GROUP PROJECT
ON EUROPEAN SOCCER

Predictive Analytics
for Data DrivenDecisions Making

Group Members:

Davide De Benedetto – 3013595

Grégoire Eugene – 3127015

SUMMARY

PREDICTIVE MODELS

Introduction

- The objective of the group project is to create different models in order to correctly classify the results of the matches in the main European soccer Leauges
- The European Soccer Database (ESD) available on Kaggle has been used; it cointains data from 2008 to 2016
- The analysis will be divided in 4 sections:
  - Data Preparation
  - Missing Values and Outliers Treatment
  - Variable Creation and Trasformation
  - Predictvie Models
- 5 different predictive models have been implemented
- Eventually, it will be compared not only the performance between the aforementioned models, but it will also be checked if the new variables created are able to increase the model accuracy in correctly predicting the results

European Soccer Database

+25k matches

+10k players

11 European Leagues

Players and Team attributes

Betting Odds

Detailed match events

Introduction

- Being the interest of the project to assess to what extent the evaulated ability of a team is a predictor of the match outcome, only the “Team Attributes” (TA) and the “Match” (M) categories have been selected from the ESD
- The TA category contains 25 variables; the attributes displayed evaulate the team performance both assigning the value of a specific skill and clustering them into either 2 or 3 classes
- The M category contains 115 variables (a sample of 25 is displayed on the right) that refer to:
  - League, season, and date
  - Opposing teams and Specific match features
  - Players information and coordinates
  - Bets evaluations
- There is a chronological mismatch in the two databases

TEAM

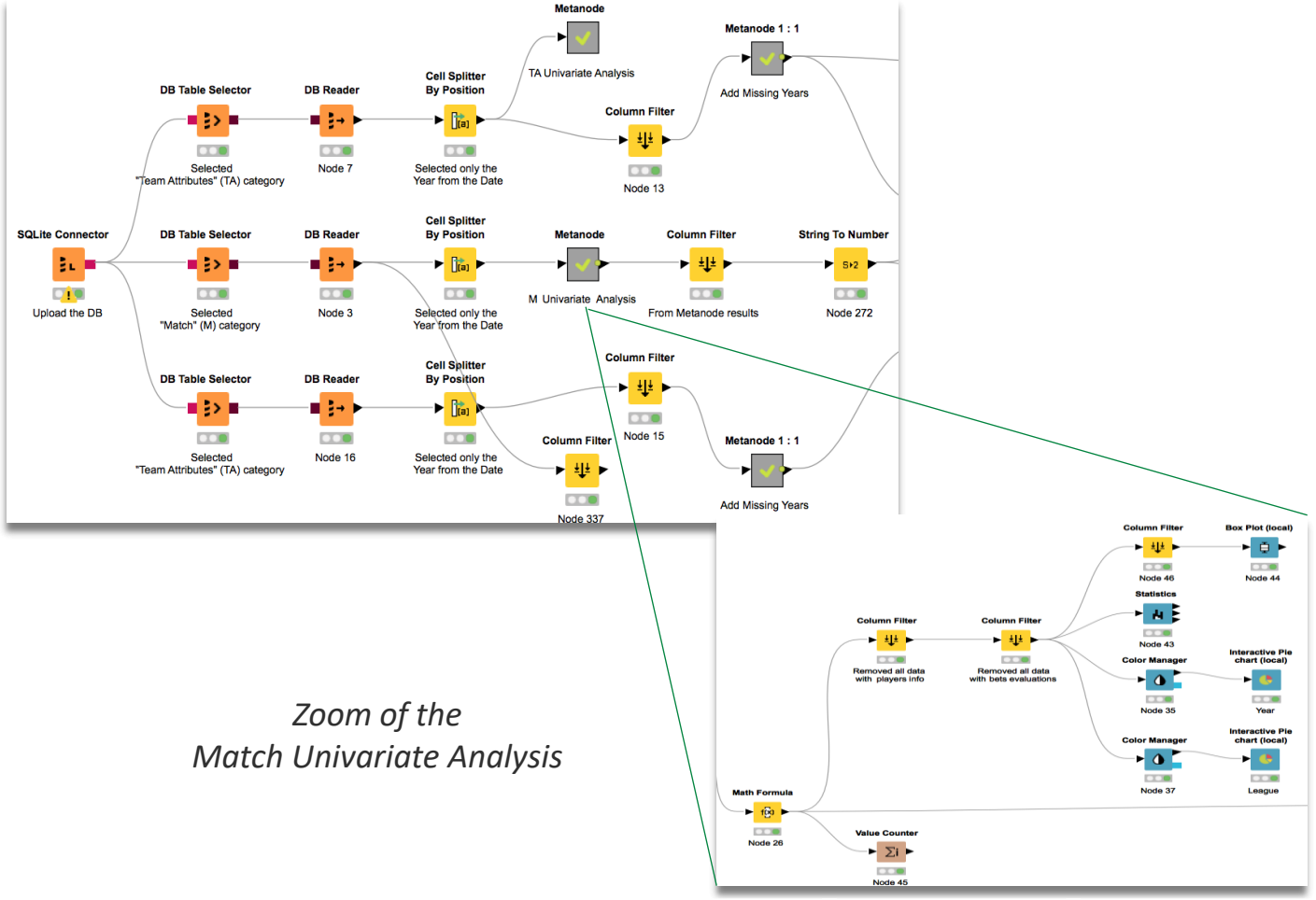
Columns: 25	Column Type	...	...	...	...	Lower Bo...	Upper Bo...	Value 0	Value 1	Value 2	Value 3	Value 4	Value 5
team_fifa_api_id	Number (integer)	1				1	112513	?	?	?	?	?	?
team_api_id	Number (integer)	2				1601	274581	?	?	?	?	?	?
id	Number (integer)	0				1	1458	?	?	?	?	?	?
defenceTeamWidthClass	String	...				?	?	Normal	Wide	Narrow	?	?	?
defenceTeamWidth	Number (integer)	...				29	73	?	?	?	?	?	?
defencePressureClass	String	...				?	?	Medium	Deep	High	?	?	?
defencePressure	Number (integer)	...				23	72	?	?	?	?	?	?
defenceDefenderLineClass	String	...				?	?	Cover	Offside Trap	?	?	?	?
defenceAggressionClass	String	...				?	?	Press	Double	Contain	?	?	?
defenceAggression	Number (integer)	...				24	72	?	?	?	?	?	?
date	String	3				?	?	2010-02-22...	2014-09-19...	2015-09-10...	2011-02-22...	2012-02-22...	2013-09-20 ...
chanceCreationShootingClass	String	...				?	?	Normal	Lots	Little	?	?	?
chanceCreationShooting	Number (integer)	...				22	80	?	?	?	?	?	?
chanceCreationPositioningClass	String	...				?	?	Organised	Free Form	?	?	?	?
chanceCreationPassingClass	String	...				?	?	Normal	Risky	Safe	?	?	?
chanceCreationPassing	Number (integer)	...				21	80	?	?	?	?	?	?
chanceCreationCrossingClass	String	...				?	?	Normal	Lots	Little	?	?	?
chanceCreationCrossing	Number (integer)	...				20	80	?	?	?	?	?	?
buildUpPlaySpeedClass	String	5				?	?	Balanced	Fast	Slow	?	?	?
buildUpPlaySpeed	Number (integer)	4				20	80	?	?	?	?	?	?
buildUpPlayPositioningClass	String	...				?	?	Organised	Free Form	?	?	?	?
buildUpPlayPassingClass	String	9				?	?	Mixed	Long	Short	?	?	?
buildUpPlayPassing	Number (integer)	8				20	80	?	?	?	?	?	?
buildUpPlayDribblingClass	String	7				?	?	Little	Normal	Lots	?	?	?
buildUpPlayDribbling	Number (integer)	6				24	77	?	?	?	?	?	?

MATCH

id	Number (integer)	0	1	25,979	?	?	?	?	?	?	?	?
country_id	Number (integer)	1	1	24,558	?	?	?	?	?	?	?	?
league_id	Number (integer)	2	1	24,558	?	?	?	?	?	?	?	?
season	String	3	?	?	2008/2009	2009/2010	2010/2011	2011/2012	2012/2013	2013/2014	2014/2015	2015/2016
stage	Number (integer)	4	1	38	?	?	?	?	?	?	?	?
date	String	5	?	?	?	?	?	?	?	?	?	?
match_api_id	Number (integer)	6	483,129	2,216,672	?	?	?	?	?	?	?	?
home_team_api_id	Number (integer)	7	1,601	274,581	?	?	?	?	?	?	?	?
away_team_api_id	Number (integer)	8	1,601	274,581	?	?	?	?	?	?	?	?
home_team_goal	Number (integer)	9	0	10	?	?	?	?	?	?	?	?
away_team_goal	Number (integer)	...	0	9	?	?	?	?	?	?	?	?
home_player_X1	Number (integer)	...	0	2	?	?	?	?	?	?	?	?
away_player_X1	Number (integer)	...	1	6	?	?	?	?	?	?	?	?
home_player_Y1	Number (integer)	...	0	3	?	?	?	?	?	?	?	?
away_player_Y1	Number (integer)	...	1	3	?	?	?	?	?	?	?	?
home_player_1	Number (integer)	...	2,984	698,273	?	?	?	?	?	?	?	?
away_player_1	Number (integer)	...	2,796	698,273	?	?	?	?	?	?	?	?
goal	String	...	?	?	?	?	?	?	?	?	?	?
shoton	String	...	?	?	?	?	?	?	?	?	?	?
shotoff	String	...	?	?	?	?	?	?	?	?	?	?
foulcommit	String	...	?	?	?	?	?	?	?	?	?	?
card	String	...	?	?	?	?	?	?	?	?	?	?
cross	String	...	?	?	?	?	?	?	?	?	?	?
corner	String	...	?	?	?	?	?	?	?	?	?	?
possession	String	...	?	?	?	?	?	?	?	?	?	?
B365H	Number (double)	...	1.04	26	?	?	?	?	?	?	?	?

Data Preparation & Analysis

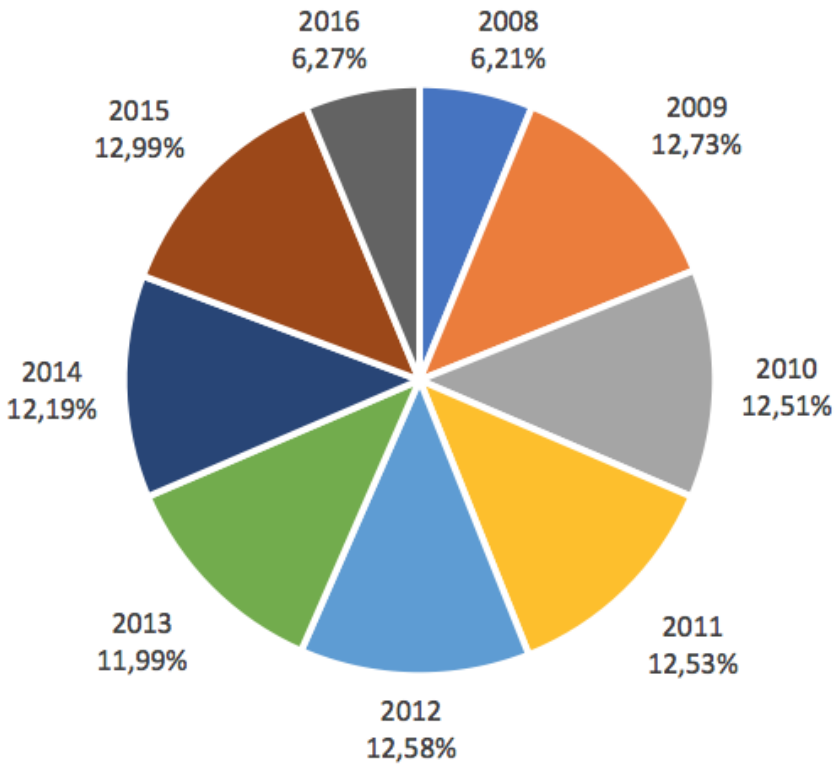
- Univariate analysis has been run for both TA and M:
  - ❖ buildUpPlayDribbling has been excluded from TA, because there were 969 missing values in 1458 observations
  - ❖ Bets evaluations and players information have been removed from M, as they were of no interest for the analysis
  - ❖ All features collected during the match (e.g. yellow cards and fouls) have not been considered, because they cannot be used to make predictions
- The pie charts in the M Univariate Analysis show that the matches are almost equally distributed among the 11 leagues and among the 9 years considered
- It appears that the Super League and 2008 are respectively the least represented league and year



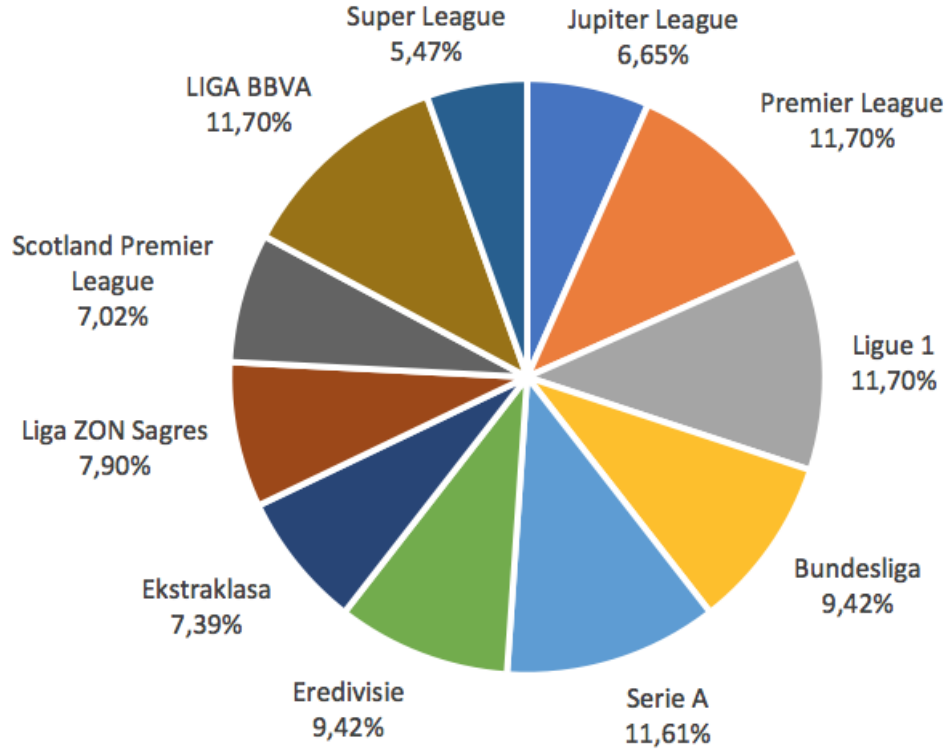
Zoom of the Match Univariate Analysis

Data Preparation & Analysis

PIE CHART – MATCHES PER YEAR

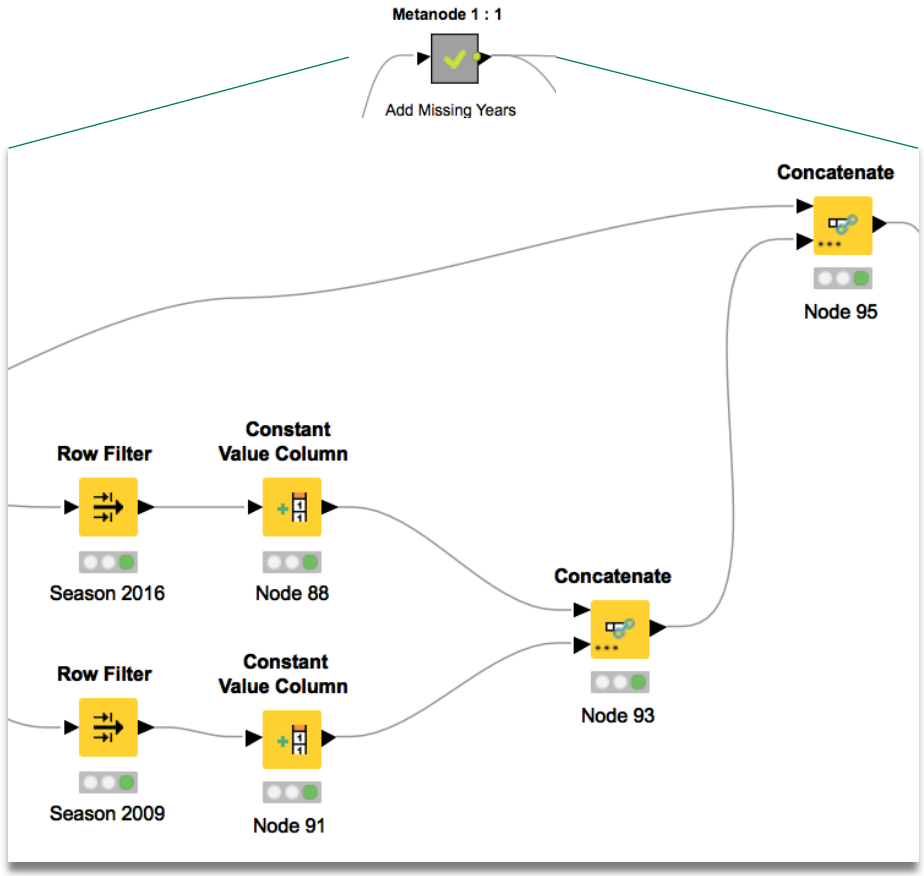


PIE CHART – MATCHES PER LEAGUE



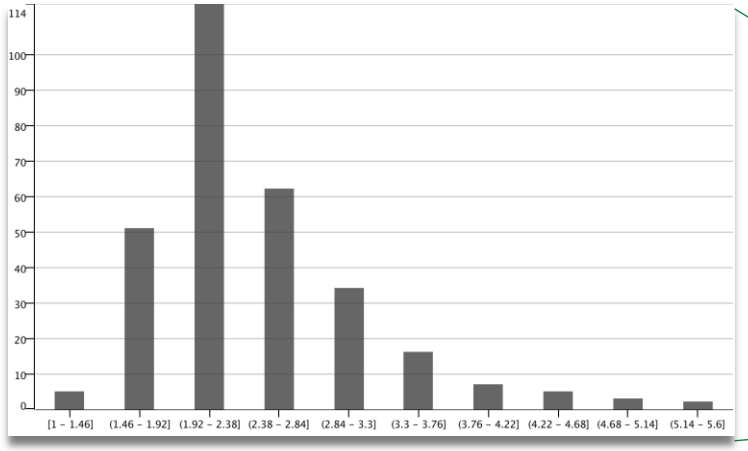
Data Preparation & Analysis

- In the first part of the workflow, there are 2 metanodes called *Add Missing Years;* they have the same internal architecture
- As highlighted in slide 4, there is a chronological mismatch between TA and M categories:
  - M variables are collected between 2008 and 2016
  - TA variables are collected between 2010 and 2015
- The aim of the metanodes is to control the mismatch
- All TA referred to 2015 have been replicated and assigned to 2016
- All TA referred to 2010 have been replicated and assigned to 2009
- Missing TA for 2008 have not been replaced with data from the closest year, because a 2 year time gap has not been deemed sufficiently shorty for the substitution

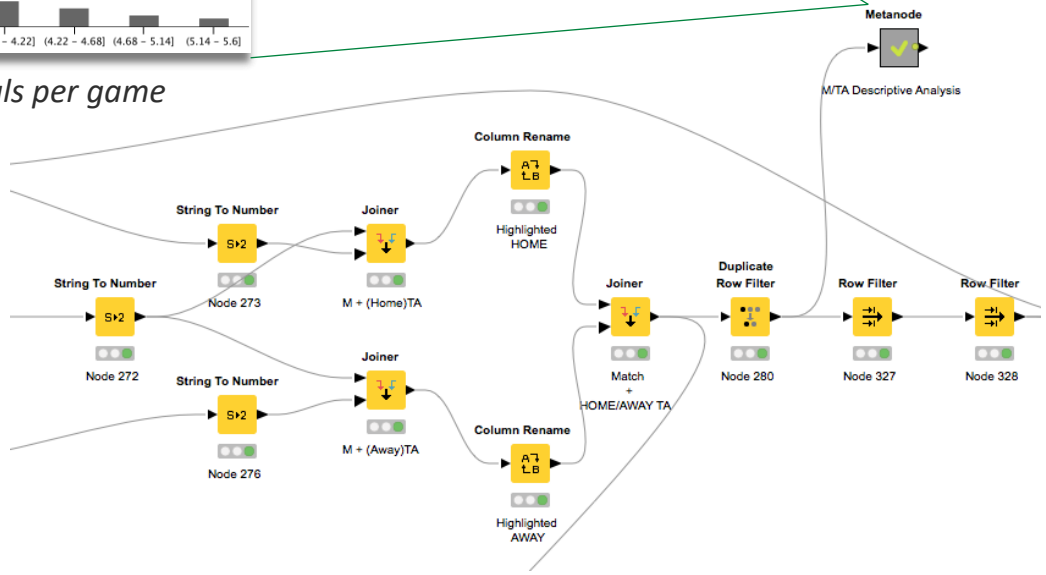


Data Preparation & Analysis

- A new variable has been created to assess the result of the game:
  - Target:
    Home Team Goals - Away Team Goals
- Two main tables have been created:
  - The first displays M selected variables along Home team TA
  - The second displays M selected variables along Away team TA
- The two tables have been merged and it has been run a descriptive analysis to display general features of the new table
- Finally, data from 2008 have been deleted as there were no TA available for those years
- Radar plots refer only to the Home team and have been built starting from the mean of each TA between 2009 and 2015

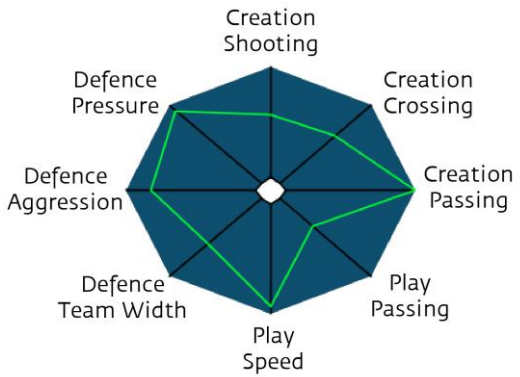


Histogram – Average team goals per game

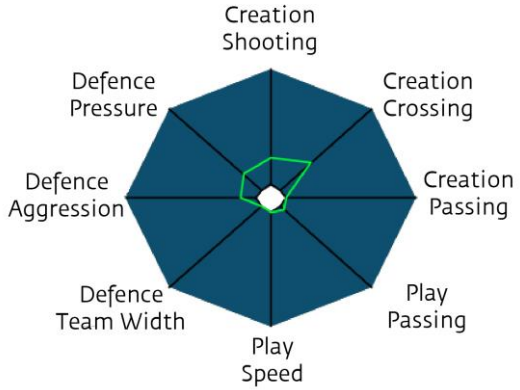


Data Preparation & Analysis

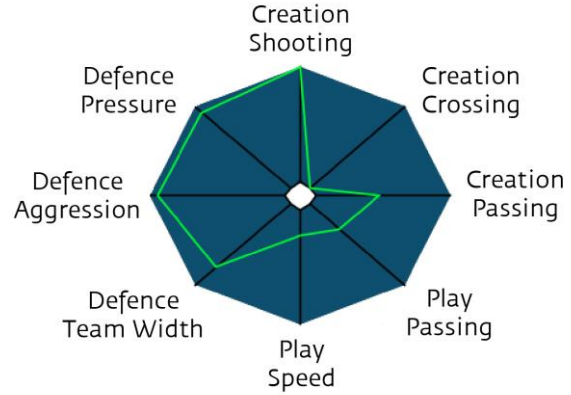
BUNDESLIGA



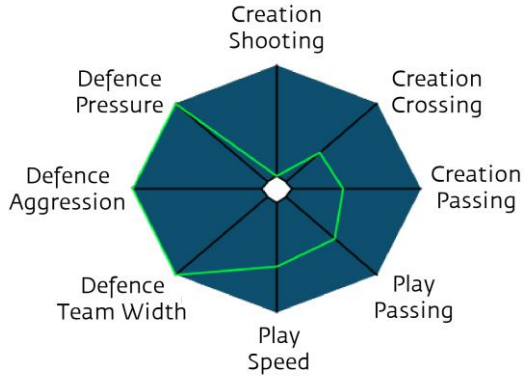
EREDIVISE



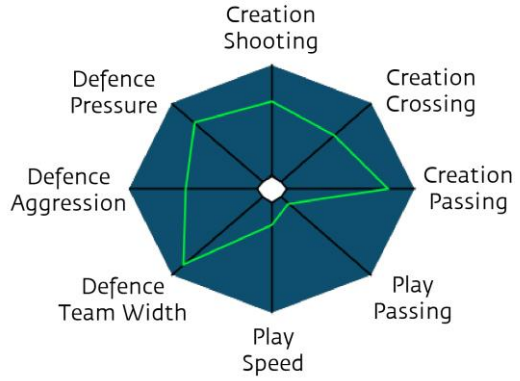
ESTRAKLASA



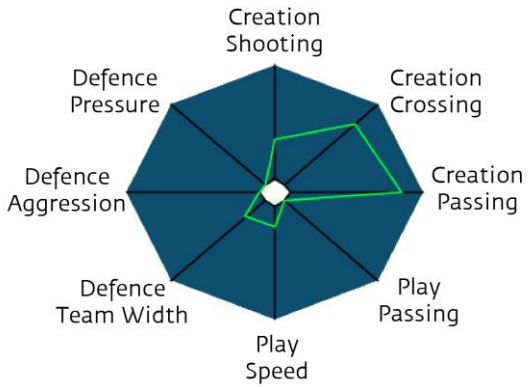
JUPITER LEAGUE



LIGA BBVA

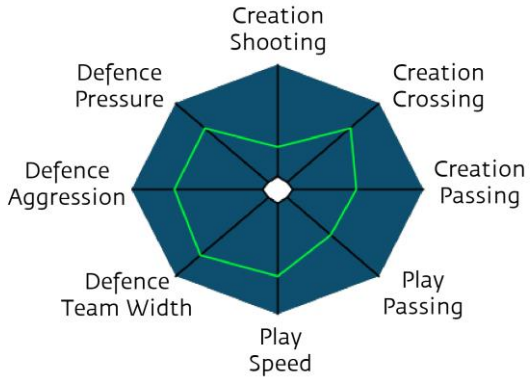


LIGA ZON SAGRES

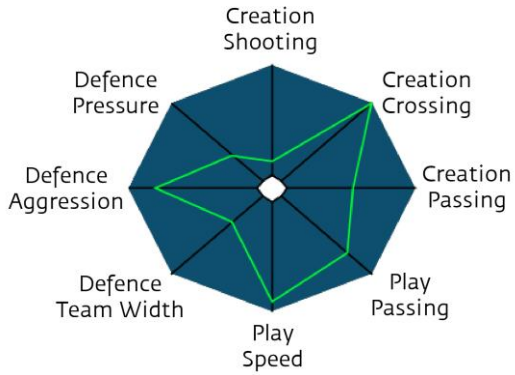


Data Preparation & Analysis

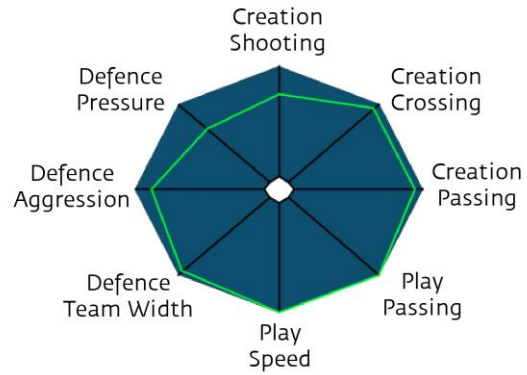
LIGUE 1



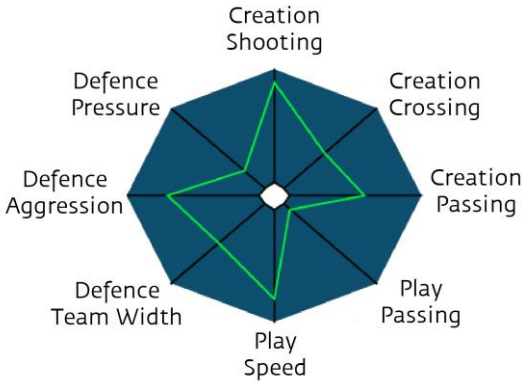
PREMIER LEAGUE



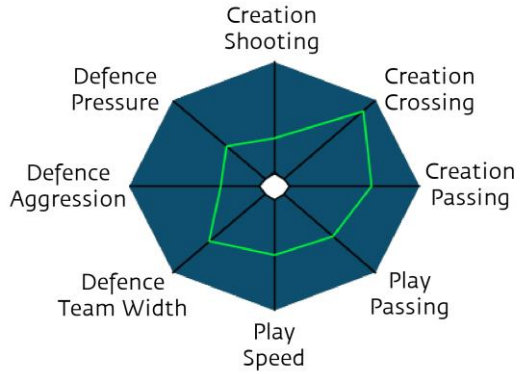
SCOTLAND PREMIER LEAGUE



SERIE A



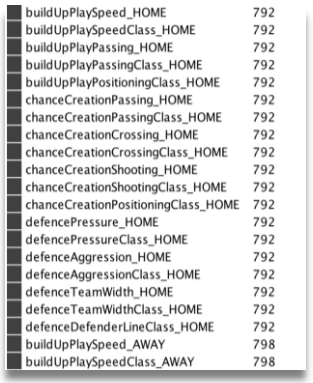
SUPER LEAGUE



Missing Values and Outliers

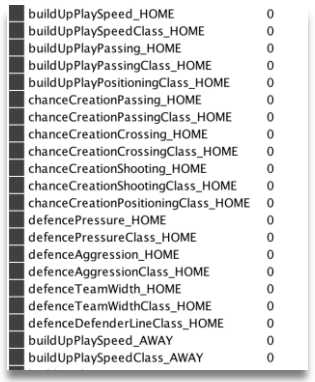
- The following treatment has been applied for missing values:
  - Numerical missing values have been replaced with the mean of the attribute of every year
  - Categorical missing values have been replaced with the mode of the attribute of every year
- As displayed in the box plot available in the M Univariate Analysis, there are some outliers in the categories *Home Team Goals, Away Team Goals,* and *Target*
- The following values have been considered outliers:
  - *Home team goals* >= 8
  - *Away Team Goals* >= 8
  - |*Target* | >= 7
- Since the number of the outliers was not significant (0.18% of all observations), they have been eliminated

MISSING VALUES

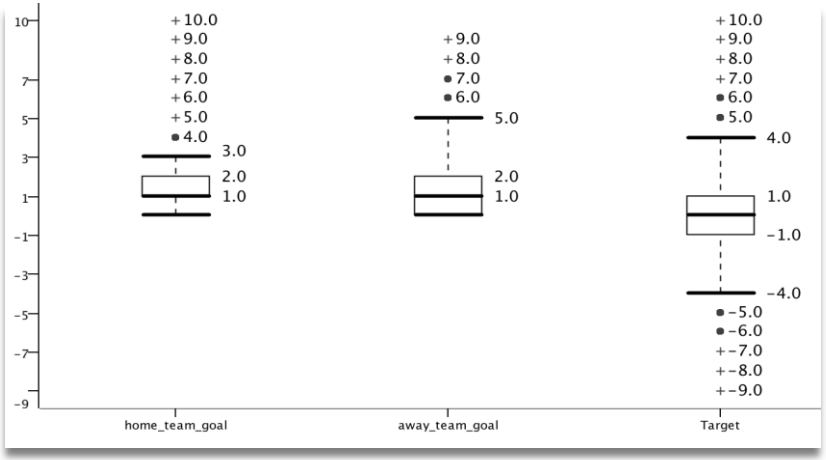


Treatment

MISSING VALUES

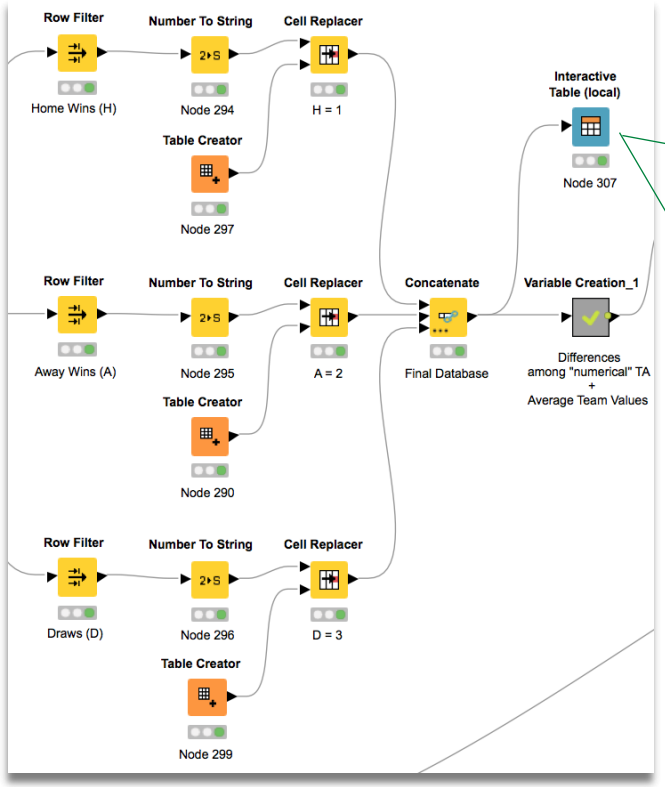


OUTLIERS



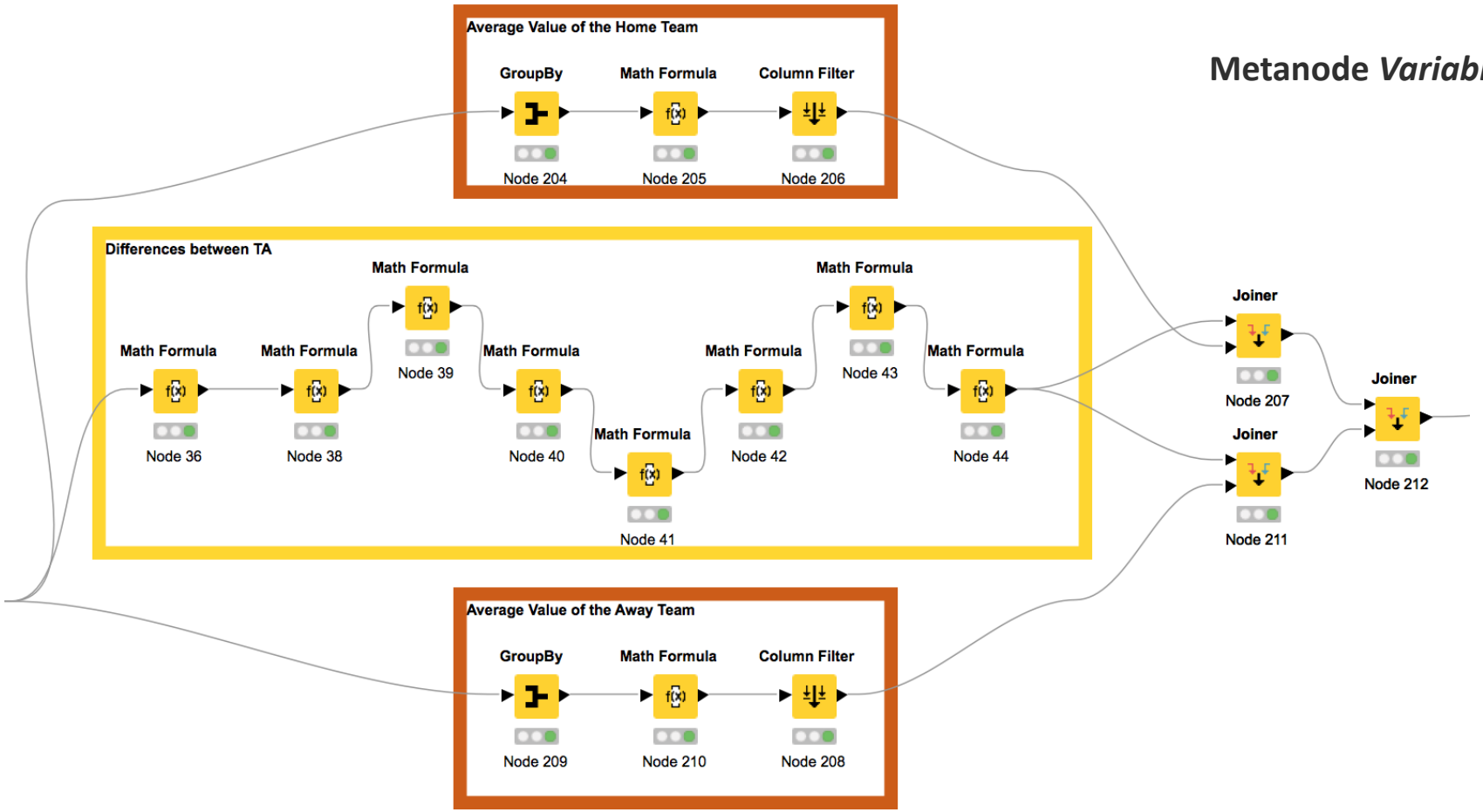
New Varibales & Transformation

- First of all, we have recoded the variable created in the ***Data Preparation and Analysis*** section, namely the *Target:*
  - If the Home team wins, it has been assigned 1
  - If the Away team wins, it has been assigned 2
  - If there is a draw, it has been assigned 3
- In the metanode *Variable Creation_1,* 10 new variables have been created:
  - 2 variables that show the Average Value of a team in a given year (Average Value Home and Average Value Away)
  - 8 new variables computed as the difference between the Home team TA and the corresponding Away team TA (TeamAttribute_DIFF)
- The architecture of the metanode is displayed in the following slide



Row ID	TrueD...	Target	buildU...	buildU...
Row9996_...	2015	1	71	Fast
Row9995_...	2015	1	45	Balanced
Row9994_...	2015	1	55	Balanced
Row9993_...	2015	1	47	Balanced
Row9992_...	2015	3	59	Balanced
Row9991_...	2015	2	51	Balanced
Row9990_...	2015	2	59	Balanced
Row998_R...	2012	1	50	Balanced
Row9989_...	2015	1	42	Balanced
Row9988_...	2015	1	68	Fast
Row9987_...	2015	1	31	Slow
Row9986_...	2015	2	58	Balanced
Row9985_...	2015	3	59	Balanced
Row9984_...	2015	1	60	Balanced
Row9983_...	2015	3	69	Fast
Row9982_...	2015	2	48	Balanced
Row9981_...	2015	2	59	Balanced
Row9980_...	2015	3	31	Slow
Row997_R...	2012	2	55	Balanced
Row9979_...	2015	2	73	Fast
Row9978_...	2015	1	71	Fast
Row9977_...	2015	1	45	Balanced
Row9976_...	2015	2	47	Balanced

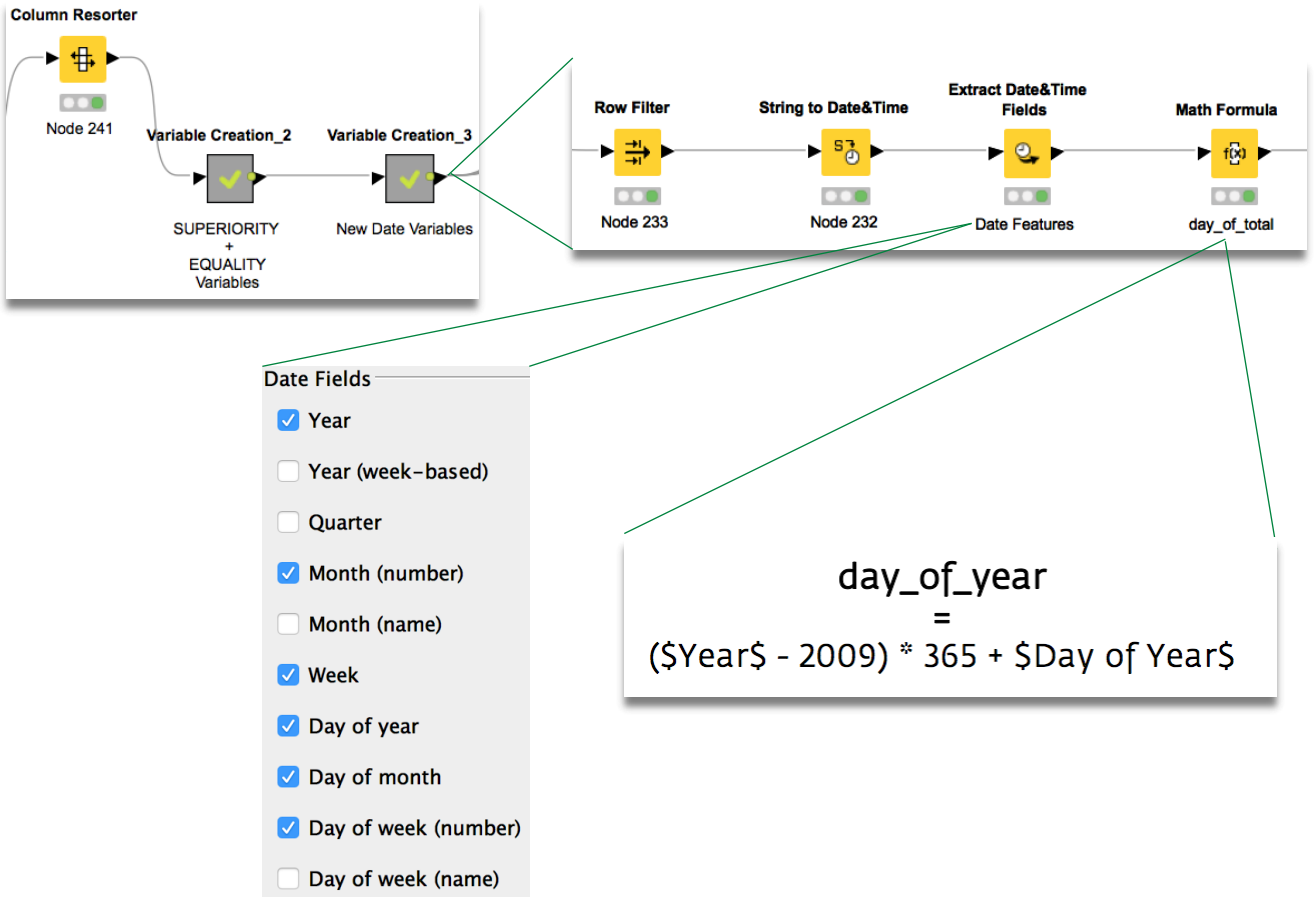
New Varibales & Transformation



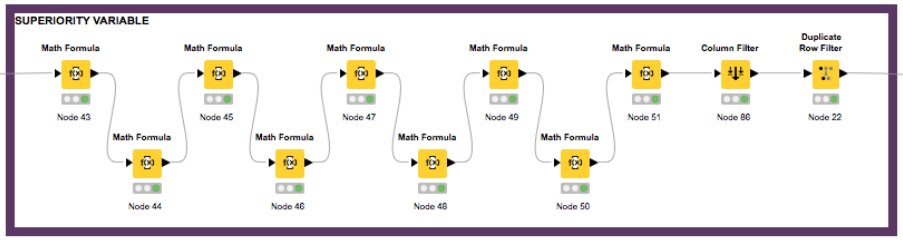
Metanode *Variable Creation_1*

New Varibales & Transformation

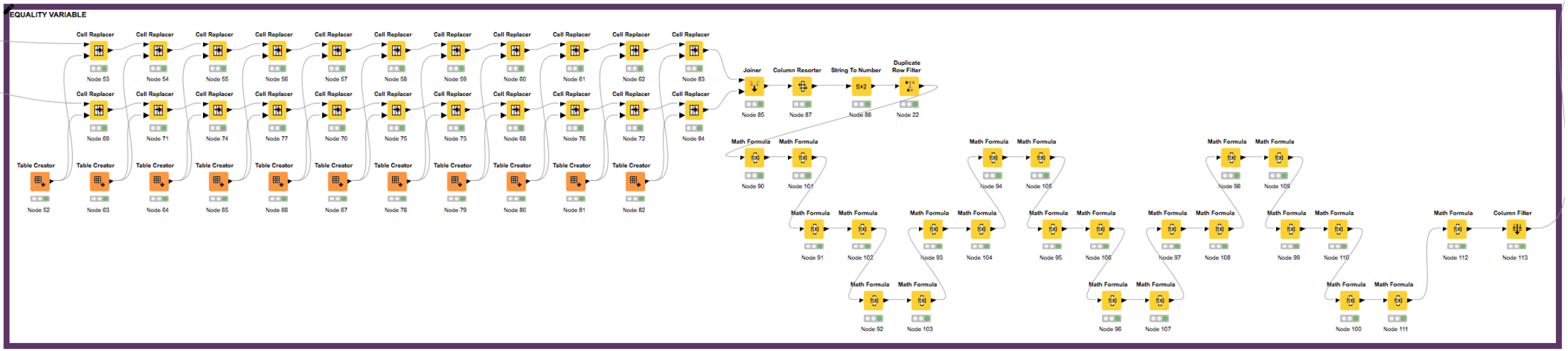
- In the metanode *Variable Creation_2,* 2 new variables have been created
  - *Superiority:* it counts how many TeamAttribute_DIFF (the variables created in the previous metanode) are positive, or rather in how many instances the Home team was better than the Away team for each skill
  - *Equality:* it counts in how many instances the two teams had the same TA when it was expressed as a class (it has been necessary to substitute all the different classes with numbers)
- The architecture of the metanode *Variable Creation_2* is displayed in the following slide
- In the metanode *Variable Creation_3,* 6 new variables have been created
  - The 6 variables displayed in the zoom "Data fields"
  - *day_of_total:* it counts the day in which the match has been played starting from January 1st 2009



New Varibales & Transformation



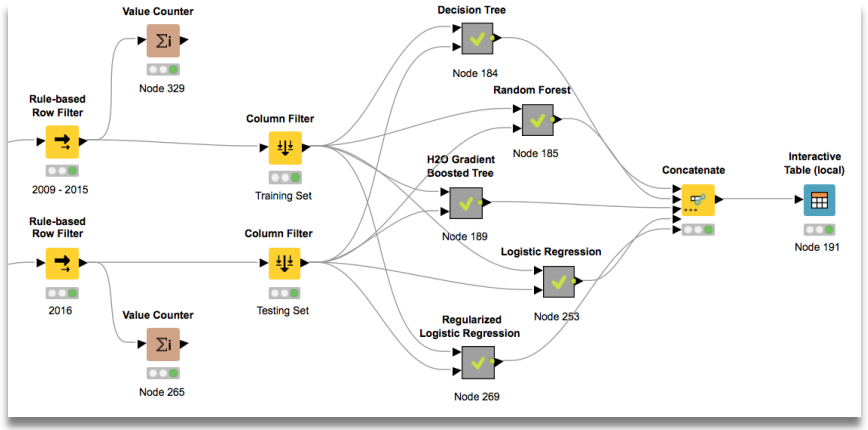
Metanode *Variable Creation_3*



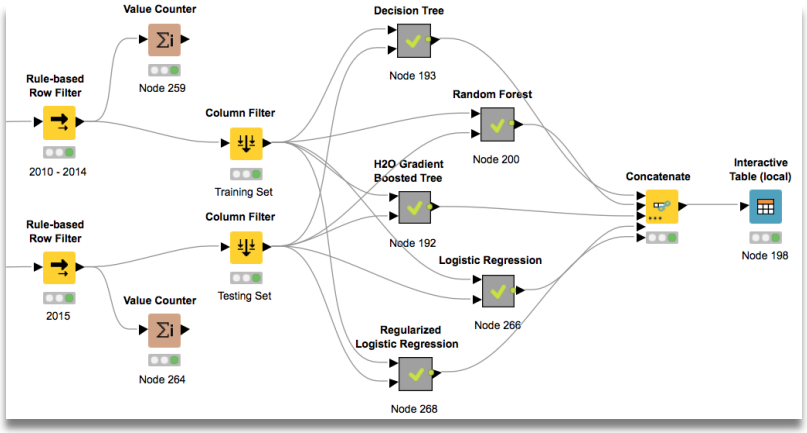
Predictive Models

- The fourth and last section of this chapter is divided in two parts:
  - ***Baseline Model:*** the databse from which the models will be built does not include the transformation described in the previous sections
  - ***Predictive model:*** the databse from which the models will be built includes all the treatments described in the previous slide
- For each version of the predictive models, 6 different models will be created
  - *Decision Tree*
  - *Random Forest*
  - *H2O Gradient Boosteed Tree*
  - *Logistic Regression*
  - *Regularized Logistic Regression – Gauss*
  - *Regularized Logistic Regression - Laplace*

BASELINE MODEL

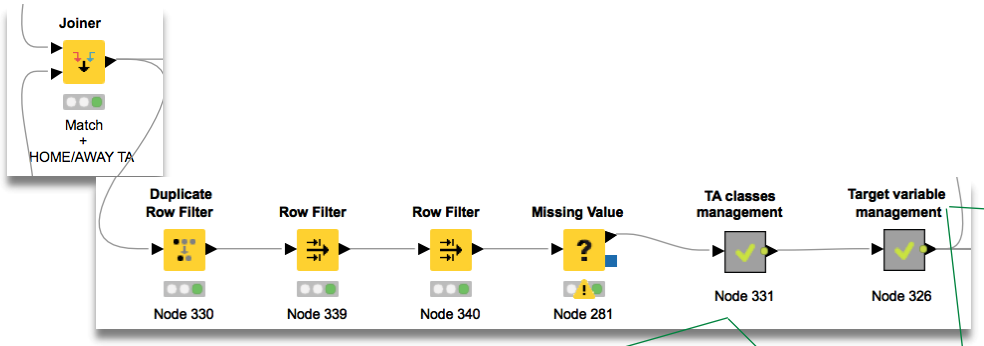


PREDICTIVE MODEL

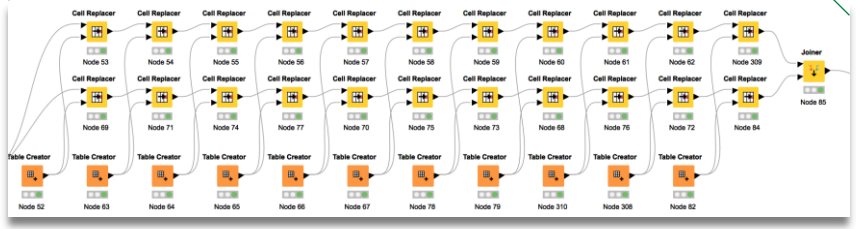


Predictive Models

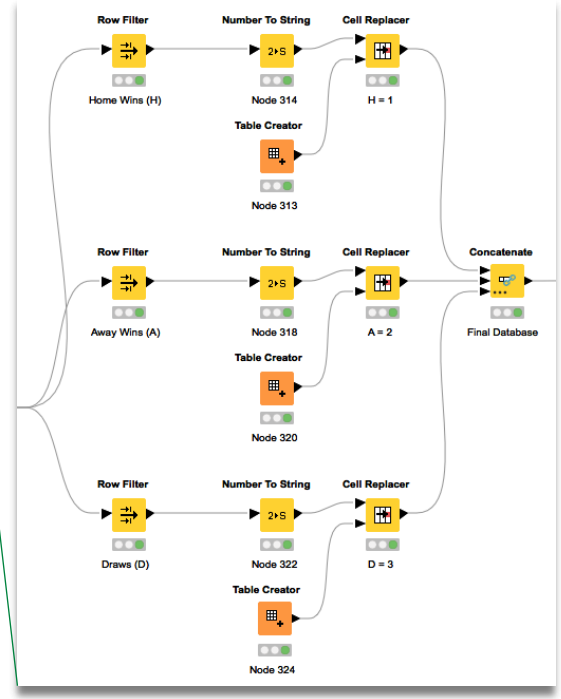
- The ***Baseline Model*** will be first analyzed
- The database is built starting from the Joiner node *Match + HOME/AWAY TA* presented in the section Data Preparation & Analysis
- Starting from this node, all data from 2008, 2009, and 2016 have been filtered out (they were not available in the ESD) and all missing values have been deleted
- Furthermore, all TA expressed in classes have been transformed into numbers, and the target variable described in slide 13 has been replicated
- The objective of this analysis is the creation of a baseline model in order to evaluate the accuaracy gains obtained in the ***Predictive Models***



Zoom of the TA classes management

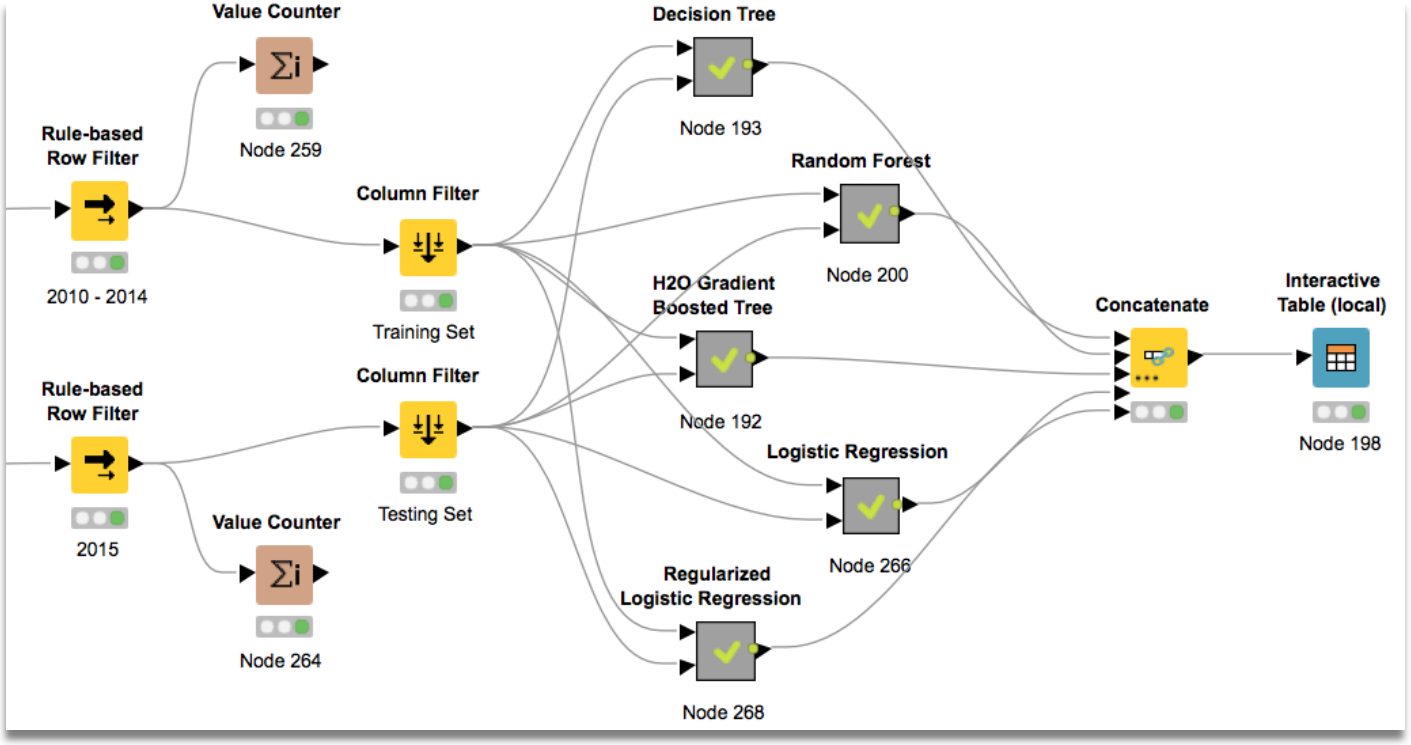


Zoom of the Target variable management



Predictive Models

- The database contains 18845 observations and 41 variables
- The purpose of the analysis is to create models that can predict the results of the matches played in the last available year (2015)
- Therefore, the databse has been divided into:
  - *Training Set:* it cointains all the matches played between 2010 and 2014 (15594 observations)
  - *Testing Set:* it contains all the matches played in 2015 (it contains 3251 observations)
- Only a few variables have been filtered out because they were of no relevance for the prediction (e.g. Home team goals)



**Each model has been assigned the same optimized parameters employed in Predictive Model*

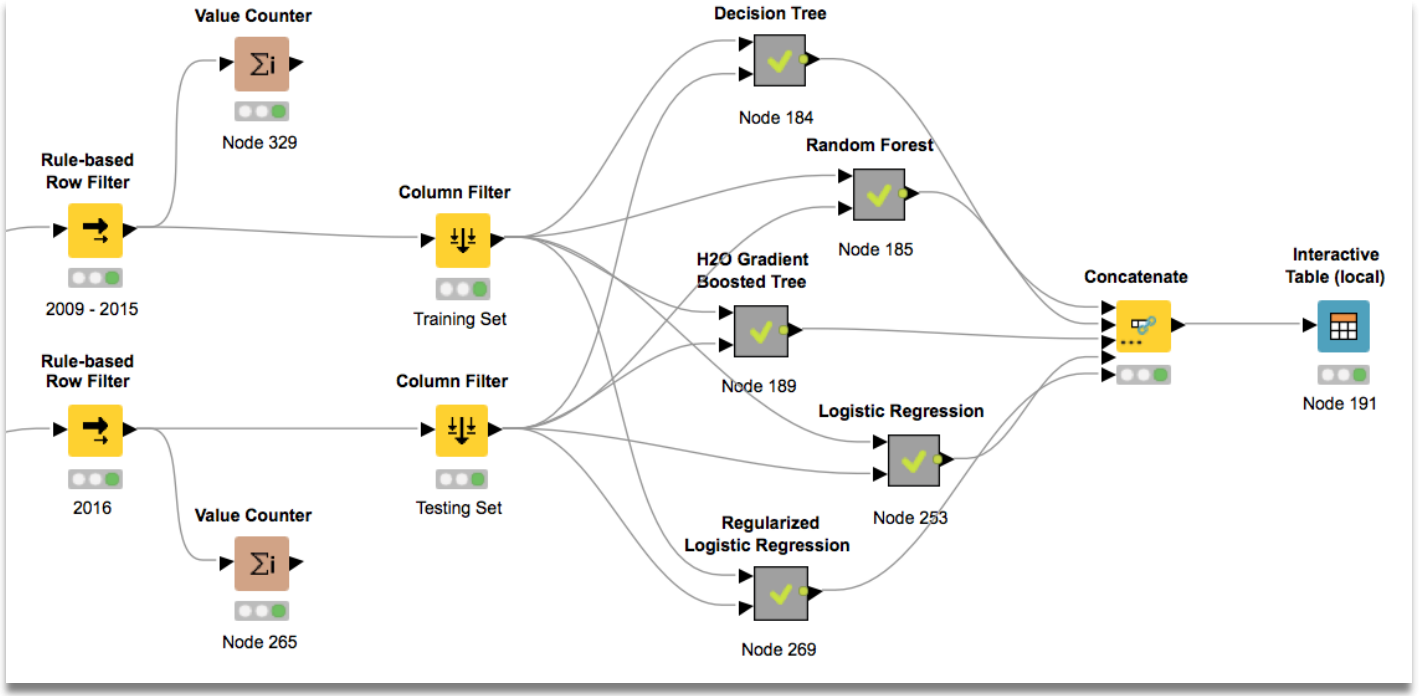
Predictive Models

- Overall, the *H2O Gradient Boosted Tree* is the model that reaches the highest level of accuracy
- The *Regularized Logistic Regression – Laplace* is the most accurate model in predicting when the home team wins
- The *Decision Tree* is the most accurate model in predicting when the away team wins and when there will be a draw
- None of the *Logistic Regression* models is able to predict when there will be a draw
- buildUpPlaySpeed_AWAY, buildUpPlayPassing_AWAY, chanceCreationShooting_HOME, and buildUpPlayPassing_HOME are among the top 10 predictors for both the *Random Forest* and the *H2O Gradient Boosted Tree*

Model	Target	Sensitivity	Specificity	Accuracy
Decision Tree	1	0,6579	0,4473	
Decision Tree	2	0,3146	0,7722	
Decision Tree	3	0,1540	0,8554	
Decision Tree	Overall	-	-	42,57%
Random Forest	1	0,8196	0,2943	
Random Forest	2	0,2844	0,8254	
Random Forest	3	0,0474	0,9672	
Random Forest	Overall	-	-	46,11%
H2O Gradient Boosted Tree	1	0,8701	0,2489	
H2O Gradient Boosted Tree	2	0,2667	0,8651	
H2O Gradient Boosted Tree	3	0,0201	0,9771	
H2O Gradient Boosted Tree	Overall	-	-	47,12%
Logistic Regression	1	0,8984	0,1829	
Logistic Regression	2	0,1958	0,8743	
Logistic Regression	3	0,0000	0,9996	
Logistic Regression	Overall	-	-	45,77%
Regularized Logistic Regression - Laplace	1	0,9136	0,1513	
Regularized Logistic Regression - Laplace	2	0,1656	0,8957	
Regularized Logistic Regression - Laplace	3	0,0000	1,0000	
Regularized Logistic Regression - Laplace	Overall	-	-	45,56%

Predictive Models

- Let's now turn to the ***Predictive Models***
- The database contains 22786 observations and 37 variables (less than the ones in baseline model)
- The purpose of the analysis is to create models that can predict the results of the matches played in the last available year (2016)
- Therefore, the databse has been divided into:
  - *Training Set:* it cointains all the matches played between 2019 and 2015 (21329 observations)
  - *Testing Set:* it contains all the matches played in 2015 (it contains 1457 observations)
- In accordance with the general principle of parsimony, some predictors have been filtered out because they were not of significative relevance (according to *H2O Gradient Boosted Tree)*



**The parameters have been optimized for each model applying optimization loop methods*

Predictive Models

- The *H2O Gradient Boosted Tree* is the model that reaches the highest level of accuracy
- The *Regularized Logistic Regression – Laplace* is the most accurate model in predicting when the home team wins
- The *Random Forest* and the *H2O Gradient Boosted Tree* are the most accurate models in predicting when the away team wins
- The *Random Forest* is the most accurate model in predicting when there will be a draw
- None of the *Logistic Regression* models is able to predict when there will be a draw
- chanceCreationPositioningClass_HOME, chanceCreationPositioningClass_AWAY, and playPassing_DIFF are among the top 10 predictors for both the *Random Forest* and the *H2O Gradient Boosted Tree*

Model	Target	Sensitivity	Specificity	Accuracy
Decision Tree	1	0,9062	0,1735	
Decision Tree	2	0,1977	0,9036	
Decision Tree	3	0,0136	0,9899	
Decision Tree	Overall	-	-	46,74%
Random Forest	1	0,7708	0,4188	
Random Forest	2	0,3227	0,8387	
Random Forest	3	0,1308	0,8780	
Random Forest	Overall	-	-	47,43%
H2O Gradient Boosted Tree	1	0,9123	0,2714	
H2O Gradient Boosted Tree	2	0,3227	0,8800	
H2O Gradient Boosted Tree	3	0,0136	0,9936	
H2O Gradient Boosted Tree	Overall	-	-	50,79%
Logistic Regression	1	0,8969	0,2069	
Logistic Regression	2	0,2682	0,8859	
Logistic Regression	3	0,0000	1,0000	
Logistic Regression	Overall	-	-	48,11%
Regularized Logistic Regression - Laplace	1	0,9246	0,1698	
Regularized Logistic Regression - Laplace	2	0,2159	0,9105	
Regularized Logistic Regression - Laplace	3	0,0000	1,0000	
Regularized Logistic Regression - Laplace	Overall	-	-	47,77%

Predictive Models

DIFFERENCES BETWEEN *PREDICTIVE MODEL* AND *BASELINE MODEL*

Model	Target	Sensitivity	Specificity	Accuracy
Decision Tree	1	0,2482	-0,2739	
Decision Tree	2	-0,1169	0,1315	
Decision Tree	3	-0,1404	0,1345	
Decision Tree	Overall			0,0417
Random Forest	1	-0,0489	0,1245	
Random Forest	2	0,0384	0,0133	
Random Forest	3	0,0834	-0,0892	
Random Forest	Overall			0,0132
H2O Gradient Boosted Tree	1	0,0422	0,0225	
H2O Gradient Boosted Tree	2	0,0561	0,0149	
H2O Gradient Boosted Tree	3	-0,0065	0,0164	
H2O Gradient Boosted Tree	Overall			0,0367
Logistic Regression	1	-0,0015	0,0240	
Logistic Regression	2	0,0723	0,0116	
Logistic Regression	3	0,0000	0,0004	
Logistic Regression	Overall			0,0234
Regularized Logistic Regression - Laplace	1	0,0110	0,0184	
Regularized Logistic Regression - Laplace	2	0,0503	0,0148	
Regularized Logistic Regression - Laplace	3	0,0000	0,0000	
Regularized Logistic Regression - Laplace	Overall			0,0221

Conclusion

PREDICTIVE MODELS

- Data manipulation, missing values and outliers treatment, as well as the creation of new variables (e.g. Equality) resulted in an improvement of the accuracy of each model:
  - the biggest improvment is registered by the Decision *Tree* (4.17 percentage points)
  - the worst improvement is registered by the *Random Forest* (1.32 percentage points)
- Overall, ***Prediction Models*** is less reliable in predicting true negatives but it is more accurate in predicting true positives
- Both ***Baseline Model*** and ***Predictive Model*** are not reliable for the forecast of draws; the most accurate model for predicting draws is the *Decision Tree* of the ***Baseline Model*** with a sensitivity of 15,40%

TIME SERIES ANALYSIS

Data Analysis

- The aim of the analysis is to predict the total attendance of a given league per gameday employing an «automatic» approach
- The European Football Games (EFG) database available on Kaggle has been used; it contains data from 2004 to 2019
- It is a similar dataset to the one used in the previous chapter, as it contains less data related to leagues but more accurate attendance data
- The target is obtained summing the attendance of each games for a given gameday
- We will try to predict the target for the last season of the dataset. The analysis will be carried via three main classes of models:
  - Exponential Smoothing
  - ARIMA
  - ARIMAX

European Football Games

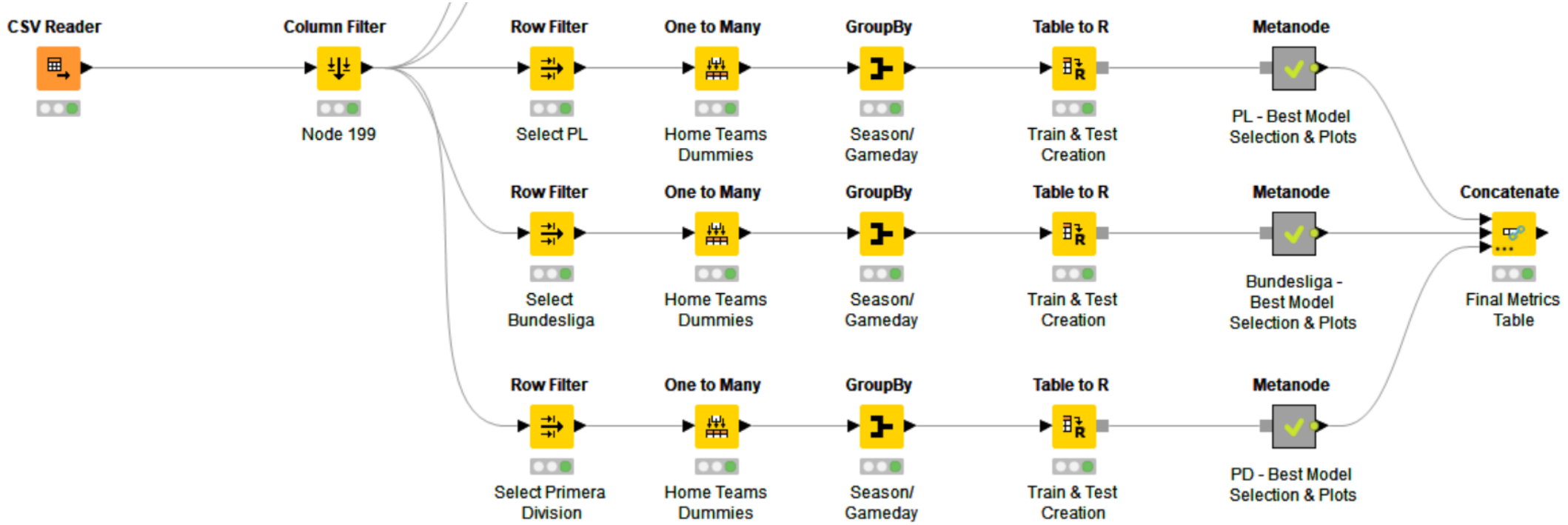
+25k matches

+100 teams

5 European Leagues

Data Analysis

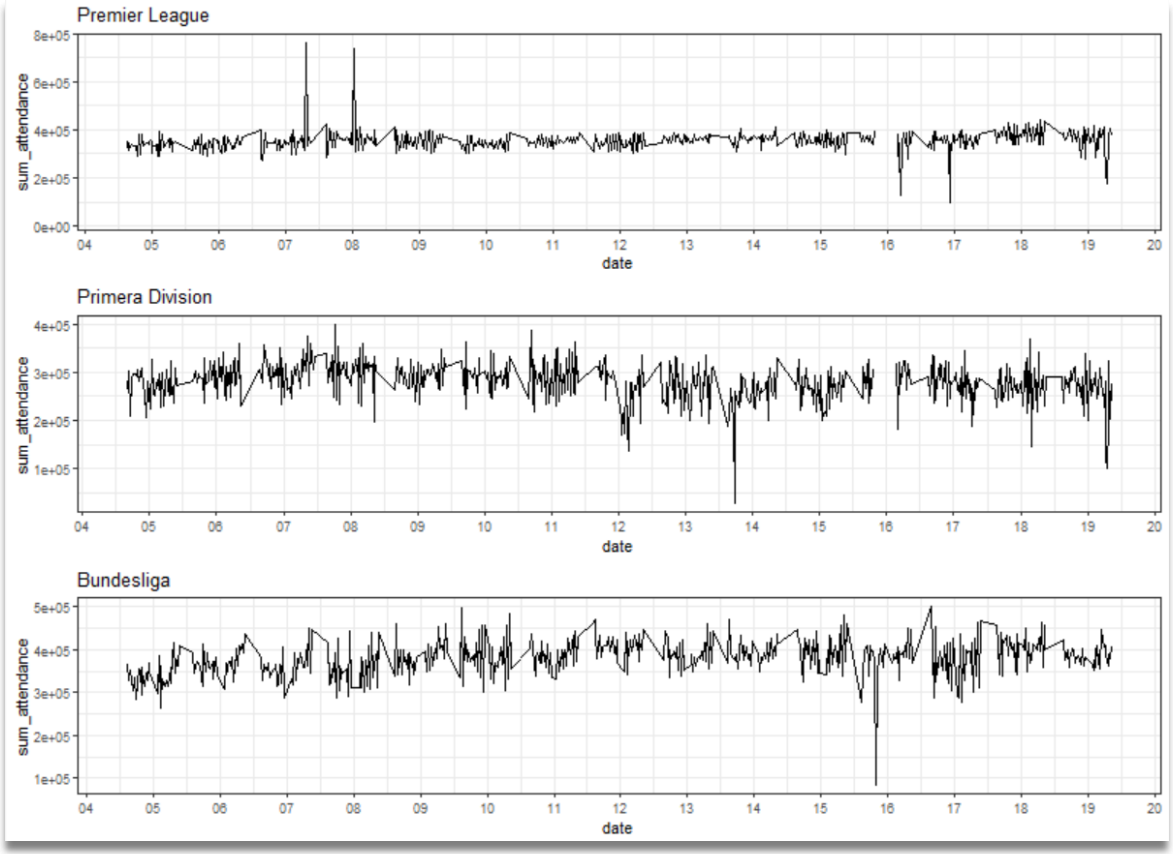
Data Preparation



Data Analysis

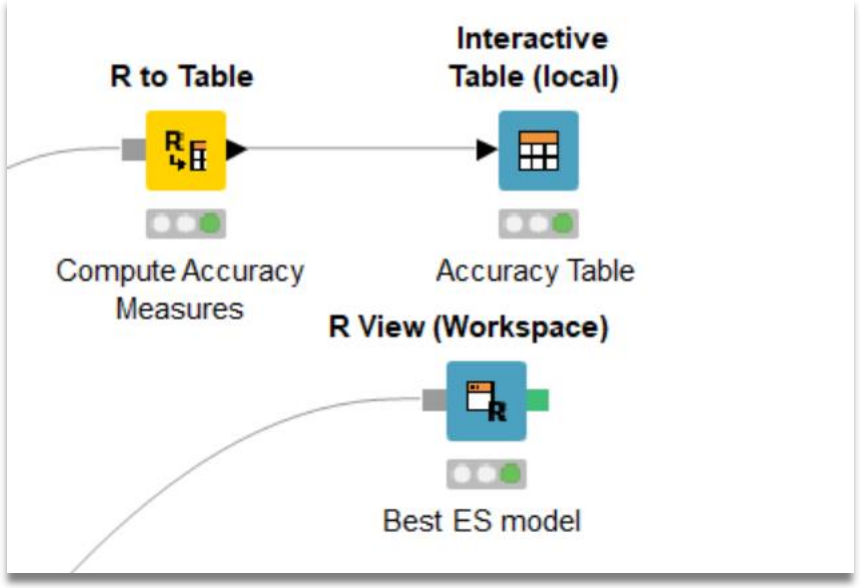
- The analysis has considered only the leagues for which we have the most non-NA values:
  - Premier League
  - Primera Division (LIGA BBVA)
  - Bundesliga
- The time plots of the different target variables are displayed for each considered league; as evident from the graphs, the target has a high variance in all 3 cases
- The missing data of 2015 for the Premier League and the Primera Division have been replaced with the R function *tsclean()*
- From these plots, it appears to be no clear trends and seasonalities; only for the Bundesliga, the average attendance per season slightly rises

TOTAL ATTENDANCE PER GAMEDAY FOR EUROPEAN LEAGUES



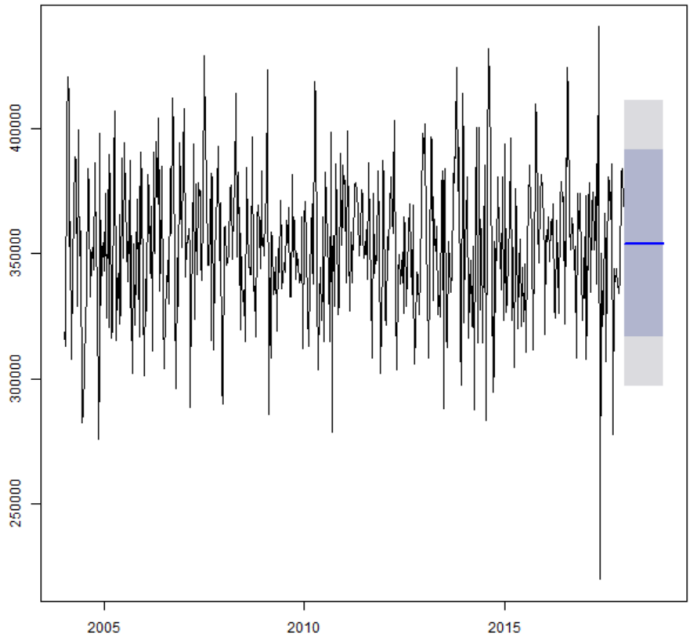
Predictions & Results

- In these 2 slides, forecast analyses will be based on ***Exponential Smoothing*** models
- The frequency of the games is equal to the number of gamedays per season; it can be either 34 or 38
- As the series have a frequency higher than 12, Holt-Winters models cannot be run in R and an alternative equivalent model has been used
- In particular, the following models will be considered:
  - Holt
  - Tbats (equivalent to Holt-Winters for data with high frequency)
  - Auto ETS
- In the following series, the best time series predictions for the 3 leagues will be displayed; however, as evident from the graphs, they are not reliable



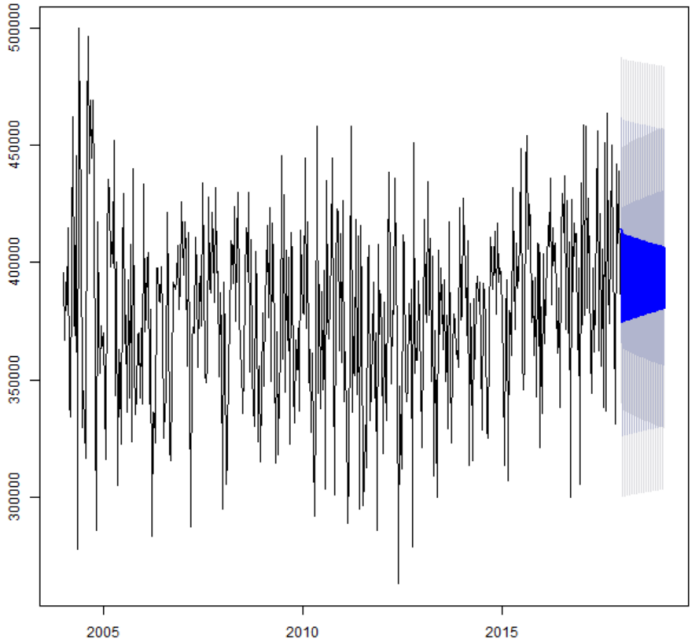
Prediction & Results

PREMIER LEAGUE



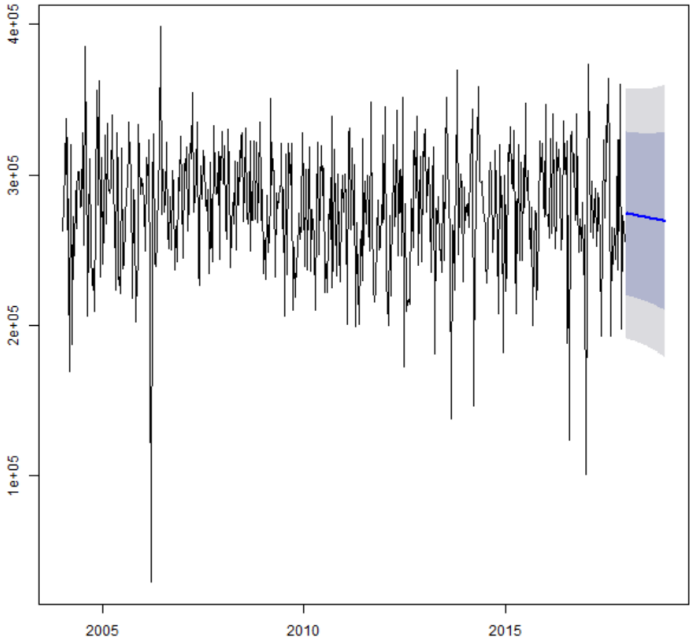
Forecasts from TBATS Model

BUNDESLIGA



Forecasts from TBATS Model

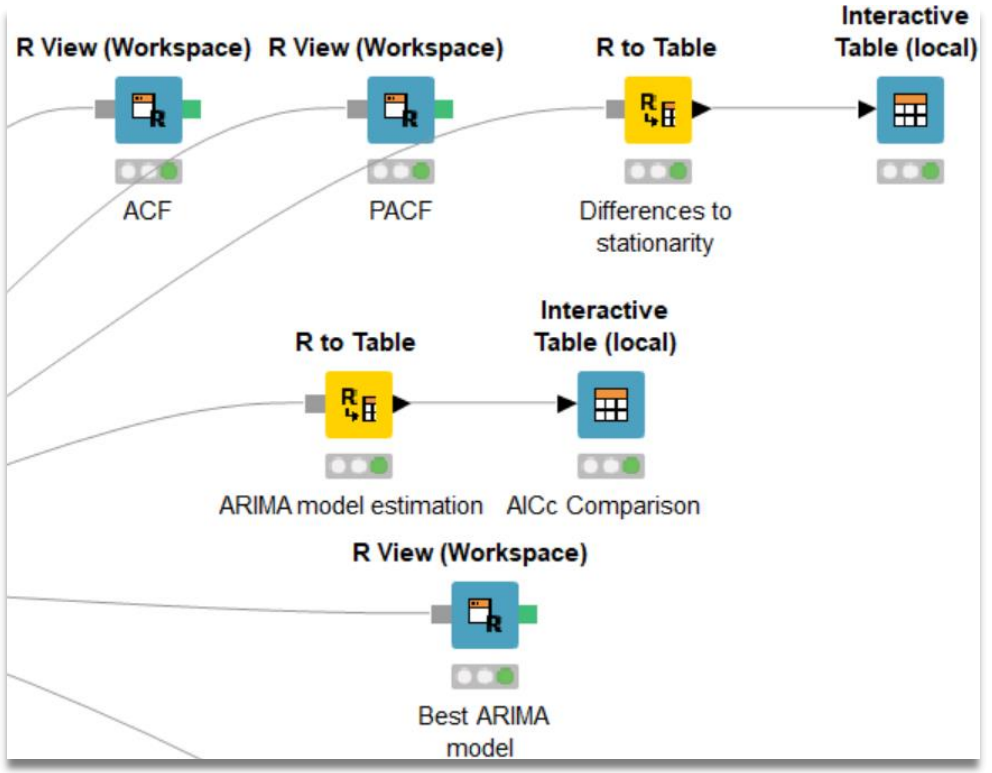
PRIMERA DIVISION



Forecasts from HOLT model

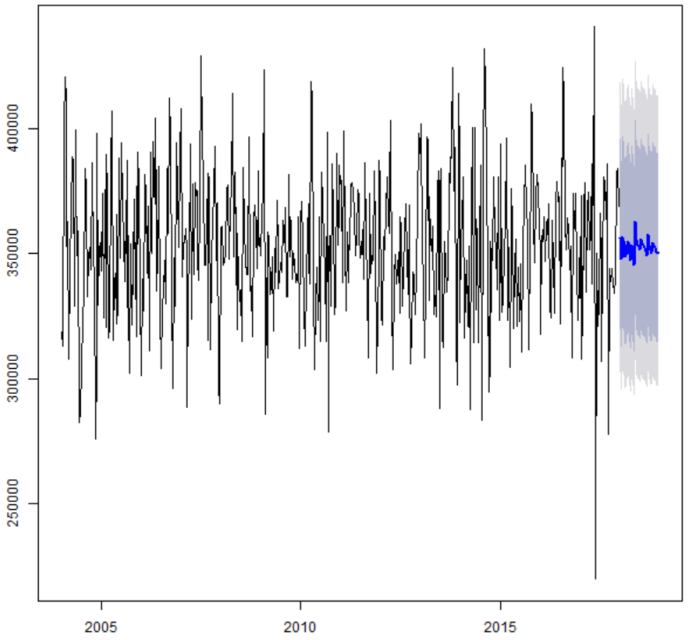
Prediction & Results

- In the following 2 slides, forecast analyses will be based on ***ARIMA*** models
- As the time plots presented in slide ___ suggest, all three series are not stationary and, considering the many irregularities, it is hard to stationarize them
- Different combinations of differencing have been tried; however, neither the ACF nor the PACF plots became similar to those observed for stationary series
- In all three cases, the auto ARIMA model coefficients are the most efficient
- However, most of the predictions are of mediocre quality, and in Spain the model is not even able to identify some seasonal variations
- The uncertainty intervals are still too large, therefore the models are still unreliable



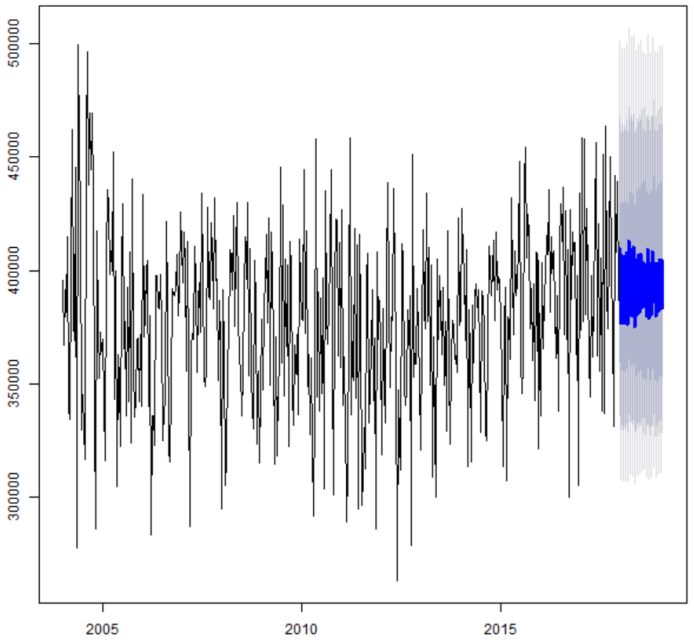
Predictions & Results

PREMIER LEAGUE



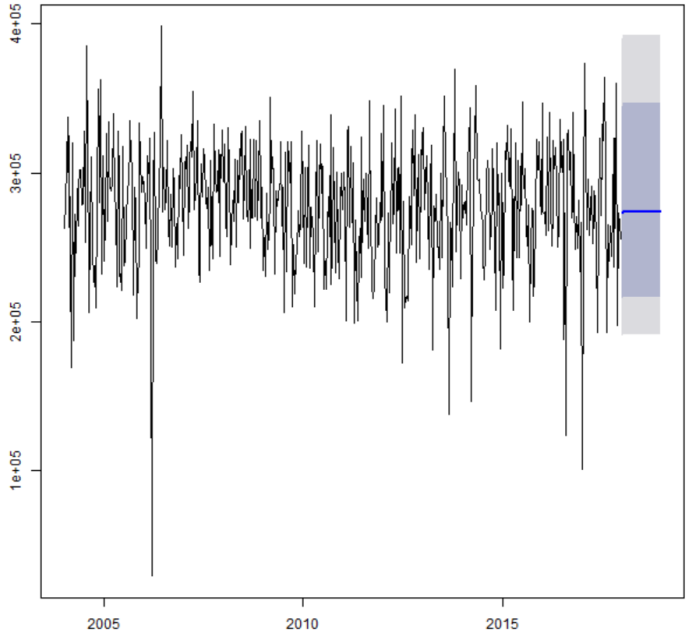
Forecasts from ARIMA (2, 0, 2)(1, 0, 1)[38]

BUNDESLIGA



Forecasts from ARIMA (3, 1, 3)(1, 0, 0)[34]

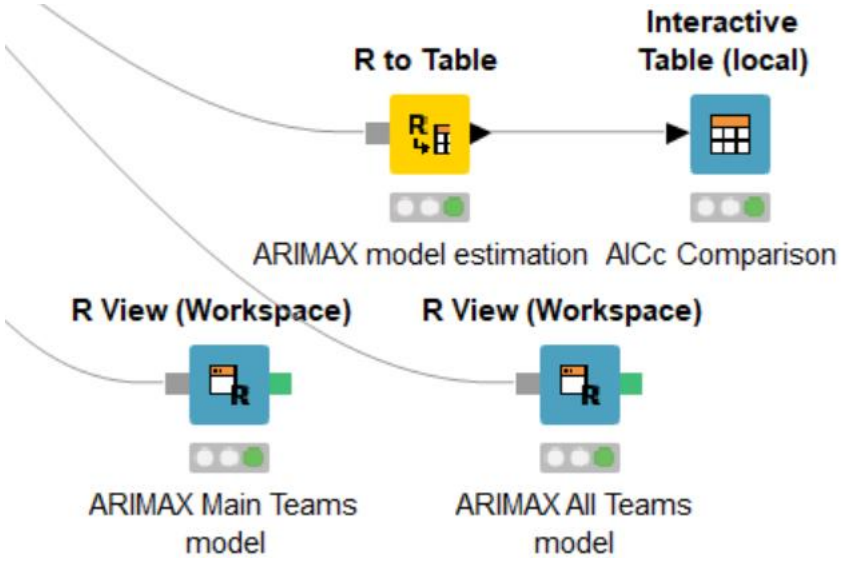
PRIMERA DIVISION



Forecasts from ARIMA (0, 0, 1)

Predictions & Results

- The unsatisfaying results suggest that most of the variation comes from elsewhere
- Therefore, it has been decided to do explanatory time series analyses to investigate the dataset; the intuition is that the team playing home at each gameday is an important predictor of the attendance at the stadium
- Two ***ARIMAX*** models have been built:
  - *Main Teams*: it only takes into account when the main teams of each championship played home
  - *All Teams*: it takes into account all teams of each League
- For England and Spain, the *All Teams ARIMAX* model is actually a Linear Regression over all the dummies
- The *All Teams* models had the best results and had narrower prediction intervals (less uncertainty of the predictions)

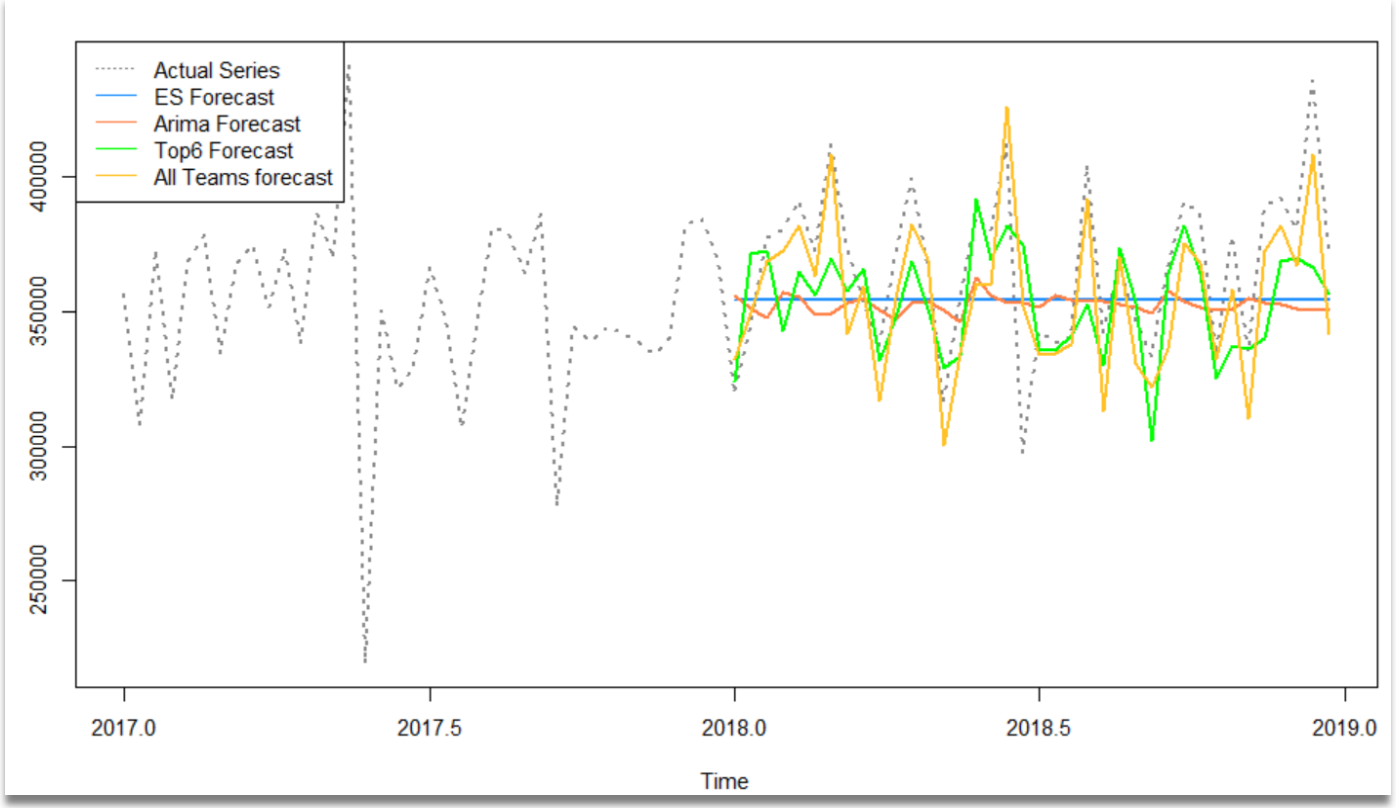


Predictions & Results

- Let's start by analysing the *Premier League*
- The two ***ARIMAX*** models are by the far the most accurate and reliable
- A large part of the variance remains still unexplained
- The *All Teams* model is a linear regression; it confirms that time-related effects are really hard to catch

Metrics	RMSE	MAPE	(p,d,q)(P,D,Q)
ES	31775	7.16	-
ARIMA	32472	7.25	-
ARIMAX - MT	27717	5.63	(3,0,3)(1,0,2)
ARIMAX - AT	19019	4.36	(0,0,0)(0,0,0)

PREMIER LEAGUE – MODEL COMPARISONS

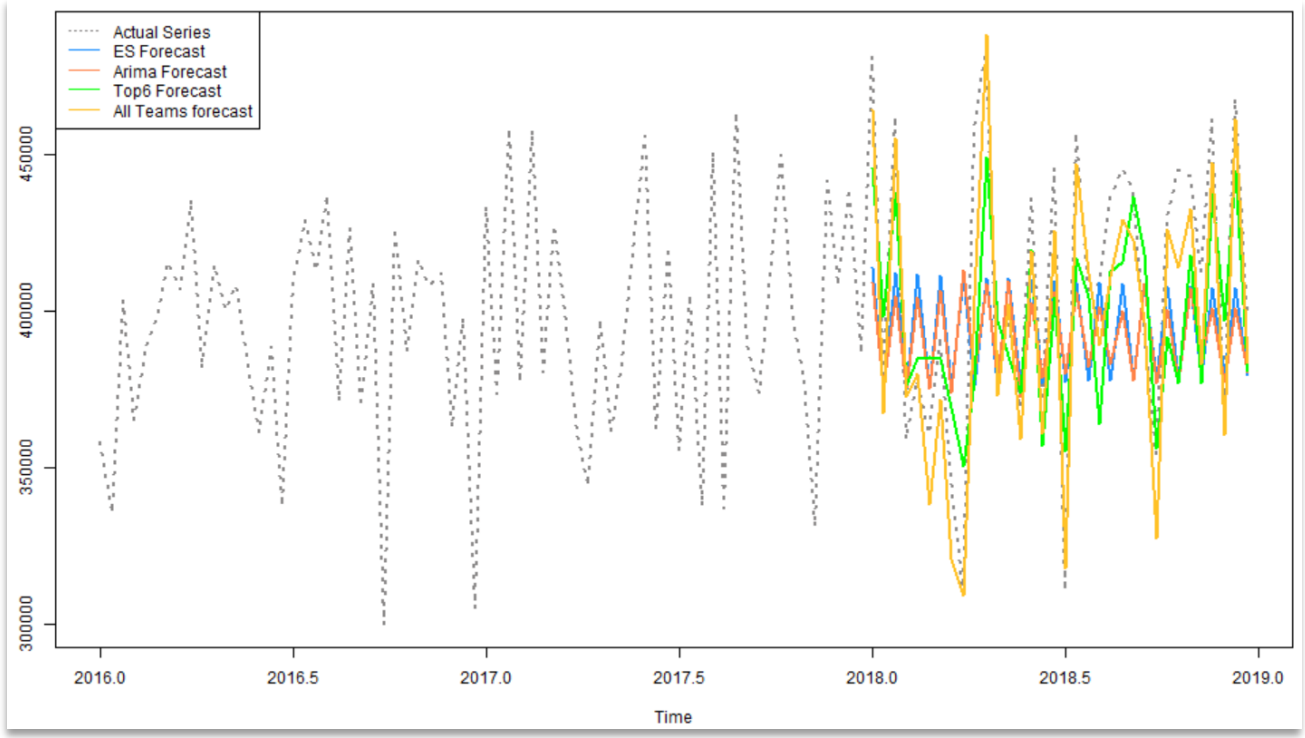


Predictions & Results

- In the *Bundesliga,* although the ***Exponential Smoothing*** and the ***ARIMA*** are able to catch some seasonality, the overall effect is often misleading: the results we get from these models are worse than for Premier League
- However, the presence of time-related effects improves a lot the ***ARIMAX*** models: Bundesliga is the league for which the predictions register the lowest errors

Metrics	RMSE	MAPE	(p,d,q)(P,D,Q)
ES	43964	8.73	-
ARIMA	45253	8.93	-
ARIMAX - MT	31607	6.47	(0,1,2)(1,0,2)
ARIMAX - AT	17949	3.48	(3,1,1)(1,0,0)

BUNDESLIGA – MODEL COMPARISONS

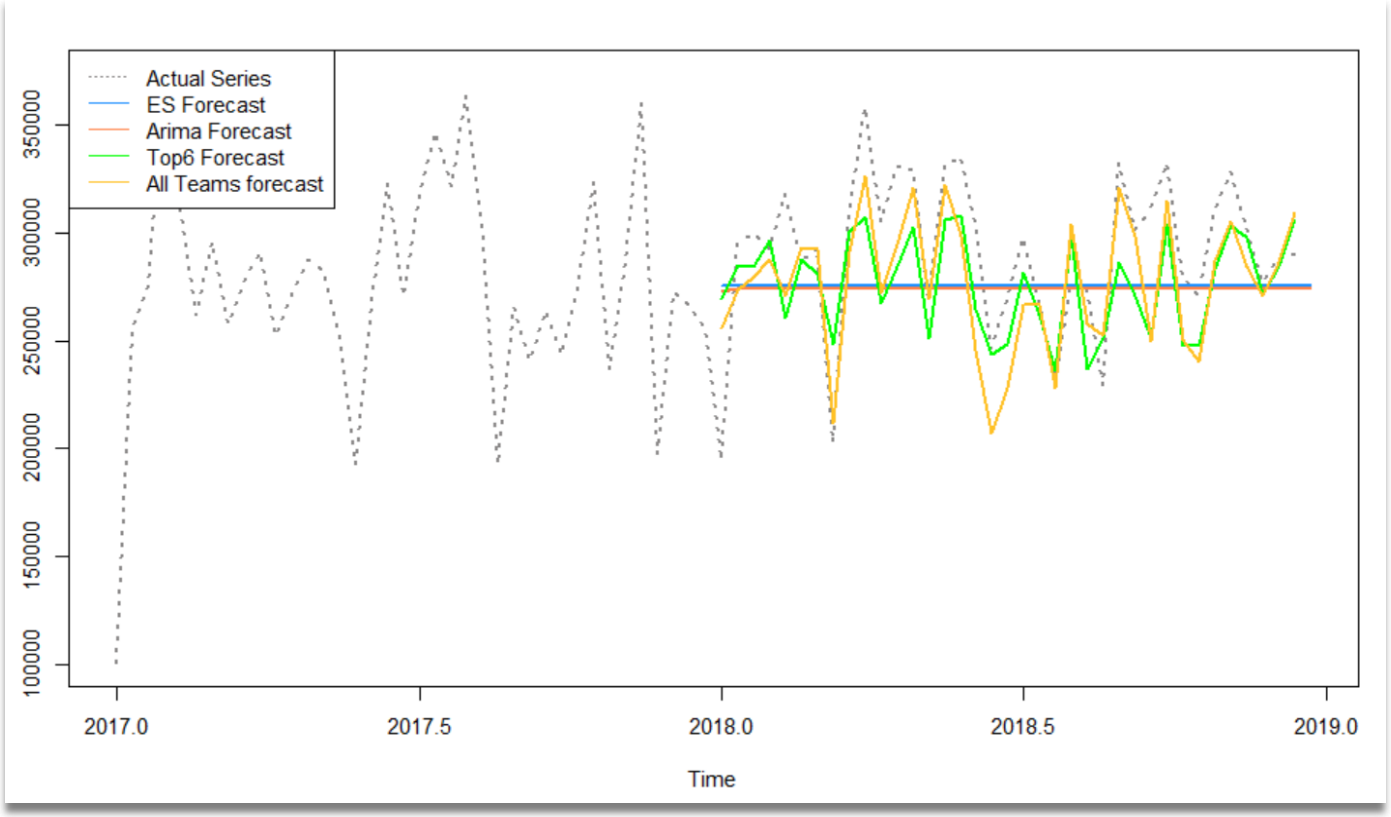


Predictions & Results

- The *Primera Division* attendance is the hardest to predict, even with the ***ARIMAX*** model
- This may be explained in light of the absence of seasonal and trend effects (ES and ARIMA predictions are flat)
- The coefficients of the ARIMAX models are close to 0, this indicates once again the absence of time-related effects

Metrics	RMSE	MAPE	(p,d,q)(P,D,Q)
ES	39098	11.14	-
ARIMA	39744	11.34	-
ARIMAX - MT	30941	8.88	(0,0,2)(0,0,2)
ARIMAX - AT	28110	7.87	(0,0,0)(0,0,0)

PRIMERA DIVISION – MODEL COMPARISONS



Conclusion

TIME SERIES ANALYSIS

- We built a useful couple of ARIMAX models, with *All Teams* more accurate, and *Main Teams* more flexible and more time-oriented
- Some factors could explain the difference of accuracy we observe between PL, Bundesliga and La Liga, such as the importance of «clashes» or stadiums with an audience more stable.
- The good results of Linear Regression showed us that we could use classic regression algorithms for this kind of problem.
- Maybe we could improve those results by using other models of this type, as Random Forests or Gradient Boosting algorithms.
- Finally, we can notice that the average audience seems positively correlated with the presence of time-related effects, and thus an attendance curve easier to predict.

Group Members:

Davide De Benedetto – 3013595

Grégoire Eugene – 3127015

Group Project on European Soccer

Predictive Analytics for Data Driven Decisions Making