Predicting the Beautiful Game

Shane McCarthy

shane.mc-carthy@ucdconnect.ie

05 April 2016

Introduction

The primary objective of this assignment was to deliver “interesting and potentially useful patterns and rules” which could be used to predict the outcomes of the remaining games in this season’s primary league. To accomplish this objective two approaches were developed, the first is a series of tree based models to predict whether or not the home team wins the game – this is effectively a binary classification task, with home team wins labelled as “1” and home team losses or draws being labelled as “0”. The second approach treats opposing teams independently and predicts the expected number of goals scored - this is effectively a multiclass classification task, again a series of tree based models were built.

When tested against the out-of-time test data both approaches performed relatively well, the best performing model for approach one (home win) was a Random Forest model which was able to correctly predict 70% of the game results. The best performing model for approach two (expected goals) was a Gradient Boosting model which was able to correctly predict the number of goals scored 52.5% of the time. A full breakdown of model performance statistics can be found in the results section of this document.

The KDD pipeline methodology was used to guide the data exploration journey moving rapidly from raw data through to predictive insights, the methodology section of this document is broken-down into the KDD pipeline steps namely Selection of Data, Pre-processing & Cleaning, Feature Selection & Extraction and Data Mining & Modelling. Interpretation & Evaluation is worthy of its own subsection and can be found under Results. Please note that no code is presented in the main body of text, this can be found in the Appendix section.

Methodology

The KDD pipeline (Fayyad, et al., 1996) approach was used in the development of the models, the unifying goal of the KDD process is to extract knowledge from data in the context of large databases. The KDD pipeline outlined in Figure 1 below is a useful tool to help guide analysts in their data exploration journey and consists of the following steps: Selection of data, Pre-processing & cleaning, Transformation, Data mining and Interpretation. The remaining methodology section is broken-down into these section.

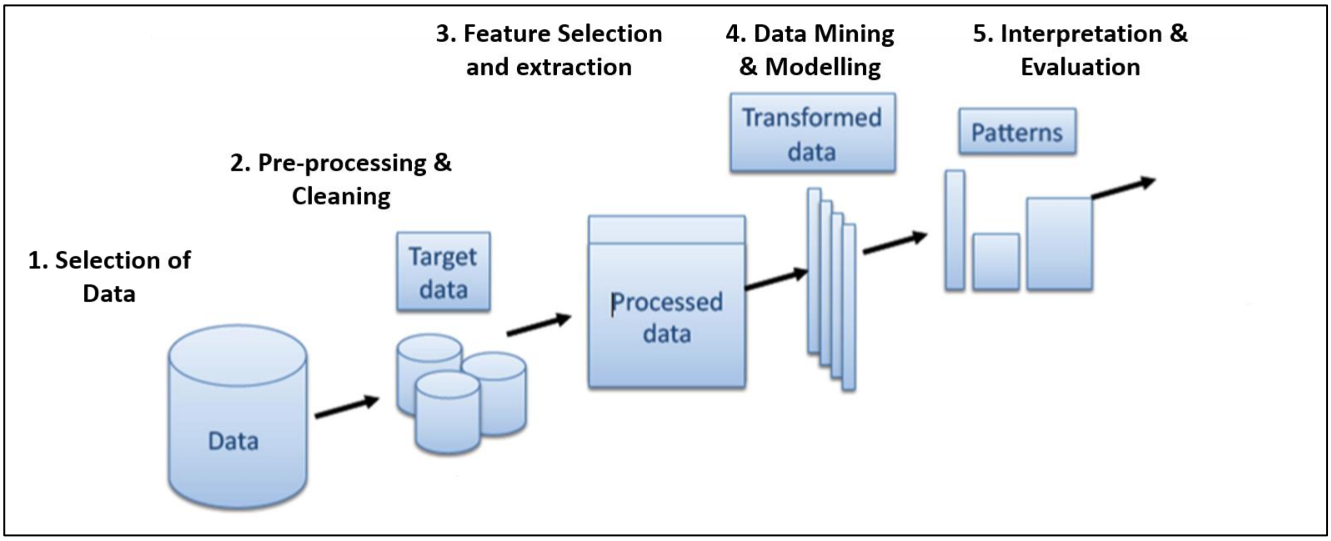

3. Feature Selection and extraction
4. Data Mining & Modelling
5. Interpretation & Evaluation
Transformed data
Patterns
2. Pre-processing & Cleaning
1. Selection of Data
Target data
Processed data
Data

Figure 1 KDD Pipeline

1. Selection of Data

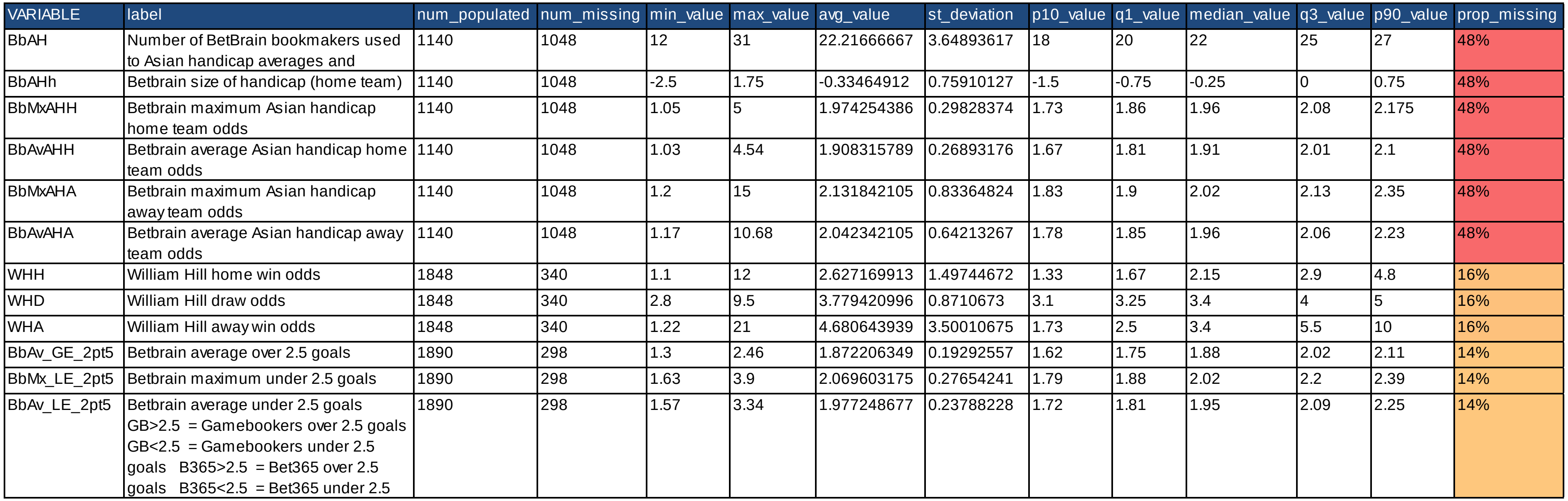
All available data available was read into the SAS environment from the six CSV files using an import loop macro[1] [csv_importer], this macro has a number of input parameters including number of files making it easily adaptably. As all variables were read in as type character, a type conversion macro [char2num] is used to ensure all variables are of the correct type. The final step of the data selection process is to read in the variable descriptions from the notes file, SAS allows us to store variable descriptions as “labels” in the table metadata, this can prove very beneficial when working with new variables.

2. Pre-processing & Cleaning

A thorough analysis of individual data attributes was preformed providing a quantitative assessment of the data quality using two data profiling macros for numeric data [univariate_num] and for categorical data [univariate_char]. In order to produce a high-quality model that can accurately predict outcomes, it is vital to explore the underlying data for the purposes of better understanding its characteristics. By gaining this insight and identifying which data to focus on, the modelling process can be highly accelerated and produce a more accurate, targeted result by avoiding what would ultimately prove to be unnecessary steps during model build.

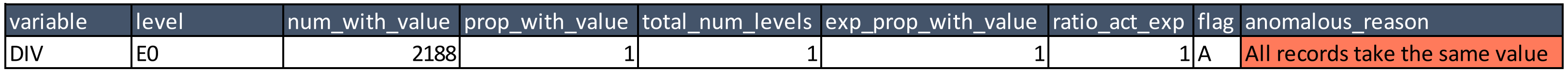
Following profiling 12 numeric variables and 1 categorical variable were dropped and not carried forward to the next stage of analysis. Details of these variables including the reasons why they were dropped can be found in Table 1 and Table 2 below.

[1] A “macro” in SAS is essentially what we would call a function in other programming languages



VARIABLE	label	num_populated	num_missing	min_value	max_value	avg_value	st_deviation	p10_value	q1_value	median_value	q3_value	p90_value	prop_missing
BbAH	Number of BetBrain bookmakers used to Asian handicap averages and	1140	1048	12	31	22.21666667	3.64893617	18	20	22	25	27	48%
BbAHh	Betbrain size of handicap (home team)	1140	1048	-2.5	1.75	-0.33464912	0.75910127	-1.5	-0.75	-0.25	0	0.75	48%
BbMxAHH	Betbrain maximum Asian handicap home team odds	1140	1048	1.05	5	1.974254386	0.29828374	1.73	1.86	1.96	2.08	2.175	48%
BbAvAHH	Betbrain average Asian handicap home team odds	1140	1048	1.03	4.54	1.908315789	0.26893176	1.67	1.81	1.91	2.01	2.1	48%
BbMxAHA	Betbrain maximum Asian handicap away team odds	1140	1048	1.2	15	2.131842105	0.83364824	1.83	1.9	2.02	2.13	2.35	48%
BbAvAHA	Betbrain average Asian handicap away team odds	1140	1048	1.17	10.68	2.042342105	0.64213267	1.78	1.85	1.96	2.06	2.23	48%
WHH	William Hill home win odds	1848	340	1.1	12	2.627169913	1.49744672	1.33	1.67	2.15	2.9	4.8	16%
WHD	William Hill draw odds	1848	340	2.8	9.5	3.779420996	0.8710673	3.1	3.25	3.4	4	5	16%
WHA	William Hill away win odds	1848	340	1.22	21	4.680643939	3.50010675	1.73	2.5	3.4	5.5	10	16%
BbAv_GE_2pt5	Betbrain average over 2.5 goals	1890	298	1.3	2.46	1.872206349	0.19292557	1.62	1.75	1.88	2.02	2.11	14%
BbMx_LE_2pt5	Betbrain maximum under 2.5 goals	1890	298	1.63	3.9	2.069603175	0.27654241	1.79	1.88	2.02	2.2	2.39	14%
BbAv_LE_2pt5	Betbrain average under 2.5 goals GB>2.5 = Gamebookers over 2.5 goals GB<2.5 = Gamebookers under 2.5 goals B365>2.5 = Bet365 over 2.5 goals B365<2.5 = Bet365 under 2.5	1890	298	1.57	3.34	1.977248677	0.23788228	1.72	1.81	1.95	2.09	2.25	14%

Table 1 Numeric Data Profiling (dropped variables only)



variable	level	num_with_value	prop_with_value	total_num_levels	exp_prop_with_value	ratio_act_exp	flag	anomalous_reason
DIV	E0	2188	1	1	1	1	A	All records take the same value

Table 2 Categorical Data Profiling (dropped variable only)

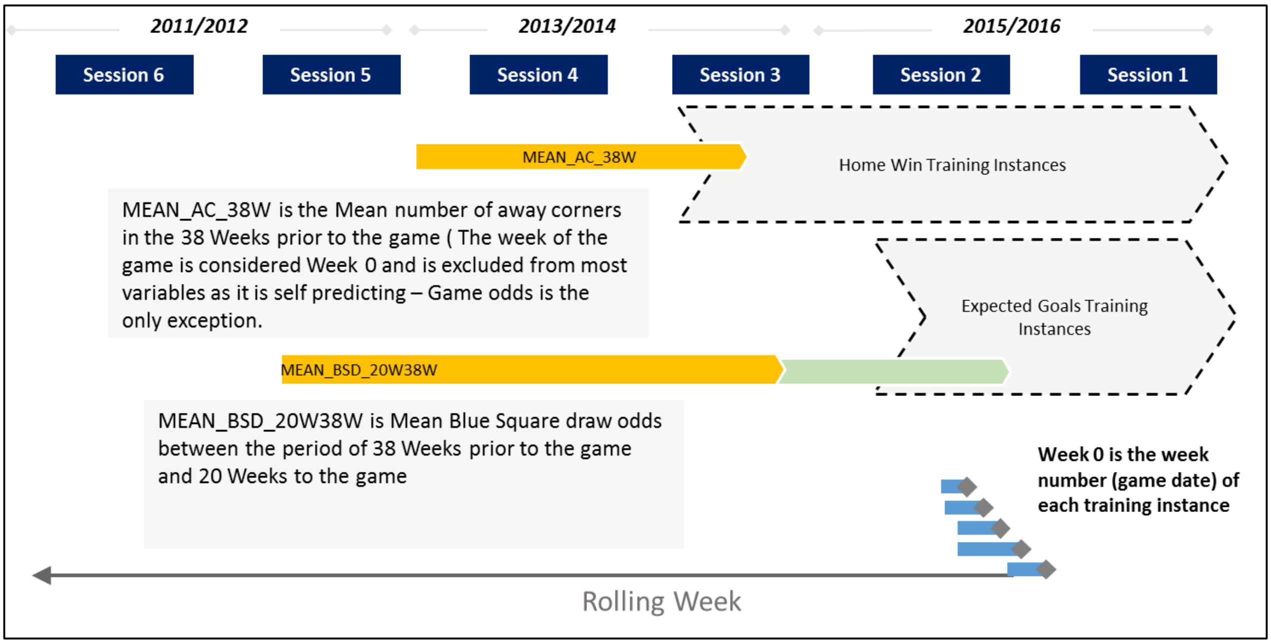
3. Feature Selection & Extraction

In this section we discuss building out the analytics base tables (ABT) which is the foundation for building a predictive model to assess probability of a home win and predicting the number of goals scored by each team. Typically an ABT provides a single connected table at a particular granularity for modelling purposes.

In total 6 ABTs were constructed for the assignment, 3 for the Home Win Model consisting of 1,2 & 3 seasons of game outcomes and 3 for the Expected Goals Model against consisting of 1,2 & 3 seasons worth of data. The Home Win Models performed best on the 3 seasons of data and the Expected Goals Models performed best on 2 seasons worth of data, reporting on all 6 ABTs is outside the scope of this assignment therefore for the remaining of this paper we'll only report on the Home Win 3 Seasons ABT (ABT_HW_3_Season) and the Expected Goal 2 Season ABT ABT_EG_2_Season.

It is important to note that although only the outcomes for 3 and 2 seasons respectfully are being considered as training instances, all the data (6 seasons) are considered as features. This is made possible by creating a rolling week variable, we do this by getting the week number for each game and combining it with the year before creating a unique week number for each particular week. Time in-between seasons is not considered, therefore if you create the feature for example - count of home wins in the last 38W for a game played on week 32 in season 1 (2016) this will include 6 weeks from season 2 (2015).

The rolling week feature creation logic is outlined in Figure 2 below, all rolling week features have a postfix of *"_nW"* or "*_nWnW"* where *n* is the week number. The time window features ("_nWnW") are particularly useful for extracting insights on how a team performed last season (_36W72W) or in the month prior to a game excluding the last 2 weeks (_2W6W). In total almost 4,000 features are created using this approach combined with a number of transformation functions including count, sum, max, min, standard deviation, ratios and ranks.



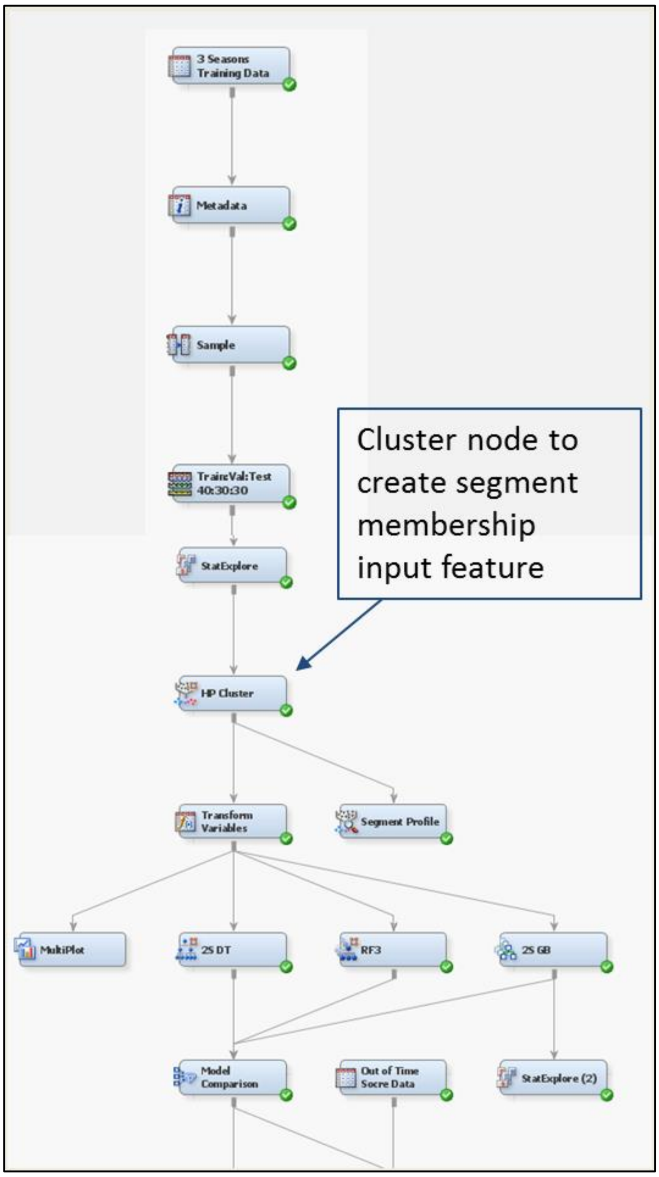
MEAN_AC_38W is the Mean number of away corners in the 38 Weeks prior to the game (The week of the game is considered Week 0 and is excluded from most variables as it is self predicting – Game odds is the only exception.

MEAN_BSD_20W38W is Mean Blue Square draw odds between the period of 38 Weeks prior to the game and 20 Weeks to the game

Week 0 is the week number (game date) of each training instance

Figure 2 Rolling week feature creation

The code to build the training universes and ABTs can all be found in appendix section: Feature Selection & Extraction Code.



Cluster node to create segment membership input feature

Figure 3 Enterprise Miner workflow Home Win Models

The next step is to import the ABTs into Enterprise Miner, the metadata node is used to ensure features are picked up as the correct type. We also use the metadata node to assign the target, in the previous step we created a binary target – 1 if the home team wins, 0 if the home team does not win.

The sample node is used to correct our target's class imbalance, on average home teams win ~55% of the time and don't win 45% of the time, this imbalance can potentially affect a models ability to learn so we simply under-sample the majority class.

The next step involves partitioning the data into train, validate and test sets 40:30:30 respectfully. A model learns from the train set, the validation set is used to tune the parameters of a model and the test set is a set of examples used only to assess the performance of a fully-trained model. It is vital to separate the test and validate sets as the error rate estimate of the final model will be biased since the validation set is used to select the final model.

The statExplore node is used to calculate the GINI importance of each feature, this is done by building a decision tree of depth one for each feature. The statExplore node provides a plot of GINI importance for all features, if a feature is has a disproportionally high GINI importance relative to other variables it may be worth investigating if this variable is self-predicting.

4. Data Mining & Modelling

Unsupervised and supervised modelling techniques are used in both the Home Win and Expected Goals model build. In this section the algorithms used are discussed at a relatively high level.

a. Unsupervised Clustering

The HP Cluster node is used to perform unsupervised K-means clustering on the input ABTs with objective of using the cluster membership as input to the predictive models. The Euclidean distance between data pairs can is measured and K-means clustering minimises the sum of squares for the distances between data and finds the corresponding cluster centroids, while k-NN rule assigns the unclassified sample to the class represented by a majority of its k number nearest neighbours in the training set.

A number of different values of K we tried before k=6 was selected as the best possible number of clusters. Figure 4 below illustrates the distance between the 6 clusters when plotted against the 1st principle and 2nd principle component.

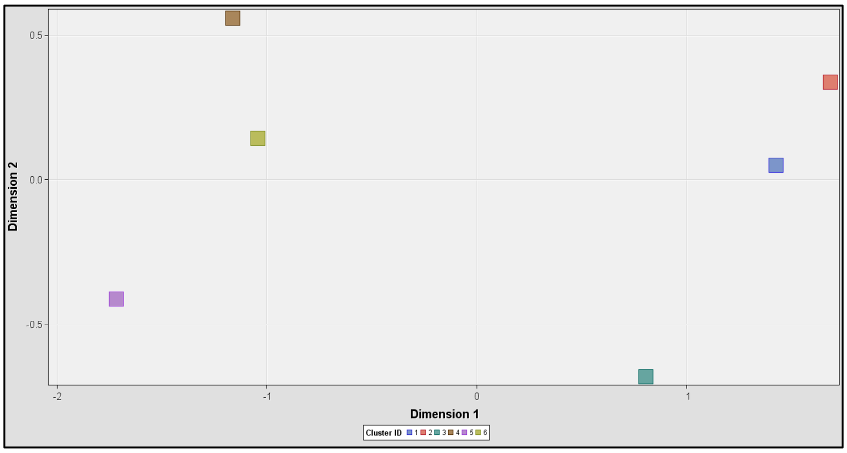


Figure 4 K-means clustering Principal Component plot

The Segment profile node is used to produce the cluster comparison plots in Figure 5 below, the population distribution is plotted in red and the cluster distribution is plotted in blue – making it easy to understand how a cluster is different to the overall population.

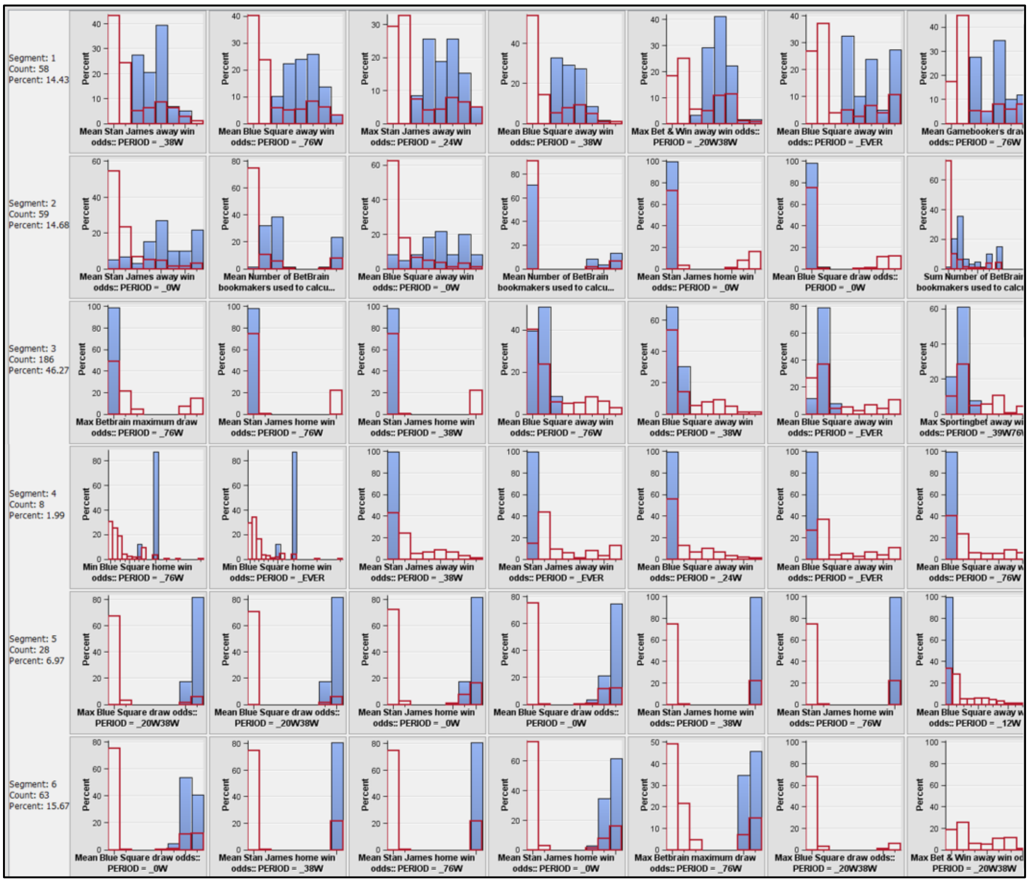


Figure 5 Segment Profile

b. Supervised learning – Decision Tree

A simple decision tree was the first model to be built for both Home Win and Expected Goals ABTs, the output if which is highly interpretable – this is partially useful for detecting issues such as self-predicting variables and future leakage. A decision tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The performance of both decision trees was relatively poor without overfitting the training data, therefore more advanced tree based models were considered.

c. Supervised learning – Random Forest

A Random Forest consists of several decision trees that differ from each other in two ways. First, the training data for a tree is a sample without replacement from all available observations. Second, the input features that are considered for splitting a node are randomly selected from all available features (Breiman, 2001). Our targets are binary and categorical therefore the posterior probabilities in the forest are the averages of the posterior probabilities of the individual trees. The node makes a second prediction by voting: the forest predicts the target category that the individual trees predict most often.

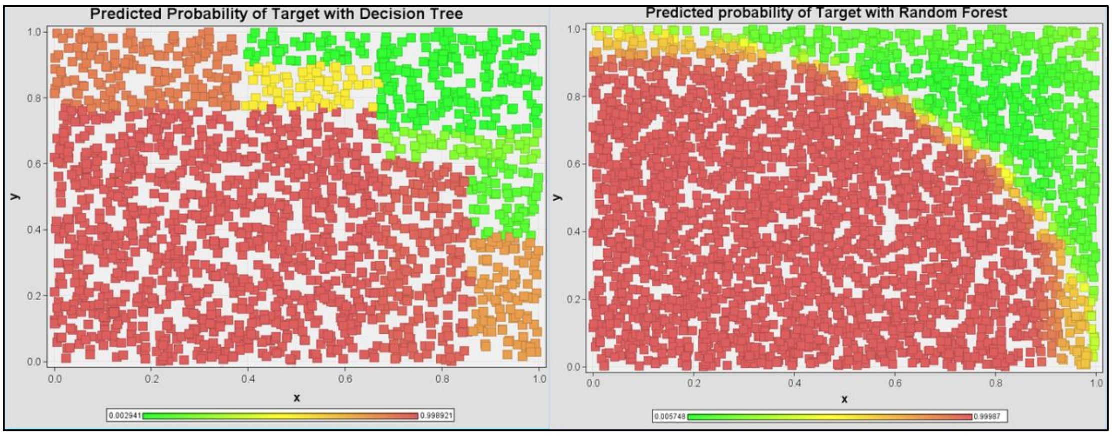
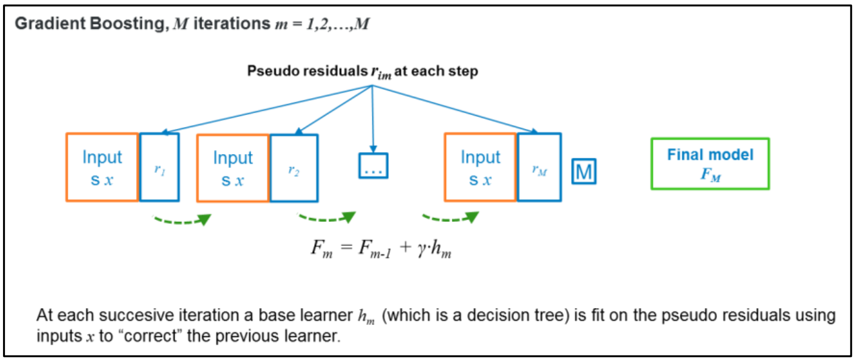


Figure 6 Probability of Decision Tree Vs Random Forest

d. Supervised learning – Gradient Boosting

The third and final model considered was Gradient boosting, this boosting approach resamples the training data several times to generate results that form a weighted average of the re-sampled data set. Tree boosting creates a series of decision trees which together form a single predictive model. A tree in the series is fit to the residual of the prediction from the earlier trees in the series. The residual is defined in terms of the derivative of a loss function (Friedman, n.d.).


Gradient Boosting, M iterations $m = 1,2,\ldots,M$
Pseudo residuals r_{im} at each step
Input s x
r_1
Input s x
r_2
...
Input s x
r_M
M
Final model F_M
$F_m = F_{m-1} + \gamma \cdot h_m$

At each succesive iteration a base learner h_m (which is a decision tree) is fit on the pseudo residuals using inputs x to "correct" the previous learner.

Figure 7 Gradient Boosting Pseudo

Results

1. Approach one – Home Win Prediction

Using the Random Forest model a Home Win misclassification rate of 33% was achieved against the holdout test data. When testing against the out-of-time test data the Model correctly predicted 14 of 20 results (misclassification rate of 30%.). Table 3 below presents the actual vs predicted results for the out-of-time test data

UID	HomeTeam	AwayTeam	Referee	ACTUAL	PREDICTED	PRED_PROB	TN	FP	FN	TP	MISCLASS
2194	Chelsea	West Ham	R Madley	0	1	0.58	0	1	0	0	1
2195	Crystal Pal	Leicester	M Jones	0	0	0.32	1	0	0	0	0
2196	Everton	Arsenal	M Clattenburg	0	0	0.38	1	0	0	0	0
2197	Swansea	Aston Villa	M Dean	1	0	0.46	0	0	1	0	1
2198	Watford	Stoke	C Pawson	0	0	0.4	1	0	0	0	0
2199	West Brom	Norwich	A Taylor	0	0	0.42	1	0	0	0	0
2200	Man City	Man United	M Oliver	0	0	0.48	1	0	0	0	0
2201	Newcastle	Sunderland	M Atkinson	0	0	0.44	1	0	0	0	0
2202	Southampton	Liverpool	R East	1	0	0.41	0	0	1	0	1
2203	Tottenham	Bournemouth	N Swarbrick	1	1	0.68	0	0	0	1	0
2204	Arsenal	Watford	A Taylor	1	1	0.8	0	0	0	1	0
2205	Aston Villa	Chelsea	N Swarbrick	0	0	0.3	1	0	0	0	0
2206	Bournemouth	Man City	R Madley	0	0	0.38	1	0	0	0	0
2207	Liverpool	Tottenham	J Moss	0	0	0.5	1	0	0	0	0
2208	Norwich	Newcastle	M Dean	1	0	0.36	0	0	1	0	1
2209	Stoke	Swansea	M Atkinson	0	0	0.43	1	0	0	0	0
2210	Sunderland	West Brom	R East	0	0	0.4	1	0	0	0	0
2211	West Ham	Crystal Pal	M Clattenburg	0	0	0.45	1	0	0	0	0
2212	Leicester	Southampton	M Oliver	1	0	0.41	0	0	1	0	1
2213	Man United	Everton	A Marriner	1	0	0.45	0	0	1	0	1

Table 3 Out-of-Time test Predicted Vs Actual

The confusion matrix for the Random Forest model is outlined below in Figure 8, the number of true positives is relatively low compared to the number of true negatives, and the majority of our misclassifications are false negatives. If we examine the probability of class “1” (PRED_PROB) in Table 3 above almost all of the false negative classifications have a probability in > .4. Adjusting the classification cut-off from .5 down to .4 will reduce the number of false negatives but will increase the number of false positives – optimising the cut-off to minimise the misclassification is a non-trivial task.

		Predicted	
		0	1
Actuals	0	TN -15	FP -1
	1	FN-5	TP -2

Figure 8 Confusion Matrix Random Forest

	Roc Index			Misclassification Rate			Cumulative Lift			Gini Coefficient		
	Train	Test	Validate	Train	Test	Validate	Train	Test	Validate	Train	Test	Validate
Model Description												
Decision Tree	0.83	0.67	0.62	0.26	0.39	0.42	1.97	1.25	1.16	0.67	0.34	0.24
Gradiant Boosting	0.81	0.74	0.69	0.28	0.34	0.33	1.8	1.75	1.48	0.62	0.48	0.38
RandomForest	0.85	0.72	0.7	0.27	0.33	0.38	2	1.68	1.74	0.69	0.44	0.39

Figure 9 Performance statistics approach one (all models)

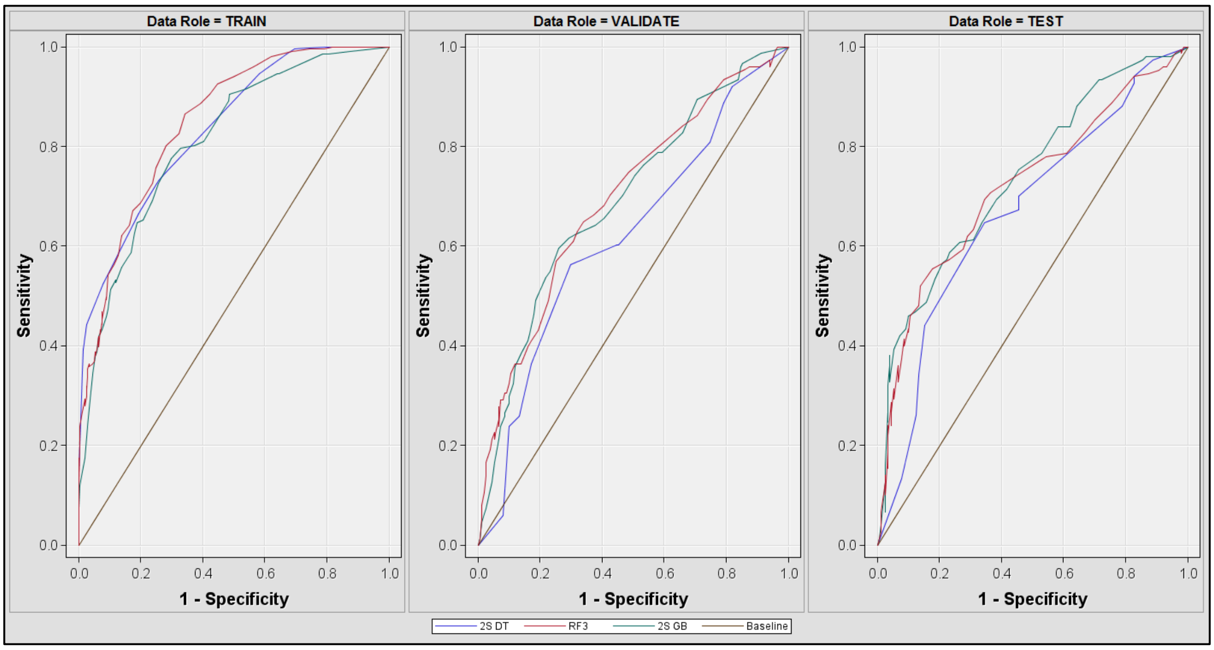


Figure 10 ROC chart approach one (all models)

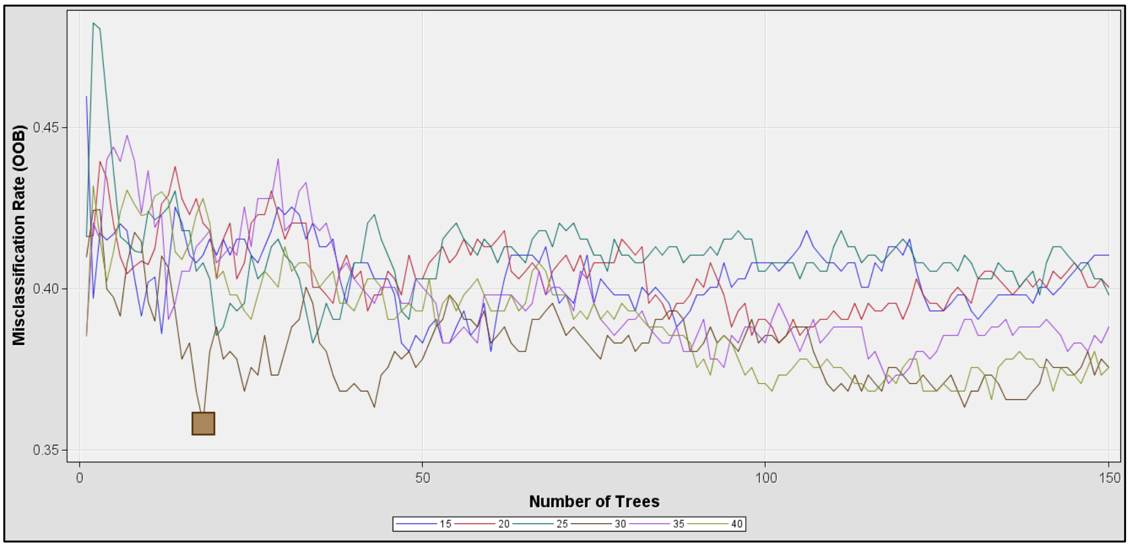
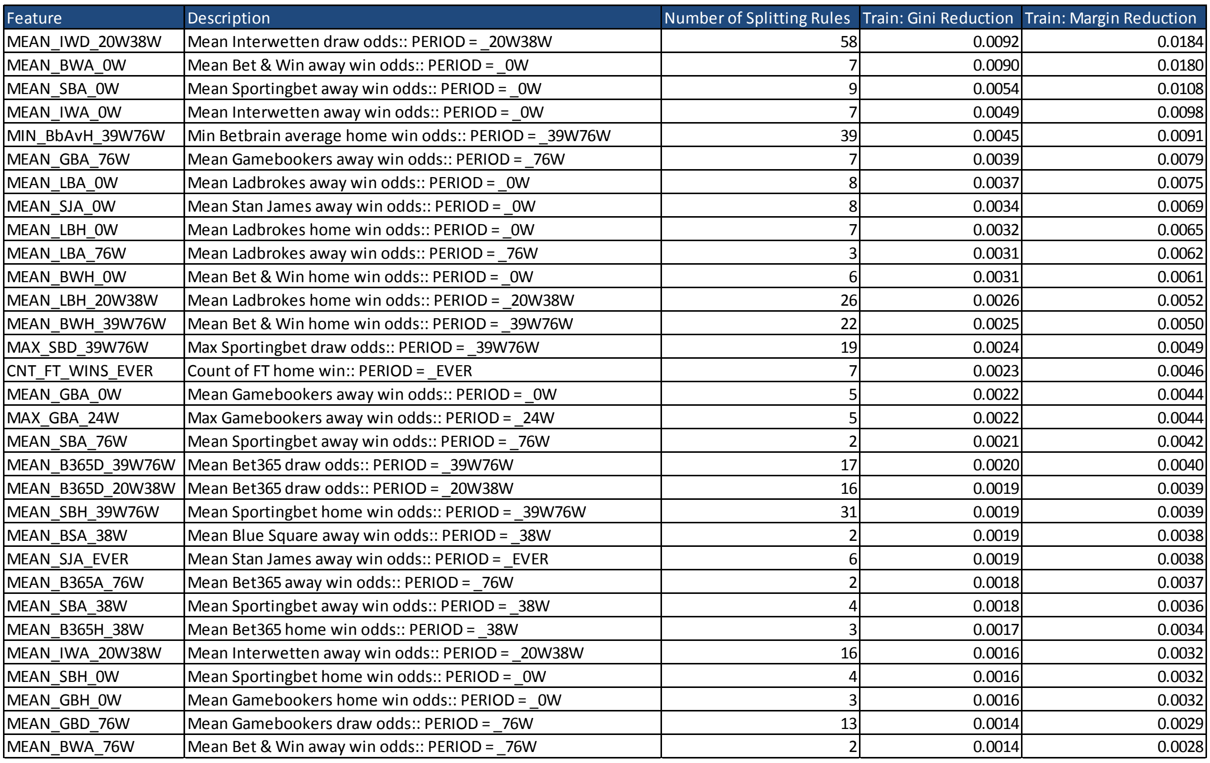
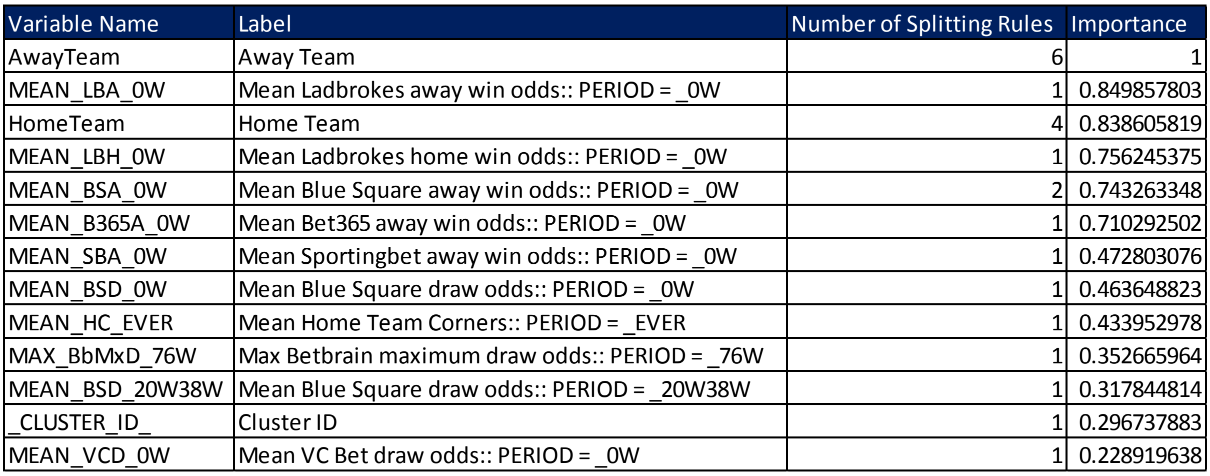


Figure 11 Optimally selecting the number of trees in the Random Forest



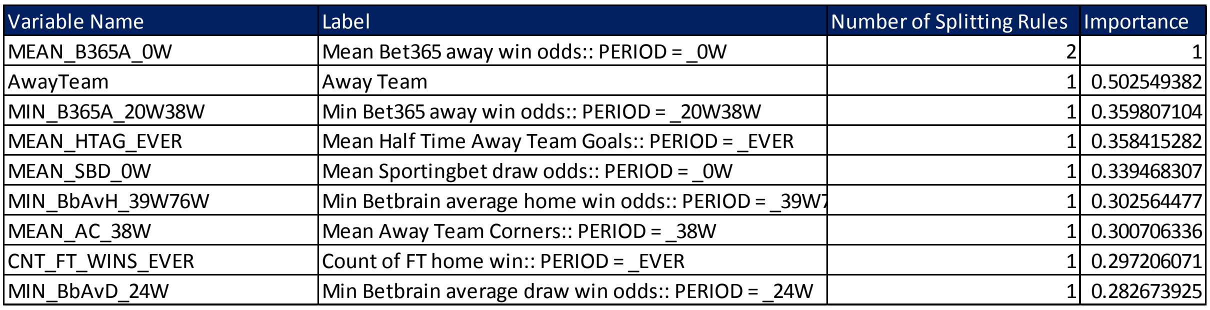
Feature	Description	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction
MEAN_IWD_20W38W	Mean Interwetten draw odds:: PERIOD = _20W38W	58	0.0092	0.0184
MEAN_BWA_0W	Mean Bet & Win away win odds:: PERIOD = _0W	7	0.0090	0.0180
MEAN_SBA_0W	Mean Sportingbet away win odds:: PERIOD = _0W	9	0.0054	0.0108
MEAN_IWA_0W	Mean Interwetten away win odds:: PERIOD = _0W	7	0.0049	0.0098
MIN_BbAvH_39W76W	Min Betbrain average home win odds:: PERIOD = _39W76W	39	0.0045	0.0091
MEAN_GBA_76W	Mean Gamebookers away win odds:: PERIOD = _76W	7	0.0039	0.0079
MEAN_LBA_0W	Mean Ladbrokes away win odds:: PERIOD = _0W	8	0.0037	0.0075
MEAN_SJA_0W	Mean Stan James away win odds:: PERIOD = _0W	8	0.0034	0.0069
MEAN_LBH_0W	Mean Ladbrokes home win odds:: PERIOD = _0W	7	0.0032	0.0065
MEAN_LBA_76W	Mean Ladbrokes away win odds:: PERIOD = _76W	3	0.0031	0.0062
MEAN_BWH_0W	Mean Bet & Win home win odds:: PERIOD = _0W	6	0.0031	0.0061
MEAN_LBH_20W38W	Mean Ladbrokes home win odds:: PERIOD = _20W38W	26	0.0026	0.0052
MEAN_BWH_39W76W	Mean Bet & Win home win odds:: PERIOD = _39W76W	22	0.0025	0.0050
MAX_SBD_39W76W	Max Sportingbet draw odds:: PERIOD = _39W76W	19	0.0024	0.0049
CNT_FT_WINS_EVER	Count of FT home win:: PERIOD = _EVER	7	0.0023	0.0046
MEAN_GBA_0W	Mean Gamebookers away win odds:: PERIOD = _0W	5	0.0022	0.0044
MAX_GBA_24W	Max Gamebookers away win odds:: PERIOD = _24W	5	0.0022	0.0044
MEAN_SBA_76W	Mean Sportingbet away win odds:: PERIOD = _76W	2	0.0021	0.0042
MEAN_B365D_39W76W	Mean Bet365 draw odds:: PERIOD = _39W76W	17	0.0020	0.0040
MEAN_B365D_20W38W	Mean Bet365 draw odds:: PERIOD = _20W38W	16	0.0019	0.0039
MEAN_SBH_39W76W	Mean Sportingbet home win odds:: PERIOD = _39W76W	31	0.0019	0.0039
MEAN_BSA_38W	Mean Blue Square away win odds:: PERIOD = _38W	2	0.0019	0.0038
MEAN_SJA_EVER	Mean Stan James away win odds:: PERIOD = _EVER	6	0.0019	0.0038
MEAN_B365A_76W	Mean Bet365 away win odds:: PERIOD = _76W	2	0.0018	0.0037
MEAN_SBA_38W	Mean Sportingbet away win odds:: PERIOD = _38W	4	0.0018	0.0036
MEAN_B365H_38W	Mean Bet365 home win odds:: PERIOD = _38W	3	0.0017	0.0034
MEAN_IWA_20W38W	Mean Interwetten away win odds:: PERIOD = _20W38W	16	0.0016	0.0032
MEAN_SBH_0W	Mean Sportingbet home win odds:: PERIOD = _0W	4	0.0016	0.0032
MEAN_GBH_0W	Mean Gamebookers home win odds:: PERIOD = _0W	3	0.0016	0.0032
MEAN_GBD_76W	Mean Gamebookers draw odds:: PERIOD = _76W	13	0.0014	0.0029
MEAN_BWA_76W	Mean Bet & Win away win odds:: PERIOD = _76W	2	0.0014	0.0028

Figure 12 Random Forest features



Variable Name	Label	Number of Splitting Rules	Importance
AwayTeam	Away Team	6	1
MEAN_LBA_0W	Mean Ladbrokes away win odds:: PERIOD = _0W	1	0.849857803
HomeTeam	Home Team	4	0.838605819
MEAN_LBH_0W	Mean Ladbrokes home win odds:: PERIOD = _0W	1	0.756245375
MEAN_BSA_0W	Mean Blue Square away win odds:: PERIOD = _0W	2	0.743263348
MEAN_B365A_0W	Mean Bet365 away win odds:: PERIOD = _0W	1	0.710292502
MEAN_SBA_0W	Mean Sportingbet away win odds:: PERIOD = _0W	1	0.472803076
MEAN_BSD_0W	Mean Blue Square draw odds:: PERIOD = _0W	1	0.463648823
MEAN_HC_EVER	Mean Home Team Corners:: PERIOD = _EVER	1	0.433952978
MAX_BbMxD_76W	Max Betbrain maximum draw odds:: PERIOD = _76W	1	0.352665964
MEAN_BSD_20W38W	Mean Blue Square draw odds:: PERIOD = _20W38W	1	0.317844814
_CLUSTER_ID_	Cluster ID	1	0.296737883
MEAN_VCD_0W	Mean VC Bet draw odds:: PERIOD = _0W	1	0.228919638

Figure 13 Gradient Boosting Features

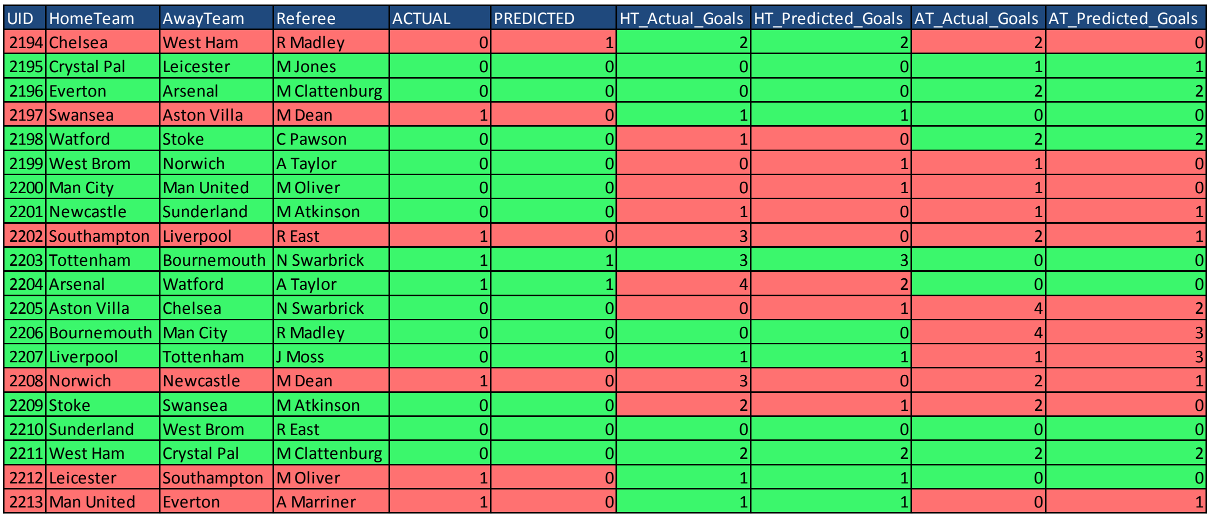


Variable Name	Label	Number of Splitting Rules	Importance
MEAN_B365A_0W	Mean Bet365 away win odds:: PERIOD = _0W	2	1
AwayTeam	Away Team	1	0.502549382
MIN_B365A_20W38W	Min Bet365 away win odds:: PERIOD = _20W38W	1	0.359807104
MEAN_HTAG_EVER	Mean Half Time Away Team Goals:: PERIOD = _EVER	1	0.358415282
MEAN_SBD_0W	Mean Sportingbet draw odds:: PERIOD = _0W	1	0.339468307
MIN_BbAvH_39W76W	Min Betbrain average home win odds:: PERIOD = _39W7	1	0.302564477
MEAN_AC_38W	Mean Away Team Corners:: PERIOD = _38W	1	0.300706336
CNT_FT_WINS_EVER	Count of FT home win:: PERIOD = _EVER	1	0.297206071
MIN_BbAvD_24W	Min Betbrain average draw win odds:: PERIOD = _24W	1	0.282673925

Figure 14 Decision Tree Features

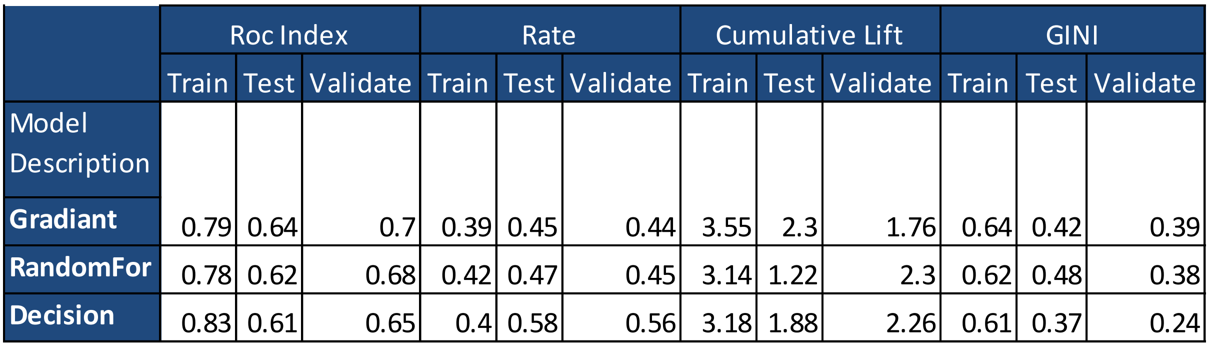
2. Approach two – Expected Goals Prediction

Using the Gradient Boosting Model an Expected Goals misclassification rate of 45% was achieved against the test data, the Model correctly predicted the number of goals scored for 21 of 40 teams in the out-of-time test resulting in a misclassification rate of 52.5%.



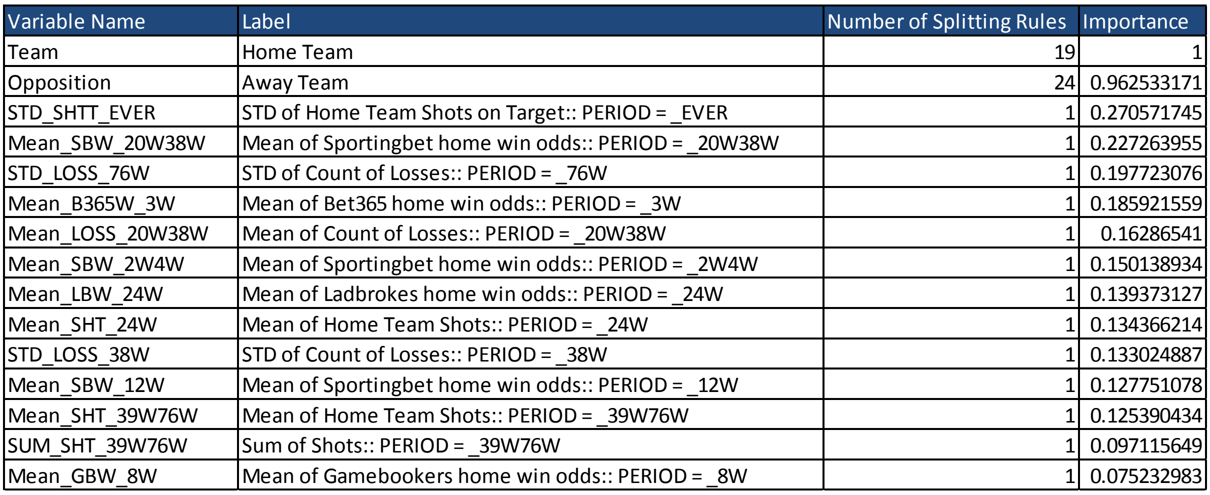
UID	HomeTeam	AwayTeam	Referee	ACTUAL	PREDICTED	HT_Actual_Goals	HT_Predicted_Goals	AT_Actual_Goals	AT_Predicted_Goals
2194	Chelsea	West Ham	R Madley	0	1	2	2	2	0
2195	Crystal Pal	Leicester	M Jones	0	0	0	0	1	1
2196	Everton	Arsenal	M Clattenburg	0	0	0	0	2	2
2197	Swansea	Aston Villa	M Dean	1	0	1	1	0	0
2198	Watford	Stoke	C Pawson	0	0	1	0	2	2
2199	West Brom	Norwich	A Taylor	0	0	0	1	1	0
2200	Man City	Man United	M Oliver	0	0	0	1	1	0
2201	Newcastle	Sunderland	M Atkinson	0	0	1	0	1	1
2202	Southampton	Liverpool	R East	1	0	3	0	2	1
2203	Tottenham	Bournemouth	N Swarbrick	1	1	3	3	0	0
2204	Arsenal	Watford	A Taylor	1	1	4	2	0	0
2205	Aston Villa	Chelsea	N Swarbrick	0	0	0	1	4	2
2206	Bournemouth	Man City	R Madley	0	0	0	0	4	3
2207	Liverpool	Tottenham	J Moss	0	0	1	1	1	3
2208	Norwich	Newcastle	M Dean	1	0	3	0	2	1
2209	Stoke	Swansea	M Atkinson	0	0	2	1	2	0
2210	Sunderland	West Brom	R East	0	0	0	0	0	0
2211	West Ham	Crystal Pal	M Clattenburg	0	0	2	2	2	2
2212	Leicester	Southampton	M Oliver	1	0	1	1	0	0
2213	Man United	Everton	A Marriner	1	0	1	1	0	1

Table 4 Out-of-Time Predicted vs Actual Goals Scored



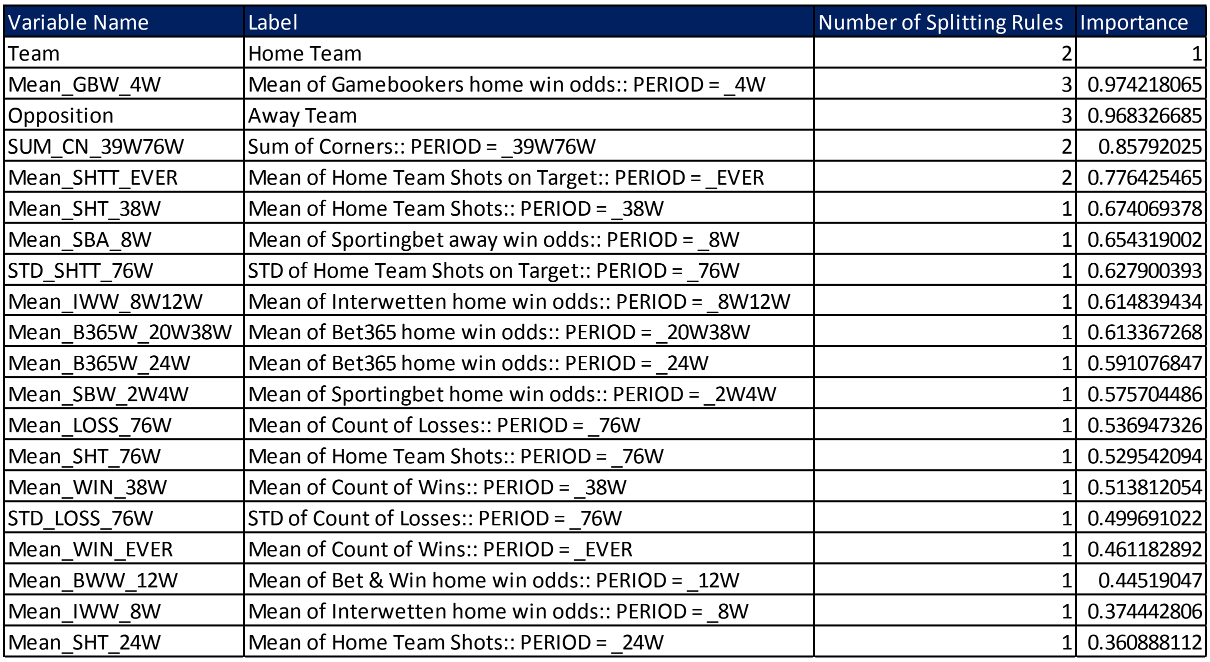
	Roc Index			Rate			Cumulative Lift			GINI		
	Train	Test	Validate	Train	Test	Validate	Train	Test	Validate	Train	Test	Validate
Model Description												
Gradiant	0.79	0.64	0.7	0.39	0.45	0.44	3.55	2.3	1.76	0.64	0.42	0.39
RandomFor	0.78	0.62	0.68	0.42	0.47	0.45	3.14	1.22	2.3	0.62	0.48	0.38
Decision	0.83	0.61	0.65	0.4	0.58	0.56	3.18	1.88	2.26	0.61	0.37	0.24

Figure 15 Performance statistics approach two (all models)



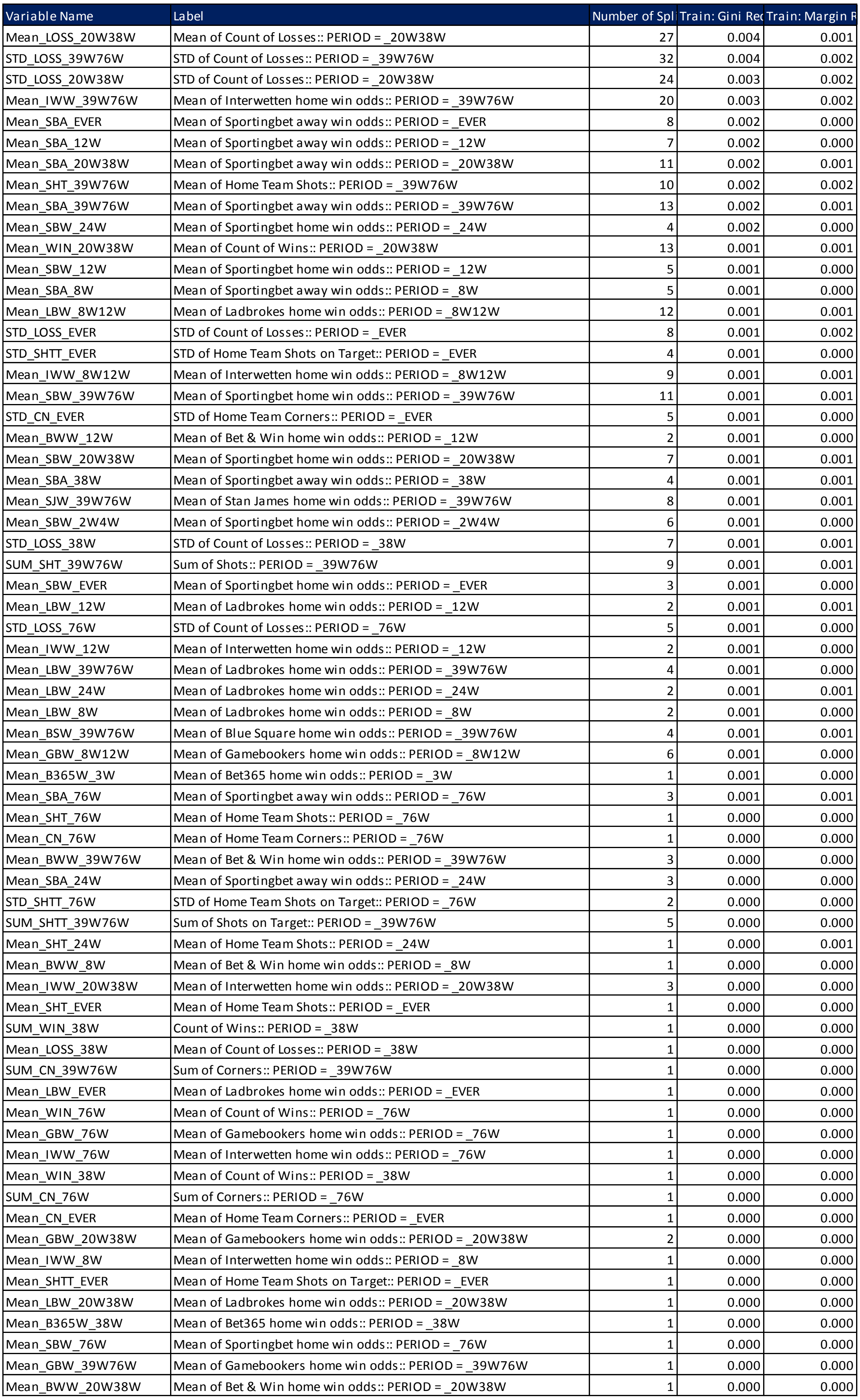
Variable Name	Label	Number of Splitting Rules	Importance
Team	Home Team	19	1
Opposition	Away Team	24	0.962533171
STD_SHTT_EVER	STD of Home Team Shots on Target:: PERIOD = _EVER	1	0.270571745
Mean_SBW_20W38W	Mean of Sportingbet home win odds:: PERIOD = _20W38W	1	0.227263955
STD_LOSS_76W	STD of Count of Losses:: PERIOD = _76W	1	0.197723076
Mean_B365W_3W	Mean of Bet365 home win odds:: PERIOD = _3W	1	0.185921559
Mean_LOSS_20W38W	Mean of Count of Losses:: PERIOD = _20W38W	1	0.16286541
Mean_SBW_2W4W	Mean of Sportingbet home win odds:: PERIOD = _2W4W	1	0.150138934
Mean_LBW_24W	Mean of Ladbrokes home win odds:: PERIOD = _24W	1	0.139373127
Mean_SHT_24W	Mean of Home Team Shots:: PERIOD = _24W	1	0.134366214
STD_LOSS_38W	STD of Count of Losses:: PERIOD = _38W	1	0.133024887
Mean_SBW_12W	Mean of Sportingbet home win odds:: PERIOD = _12W	1	0.127751078
Mean_SHT_39W76W	Mean of Home Team Shots:: PERIOD = _39W76W	1	0.125390434
SUM_SHT_39W76W	Sum of Shots:: PERIOD = _39W76W	1	0.097115649
Mean_GBW_8W	Mean of Gamebookers home win odds:: PERIOD = _8W	1	0.075232983

Figure 16 Gradient Boosting features



Variable Name	Label	Number of Splitting Rules	Importance
Team	Home Team	2	1
Mean_GBW_4W	Mean of Gamebookers home win odds:: PERIOD = _4W	3	0.974218065
Opposition	Away Team	3	0.968326685
SUM_CN_39W76W	Sum of Corners:: PERIOD = _39W76W	2	0.85792025
Mean_SHTT_EVER	Mean of Home Team Shots on Target:: PERIOD = _EVER	2	0.776425465
Mean_SHT_38W	Mean of Home Team Shots:: PERIOD = _38W	1	0.674069378
Mean_SBA_8W	Mean of Sportingbet away win odds:: PERIOD = _8W	1	0.654319002
STD_SHTT_76W	STD of Home Team Shots on Target:: PERIOD = _76W	1	0.627900393
Mean_IWW_8W12W	Mean of Interwetten home win odds:: PERIOD = _8W12W	1	0.614839434
Mean_B365W_20W38W	Mean of Bet365 home win odds:: PERIOD = _20W38W	1	0.613367268
Mean_B365W_24W	Mean of Bet365 home win odds:: PERIOD = _24W	1	0.591076847
Mean_SBW_2W4W	Mean of Sportingbet home win odds:: PERIOD = _2W4W	1	0.575704486
Mean_LOSS_76W	Mean of Count of Losses:: PERIOD = _76W	1	0.536947326
Mean_SHT_76W	Mean of Home Team Shots:: PERIOD = _76W	1	0.529542094
Mean_WIN_38W	Mean of Count of Wins:: PERIOD = _38W	1	0.513812054
STD_LOSS_76W	STD of Count of Losses:: PERIOD = _76W	1	0.499691022
Mean_WIN_EVER	Mean of Count of Wins:: PERIOD = _EVER	1	0.461182892
Mean_BWW_12W	Mean of Bet & Win home win odds:: PERIOD = _12W	1	0.44519047
Mean_IWW_8W	Mean of Interwetten home win odds:: PERIOD = _8W	1	0.374442806
Mean_SHT_24W	Mean of Home Team Shots:: PERIOD = _24W	1	0.360888112

Figure 17 Decision Tree features



Variable Name	Label	Number of Spl	Train: Gini Red	Train: Margin R
Mean_LOSS_20W38W	Mean of Count of Losses:: PERIOD = _20W38W	27	0.004	0.001
STD_LOSS_39W76W	STD of Count of Losses:: PERIOD = _39W76W	32	0.004	0.002
STD_LOSS_20W38W	STD of Count of Losses:: PERIOD = _20W38W	24	0.003	0.002
Mean_IWW_39W76W	Mean of Interwetten home win odds:: PERIOD = _39W76W	20	0.003	0.002
Mean_SBA_EVER	Mean of Sportingbet away win odds:: PERIOD = _EVER	8	0.002	0.000
Mean_SBA_12W	Mean of Sportingbet away win odds:: PERIOD = _12W	7	0.002	0.000
Mean_SBA_20W38W	Mean of Sportingbet away win odds:: PERIOD = _20W38W	11	0.002	0.001
Mean_SHT_39W76W	Mean of Home Team Shots:: PERIOD = _39W76W	10	0.002	0.002
Mean_SBA_39W76W	Mean of Sportingbet away win odds:: PERIOD = _39W76W	13	0.002	0.001
Mean_SBW_24W	Mean of Sportingbet home win odds:: PERIOD = _24W	4	0.002	0.000
Mean_WIN_20W38W	Mean of Count of Wins:: PERIOD = _20W38W	13	0.001	0.001
Mean_SBW_12W	Mean of Sportingbet home win odds:: PERIOD = _12W	5	0.001	0.000
Mean_SBA_8W	Mean of Sportingbet away win odds:: PERIOD = _8W	5	0.001	0.000
Mean_LBW_8W12W	Mean of Ladbrokes home win odds:: PERIOD = _8W12W	12	0.001	0.001
STD_LOSS_EVER	STD of Count of Losses:: PERIOD = _EVER	8	0.001	0.002
STD_SHTT_EVER	STD of Home Team Shots on Target:: PERIOD = _EVER	4	0.001	0.000
Mean_IWW_8W12W	Mean of Interwetten home win odds:: PERIOD = _8W12W	9	0.001	0.001
Mean_SBW_39W76W	Mean of Sportingbet home win odds:: PERIOD = _39W76W	11	0.001	0.001
STD_CN_EVER	STD of Home Team Corners:: PERIOD = _EVER	5	0.001	0.000
Mean_BWW_12W	Mean of Bet & Win home win odds:: PERIOD = _12W	2	0.001	0.000
Mean_SBW_20W38W	Mean of Sportingbet home win odds:: PERIOD = _20W38W	7	0.001	0.001
Mean_SBA_38W	Mean of Sportingbet away win odds:: PERIOD = _38W	4	0.001	0.001
Mean_SJW_39W76W	Mean of Stan James home win odds:: PERIOD = _39W76W	8	0.001	0.001
Mean_SBW_2W4W	Mean of Sportingbet home win odds:: PERIOD = _2W4W	6	0.001	0.000
STD_LOSS_38W	STD of Count of Losses:: PERIOD = _38W	7	0.001	0.001
SUM_SHT_39W76W	Sum of Shots:: PERIOD = _39W76W	9	0.001	0.001
Mean_SBW_EVER	Mean of Sportingbet home win odds:: PERIOD = _EVER	3	0.001	0.000
Mean_LBW_12W	Mean of Ladbrokes home win odds:: PERIOD = _12W	2	0.001	0.001
STD_LOSS_76W	STD of Count of Losses:: PERIOD = _76W	5	0.001	0.000
Mean_IWW_12W	Mean of Interwetten home win odds:: PERIOD = _12W	2	0.001	0.000
Mean_LBW_39W76W	Mean of Ladbrokes home win odds:: PERIOD = _39W76W	4	0.001	0.000
Mean_LBW_24W	Mean of Ladbrokes home win odds:: PERIOD = _24W	2	0.001	0.001
Mean_LBW_8W	Mean of Ladbrokes home win odds:: PERIOD = _8W	2	0.001	0.000
Mean_BSW_39W76W	Mean of Blue Square home win odds:: PERIOD = _39W76W	4	0.001	0.001
Mean_GBW_8W12W	Mean of Gamebookers home win odds:: PERIOD = _8W12W	6	0.001	0.000
Mean_B365W_3W	Mean of Bet365 home win odds:: PERIOD = _3W	1	0.001	0.000
Mean_SBA_76W	Mean of Sportingbet away win odds:: PERIOD = _76W	3	0.001	0.001
Mean_SHT_76W	Mean of Home Team Shots:: PERIOD = _76W	1	0.000	0.000
Mean_CN_76W	Mean of Home Team Corners:: PERIOD = _76W	1	0.000	0.000
Mean_BWW_39W76W	Mean of Bet & Win home win odds:: PERIOD = _39W76W	3	0.000	0.000
Mean_SBA_24W	Mean of Sportingbet away win odds:: PERIOD = _24W	3	0.000	0.000
STD_SHTT_76W	STD of Home Team Shots on Target:: PERIOD = _76W	2	0.000	0.000
SUM_SHTT_39W76W	Sum of Shots on Target:: PERIOD = _39W76W	5	0.000	0.000
Mean_SHT_24W	Mean of Home Team Shots:: PERIOD = _24W	1	0.000	0.001
Mean_BWW_8W	Mean of Bet & Win home win odds:: PERIOD = _8W	1	0.000	0.000
Mean_IWW_20W38W	Mean of Interwetten home win odds:: PERIOD = _20W38W	3	0.000	0.000
Mean_SHT_EVER	Mean of Home Team Shots:: PERIOD = _EVER	1	0.000	0.000
SUM_WIN_38W	Count of Wins:: PERIOD = _38W	1	0.000	0.000
Mean_LOSS_38W	Mean of Count of Losses:: PERIOD = _38W	1	0.000	0.000
SUM_CN_39W76W	Sum of Corners:: PERIOD = _39W76W	1	0.000	0.000
Mean_LBW_EVER	Mean of Ladbrokes home win odds:: PERIOD = _EVER	1	0.000	0.000
Mean_WIN_76W	Mean of Count of Wins:: PERIOD = _76W	1	0.000	0.000
Mean_GBW_76W	Mean of Gamebookers home win odds:: PERIOD = _76W	1	0.000	0.000
Mean_IWW_76W	Mean of Interwetten home win odds:: PERIOD = _76W	1	0.000	0.000
Mean_WIN_38W	Mean of Count of Wins:: PERIOD = _38W	1	0.000	0.000
SUM_CN_76W	Sum of Corners:: PERIOD = _76W	1	0.000	0.000
Mean_CN_EVER	Mean of Home Team Corners:: PERIOD = _EVER	1	0.000	0.000
Mean_GBW_20W38W	Mean of Gamebookers home win odds:: PERIOD = _20W38W	2	0.000	0.000
Mean_IWW_8W	Mean of Interwetten home win odds:: PERIOD = _8W	1	0.000	0.000
Mean_SHTT_EVER	Mean of Home Team Shots on Target:: PERIOD = _EVER	1	0.000	0.000
Mean_LBW_20W38W	Mean of Ladbrokes home win odds:: PERIOD = _20W38W	1	0.000	0.000
Mean_B365W_38W	Mean of Bet365 home win odds:: PERIOD = _38W	1	0.000	0.000
Mean_SBW_76W	Mean of Sportingbet home win odds:: PERIOD = _76W	1	0.000	0.000
Mean_GBW_39W76W	Mean of Gamebookers home win odds:: PERIOD = _39W76W	1	0.000	0.000
Mean_BWW_20W38W	Mean of Bet & Win home win odds:: PERIOD = _20W38W	1	0.000	0.000

Figure 18 Random Forest Features

Discussion & Future Work

To summarise two different approaches were developed, the first to predict home team wins and the second to predict the number of goals scored. Each approach comprised of three models (Decision Tree, Random Forest and Gradient boosting), the strongest performing model based on holdout test misclassification rate was selected as the final model for each approach and was tested against the out-of-time test data. The best performing model for approach one was a Random Forest model which was able to correctly predict 70% of the game results. The best performing model for approach two was a Gradient Boosting model which was able to correctly predict the number of goals scored 52.5% of the time.

Key finding from this study include:

- If one model does not meet the required performance benchmarks, let multiple models vote for a prediction. For example, a standalone decision tree was constructed for the home win prediction, this gave us a ROC of .67 and a misclassification rate of .39 on the holdout test data while the Random Forest model gave us a ROC of .72 and a misclassification rate of .33. Because the Random Forest algorithm uses bootstrap aggregation the variance of the data is reduced resulting in increased precision. Similarly the Gradient boosting model reduces bias ultimately leading to increased accuracy.
- Class imbalance hinders the ability of a model to learn particularly on relatively small training sets. This is a topical area of research with a number of empirical studies arguing that the class imbalance is a relative problem that depends on
  (1) the degree of the class imbalance
  (2) the complexity of the concept represented by the data
  (3) the overall size of the training set
  (4) the classifier involved.
- Training instances have a big impact on model performance, 1,2 and 3 seasons worth of training instances were considered for both approaches. Given that volatility of premiership teams with transfers, new owners and new management changing every second season it was initially expected that just one season worth of data would be the optimum for training the models. However, it transpired that 3 seasons worth of data would deliver the best models for approach one and 2 seasons worth of data would work best for approach two.

Data permitting other interesting areas of research include how social media sentiment analysis may be used as a predictor of victory and how graph theory could be used to model team performance.

Bibliography

Breiman, L., 2001. Random forests. *Machine learning,* Volume 45(1), pp. 5-32.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* Volume 39(11), pp. 27-34.

Friedman, J., n.d. Greedy function approximation: a gradient boosting machine. *Annals of statistics,* pp. 1189-1232.

Appendix

Selection of Data Code

```
* PROJECT:           Assignment 2: Sports Analytics
* NAME:              Read_in_data.sas
* AUTHOR:            Shane McCarthy
* EMAIL:             shane.mc-carthy@ucdconnect.ie
* DATE CREATED:      12/03/16
* PURPOSE:           This script reads in the data from csv files, ensures all variables are
of correct type and added teh variable
*                                      descriptions to the metadata as labels
;

*I've downloaded all the required data to here;
libname wd "C:\Users\shane.mc.carthy\Dropbox\Masters\Semester2\MIS40970 Data Mining for Bus
Analytics\Assignments\Assignment2";
%let path = C:\Users\shane.mc.carthy\Dropbox\Masters\Semester2\MIS40970 Data Mining for Bus
Analytics\Assignments\Assignment2\;


* The %csv_importer macro, loops through n files using proc import stacking them together;
%macro csv_importer(path_loc=/*path location of csv files*/
                                      ,num_file=/*number of csv files to iterate through*/
                                      );
%DO n=1 %TO &num_file.;

%Put *** Reading in &n. of &num_file. csv files called: "20%eval(9+&n.)_%eval(10+&n.).csv";
               proc import
               DATAFILE= "&path_loc.20%eval(9+&n.)_%eval(10+&n.).csv"
       out=work.part0&n
       dbms=csv
       replace;
       getnames=no;
               run;
%end;

* the table header has been stored as row 1(_N_=1), we only need this in the 1st of n files;
%DO n=2 %TO &num_file.;
       data work.part0&n.;
       set work.part0&n.;
       if _N_=1 then delete;
       run;
%end;

* stack n files and create a base table;
       data base_table;
       set part01-part06;
       run;

* now move the column headers from row 1, we use proc transpose to get
  the current column names VAR1-VAR71 and the actual column names;
       proc transpose data=base_table(obs=1) out=temp;
       var _all_;
       run;

* because SAS has a strict column naming convention we
  need to rename some of the actual coulmn names first;
```

```
        data temp;
        set temp;
        if COL1 ="BbMx>2.5" then COL1= "BbMx_GE_2pt5";
        if COL1 ="BbAv>2.5" then COL1= "BbAv_GE_2pt5";
        if COL1 ="BbMx<2.5" then COL1= "BbMx_LE_2pt5";
        if COL1 ="BbAv<2.5" then COL1= "BbAv_LE_2pt5";
        run;

* create a macro variable called "rename" which contains all the rename
  statements required;
        proc sql noprint;
                select catx('=',_name_,col1) into :rename separated by ' '
         from temp;
        quit;

* rename the columns on the base table, note we starte from row 2;
        data WORK.BASE_TABLE;
        set  WORK.BASE_TABLE(firstobs=2 rename=(&rename));
        run;

* delete intermediate tables;
        proc datasets library= work;
                delete  part01-part06
                                temp;
        run;

%mend;
*call the macro, we've currently 6 input files;
%csv_importer(path_loc=&path.,num_file=6);

* now all variables have been read into the base table as type charater,
 we need to convert numeric variables back to type numeric
 the char2num macro converts all variables from type CHAR to type NUM unless
 excluded using the excl_vars parameter ;

%macro char2num(libin=/* input library*/,
                                dsin= /* input dataset */,
                                libout=/* output library */,
                                dsout= /* output dataset */,
                                excl_vars=/* variables that you do NOT want to convert,
seperated by |*/
                                );

*we use proc sql here to extract the base table metadata from dictionary.columns
        prxmatch is an regular-expression function, we use it here to exclude true charater
variables;
        proc sql ;
                CREATE TABLE VAR_LIST as
                        select  name AS VARIABLE
                                        ,type
                                from dictionary.columns
                                where libname=upcase("&LIBIN.")
                                        and memname=upcase("&DSIN.")
                                        and
prxmatch(cat("m/",upcase("&excl_vars."),"/oi"),upcase(name))=0;
        quit;

        *Count number of vars we want to convert and store value in NVAR, this is used for
loop later;
        PROC SQL noprint;
                SELECT COUNT(VARIABLE) INTO :NVAR FROM VAR_LIST;
        QUIT;

        *Create a series of macro variables containing the names of variables we want to
convert;
        DATA _NULL_;
                length ii $4.;
                SET VAR_LIST end=last;
                        i+1;
                        ii=LEFT(PUT(i,4.));
                        call symputx('var'||ii, LEFT(VARIABLE));
                        IF last THEN call symputx('NVAR', TRIM(LEFT(_N_)));
        RUN;

        *this block of code actually converts the variavle from type CHAR to type NUM,
        SAS does not like us changing the type without renaming the variable so we create
        an intermediate temp variable, we then delete the variable of type CHAR, before
renaming the temp variable;

        DATA TEMP_CHARS;
                SET &LIBIN..&DSIN.;

                %DO X=1 %TO &NVAR.;
                        temp&X. =INPUT(&&VAR&X.,5.);
                %end;

        %DO X=1 %TO &NVAR.;
                        drop &&VAR&X.;
                %end;
        RUN;

        DATA &LIBOUT..&DSOUT.;
                SET TEMP_CHARS;

                %DO X=1 %TO &NVAR.;
                        RENAME temp&X.= &&VAR&X.;
                %END;
        RUN;


        * delete intermediate tables;
        PROC DATASETS LIBRARY= WORK;
                DELETE  VAR_LIST
                        TEMP_CHARS;
        RUN;
%mend char2num;
%char2num(libin=WORK,dsin=BASE_TABLE,libout=WORK,dsout=BASE_TABLE,excl_vars=Div|Date|HomeTeam|
AwayTeam|FTR|HTR|Referee);


        *date is still of type CHAR, we need to convert this to the correct date format;
        DATA BASE_TABLE;
        Format ID $Char8.;
        Format Match_Date Date9.;
        SET BASE_TABLE;

        ID=put(_N_,z4.);
        Match_Date=input(strip(date),ddmmyy11.);


        drop date;
        run;

        *Add variable desciptions as lables, this is stored as metadata and will be useful
later on for graphs...etc.;
        data BASE_TABLE;
        set BASE_TABLE;
        label
                Div ="League Division"
                Match_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 (H=Home Win, D=Draw, A=Away Win)"
                HTHG  = "Half Time Home Team Goals"
                HTAG  = "Half Time Away Team Goals"
                HTR  = "Half Time Result (H=Home Win, D=Draw, A=Away Win)"
                Attendance  = "Crowd Attendance"
                Referee  = "Match Referee"
                HS  = "Home Team Shots"
                AS  = "Away Team Shots"
                HST  = "Home Team Shots on Target"
                AST  = "Away Team Shots on Target"
                HHW  = "Home Team Hit Woodwork"
                AHW  = "Away Team Hit Woodwork"
                HC  = "Home Team Corners"
                AC  = "Away Team Corners"
                HF  = "Home Team Fouls Committed"
                AF  = "Away Team Fouls Committed"
                HO  = "Home Team Offsides"
                AO  = "Away Team Offsides"
                HY  = "Home Team Yellow Cards"
```

```
               AY  = "Away Team Yellow Cards"
               HR  = "Home Team Red Cards"
               AR  = "Away Team Red Cards"
               HBP  = "Home Team Bookings Points (10 = yellow, 25 = red)"
               ABP  = "Away Team Bookings Points (10 = yellow, 25 = red)"
               B365H  = "Bet365 home win odds"
               B365D  = "Bet365 draw odds"
               B365A  = "Bet365 away win odds"
               BSH  = "Blue Square home win odds"
               BSD  = "Blue Square draw odds"
               BSA  = "Blue Square away win odds"
               BWH  = "Bet & Win home win odds"
               BWD  = "Bet & Win draw odds"
               BWA  = "Bet & Win away win odds"
               GBH  = "Gamebookers home win odds"
               GBD  = "Gamebookers draw odds"
               GBA  = "Gamebookers away win odds"
               IWH  = "Interwetten home win odds"
               IWD  = "Interwetten draw odds"
               IWA  = "Interwetten away win odds"
               LBH  = "Ladbrokes home win odds"
               LBD  = "Ladbrokes draw odds"
               LBA  = "Ladbrokes away win odds"
               PSH  = "Pinnacle Sports home win odds"
               PSD  = "Pinnacle Sports draw odds"
               PSA  = "Pinnacle Sports away win odds"
               SOH  = "Sporting Odds home win odds"
               SOD  = "Sporting Odds draw odds"
               SOA  = "Sporting Odds away win odds"
               SBH  = "Sportingbet home win odds"
               SBD  = "Sportingbet draw odds"
               SBA  = "Sportingbet away win odds"
               SJH  = "Stan James home win odds"
               SJD  = "Stan James draw odds"
               SJA  = "Stan James away win odds"
               SYH  = "Stanleybet home win odds"
               SYD  = "Stanleybet draw odds"
               SYA  = "Stanleybet away win odds"
               VCH  = "VC Bet home win odds"
               VCD  = "VC Bet draw odds"
               VCA  = "VC Bet away win odds"
               WHH  = "William Hill home win odds"
               WHD  = "William Hill draw odds"
               WHA  = "William Hill 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 win 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 goals
(total goals) averages and maximums"
               BbMx_GE_2pt5  = "Betbrain maximum over 2.5 goals"
               BbAv_GE_2pt5  = "Betbrain average over 2.5 goals"
               BbMx_LE_2pt5  = "Betbrain maximum under 2.5 goals"
               BbAv LE 2pt5  = "Betbrain average under 2.5 goals"
               GB>2.5  = "Gamebookers over 2.5 goals"
               GB<2.5  = "Gamebookers under 2.5 goals"
               B365>2.5  = "Bet365 over 2.5 goals"
               B365<2.5  = "Bet365 under 2.5 goals"
               BbAH  = "Number of BetBrain bookmakers used to 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"
               BbMxAHA  = "Betbrain maximum Asian handicap away team odds"
               BbAvAHA  = "Betbrain average Asian handicap away team odds"
               GBAHH  = "Gamebookers Asian handicap home team odds"
               GBAHA  = "Gamebookers Asian handicap away team odds"
               GBAH  = "Gamebookers size of handicap (home team)"
               LBAHH  = "Ladbrokes Asian handicap home team odds"
               LBAHA  = "Ladbrokes Asian handicap away team odds"
               LBAH  = "Ladbrokes size of handicap (home team)"
               B365AHH  = "Bet365 Asian handicap home team odds"
               B365AHA  = "Bet365 Asian handicap away team odds"
               B365AH   = "Bet365 size of handicap (home team)"
;
run;

*end of script;
```

Pre-processing & Cleaning Code

```
* PROJECT:           Assignment 2: Sports Analytics
* NAME:              Univarate_Analysis.sas
* AUTHOR:            Shane McCarthy
* EMAIL:             shane.mc-carthy@ucdconnect.ie
* DATE CREATED:      12/03/16
* PURPOSE:           This script performs univariate analysis on all variables and drops
variables with data quility issues
;


%macro univariate_num(libin=,dsin=,libout=,dsout=);

*we use proc sql here to extract the base table metadata from dictionary.columns, selecting
all variables
 of type NUM that are not Dates;
       proc sql ;
               CREATE TABLE VAR_LIST as
                       select
                               name AS VARIABLE
                               ,label
                               from dictionary.columns
                               where libname=upcase("&LIBIN.")
                                       and memname=upcase("&DSIN.")
                                       and type ="num"
                                       and format ^="DATE9.";
       quit;

       *Count number of numeric vars we want to process;
       PROC SQL noprint;
               SELECT COUNT(VARIABLE) INTO :NVAR FROM VAR_LIST;
       QUIT;

       PROC SQL noprint;
               SELECT COUNT(*) INTO :NOBS FROM &LIBIN..&DSIN.;
       QUIT;

       *Create a series of macro variables containing the names of variables we want to
process;
       DATA _NULL_;
               length ii $4.;
               SET VAR_LIST end=last;
                       i+1;
                       ii=LEFT(PUT(i,4.));
                       call symputx('var'||ii, LEFT(VARIABLE));
                       IF last THEN call symputx('NVAR', TRIM(LEFT(_N_)));
       RUN;

       *loop through all selected variables using proc means to claculate stats;
       %DO X=1 %TO &NVAR.;
       proc means data = base_table noprint nway missing;
        var &&VAR&X.;
    output out = &&VAR&X.(drop=_:)
        n=num_populated
        nmiss = num_missing
        min = min_value
        max = max_value
        mean = avg_value
        std = st_deviation
        p10 = p10_value
        q1 = q1_value
        median = median_value
        q3 = q3_value
        p90 = p90_value
        ;
       run;
```

```
        %end;

        *stack stats together into one table;
        DATA stacked;
        SET
                %DO X = 1 %TO &NVAR.;
                        &&VAR&X.
                %END;
        ;
        RUN;

        *merge with variable name and label;
        Data &libout..&dsout.;
        merge VAR_LIST stacked;
    length flag $1 anomalous_reason $100;
        prop_missing = num_missing/&NOBS;

    *logic to flag anomalous variables for deeper inspection;
    if prop_missing = 1 then do;
        flag = 'A';
        anomalous_reason = '100% missing values';
    end;
    else if min_value = max_value then do;
        flag = 'A';
        anomalous_reason = 'All records take same value';
    end;
    else if prop_missing > 0.9 then do;
        flag = 'B';
        anomalous_reason = '>90% missing values';
    end;
    else if min_value = p90_value then do;
        flag = 'B';
        anomalous_reason = '>90% records take minimum value';
    end;
    else if max_value = p10_value then do;
        flag = 'B';
        anomalous_reason = '>90% records take maximum value';
    end;
    else if prop_missing > 0.75 then do;
        flag = 'C';
        anomalous_reason = '>75% missing values';
    end;
    else if p10_value = p90_value and p90_value ne . then do;
        flag = 'C';
        anomalous_reason = '>80% records take same value';
    end;
    else if min_value = q3_value then do;
        flag = 'C';
        anomalous_reason = '>75% records take minimum value';
    end;
    else if max_value = q1_value then do;
        flag = 'C';
        anomalous_reason = '>75% records take maximum value';
    end;
    else if prop_missing > 0.5 then do;
        flag = 'D';
        anomalous_reason = '>50% missing values';
    end;
    else if min_value = median_value then do;
        flag = 'D';
        anomalous_reason = '>50% records take minimum value';
    end;
    else if max_value = median_value then do;
        flag = 'D';
        anomalous_reason = '>50% records take maximum value';
    end;
    else if q1_value = q3_value and q3_value ne . then do;
        flag = 'D';
        anomalous_reason = '>50% records take same value';
    end;

        format sum_: min_: max_: avg_: st_: q1_: q3_: median_: p10_: p90_: prop_missing: 3.2;
        run;

        *sort;
        proc sort data=&libout..&dsout.; BY DESCENDING prop_missing DESCENDING
anomalous_reason; RUN;
```

```
        *Delete intermediate tables stacked;
        PROC DATASETS LIBRARY= WORK;
        DELETE
            %DO X = 1 %TO &NVAR.;
                                &&VAR&X.
            %END;
                stacked
                VAR_LIST

            ;
        RUN;


%mend;
%univariate_num(libin=WORK,dsin=BASE_TABLE,libout=WORK,dsout=UNIVAR_NUM_OUTPUT)


%macro univariate_char(libin=,dsin=,libout=,dsout=,report_levels=,excl_vars=);

*we use proc sql here to extract the base table metadata from dictionary.columns, selecting
all variables
 of type NUM that are not Dates;
        proc sql ;
                CREATE TABLE VAR_LIST as
                        select
                                name AS VARIABLE
                                ,label
                                from dictionary.columns
                                where libname=upcase("&LIBIN.")
                                        and memname=upcase("&DSIN.")
                                        and type ="char"
                                        and format ^="DATE9."
                                        and
prxmatch(cat("m/",upcase("&excl_vars."),"/oi"),upcase(name))=0;
        quit;

        *Count number of numeric vars we want to process;
        PROC SQL noprint;
                SELECT COUNT(VARIABLE) INTO :NVAR FROM VAR_LIST;
        QUIT;

        PROC SQL noprint;
                SELECT COUNT(*) INTO :NOBS FROM &LIBIN..&DSIN.;
        QUIT;

        *Create a series of macro variables containing the names of variables we want to
process;
        DATA _NULL_;
                length ii $4.;
                SET VAR_LIST end=last;
                       i+1;
                       ii=LEFT(PUT(i,4.));
                       call symputx('var'||ii, LEFT(VARIABLE));
                       IF last THEN call symputx('NVAR', TRIM(LEFT(_N_)));
        RUN;

        %DO X=1 %TO &NVAR.;

        proc sql;
            create table &&VAR&X. as
                select         &&VAR&X. as Level
                        ,count(*) as num_with_value  /*Calculate the proportion of records
taking each value*/
                from &libin..&dsin.
                group by &&VAR&X.
                order by num_with_value desc  /*Order by descending proportion*/
                ;
        quit;

        data  &&VAR&X.;
    length variable $32 level $100;
    format variable $32. level $100.;
    set  &&VAR&X.  end = final;
/*    Just take first 8 and / or missing value values. Since ordered previously by
descending*/
/*    proportion, the first 8 represent the 8 most frequent values*/
```

```
    variable = upcase("&&VAR&X.");
    prop_with_value = num_with_value/&NOBS;
    if _n_ le &report_levels or strip(level) in ('','.') then output;
    if final then call symput('num_levels',put(_n_,8.));  /*Put the total number of possible
levels in to a macro variable*/
       run;

       data &&VAR&X.;
       set &&VAR&X.;
       total_num_levels = &num_levels;
    exp_prop_with_value = 1/&num_levels;
    ratio_act_exp = prop_with_value / exp_prop_with_value;

       if prop_with_value = 1 then do;
        flag = 'A';
        anomalous_reason = 'All records take the same value';
    end;
    else if prop_with_value ge 0.9 then do;
        flag = 'B';
        anomalous_reason = '>90% of records take the same value';
    end;
    else if prop_with_value ge 0.75 then do;
        flag = 'C';
        anomalous_reason = '>75% of records take the same value';
    end;
    else if prop_with_value ge 0.50 then do;
        flag = 'D';
        anomalous_reason = '>50% of records take the same value';
    end;
       run;

       %end;


               DATA &libout..&dsout.;
               SET
               %DO X = 1 %TO &NVAR.;
                       &&VAR&X.
               %END;
       ;
       RUN;


       *sort;
       proc sort data=&libout..&dsout.; BY  DESCENDING anomalous_reason; RUN;

       *Delete intermediate tables stacked;
       PROC DATASETS LIBRARY= WORK;
       DELETE
           %DO X = 1 %TO &NVAR.;
                              &&VAR&X.
           %END;
               VAR_LIST

           ;
       RUN;


%mend univariate_char;
%univariate_char(libin=WORK,dsin=BASE_TABLE,libout=WORK,dsout=UNIVAR_CHAR_OUTPUT,report_levels
=100,excl_vars=ID)



*Following the univariate analysis of both numeric and categorical variables the follow
variables are exclueded;
*BbAH BbAHh BbMxAHH BbAvAHH BbMxAHA BbAvAHA ~ no data since 19May13;
*DIV ~ all values are the same;
* The fllowing all have missing data from 15Sep14 - 24May15, because this data MAY be usefulr
  we'll keep them for now WHH WHD WHA BbAv_GE_2pt5 BbMx_LE_2pt5 BbAv_LE_2pt5;

data wd.base_table;
set base_table;
Drop
BbAH
BbAHh
```

```
BbMxAHH
BbAvAHH
BbMxAHA
BbAvAHA
WHH
WHDDIV
WHA
BbAv_GE_2pt5
BbMx_LE_2pt5
BbAv_LE_2pt5;


run;
```

Feature Selection & Extraction Code

e. Define Universes

```
* PROJECT:            Assignment 2: Sports Analytics
* NAME:               Define_training_universe.sas
* AUTHOR:             Shane McCarthy
* EMAIL:              shane.mc-carthy@ucdconnect.ie
* DATE CREATED:       13/03/16
* PURPOSE:            This script performs univariate analysis on all variables and
drops variables with data quility issues
;



*******************************************************
* Step1: Define the Home Win training universe
*
*******************************************************;


data base_table;
set  wd.base_table;

*Label the seasons, the season runs from Aug to May
01 the current season;
if match_date GE "01Aug2010"d and match_date LE "30May2011"d then Season = 06;
else if match_date GE "01Aug2011"d and match_date LE "30May2012"d then Season = 05;
else if match_date GE "01Aug2012"d and match_date LE "30May2013"d then Season = 04;
else if match_date GE "01Aug2013"d and match_date LE "30May2014"d then Season = 03;
else if match_date GE "01Aug2014"d and match_date LE "30May2015"d then Season = 02;
else if match_date GE "01Aug2015"d and match_date LE "30May2016"d then Season = 01;

*Get the calender week number and concatenate with year, this is used later to form
a rolling week ;
Cal_Week_num=cat(year(match_date),put(week(match_date-1,"V"),z2.));
run;


*next create a premiership week number;
proc sort data=base_table; by descending Season Cal_Week_num ; run;

* create a premiership week number for fixtures;
data base_table;
       set  base_table;
       by  descending Season Cal_Week_num;
       retain Prem_week_number 0;

       if first.Season then
              do;
                     Prem_week_number=1;
              end;
       else if first.Cal_Week_num then
             do;
                    Prem_week_number+1;
             end;

       Year_Week =cat(year(match_date),put(Prem_week_number,z2.));
             *drop Cal_Week_num;

run;

*
we don't want to train our model on all the data, given the variability of teams
across the seasons
with transfers, new managers...etc.
Here we define our training cases
Attempt one, just use one FULL season to train our model ~UNIVERSE A
Attempt two, take two full seasons to train our model ~UNIVERSE B;
proc sort data=work.base_table; by  ID  ;run;
data base_table;
set base_table;
by id Year_Week;

if first.id then do;
cnt+1;
end;

if cnt GE 760 and cnt LT 1049 then TRAIN_UNIVERSE = "D";
else if  cnt GE 1049 and cnt LT 1429 then TRAIN_UNIVERSE = "C";
else if cnt GE 1429 and cnt LT 1809 then TRAIN_UNIVERSE = "B";
else if cnt GE 1809 then TRAIN_UNIVERSE = "A";

drop cnt;
run;

proc sort data=work.base_table; by  Year_Week ;run;

*
Here we define our target
Attempt one - home team to win (0=LOSS,1=WIN,2=DRAW);
data base_table;
set base_table;
by Year_Week;

if first.Year_Week then do;
Rolling_week+1;
end;

*define our first target, home team winning, drawing, losing;
if FTHG > FTAG then TARGET_HW = 1;
else if FTHG < FTAG then TARGET_HW = 0;
else if FTHG = FTAG then TARGET_HW = 0;

run;

*one FULL season to train our model ~UNIVERSE A;
data wd.UNIVERSE_A;
set base_table (where=(TRAIN_UNIVERSE = "A"));
Keep ID MATCH_DATE Rolling_week HomeTeam AwayTeam Referee TARGET_HW;
rename Rolling_week = Match_week;
run;

*two full seasons to train our model ~UNIVERSE B;
data wd.UNIVERSE_B;
set base_table (where=(TRAIN_UNIVERSE in ("A","B")));
Keep ID MATCH_DATE Rolling_week HomeTeam AwayTeam Referee  TARGET_HW;
rename Rolling_week = Match_week;
run;
```

```
data wd.UNIVERSE_Z;
set base_table (where=(TRAIN_UNIVERSE in ("A","B","C")));
Keep ID MATCH_DATE Rolling_week HomeTeam AwayTeam Referee  TARGET_HW;
rename Rolling_week = Match_week;
run;

data wd.UNIVERSE_X;
set base_table (where=(TRAIN_UNIVERSE in ("A","B","C","D")));
Keep ID MATCH_DATE Rolling_week HomeTeam AwayTeam Referee  TARGET_HW;
rename Rolling_week = Match_week;
run;



*******************************************************
* Step2 Define the goals scored universe
*
*******************************************************;

data home;
set base_table;

keep Rolling_week  MATCH_DATE TRAIN_UNIVERSE homeTeam AwayTeam FTHG Referee
    WIN_CNT LOSS_CNT DRAW_CNT HTWIN_CNT HTLOSS_CNT HTDRAW_CNT BSD BSH VCD VCH SJD
SJH WHD SBA SBD SBH LBD LBH IWD IWH GBD GBH BWD BWH B365D B365H HR HY HC HF HST HS
HTHG
;

if FTR = "H" THEN WIN_CNT=1;
else WIN_CNT=0;

if FTR = "A" THEN LOSS_CNT=1;
else LOSS_CNT=0;

if FTR = "D" THEN DRAW_CNT=1;
else DRAW_CNT=0;

if HTR = "H" THEN HTWIN_CNT=1;
else HTWIN_CNT=0;

if HTR = "A" THEN HTLOSS_CNT=1;
else HTLOSS_CNT=0;

if HTR = "D" THEN HTDRAW_CNT=1;
else HTDRAW_CNT=0;



rename homeTeam=Team;
rename AwayTeam=Opposition;
rename FTHG=TARGET_GOAL;
rename HTHG=HT;
rename HS=SHT;
rename HST=SHTT;
rename HF=FC;
rename HC=CN;
rename HY=YC;
rename HR=RC;
rename B365H=B365W;
rename B365D=B365D;
rename BWH=BWW;
rename BWD=BWD;
rename GBH=GBW;
rename GBD=GBD;
rename IWH=IWW;
rename IWD=IWD;
rename LBH=LBW;
rename LBD=LBD;
rename SBH=SBW;
rename SBD=SBD;
rename SBA=SBW;
rename WHD=WHD;
rename SJH=SJW;
rename SJD=SJD;
rename VCH=VCW;
rename VCD=VCD;
rename BSH=BSW;
rename BSD=BSD;
rename BbMxH=BbW;
rename BbAvH=BbW;
rename BbMxA=BbW;
rename BbAvA=BbW;


run;

data away;
set base_table ;

keep Rolling_week MATCH_DATE TRAIN_UNIVERSE homeTeam AwayTeam FTAG Referee
WIN_CNT LOSS_CNT DRAW_CNT HTWIN_CNT HTLOSS_CNT HTDRAW_CNT BSA BSD VCA VCD SJA SJD
WHD SBD LBA LBD IWA IWD GBA GBD BWA BWD B365A B365D AR AY AC AF AST AS HTAG
;


if FTR = "A" THEN WIN_CNT=1;
else WIN_CNT=0;

if FTR = "H" THEN LOSS_CNT=1;
else LOSS_CNT=0;

if FTR = "D" THEN DRAW_CNT=1;
else DRAW_CNT=0;

if HTR = "A" THEN HTWIN_CNT=1;
else HTWIN_CNT=0;

if HTR = "H" THEN HTLOSS_CNT=1;
else HTLOSS_CNT=0;

if HTR = "D" THEN HTDRAW_CNT=1;
else HTDRAW_CNT=0;

rename AwayTeam=Team;
rename homeTeam=Opposition;
rename FTAG=TARGET_GOAL;
rename HTAG=HT;
rename AS=SHT;
rename AST=SHTT;
rename AF=FC;
rename AC=CN;
rename AY=YC;
rename AR=RC;
rename B365D=B365D;
rename B365A=B365W;
rename BWD=BWD;
rename BWA=BWW;
rename GBD=GBD;
rename GBA=GBW;
rename IWD=IWD;
rename IWA=IWW;
rename LBD=LBD;
rename LBA=LBW;
rename SBD=SBD;
rename WHD=WHD;
rename SJD=SJD;
rename SJA=SJW;
rename VCD=VCD;
rename VCA=VCW;
rename BSD=BSD;
rename BSA=BSW;
rename BbMxA=BbW;
rename BbAvA=BbW;

run;

data full;
retain UID;
set home away;

UID = _N_;
run;

*one FULL season to train our model ~UNIVERSE A;
data wd.UNIVERSE_C;
set full (where=(TRAIN_UNIVERSE = "A"));
Keep UID MATCH_DATE Rolling_week TEAM Opposition Referee TARGET_GOAL;
rename Rolling_week = Match_week;
run;

*two full seasons to train our model ~UNIVERSE B;
data wd.UNIVERSE_D;
set full (where=(TRAIN_UNIVERSE IN ("A","B")));
Keep UID MATCH_DATE Rolling_week TEAM Opposition Referee TARGET_GOAL;
rename Rolling_week = Match_week;
run;

********************************************************
* Step3 Form Base Tables
*
********************************************************;


*Universe A base table;

PROC SQL;
   CREATE TABLE base_table_a AS
   SELECT
               u.ID as UID
               ,u.MATCH_DATE
               ,u.TARGET_HW
               ,u.Match_week
               ,bt.rolling_week
               ,bt.*
      FROM wd.UNIVERSE_A u
         inner join  base_table bt on bt.HomeTeam=u.HomeTeam
         order by UID;
QUIT;

*Universe B base table;

PROC SQL;
   CREATE TABLE base_table_b AS
   SELECT
               u.ID as UID
               ,u.MATCH_DATE
               ,u.TARGET_HW
               ,u.Match_week
               ,bt.rolling_week
               ,bt.*
      FROM wd.UNIVERSE_b u
         inner join  base_table bt on bt.HomeTeam=u.HomeTeam
         order by UID;
QUIT;
```

```
PROC SQL;
   CREATE TABLE base_table_z AS
   SELECT
               u.ID as UID
               ,u.MATCH_DATE
               ,u.TARGET_HW
               ,u.Match_week
               ,bt.rolling_week
               ,bt.*
      FROM wd.UNIVERSE_z u
         inner join  base_table bt on bt.HomeTeam=u.HomeTeam
         order by UID;
QUIT;

PROC SQL;
   CREATE TABLE base_table_x AS
   SELECT
               u.ID as UID
               ,u.MATCH_DATE
               ,u.TARGET_HW
               ,u.Match_week
               ,bt.rolling_week
               ,bt.*
      FROM wd.UNIVERSE_x u
         inner join  base_table bt on bt.HomeTeam=u.HomeTeam
         order by UID;
QUIT;

PROC SQL;
   CREATE TABLE base_table_c AS
   SELECT
               u.UID
               ,u.MATCH_DATE
               ,u.TARGET_GOAL
               ,u.Match_week
               ,bt.rolling_week
               ,bt.*
      FROM wd.UNIVERSE_c u
         inner join  full bt on bt.TEAM=u.TEAM
         order by UID;
QUIT;

PROC SQL;
   CREATE TABLE base_table_d AS
   SELECT
               u.UID
               ,u.MATCH_DATE
               ,u.TARGET_GOAL
               ,u.Match_week
               ,bt.rolling_week
               ,bt.*
      FROM wd.UNIVERSE_d u
         inner join  full bt on bt.TEAM=u.TEAM
         order by UID;
QUIT;


*******************************************************
* Step4 Define time bins for base tables
*
*******************************************************;
*calculate the time since previous games (in weeks) w.r.t the games in our training univesre;
*we can't use games in the futre (w.r.t our training games) to predict the outcome of a game so remove these;
*now create time bands, these wll be used to aggregate data and tranpose the data;
```

```
%macro timeBand(libin=,dsin=,libout=,dsout=);

data &libout..&dsout.;
set &libin..&dsin.;

       time_diff = Match_week -rolling_week;

       if TIME_DIFF < 0 then delete;

       IF TIME_DIFF =0 THEN BIN_PLAY_0W = 1; ELSE BIN_PLAY_0W = 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 1 THEN BIN_PLAY_1W = 1; ELSE BIN_PLAY_1W
= 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 2 THEN BIN_PLAY_2W = 1; ELSE BIN_PLAY_2W
= 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 3 THEN BIN_PLAY_3W = 1; ELSE BIN_PLAY_3W
= 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 4 THEN BIN_PLAY_4W = 1; ELSE BIN_PLAY_4W
= 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 8 THEN BIN_PLAY_8W = 1; ELSE BIN_PLAY_8W
= 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 12 THEN BIN_PLAY_12W = 1; ELSE
BIN_PLAY_12W = 0 ;
/*     IF TIME_DIFF GT 0 AND TIME_DIFF LE 16 THEN BIN_PLAY_16W = 1; ELSE
BIN_PLAY_16W = 0 ;*/
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 24 THEN BIN_PLAY_24W = 1; ELSE
BIN_PLAY_24W = 0 ;
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 38 THEN BIN_PLAY_38W = 1; ELSE
BIN_PLAY_38W = 0 ;
/*     IF TIME_DIFF GT 0 AND TIME_DIFF LE 48 THEN BIN_PLAY_48W = 1; ELSE
BIN_PLAY_48W = 0 ;*/
       IF TIME_DIFF GT 0 AND TIME_DIFF LE 76 THEN BIN_PLAY_76W = 1; ELSE
BIN_PLAY_76W = 0 ;
       IF TIME_DIFF GT 0 THEN BIN_PLAY_EVER = 1; ELSE BIN_PLAY_EVER = 0 ;

/*     IF TIME_DIFF GE 0 AND  TIME_DIFF LE 1 THEN BIN_PLAY_0W1W =1; ELSE
BIN_PLAY_0W1W=0;*/
       IF TIME_DIFF GE 2 AND  TIME_DIFF LE 4 THEN BIN_PLAY_2W4W =1; ELSE
BIN_PLAY_2W4W=0;
/*     IF TIME_DIFF GE 5 AND  TIME_DIFF LE 7 THEN BIN_PLAY_5W7W =1; ELSE
BIN_PLAY_5W7W=0;*/
       IF TIME_DIFF GE 8 AND  TIME_DIFF LE 12 THEN BIN_PLAY_8W12W =1; ELSE
BIN_PLAY_8W12W=0;
/*     IF TIME_DIFF GE 13 AND  TIME_DIFF LE 19 THEN BIN_PLAY_13W19W =1; ELSE
BIN_PLAY_13W19W=0;*/
       IF TIME_DIFF GE 20 AND  TIME_DIFF LE 38 THEN BIN_PLAY_20W38W =1; ELSE
BIN_PLAY_20W38W=0;
       IF TIME_DIFF GE 39 AND  TIME_DIFF LE 76 THEN BIN_PLAY_39W76W =1; ELSE
BIN_PLAY_39W76W=0;

run;

%mend timeBand;
%timeBand(libin=WORK,dsin=base_table_a,libout=WORK,dsout=base_table_a);
%timeBand(libin=WORK,dsin=base_table_b,libout=WORK,dsout=base_table_b);
%timeBand(libin=WORK,dsin=base_table_c,libout=WORK,dsout=base_table_c);
%timeBand(libin=WORK,dsin=base_table_d,libout=WORK,dsout=base_table_d);
%timeBand(libin=WORK,dsin=base_table_z,libout=WORK,dsout=base_table_z);
%timeBand(libin=WORK,dsin=base_table_x,libout=WORK,dsout=base_table_x);



*create date week flags for timeline;


PROC SQL;
   CREATE TABLE class_imbalance AS
   SELECT Season
```

```
            ,(COUNT(TARGET_HW)) AS CNT_GAMES
                  ,SUM(CASE WHEN TARGET_HW=0 THEN 1 ELSE 0 END) AS CNT_0
                  ,SUM(CASE WHEN TARGET_HW=1 THEN 1 ELSE 0 END) AS CNT_1
                  ,SUM(CASE WHEN TARGET_HW=2 THEN 1 ELSE 0 END) AS CNT_2
                  ,calculated CNT_0/calculated CNT_GAMES as CNT_0_PREC
                  ,calculated CNT_1/calculated CNT_GAMES as CNT_1_PREC
                  ,calculated CNT_2/calculated CNT_GAMES as CNT_2_PREC
      FROM WORK.BASE_TABLE
      GROUP BY Season
      ORDER BY Season;
QUIT;
```

f. Build Home Win ABTs

```
* PROJECT:           Assignment 2: Sports Analytics
* NAME:              Build ABTs.sas
* AUTHOR:            Shane McCarthy
* EMAIL:             shane.mc-carthy@ucdconnect.ie
* DATE CREATED:      22/03/16
* PURPOSE:           This script builds the ABT (analytics base table for
modelling)
;



%MACRO periodLOOP(libin = /*input library*/
                        ,dsin=/*input dataset*/
                        ,libout=/*output library*/
                        ,dsout=/*output dataset*/
                        ,universe=/*universe name*/
                        ,Ppostfix=/*dist to the period tag postfix on period
vars (_1M)*/);


/*Get period names*/
PROC SQL;
             CREATE TABLE PERD_LST as
                    SELECT
                             name AS PERIODS
                    FROM
                          dictionary.columns
                    WHERE
                          libname= "&LIBIN"
                          AND memname = "&DSIN"
                          and prxmatch(cat("m/BIN_PLAY/oi"),upcase(name)) >0;
;
QUIT;

/*     Count number oF PERIODS */
       PROC SQL NOPRINT;
             SELECT COUNT(PERIODS) INTO :N_PERIODS FROM PERD_LST;
QUIT;

/*     Creates macro variables containing period names*/
       DATA _NULL_;
             length ii $20.;
             SET PERD_LST end=last;
                   i+1;
                   ii=LEFT(PUT(i,20.));
                   call symputx('PERIOD'||ii, LEFT(PERIODS));
                   IF last THEN call symputx('N_PERIODS', TRIM(LEFT(_N_)));
       RUN;
/*Check the value stored*/

%PUT ***NOTE: &N_PERIODS. periods have been read into the macro var "PERIOD" ;

/*Loop through each period */
```

```
%DO X=1 %TO &N_PERIODS.;


       %PUT ***NOTE: STARTING NUMBER &X. OF &N_PERIODS, PERIOD = &&PERIOD&X.;

/****STEP 1: Aggregate functions by period */



PROC SQL;
CREATE TABLE &&PERIOD&X. AS
             SELECT
             UID
             ,substr("&&PERIOD&X.",9) AS PERIOD LENGTH = 15
             ,sum(case when FTR = "H" then 1 else 0 end) as CNT_FT_WINS label
"Count of FT home win"
             ,sum(case when FTR = "A" then 1 else 0 end) as CNT_FT_LOSS label
"Count of FT home loss"
             ,sum(case when FTR = "D" then 1 else 0 end) as CNT_FT_DRAW label
"Count of FT home draw"
             ,sum(case when HTR = "H" then 1 else 0 end) as CNT_HT_WINS label
"Count of HT home win"
             ,sum(case when HTR = "A" then 1 else 0 end) as CNT_HT_LOSS label
"Count of HT home loss"
             ,sum(case when HTR = "D" then 1 else 0 end) as CNT_HT_DRAW label
"Count of HT home draw"
             ,SUM(FTHG) AS CNT_GOALS_SCORED label "Count of goals scored"
             ,SUM(FTAG) AS CNT_GOALS_SUCC label "Count of goals succeeded"
             ,sum(case when AwayTeam ="Arsenal " then 1 else 0 end) as CNT_OPP_1
label "Opposition is Arsenal"
             ,sum(case when AwayTeam ="Aston Villa " then 1 else 0 end) as
CNT_OPP_2 label "Opposition is Aston Villa"
             ,sum(case when AwayTeam ="Birmingham " then 1 else 0 end) as CNT_OPP_3
label "Opposition is Birmingham"
             ,sum(case when AwayTeam ="Blackburn " then 1 else 0 end) as CNT_OPP_4
label "Opposition is Blackburn"
             ,sum(case when AwayTeam ="Blackpool " then 1 else 0 end) as CNT_OPP_5
label "Opposition is Blackpool"
             ,sum(case when AwayTeam ="Bolton " then 1 else 0 end) as CNT_OPP_6
label "Opposition is Bolton"
             ,sum(case when AwayTeam ="Bournemouth " then 1 else 0 end) as
CNT_OPP_7 label "Opposition is Bournemouth"
             ,sum(case when AwayTeam ="Burnley " then 1 else 0 end) as CNT_OPP_8
label "Opposition is Burnley"
             ,sum(case when AwayTeam ="Cardiff " then 1 else 0 end) as CNT_OPP_9
label "Opposition is Cardiff"
             ,sum(case when AwayTeam ="Chelsea " then 1 else 0 end) as CNT_OPP_10
label "Opposition is Chelsea"
             ,sum(case when AwayTeam ="Crystal Pal " then 1 else 0 end) as
CNT_OPP_11 label "Opposition is Crystal Pal"
             ,sum(case when AwayTeam ="Everton " then 1 else 0 end) as CNT_OPP_12
label "Opposition is Everton"
             ,sum(case when AwayTeam ="Fulham " then 1 else 0 end) as CNT_OPP_13
label "Opposition is Fulham"
             ,sum(case when AwayTeam ="Hull " then 1 else 0 end) as CNT_OPP_14
label "Opposition is Hull"
             ,sum(case when AwayTeam ="Leicester " then 1 else 0 end) as CNT_OPP_15
label "Opposition is Leicester"
             ,sum(case when AwayTeam ="Liverpool " then 1 else 0 end) as CNT_OPP_16
label "Opposition is Liverpool"
             ,sum(case when AwayTeam ="Man City " then 1 else 0 end) as CNT_OPP_17
label "Opposition is Man City"
             ,sum(case when AwayTeam ="Man United " then 1 else 0 end) as
CNT_OPP_18 label "Opposition is Man United"
             ,sum(case when AwayTeam ="Newcastle " then 1 else 0 end) as CNT_OPP_19
label "Opposition is Newcastle"
```

```
             ,sum(case when AwayTeam ="Norwich " then 1 else 0 end) as CNT_OPP_20
label "Opposition is Norwich"
             ,sum(case when AwayTeam ="QPR " then 1 else 0 end) as CNT_OPP_21 label
"Opposition is QPR"
             ,sum(case when AwayTeam ="Reading " then 1 else 0 end) as CNT_OPP_22
label "Opposition is Reading"
             ,sum(case when AwayTeam ="Southampton " then 1 else 0 end) as
CNT_OPP_23 label "Opposition is Southampton"
             ,sum(case when AwayTeam ="Stoke " then 1 else 0 end) as CNT_OPP_24
label "Opposition is Stoke"
             ,sum(case when AwayTeam ="Sunderland " then 1 else 0 end) as
CNT_OPP_25 label "Opposition is Sunderland"
             ,sum(case when AwayTeam ="Swansea " then 1 else 0 end) as CNT_OPP_26
label "Opposition is Swansea"
             ,sum(case when AwayTeam ="Tottenham " then 1 else 0 end) as CNT_OPP_27
label "Opposition is Tottenham"
             ,sum(case when AwayTeam ="Watford " then 1 else 0 end) as CNT_OPP_28
label "Opposition is Watford"
             ,sum(case when AwayTeam ="West Brom " then 1 else 0 end) as CNT_OPP_29
label "Opposition is West Brom"
             ,sum(case when AwayTeam ="West Ham " then 1 else 0 end) as CNT_OPP_30
label "Opposition is West Ham"
             ,sum(case when AwayTeam ="Wigan " then 1 else 0 end) as CNT_OPP_31
label "Opposition is Wigan"
             ,sum(case when AwayTeam ="Wolves " then 1 else 0 end) as CNT_OPP_32
label "Opposition is Wolves"
             ,sum(case when Referee ="A Marriner " then 1 else 0 end) as CNT_REF_1
label "REF  is A Marriner"
             ,sum(case when Referee ="A Taylor " then 1 else 0 end) as CNT_REF_2
label "REF  is A Taylor"
             ,sum(case when Referee ="C Foy " then 1 else 0 end) as CNT_REF_3 label
"REF  is C Foy"
             ,sum(case when Referee ="C Pawson " then 1 else 0 end) as CNT_REF_4
label "REF  is C Pawson"
             ,sum(case when Referee ="G Scott " then 1 else 0 end) as CNT_REF_5
label "REF  is G Scott"
             ,sum(case when Referee ="H Webb " then 1 else 0 end) as CNT_REF_6
label "REF  is H Webb"
             ,sum(case when Referee ="J Moss " then 1 else 0 end) as CNT_REF_7
label "REF  is J Moss"
             ,sum(case when Referee ="K Friend " then 1 else 0 end) as CNT_REF_8
label "REF  is K Friend"
             ,sum(case when Referee ="K Stroud " then 1 else 0 end) as CNT_REF_9
label "REF  is K Stroud"
             ,sum(case when Referee ="L Mason " then 1 else 0 end) as CNT_REF_10
label "REF  is L Mason"
             ,sum(case when Referee ="L Probert " then 1 else 0 end) as CNT_REF_11
label "REF  is L Probert"
             ,sum(case when Referee ="M Atkinson " then 1 else 0 end) as CNT_REF_12
label "REF  is M Atkinson"
             ,sum(case when Referee ="M Clattenbu " then 1 else 0 end) as
CNT_REF_13 label "REF  is M Clattenbu"
             ,sum(case when Referee ="M Clattenburg " then 1 else 0 end) as
CNT_REF_14 label "REF  is M Clattenburg"
             ,sum(case when Referee ="M Dean " then 1 else 0 end) as CNT_REF_15
label "REF  is M Dean"
             ,sum(case when Referee ="M Halsey " then 1 else 0 end) as CNT_REF_16
label "REF  is M Halsey"
             ,sum(case when Referee ="M Jones " then 1 else 0 end) as CNT_REF_17
label "REF  is M Jones"
             ,sum(case when Referee ="M Oliver " then 1 else 0 end) as CNT_REF_18
label "REF  is M Oliver"
             ,sum(case when Referee ="N Swarbrick " then 1 else 0 end) as
CNT_REF_19 label "REF  is N Swarbrick"
             ,sum(case when Referee ="P Dowd " then 1 else 0 end) as CNT_REF_20
label "REF  is P Dowd"
             ,sum(case when Referee ="P Tierney " then 1 else 0 end) as CNT_REF_21
label "REF  is P Tierney"
```

```
             ,sum(case when Referee ="P Walton " then 1 else 0 end) as CNT_REF_22
label "REF  is P Walton"
             ,sum(case when Referee ="R East " then 1 else 0 end) as CNT_REF_23
label "REF  is R East"
             ,sum(case when Referee ="R Madley " then 1 else 0 end) as CNT_REF_24
label "REF  is R Madley"
             ,sum(case when Referee ="S Attwell " then 1 else 0 end) as CNT_REF_25
label "REF  is S Attwell"
             ,sum(case when Referee ="S Hooper " then 1 else 0 end) as CNT_REF_26
label "REF  is S Hooper"
             ,sum(FTHG) AS SUM_FTHG label "Sum Full Time Home Team Goals"
             ,sum(FTAG) AS SUM_FTAG label "Sum Full Time Away Team Goals"
             ,sum(HTHG) AS SUM_HTHG label "Sum Half Time Home Team Goals"
             ,sum(HTAG) AS SUM_HTAG label "Sum Half Time Away Team Goals"
             ,sum(HS) AS SUM_HS label "Sum Home Team Shots"
             ,sum(AS) AS SUM_AS label "Sum Away Team Shots"
             ,sum(HST) AS SUM_HST label "Sum Home Team Shots on Target"
             ,sum(AST) AS SUM_AST label "Sum Away Team Shots on Target"
             ,sum(HF) AS SUM_HF label "Sum Home Team Fouls Committed"
             ,sum(AF) AS SUM_AF label "Sum Away Team Fouls Committed"
             ,sum(HC) AS SUM_HC label "Sum Home Team Corners"
             ,sum(AC) AS SUM_AC label "Sum Away Team Corners"
             ,sum(HY) AS SUM_HY label "Sum Home Team Yellow Cards"
             ,sum(AY) AS SUM_AY label "Sum Away Team Yellow Cards"
             ,sum(HR) AS SUM_HR label "Sum Home Team Red Cards"
             ,sum(AR) AS SUM_AR label "Sum Away Team Red Cards"
             ,sum(Bb1X2) AS SUM_Bb1X2 label "Sum Number of BetBrain bookmakers used
to calculate match odds averages and maximums"
             ,sum(BbOU) AS SUM_BbOU label "Sum Number of BetBrain bookmakers used
to calculate over/under 2.5 goals (total goals) averages and maximums"
             ,MEAN(FTHG) AS MEAN_FTHG label "Mean Full Time Home Team Goals"
             ,MEAN(FTAG) AS MEAN_FTAG label "Mean Full Time Away Team Goals"
             ,MEAN(HTHG) AS MEAN_HTHG label "Mean Half Time Home Team Goals"
             ,MEAN(HTAG) AS MEAN_HTAG label "Mean Half Time Away Team Goals"
             ,MEAN(HS) AS MEAN_HS label "Mean Home Team Shots"
             ,MEAN(AS) AS MEAN_AS label "Mean Away Team Shots"
             ,MEAN(HST) AS MEAN_HST label "Mean Home Team Shots on Target"
             ,MEAN(AST) AS MEAN_AST label "Mean Away Team Shots on Target"
             ,MEAN(HF) AS MEAN_HF label "Mean Home Team Fouls Committed"
             ,MEAN(AF) AS MEAN_AF label "Mean Away Team Fouls Committed"
             ,MEAN(HC) AS MEAN_HC label "Mean Home Team Corners"
             ,MEAN(AC) AS MEAN_AC label "Mean Away Team Corners"
             ,MEAN(HY) AS MEAN_HY label "Mean Home Team Yellow Cards"
             ,MEAN(AY) AS MEAN_AY label "Mean Away Team Yellow Cards"
             ,MEAN(HR) AS MEAN_HR label "Mean Home Team Red Cards"
             ,MEAN(AR) AS MEAN_AR label "Mean Away Team Red Cards"
             ,MEAN(B365H) AS MEAN_B365H label "Mean Bet365 home win odds"
             ,MEAN(B365D) AS MEAN_B365D label "Mean Bet365 draw odds"
             ,MEAN(B365A) AS MEAN_B365A label "Mean Bet365 away win odds"
             ,MEAN(BWH) AS MEAN_BWH label "Mean Bet & Win home win odds"
             ,MEAN(BWD) AS MEAN_BWD label "Mean Bet & Win draw odds"
             ,MEAN(BWA) AS MEAN_BWA label "Mean Bet & Win away win odds"
             ,MEAN(GBH) AS MEAN_GBH label "Mean Gamebookers home win odds"
             ,MEAN(GBD) AS MEAN_GBD label "Mean Gamebookers draw odds"
             ,MEAN(GBA) AS MEAN_GBA label "Mean Gamebookers away win odds"
             ,MEAN(IWH) AS MEAN_IWH label "Mean Interwetten home win odds"
             ,MEAN(IWD) AS MEAN_IWD label "Mean Interwetten draw odds"
             ,MEAN(IWA) AS MEAN_IWA label "Mean Interwetten away win odds"
             ,MEAN(LBH) AS MEAN_LBH label "Mean Ladbrokes home win odds"
             ,MEAN(LBD) AS MEAN_LBD label "Mean Ladbrokes draw odds"
             ,MEAN(LBA) AS MEAN_LBA label "Mean Ladbrokes away win odds"
             ,MEAN(SBH) AS MEAN_SBH label "Mean Sportingbet home win odds"
             ,MEAN(SBD) AS MEAN_SBD label "Mean Sportingbet draw odds"
             ,MEAN(SBA) AS MEAN_SBA label "Mean Sportingbet away win odds"
             ,MEAN(WHD) AS MEAN_WHD label "Mean William Hill draw odds"
             ,MEAN(SJH) AS MEAN_SJH label "Mean Stan James home win odds"
             ,MEAN(SJD) AS MEAN_SJD label "Mean Stan James draw odds"
             ,MEAN(SJA) AS MEAN_SJA label "Mean Stan James away win odds"
             ,MEAN(VCH) AS MEAN_VCH label "Mean VC Bet home win odds"
             ,MEAN(VCD) AS MEAN_VCD label "Mean VC Bet draw odds"
             ,MEAN(VCA) AS MEAN_VCA label "Mean VC Bet away win odds"
             ,MEAN(BSH) AS MEAN_BSH label "Mean Blue Square home win odds"
             ,MEAN(BSD) AS MEAN_BSD label "Mean Blue Square draw odds"
             ,MEAN(BSA) AS MEAN_BSA label "Mean Blue Square away win odds"
             ,MEAN(Bb1X2) AS MEAN_Bb1X2 label "Mean Number of BetBrain bookmakers
used to calculate match odds averages and maximums"
             ,MEAN(BbMxH) AS MEAN_BbMxH label "Mean Betbrain maximum home win odds"
             ,MEAN(BbAvH) AS MEAN_BbAvH label "Mean Betbrain average home win odds"
             ,MEAN(BbMxD) AS MEAN_BbMxD label "Mean Betbrain maximum draw odds"
             ,MEAN(BbAvD) AS MEAN_BbAvD label "Mean Betbrain average draw win odds"
             ,MEAN(BbMxA) AS MEAN_BbMxA label "Mean Betbrain maximum away win odds"
             ,MEAN(BbAvA) AS MEAN_BbAvA label "Mean Betbrain average away win odds"
             ,MEAN(BbOU) AS MEAN_BbOU label "Mean Number of BetBrain bookmakers
used to calculate over/under 2.5 goals (total goals) averages and maximums"
             ,MEAN(BbMx_GE_2pt5) AS MEAN_BbMx_GE_2pt5 label "Mean Betbrain maximum
over 2.5 goals"
             ,MAX(FTHG) AS MAX_FTHG label "Max Full Time Home Team Goals"
             ,MAX(FTAG) AS MAX_FTAG label "Max Full Time Away Team Goals"
             ,MAX(HTHG) AS MAX_HTHG label "Max Half Time Home Team Goals"
             ,MAX(HTAG) AS MAX_HTAG label "Max Half Time Away Team Goals"
             ,MAX(HS) AS MAX_HS label "Max Home Team Shots"
             ,MAX(AS) AS MAX_AS label "Max Away Team Shots"
             ,MAX(HST) AS MAX_HST label "Max Home Team Shots on Target"
             ,MAX(AST) AS MAX_AST label "Max Away Team Shots on Target"
             ,MAX(HF) AS MAX_HF label "Max Home Team Fouls Committed"
             ,MAX(AF) AS MAX_AF label "Max Away Team Fouls Committed"
             ,MAX(HC) AS MAX_HC label "Max Home Team Corners"
             ,MAX(AC) AS MAX_AC label "Max Away Team Corners"
             ,MAX(HY) AS MAX_HY label "Max Home Team Yellow Cards"
             ,MAX(AY) AS MAX_AY label "Max Away Team Yellow Cards"
             ,MAX(HR) AS MAX_HR label "Max Home Team Red Cards"
             ,MAX(AR) AS MAX_AR label "Max Away Team Red Cards"
             ,MAX(B365H) AS MAX_B365H label "Max Bet365 home win odds"
             ,MAX(B365D) AS MAX_B365D label "Max Bet365 draw odds"
             ,MAX(B365A) AS MAX_B365A label "Max Bet365 away win odds"
             ,MAX(BWH) AS MAX_BWH label "Max Bet & Win home win odds"
             ,MAX(BWD) AS MAX_BWD label "Max Bet & Win draw odds"
             ,MAX(BWA) AS MAX_BWA label "Max Bet & Win away win odds"
             ,MAX(GBH) AS MAX_GBH label "Max Gamebookers home win odds"
             ,MAX(GBD) AS MAX_GBD label "Max Gamebookers draw odds"
             ,MAX(GBA) AS MAX_GBA label "Max Gamebookers away win odds"
             ,MAX(IWH) AS MAX_IWH label "Max Interwetten home win odds"
             ,MAX(IWD) AS MAX_IWD label "Max Interwetten draw odds"
             ,MAX(IWA) AS MAX_IWA label "Max Interwetten away win odds"
             ,MAX(LBH) AS MAX_LBH label "Max Ladbrokes home win odds"
             ,MAX(LBD) AS MAX_LBD label "Max Ladbrokes draw odds"
             ,MAX(LBA) AS MAX_LBA label "Max Ladbrokes away win odds"
             ,MAX(SBH) AS MAX_SBH label "Max Sportingbet home win odds"
             ,MAX(SBD) AS MAX_SBD label "Max Sportingbet draw odds"
             ,MAX(SBA) AS MAX_SBA label "Max Sportingbet away win odds"
             ,MAX(WHD) AS MAX_WHD label "Max William Hill draw odds"
             ,MAX(SJH) AS MAX_SJH label "Max Stan James home win odds"
             ,MAX(SJD) AS MAX_SJD label "Max Stan James draw odds"
             ,MAX(SJA) AS MAX_SJA label "Max Stan James away win odds"
             ,MAX(VCH) AS MAX_VCH label "Max VC Bet home win odds"
             ,MAX(VCD) AS MAX_VCD label "Max VC Bet draw odds"
             ,MAX(VCA) AS MAX_VCA label "Max VC Bet away win odds"
             ,MAX(BSH) AS MAX_BSH label "Max Blue Square home win odds"
             ,MAX(BSD) AS MAX_BSD label "Max Blue Square draw odds"
             ,MAX(BSA) AS MAX_BSA label "Max Blue Square away win odds"
             ,MAX(Bb1X2) AS MAX_Bb1X2 label "Max Number of BetBrain bookmakers used
to calculate match odds averages and maximums"
             ,MAX(BbMxH) AS MAX_BbMxH label "Max Betbrain maximum home win odds"
             ,MAX(BbAvH) AS MAX_BbAvH label "Max Betbrain average home win odds"
             ,MAX(BbMxD) AS MAX_BbMxD label "Max Betbrain maximum draw odds"
             ,MAX(BbAvD) AS MAX_BbAvD label "Max Betbrain average draw win odds"
```

```
              ,MAX(BbMxA) AS MAX_BbMxA label "Max Betbrain maximum away win odds"
              ,MAX(BbAvA) AS MAX_BbAvA label "Max Betbrain average away win odds"
              ,MAX(BbOU) AS MAX_BbOU label "Max Number of BetBrain bookmakers used
to calculate over/under 2.5 goals (total goals) averages and maximums"
              ,MAX(BbMx_GE_2pt5) AS MAX_BbMx_GE_2pt5 label "Max Betbrain maximum
over 2.5 goals"
              ,MIN(FTHG) AS MIN_FTHG label "Min Full Time Home Team Goals"
              ,MIN(FTAG) AS MIN_FTAG label "Min Full Time Away Team Goals"
              ,MIN(HTHG) AS MIN_HTHG label "Min Half Time Home Team Goals"
              ,MIN(HTAG) AS MIN_HTAG label "Min Half Time Away Team Goals"
              ,MIN(HS) AS MIN_HS label "Min Home Team Shots"
              ,MIN(AS) AS MIN_AS label "Min Away Team Shots"
              ,MIN(HST) AS MIN_HST label "Min Home Team Shots on Target"
              ,MIN(AST) AS MIN_AST label "Min Away Team Shots on Target"
              ,MIN(HF) AS MIN_HF label "Min Home Team Fouls Committed"
              ,MIN(AF) AS MIN_AF label "Min Away Team Fouls Committed"
              ,MIN(HC) AS MIN_HC label "Min Home Team Corners"
              ,MIN(AC) AS MIN_AC label "Min Away Team Corners"
              ,MIN(HY) AS MIN_HY label "Min Home Team Yellow Cards"
              ,MIN(AY) AS MIN_AY label "Min Away Team Yellow Cards"
              ,MIN(HR) AS MIN_HR label "Min Home Team Red Cards"
              ,MIN(AR) AS MIN_AR label "Min Away Team Red Cards"
              ,MIN(B365H) AS MIN_B365H label "Min Bet365 home win odds"
              ,MIN(B365D) AS MIN_B365D label "Min Bet365 draw odds"
              ,MIN(B365A) AS MIN_B365A label "Min Bet365 away win odds"
              ,MIN(BWH) AS MIN_BWH label "Min Bet & Win home win odds"
              ,MIN(BWD) AS MIN_BWD label "Min Bet & Win draw odds"
              ,MIN(BWA) AS MIN_BWA label "Min Bet & Win away win odds"
              ,MIN(GBH) AS MIN_GBH label "Min Gamebookers home win odds"
              ,MIN(GBD) AS MIN_GBD label "Min Gamebookers draw odds"
              ,MIN(GBA) AS MIN_GBA label "Min Gamebookers away win odds"
              ,MIN(IWH) AS MIN_IWH label "Min Interwetten home win odds"
              ,MIN(IWD) AS MIN_IWD label "Min Interwetten draw odds"
              ,MIN(IWA) AS MIN_IWA label "Min Interwetten away win odds"
              ,MIN(LBH) AS MIN_LBH label "Min Ladbrokes home win odds"
              ,MIN(LBD) AS MIN_LBD label "Min Ladbrokes draw odds"
              ,MIN(LBA) AS MIN_LBA label "Min Ladbrokes away win odds"
              ,MIN(SBH) AS MIN_SBH label "Min Sportingbet home win odds"
              ,MIN(SBD) AS MIN_SBD label "Min Sportingbet draw odds"
              ,MIN(SBA) AS MIN_SBA label "Min Sportingbet away win odds"
              ,MIN(WHD) AS MIN_WHD label "Min William Hill draw odds"
              ,MIN(SJH) AS MIN_SJH label "Min Stan James home win odds"
              ,MIN(SJD) AS MIN_SJD label "Min Stan James draw odds"
              ,MIN(SJA) AS MIN_SJA label "Min Stan James away win odds"
              ,MIN(VCH) AS MIN_VCH label "Min VC Bet home win odds"
              ,MIN(VCD) AS MIN_VCD label "Min VC Bet draw odds"
              ,MIN(VCA) AS MIN_VCA label "Min VC Bet away win odds"
              ,MIN(BSH) AS MIN_BSH label "Min Blue Square home win odds"
              ,MIN(BSD) AS MIN_BSD label "Min Blue Square draw odds"
              ,MIN(BSA) AS MIN_BSA label "Min Blue Square away win odds"
              ,MIN(Bb1X2) AS MIN_Bb1X2 label "Min Number of BetBrain bookmakers used
to calculate match odds averages and maximums"
              ,MIN(BbMxH) AS MIN_BbMxH label "Min Betbrain maximum home win odds"
              ,MIN(BbAvH) AS MIN_BbAvH label "Min Betbrain average home win odds"
              ,MIN(BbMxD) AS MIN_BbMxD label "Min Betbrain maximum draw odds"
              ,MIN(BbAvD) AS MIN_BbAvD label "Min Betbrain average draw win odds"
              ,MIN(BbMxA) AS MIN_BbMxA label "Min Betbrain maximum away win odds"
              ,MIN(BbAvA) AS MIN_BbAvA label "Min Betbrain average away win odds"
              ,MIN(BbOU) AS MIN_BbOU label "Min Number of BetBrain bookmakers used
to calculate over/under 2.5 goals (total goals) averages and maximums"
              ,MIN(BbMx_GE_2pt5) AS MIN_BbMx_GE_2pt5 label "Min Betbrain maximum
over 2.5 goals"
              from  &libin..&dsin.
              WHERE &&PERIOD&X.= 1
              group by 1,2;

       %END;
```

```
/****STEP 2: Stack periods back together and delete intermediate tables once
stacked */

DATA periods_stacked;
       SET
             %DO X = 1 %TO &N_PERIODS.;
                    &&PERIOD&X.
             %END;
       ;
RUN;

/*Delete intermediate tables once stacked*/
PROC DATASETS LIBRARY= WORK;
DELETE
    %DO X = 1 %TO &N_PERIODS.;
                    &&PERIOD&X.
    %END;
    ;
RUN;



/****STEP 3: Rank values by period into decile groups (0-9)*/

PROC SORT DATA=periods_stacked out=periods_stacked; BY PERIOD ; RUN;




       PROC RANK DATA = periods_stacked
       GROUPS=10
       TIES=MEAN
       OUT=periods_stacked_rnk;
       BY PERIOD;

       VAR
             CNT_FT_WINS
             CNT_FT_LOSS
             CNT_FT_DRAW
             CNT_HT_WINS
             CNT_HT_LOSS
             CNT_HT_DRAW
             SUM_FTHG
             SUM_FTAG
             SUM_HTHG
             SUM_HTAG
             SUM_HS
             SUM_AS
             SUM_HST
             SUM_AST
             SUM_HF
             SUM_AF
             SUM_HC
             SUM_AC
             SUM_HY
             SUM_AY
             SUM_HR
             SUM_AR



             ;
       RANKS
             CNT_FT_WINS_RNK
             CNT_FT_LOSS_RNK
             CNT_FT_DRAW_RNK
             CNT_HT_WINS_RNK
             CNT_HT_LOSS_RNK
```

```
              CNT_HT_DRAW_RNK
              SUM_FTHG_RNK
              SUM_FTAG_RNK
              SUM_HTHG_RNK
              SUM_HTAG_RNK
              SUM_HS_RNK
              SUM_AS_RNK
              SUM_HST_RNK
              SUM_AST_RNK
              SUM_HF_RNK
              SUM_AF_RNK
              SUM_HC_RNK
              SUM_AC_RNK
              SUM_HY_RNK
              SUM_AY_RNK
              SUM_HR_RNK
              SUM_AR_RNK
;
       RUN;


PROC SQL;
       SELECT
              name AS variables
       into :trans_var SEPARATED BY " "
              FROM
                     dictionary.columns
              WHERE
                     libname= "WORK"
                     AND memname = "PERIODS_STACKED_RNK"
                     and prxmatch(cat("m/UID|PERIOD/oi"),upcase(name)) =0;
       ;
QUIT;

%put &trans_var.;


/****STEP 5: Tranpose by all vars period */

%MultiTranspose
(out= data_transposed
,data=periods_stacked_rnk
,vars=  &trans_var.
,by= UID
,pivot=PERIOD
);



PROC SQL;
create table temp as
       SELECT
               u.TARGET_HW
               ,u.HomeTeam
               ,u.AwayTeam
               ,u.Referee
               ,dt.*

              FROM &UNIVERSE. u
              inner join DATA_TRANSPOSED dt on dt.UID=u.ID;

       ;
QUIT;

data &libout..&dsout. ;
set temp ;

drop
```

```
CNT_FT_WINS_0W CNT_FT_LOSS_0W CNT_FT_DRAW_0W CNT_HT_WINS_0W CNT_HT_LOSS_0W
CNT_HT_DRAW_0W CNT_GOALS_SCORED_0W CNT_GOALS_SUCC_0W SUM_FTHG_0W
SUM_FTAG_0W SUM_HTHG_0W SUM_HTAG_0W SUM_HS_0W SUM_AS_0W SUM_HST_0W SUM_AST_0W
SUM_HF_0W SUM_AF_0W SUM_HC_0W SUM_AC_0W SUM_HY_0W SUM_AY_0W SUM_HR_0W
SUM_AR_0W MEAN_FTHG_0W MEAN_FTAG_0W MEAN_HTHG_0W MEAN_HTAG_0W MEAN_HS_0W MEAN_AS_0W
MEAN_HST_0W MEAN_AST_0W MEAN_HF_0W MEAN_AF_0W MEAN_HC_0W MEAN_AC_0W
MEAN_HY_0W MEAN_AY_0W MEAN_HR_0W MEAN_AR_0W MAX_FTHG_0W MAX_FTAG_0W MAX_HTHG_0W
MAX_HTAG_0W MAX_HS_0W MAX_AS_0W MAX_HST_0W MAX_AST_0W MAX_HF_0W
MAX_AF_0W MAX_HC_0W MAX_AC_0W MAX_HY_0W MAX_AY_0W MAX_HR_0W MAX_AR_0W MAX_B365H_0W
MAX_B365D_0W MAX_B365A_0W MAX_BWH_0W MAX_BWD_0W MAX_BWA_0W
MAX_GBH_0W MAX_GBD_0W MAX_GBA_0W MAX_IWH_0W MAX_IWD_0W MAX_IWA_0W MAX_LBH_0W
MAX_LBD_0W MAX_LBA_0W MAX_SBH_0W MAX_SBD_0W MAX_SBA_0W
MAX_WHD_0W MAX_SJH_0W MAX_SJD_0W MAX_SJA_0W MAX_VCH_0W MAX_VCD_0W MAX_VCA_0W
MAX_BSH_0W MAX_BSD_0W MAX_BSA_0W MAX_Bb1X2_0W MAX_BbMxH_0W
MAX_BbAvH_0W MAX_BbMxD_0W MAX_BbAvD_0W MAX_BbMxA_0W MAX_BbAvA_0W MAX_BbOU_0W
MAX_BbMx_GE_2pt5_0W MIN_FTHG_0W MIN_FTAG_0W MIN_HTHG_0W MIN_HTAG_0W
MIN_HS_0W MIN_AS_0W MIN_HST_0W MIN_AST_0W MIN_HF_0W MIN_AF_0W MIN_HC_0W MIN_AC_0W
MIN_HY_0W MIN_AY_0W MIN_HR_0W MIN_AR_0W MIN_B365H_0W MIN_B365D_0W
MIN_B365A_0W MIN_BWH_0W MIN_BWD_0W MIN_BWA_0W MIN_GBH_0W MIN_GBD_0W MIN_GBA_0W
MIN_IWH_0W MIN_IWD_0W MIN_IWA_0W MIN_LBH_0W MIN_LBD_0W MIN_LBA_0W
MIN_SBH_0W MIN_SBD_0W MIN_SBA_0W MIN_WHD_0W MIN_SJH_0W MIN_SJD_0W MIN_SJA_0W
MIN_VCH_0W MIN_VCD_0W MIN_VCA_0W MIN_BSH_0W MIN_BSD_0W MIN_BSA_0W
MIN_Bb1X2_0W MIN_BbMxH_0W MIN_BbAvH_0W MIN_BbMxD_0W MIN_BbAvD_0W MIN_BbMxA_0W
MIN_BbAvA_0W MIN_BbOU_0W MIN_BbMx_GE_2pt5_0W
CNT_FT_WINS_RNK_0W CNT_FT_LOSS_RNK_0W CNT_FT_DRAW_RNK_0W CNT_HT_WINS_RNK_0W
CNT_HT_LOSS_RNK_0W CNT_HT_DRAW_RNK_0W
SUM_FTHG_RNK_0W SUM_FTAG_RNK_0W SUM_HTHG_RNK_0W SUM_HTAG_RNK_0W SUM_HS_RNK_0W
SUM_AS_RNK_0W SUM_HST_RNK_0W
SUM_AST_RNK_0W SUM_HF_RNK_0W SUM_AF_RNK_0W SUM_HC_RNK_0W SUM_AC_RNK_0W
SUM_HY_RNK_0W SUM_AY_RNK_0W SUM_HR_RNK_0W  SUM_AR_RNK_0W;
run;

%MEND periodLOOP;

%periodLOOP(libin=WORK ,dsin=BASE_TABLE_A,
libout=WD,dsout=ABT_A1_HOME_WIN,universe=WD.UNIVERSE_A, Ppostfix=12);
%periodLOOP(libin=WORK ,dsin=BASE_TABLE_B,
libout=WD,dsout=ABT_B1_HOME_WIN,universe=WD.UNIVERSE_B, Ppostfix=12);
%periodLOOP(libin=WORK ,dsin=BASE_TABLE_Z,
libout=WD,dsout=ABT_Z_HOME_WIN,universe=WD.UNIVERSE_Z, Ppostfix=12);
%periodLOOP(libin=WORK ,dsin=BASE_TABLE_X,
libout=WD,dsout=ABT_X_HOME_WIN,universe=WD.UNIVERSE_X, Ppostfix=12);
```

g. Build Expected Goals ABT

```
* PROJECT:           Assignment 2: Sports Analytics
* NAME:              Build ABTs.sas
* AUTHOR:            Shane McCarthy
* EMAIL:             shane.mc-carthy@ucdconnect.ie
* DATE CREATED:      22/03/16
* PURPOSE:           This script builds the ABT (analytics base table for
modelling)
;



%MACRO periodLOOP2(libin = /*input library*/
                        ,dsin=/*input dataset*/
                        ,libout=/*output library*/
                        ,dsout=/*output dataset*/
                        ,universe=/*universe name*/
                        ,Ppostfix=/*dist to the period tag postfix on period
vars (_1M)*/);


/*Get period names*/
PROC SQL;
```

```
             CREATE TABLE PERD_LST as
                    SELECT
                             name AS PERIODS
                    FROM
                           dictionary.columns
                    WHERE
                           libname= "&LIBIN"
                           AND memname = "&DSIN"
                           and prxmatch(cat("m/BIN_PLAY/oi"),upcase(name)) >0;
;
QUIT;

/*     Count number oF PERIODS */
       PROC SQL NOPRINT;
             SELECT COUNT(PERIODS) INTO :N_PERIODS FROM PERD_LST;
QUIT;

/*     Creates macro variables containing period names*/
       DATA _NULL_;
             length ii $20.;
             SET PERD_LST end=last;
                    i+1;
                    ii=LEFT(PUT(i,20.));
                    call symputx('PERIOD'||ii, LEFT(PERIODS));
                    IF last THEN call symputx('N_PERIODS', TRIM(LEFT(_N_)));
       RUN;
/*Check the value stored*/

%PUT ***NOTE: &N_PERIODS. periods have been read into the macro var "PERIOD" ;

/*Loop through each period */

%DO X=1 %TO &N_PERIODS.;


       %PUT ***NOTE: STARTING NUMBER &X. OF &N_PERIODS, PERIOD = &&PERIOD&X.;

/****STEP 1: Aggregate functions by period */

PROC SQL;
CREATE TABLE &&PERIOD&X. AS
             SELECT
             UID
             ,substr("&&PERIOD&X.",9) AS PERIOD LENGTH = 15
,SUM(WIN_CNT) AS SUM_WIN label "Count of Wins "
,SUM(LOSS_CNT) AS SUM_LOSS label "Count of Losses"
,SUM(DRAW_CNT) AS SUM_DRAW label "Count of Draws"
,SUM(HTWIN_CNT) AS SUM_HTWIN label "Count of  HT Wins "
,SUM(HTLOSS_CNT) AS SUM_HTLOSS label "Count of HT Losses"
,SUM(HTDRAW_CNT) AS SUM_HTDRAW label "Count of HT Draws"
,SUM(HT) AS SUM_HT label "Sum of Half Time Goals"
,SUM(SHT) AS SUM_SHT label "Sum of Shots"
,SUM(SHTT) AS SUM_SHTT label "Sum of Shots on Target"
,SUM(FC) AS SUM_FC label "Sum of Fouls Committed"
,SUM(CN) AS SUM_CN label "Sum of Corners"
,SUM(YC) AS SUM_YC label "Sum of Yellow Cards"
,SUM(RC) AS SUM_RC label "Sum of Red Cards"

,Mean(HT) AS Mean_HT label "Mean of Half Time Home Team Goals"
,Mean(SHT) AS Mean_SHT label "Mean of Home Team Shots"
,Mean(SHTT) AS Mean_SHTT label "Mean of Home Team Shots on Target"
,Mean(FC) AS Mean_FC label "Mean of Home Team Fouls Committed"
,Mean(CN) AS Mean_CN label "Mean of Home Team Corners"
,Mean(YC) AS Mean_YC label "Mean of Home Team Yellow Cards"
,Mean(RC) AS Mean_RC label "Mean of Home Team Red Cards"
,Mean(B365W) AS Mean_B365W label "Mean of Bet365 home win odds"
,Mean(B365D) AS Mean_B365D label "Mean of Bet365 draw odds"
,Mean(BWW) AS Mean_BWW label "Mean of Bet & Win home win odds"
```

```
,Mean(BWD) AS Mean_BWD label "Mean of Bet & Win draw odds"
,Mean(GBW) AS Mean_GBW label "Mean of Gamebookers home win odds"
,Mean(GBD) AS Mean_GBD label "Mean of Gamebookers draw odds"
,Mean(IWW) AS Mean_IWW label "Mean of Interwetten home win odds"
,Mean(IWD) AS Mean_IWD label "Mean of Interwetten draw odds"
,Mean(LBW) AS Mean_LBW label "Mean of Ladbrokes home win odds"
,Mean(LBD) AS Mean_LBD label "Mean of Ladbrokes draw odds"
,Mean(SBW) AS Mean_SBW label "Mean of Sportingbet home win odds"
,Mean(SBD) AS Mean_SBD label "Mean of Sportingbet draw odds"
,Mean(SBA) AS Mean_SBA label "Mean of Sportingbet away win odds"
,Mean(WHD) AS Mean_WHD label "Mean of William Hill draw odds"
,Mean(SJW) AS Mean_SJW label "Mean of Stan James home win odds"
,Mean(SJD) AS Mean_SJD label "Mean of Stan James draw odds"
,Mean(VCW) AS Mean_VCW label "Mean of VC Bet home win odds"
,Mean(VCD) AS Mean_VCD label "Mean of VC Bet draw odds"
,Mean(BSW) AS Mean_BSW label "Mean of Blue Square home win odds"
,Mean(BSD) AS Mean_BSD label "Mean of Blue Square draw odds"
,Mean(WIN_CNT) AS Mean_WIN label "Mean of Count of Wins "
,Mean(LOSS_CNT) AS Mean_LOSS label "Mean of Count of Losses"
,Mean(DRAW_CNT) AS Mean_DRAW label "Mean of Count of Draws"
,Mean(HTWIN_CNT) AS Mean_HTWIN label "Mean of Count of  HT Wins "
,Mean(HTLOSS_CNT) AS Mean_HTLOSS label "Mean of Count of HT Losses"
,Mean(HTDRAW_CNT) AS Mean_HTDRAW label "Mean of Count of HT Draws"

,STD(HT) AS STD_HT label "STD of Half Time Home Team Goals"
,STD(SHT) AS STD_SHT label "STD of Home Team Shots"
,STD(SHTT) AS STD_SHTT label "STD of Home Team Shots on Target"
,STD(FC) AS STD_FC label "STD of Home Team Fouls Committed"
,STD(CN) AS STD_CN label "STD of Home Team Corners"
,STD(YC) AS STD_YC label "STD of Home Team Yellow Cards"
,STD(RC) AS STD_RC label "STD of Home Team Red Cards"
,STD(WIN_CNT) AS STD_WIN label "STD of Count of Wins "
,STD(LOSS_CNT) AS STD_LOSS label "STD of Count of Losses"
,STD(DRAW_CNT) AS STD_DRAW label "STD of Count of Draws"
,STD(HTWIN_CNT) AS STD_HTWIN label "STD of Count of  HT Wins "
,STD(HTLOSS_CNT) AS STD_HTLOSS label "STD of Count of HT Losses"
,STD(HTDRAW_CNT) AS STD_HTDRAW label "STD of Count of HT Draws"

,MAX(HT) AS MAX_HT label "MAX of Half Time Home Team Goals"
,MAX(SHT) AS MAX_SHT label "MAX of Home Team Shots"
,MAX(SHTT) AS MAX_SHTT label "MAX of Home Team Shots on Target"
,MAX(FC) AS MAX_FC label "MAX of Home Team Fouls Committed"
,MAX(CN) AS MAX_CN label "MAX of Home Team Corners"
,MAX(YC) AS MAX_YC label "MAX of Home Team Yellow Cards"
,MAX(RC) AS MAX_RC label "MAX of Home Team Red Cards"
,MAX(WIN_CNT) AS MAX_WIN label "MAX of Count of Wins "
,MAX(LOSS_CNT) AS MAX_LOSS label "MAX of Count of Losses"
,MAX(DRAW_CNT) AS MAX_DRAW label "MAX of Count of Draws"
,MAX(HTWIN_CNT) AS MAX_HTWIN label "MAX of Count of  HT Wins "
,MAX(HTLOSS_CNT) AS MAX_HTLOSS label "MAX of Count of HT Losses"
,MAX(HTDRAW_CNT) AS MAX_HTDRAW label "MAX of Count of HT Draws"

,MIN(HT) AS MIN_HT label "MIN of Half Time Home Team Goals"
,MIN(SHT) AS MIN_SHT label "MIN of Home Team Shots"
,MIN(SHTT) AS MIN_SHTT label "MIN of Home Team Shots on Target"
,MIN(FC) AS MIN_FC label "MIN of Home Team Fouls Committed"
,MIN(CN) AS MIN_CN label "MIN of Home Team Corners"
,MIN(YC) AS MIN_YC label "MIN of Home Team Yellow Cards"
,MIN(RC) AS MIN_RC label "MIN of Home Team Red Cards"

,MIN(WIN_CNT) AS MIN_WIN label "MIN of Count of Wins "
,MIN(LOSS_CNT) AS MIN_LOSS label "MIN of Count of Losses"
,MIN(DRAW_CNT) AS MIN_DRAW label "MIN of Count of Draws"
,MIN(HTWIN_CNT) AS MIN_HTWIN label "MIN of Count of  HT Wins "
,MIN(HTLOSS_CNT) AS MIN_HTLOSS label "MIN of Count of HT Losses"
,MIN(HTDRAW_CNT) AS MIN_HTDRAW label "MIN of Count of HT Draws"
```

```
                from  &libin..&dsin.
                WHERE &&PERIOD&X.= 1
                group by 1,2;

        %END;

/****STEP 2: Stack periods back together and delete intermediate tables once
stacked */

DATA periods_stacked;
        SET
                %DO X = 1 %TO &N_PERIODS.;
                        &&PERIOD&X.
                %END;
        ;
RUN;

/*Delete intermediate tables once stacked*/
PROC DATASETS LIBRARY= WORK;
DELETE
    %DO X = 1 %TO &N_PERIODS.;
                        &&PERIOD&X.
    %END;
    ;
RUN;



/****STEP 3: Rank values by period into decile groups (0-9)*/

PROC SORT DATA=periods_stacked out=periods_stacked; BY PERIOD ; RUN;


        PROC RANK DATA = periods_stacked
        GROUPS=10
        TIES=MEAN
        OUT=periods_stacked_rnk;
        BY PERIOD;

        VAR
        SUM_WIN
        SUM_LOSS
        SUM_DRAW
        SUM_HTWIN
        SUM_HTLOSS
        SUM_HTDRAW
        SUM_HT
        SUM_SHT
        SUM_SHTT
        SUM_FC
        SUM_CN
        SUM_YC
        SUM_RC

                ;
        RANKS
        SUM_WIN_RNK
        SUM_LOSS_RNK
        SUM_DRAW_RNK
        SUM_HTWIN_RNK
        SUM_HTLOSS_RNK
        SUM_HTDRAW_RNK
        SUM_HT_RNK
        SUM_SHT_RNK
        SUM_SHTT_RNK
        SUM_FC_RNK
        SUM_CN_RNK
```

```
       SUM_YC_RNK
       SUM_RC_RNK

;
       RUN;

data periods_stacked_rnk;
set periods_stacked_rnk (where=(Period ^="_0W"));
run;


PROC SQL;
       SELECT
              name AS variables
       into :trans_var SEPARATED BY " "
              FROM
                     dictionary.columns
              WHERE
                     libname= "WORK"
                     AND memname = "PERIODS_STACKED_RNK"
                     and prxmatch(cat("m/UID|PERIOD/oi"),upcase(name)) =0;
       ;
QUIT;

%put &trans_var.;


/****STEP 5: Tranpose by all vars period */

%MultiTranspose
(out= data_transposed
,data=periods_stacked_rnk
,vars=  &trans_var.
,by= UID
,pivot=PERIOD
);


PROC SQL;
create table temp as
       SELECT
          u.Team
          ,u.Opposition
          ,u.Referee
          ,u.TARGET_GOAL
                ,dt.*

              FROM &UNIVERSE. u
              inner join DATA_TRANSPOSED dt on dt.UID=u.UID;

       ;
QUIT;

data &libout..&dsout. ;
set temp ;
run;
%MEND periodLOOP2;

%periodLOOP(libin=WORK ,dsin=BASE_TABLE_C,
libout=WD,dsout=ABT_C_GOAL_SCORED,universe=WD.UNIVERSE_C, Ppostfix=12);


%periodLOOP(libin=WORK ,dsin=BASE_TABLE_D,
libout=WD,dsout=ABT_D_GOAL_SCORED,universe=WD.UNIVERSE_D, Ppostfix=12);



%put &trans_var.;
```

Additional Plots

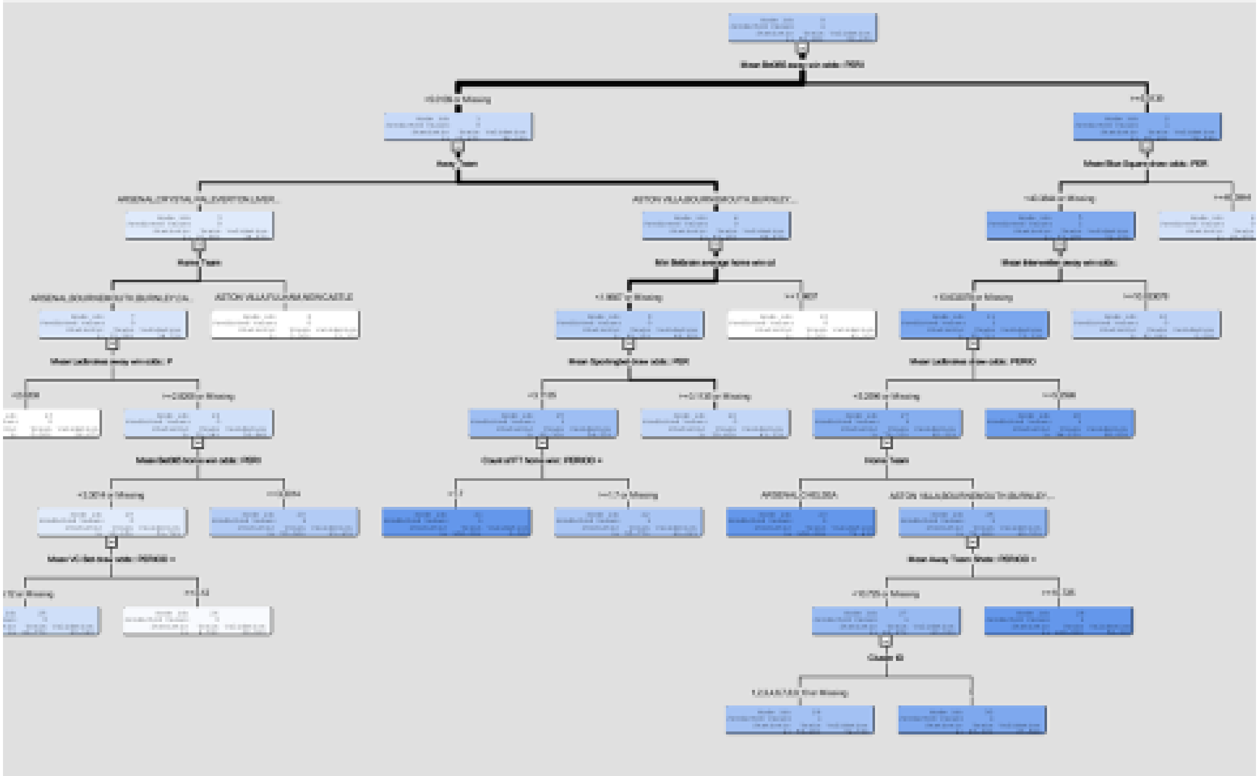

Away Team
ARSENAL,CRYSTAL PAL,EVERTON,LIVER...
ASTON VILLA,BOURNEMOUTH,BURNLEY...
Home Team
ASTON VILLA,FULHAM,NEWCASTLE
Home Team
ARSENAL,CHELSEA
ASTON VILLA,BOURNEMOUTH,BURNLEY...

Figure 19 Home Win Decision Tree Plot

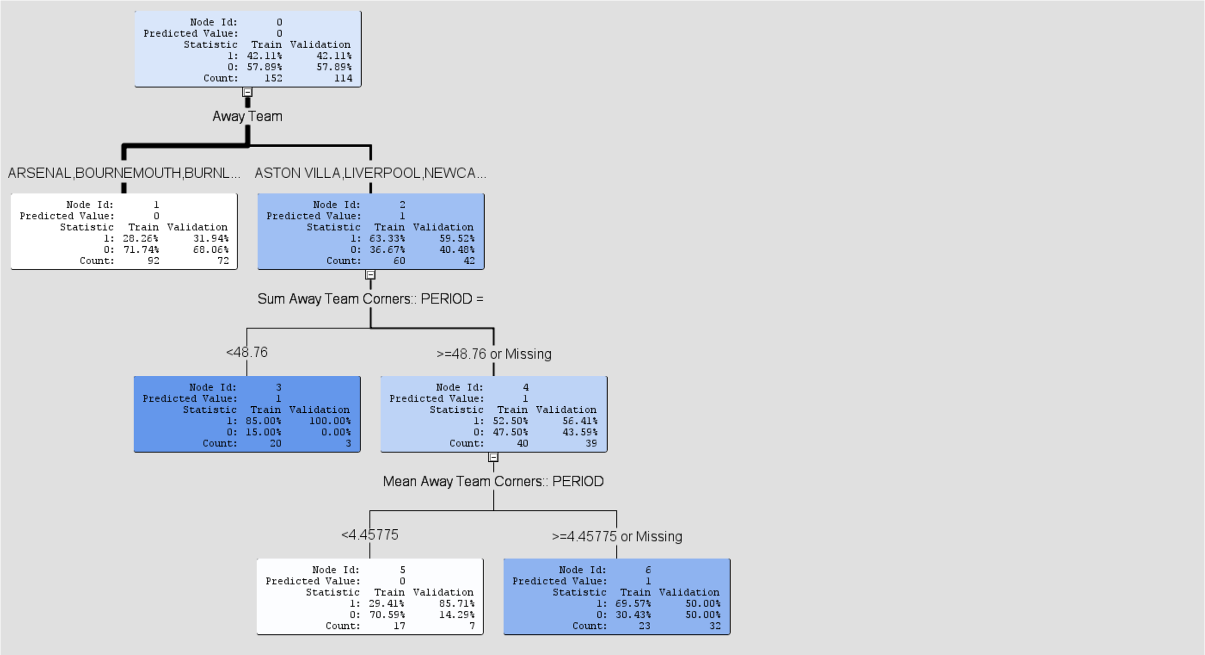


Figure 20 Expected Goals Decision Tree Plot

Decision Tree Rules – Approach one

--

NODE = 6

--

(Min Betbrain average draw win odds:: PERIOD = _24W >=2.0077)

AND (Mean Bet365 away win odds:: PERIOD = _0W >=5.0139)

PREDICTED VALUE IS 1

PREDICTED 1 = 0.9362(44/47)

PREDICTED 0 = 0.06383(3/47)

--

NODE = 19

--

MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W <2.4135)

AND (Mean Away Team Corners:: PERIOD = _38W >=5.36)

AND MISSING(Min Bet365 away win odds:: PERIOD = _20W38W) OR (Min Bet365 away win odds:: PERIOD = _20W38W >=1.4564)

AND MISSING(Mean Sportingbet draw odds:: PERIOD = _0W) OR (Mean Sportingbet draw odds:: PERIOD = _0W >=3.1135)

AND MISSING(Min Betbrain average home win odds:: PERIOD = _39W76W) OR (Min Betbrain average home win odds:: PERIOD = _39W76W <1.9607)

```
AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 0

   PREDICTED 1 = 0( 0/12)

   PREDICTED 0 = 1( 12/12)

*------------------------------------------------------------*

NODE = 15

*------------------------------------------------------------*

(Min Bet365 away win odds:: PERIOD = _20W38W <1.4564)

AND MISSING(Mean Sportingbet draw odds:: PERIOD = _0W) OR (Mean Sportingbet draw odds:: PERIOD =
_0W >=3.1135)

AND MISSING(Min Betbrain average home win odds:: PERIOD = _39W76W) OR (Min Betbrain average home
win odds:: PERIOD = _39W76W <1.9607)

AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 1

   PREDICTED 1 = 0.8182( 9/11)

   PREDICTED 0 = 0.1818( 2/11)

*------------------------------------------------------------*

NODE = 13

*------------------------------------------------------------*

(Count of FT home win:: PERIOD = _EVER <1.7)

AND (Mean Sportingbet draw odds:: PERIOD = _0W <3.1135)

AND MISSING(Min Betbrain average home win odds:: PERIOD = _39W76W) OR (Min Betbrain average home
win odds:: PERIOD = _39W76W <1.9607)

AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 1

   PREDICTED 1 = 1( 9/9)

   PREDICTED 0 = 0( 0/9)
```

```
*------------------------------------------------------------*

NODE = 14

*------------------------------------------------------------*

MISSING(Count of FT home win:: PERIOD = _EVER) OR (Count of FT home win:: PERIOD = _EVER >=1.7)

AND (Mean Sportingbet draw odds:: PERIOD = _0W <3.1135)

AND MISSING(Min Betbrain average home win odds:: PERIOD = _39W76W) OR (Min Betbrain average home
win odds:: PERIOD = _39W76W <1.9607)

AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

  PREDICTED VALUE IS 1

  PREDICTED 1 = 0.5577( 29/52)

  PREDICTED 0 = 0.4423( 23/52)

*------------------------------------------------------------*

NODE = 20

*------------------------------------------------------------*

(Mean Bet365 away win odds:: PERIOD = _0W >=2.4135)

AND (Mean Away Team Corners:: PERIOD = _38W >=5.36)

AND MISSING(Min Bet365 away win odds:: PERIOD = _20W38W) OR (Min Bet365 away win odds:: PERIOD =
_20W38W >=1.4564)

AND MISSING(Mean Sportingbet draw odds:: PERIOD = _0W) OR (Mean Sportingbet draw odds:: PERIOD =
_0W >=3.1135)

AND MISSING(Min Betbrain average home win odds:: PERIOD = _39W76W) OR (Min Betbrain average home
win odds:: PERIOD = _39W76W <1.9607)

AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

  PREDICTED VALUE IS 0

  PREDICTED 1 = 0.25( 2/8)

  PREDICTED 0 = 0.75( 6/8)

*------------------------------------------------------------*

NODE = 7

*------------------------------------------------------------*

MISSING(Mean Half Time Away Team Goals:: PERIOD = _EVER) OR (Mean Half Time Away Team Goals:: PERIOD
= _EVER <0.56)
```

```
AND (Away Team IS ONE OF ARSENAL, CRYSTAL PAL, EVERTON, LIVERPOOL, MAN UNITED, TOTTENHAM)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 0

   PREDICTED 1 = 0.3182( 21/66)

   PREDICTED 0 = 0.6818( 45/66)

*------------------------------------------------------------*

NODE = 17

*------------------------------------------------------------*

MISSING(Mean Away Team Corners:: PERIOD = _38W) OR (Mean Away Team Corners:: PERIOD = _38W <5.36)

AND MISSING(Min Bet365 away win odds:: PERIOD = _20W38W) OR (Min Bet365 away win odds:: PERIOD =
_20W38W >=1.4564)

AND MISSING(Mean Sportingbet draw odds:: PERIOD = _0W) OR (Mean Sportingbet draw odds:: PERIOD =
_0W >=3.1135)

AND MISSING(Min Betbrain average home win odds:: PERIOD = _39W76W) OR (Min Betbrain average home
win odds:: PERIOD = _39W76W <1.9607)

AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 0

   PREDICTED 1 = 0.4143( 29/70)

   PREDICTED 0 = 0.5857( 41/70)

*------------------------------------------------------------*

NODE = 5

*------------------------------------------------------------*

MISSING(Min Betbrain average draw win odds:: PERIOD = _24W) OR (Min Betbrain average draw win odds::
PERIOD = _24W <2.0077)

AND (Mean Bet365 away win odds:: PERIOD = _0W >=5.0139)

   PREDICTED VALUE IS 1

   PREDICTED 1 = 0.7215( 57/79)

   PREDICTED 0 = 0.2785( 22/79)

*------------------------------------------------------------*

NODE = 10

*------------------------------------------------------------*

(Min Betbrain average home win odds:: PERIOD = _39W76W >=1.9607)
```

```
AND MISSING(Away Team) OR (Away Team IS ONE OF ASTON VILLA, BOURNEMOUTH, BURNLEY, CARDIFF,
CHELSEA, FULHAM, HULL, LEICESTER, MAN CITY, NEWCASTLE, NORWICH, QPR, READING, SOUTHAMPTON,
STOKE, SUNDERLAND, SWANSEA, WATFORD, WEST BROM, WEST HAM, WIGAN)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 0

   PREDICTED 1 = 0( 0/7)

   PREDICTED 0 = 1( 7/7)

*------------------------------------------------------------*

NODE = 8

*------------------------------------------------------------*

(Mean Half Time Away Team Goals:: PERIOD = _EVER >=0.56)

AND (Away Team IS ONE OF ARSENAL, CRYSTAL PAL, EVERTON, LIVERPOOL, MAN UNITED, TOTTENHAM)

AND MISSING(Mean Bet365 away win odds:: PERIOD = _0W) OR (Mean Bet365 away win odds:: PERIOD = _0W
<5.0139)

   PREDICTED VALUE IS 0

   PREDICTED 1 = 0.02439( 1/41)

   PREDICTED 0 = 0.9756( 40/41)
```