoming more exact.



Finally, the below image illustrates our full Top-Down Induction of Decision Tree implementation based on our dataset of statistic for 64 teams which are classified to a stage reached throughout 4 different UEFA European Championship Tournaments.



5. Results

Detailed in the section above, is how we implemented our own Decision Tree algorithm. Training this algorithm on 80% of our data and testing it on the other 20% of our data enabled us to achieve an accuracy of 44%. This was the mean accuracy from the iterations of our calculations. The accuracy we attained was low and was not what we were not expecting after executing the relevant data transformation, data integration and decision tree techniques.

When we ran our calculations, the countries in our dataset is passed through our decision tree and depending on their statistics, they either go down a True of False path to take them to the resulting child node. This ensures that our accuracy is not inflated and it also displays the robustness of our solution. We feel that this is a decent accuracy score considering how few rows we had in our dataset. We understood the disadvantages of choosing such a small dataset, however, we felt that the number of attributes made available to us, helped us to battle these disadvantages.

We noticed some enormous differences between the results of the stages our algorithm returned for some countries and the actual stages that the country reached. For example, our algorithm returned that Portugal would be placed in the group stages based on their statistics, but Portugal actually won the EUROs in 2016.

Another difference we noticed was that when we passed Turkey through our algorithm, the result we received was that Turkey would end up in the final but in reality, Turkey exited the EUROs in the group stages. We observed that this happened because of Greece in our training data. Greece won the EUROs in 2004 despite having very poor statistics entering the competition. We didn't remove Greece from the training set which meant we could spot another potential Greece sort if they were to emerge in the future. Seeing a Turkey had poor statistics, we observed and concluded that was the reason for this

An example that was accurate was the position that Spain attained in the competition. Both our algorithm and the actual stage reached showed that using their statistics we were able to correctly predict Spain's position in the competition.

We believe that the reason for this was due to having only 4 EUROs competition as our training set. If we had more than 4, say the last 10 European Championships, we would have a more detailed statistical advantage that would help increase our overall accuracy.

It was somewhat disheartening to discover that many of the attributes which were indicative of an elite country could not play a part in the country going far in the competition. The fact that our

prediction algorithm was inaccurate for more than half of the time indicates that there are other factors at play. It is possible for a country who fits our model of an elite country may have not been able to handle the pressure of playing at such a big tournament or simply underperformed at the time of the tournament and therefore exited early. Similarly, a country who fit our model of a lower country may have been able embrace the pressure and overperformed at the tournament. We also didn't rule out the possibility that some countries would have been favoured with controversial decisions that would've seen them go farther than they were meant to in the competition. This is simply too much scope in a number of the key attributes in our dataset.

We believe that we performed as well as possible given the context. We outperformed other algorithms using our own decision tree and in doing so, we displayed our deep understanding of the algorithm and the nature of our dataset.

6. Discoveries & Conclusion

Throughout our Data Mining project, we made a lot of discoveries and learned a lot, which proved many of our initial assumptions to be false.

Creating our own dataset and then transforming it involved a lot of work for us which meant we could not create a very large dataset as we would have liked. We assumed that our original attributes would be enough to form predictions through a classification method. However, all of our attributes were numeric, and they were more spread then we originally assumed.

Data Transformation

During our Data Transformation process, we focused on increasing the number of attributes we had, and then finding any relations between these. We gather information on teams who played **difficult matches** as these would produce different results to easier opposition.

Having used this in our predictions we learned we should have also looked at teams who played against **very poor ranked teams**, to account for the amount of goals they may have scored against these.

Focussing on **only one** of the two theory's above we felt **skewed our data** and attribute selection process **towards defensive statistics**, rather than fully crediting goals and wins.

Algorithm choice

Having analysed our Dataset, we thought that if we had of preformed clustering and transformed the numeric data into ranges or classes it would have skewed the results and not been accurate.

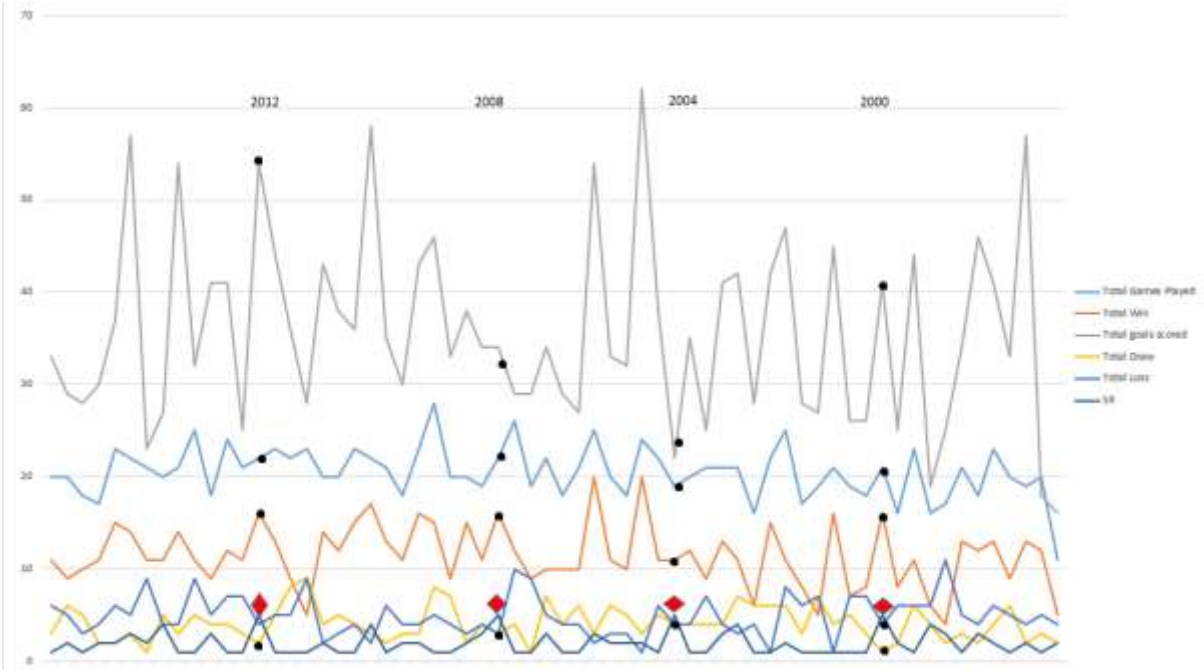
Ultimately, we were satisfied with our Top-Down Induction of Decision Tree Classification Model, as we felt it was best suited to our data and our goal.

Attribute selection

We spent a large amount of our time trying to identify which attributes to use in our model. We used various techniques such as calculating **correlations** between attributes and using the **χ2(chi-square)**

test. This test calculates if attribute A causes a change in attribute B. For example, playing more games means you will win more matches / score more goals.

Below is a graph which illustrates statistics for each of the teams who played in Euro 2000,2004,2008,2012.



We were able to narrow our attribute selection down, but wanted some sort of evolution that these were the correct ones to use. As mentioned previously we used online software, which backed our selections. This proved to us that the calculations we preformed were indeed to be very useful in our selections for our model.

	A	B	C	D	E	F	G
	Tournam	Team	CorrelationTotalGames/wins	CorrelationTotalGames/loses	CorrelationGamesPlayed/GoalsScored	Total Games Played	Games played vs Norma
	2012	Croatia	0.15396444	-0.137573837	0.176410086	20	20
	2012	Czech Rep	0.601860991	0.137573837	0.355620331	20	17
	2012	Denmark	1.089277826	1.982681766	1.154160103	18	13
	2012	England	0.588687563	1.578052834	1.188420885	17	15
	2012	France	1.352746386	0.250869938	0.005106174	23	19
	2012	Germany	0.456953283	-0.12138868	0.793104166	22	18
	2012	Greece	0.009056732	-0.05664805	0.036731512	21	19
	2012	Italy	0.15396444	0.41272151	0.445225454	20	16
	2012	Netherlar	-0.030463552	0.024277736	-0.044967277	21	17
	2012	Poland	-0.5705741	3.568827179	-0.819788045	25	21
	2012	Portugal	1.734775797	0.396536353	-0.524618228	18	16
	2012	Republic	0.193484723	1.141053588	0.50320524	24	20
	2012	Russia	0.009056732	-0.024277736	0.031460622	21	19
	2012	Spain	0.852156123	-0.364166039	0.674509151	22	18
	2012	Sweden	0.535993851	-0.250869938	0.576997694	23	22
	2012	Ukraine	-0.531053816	-0.12138868	-0.037060942	22	17
	Avg		0.412492964	0.526017611	0.282157308	21.0625	
	StandardDev					2.205107707	
	SUM						
	2008	Austria				23	
	2008	Croatia				20	
	2008	Czech Republic				20	
	2008	France				23	
	2008	Germany				22	
	2008	Greece				21	
	2008	Italy				18	

Finally

We believe that given a **larger dataset** we would be able to create a more accurate solid model which we could **evaluate and tweak** as we tested it out, as having detailed attributes just was not enough considering the context of what we were aiming to predict.