Project Report: Predicting Soccer Match Results

11663 B Applied Machine Learning - Section B
Name: Frank Yue Ying | Email: yying2@andrew.cmu.edu | Date: 2022-05-06

Abstract

Sports analytics is crucial for today's professional football(soccer) teams to improve their performances. Previous analysis only focuses on summary statistics without considering the "flow" of the game in their framework. Building on top of two independent studies, this project utilized text commentary data collected from Kaggle in order to predict the result of the game. This project also used error analysis to help with data manipulation, and tuning to identify the best hyperparameter for a logistic regression model. Results from this project shows that game event data offers insights to support better prediction in soccer matches, but a better framework to capture the "flow" is needed.

1-Introduction

The field of sports analytics can be dated back to 1858 with the invention of box score in baseball, which recorded player's statistics in a tabular format.[1] From 2000 onward, we have seen stories like the Moneyball attracted broad attention, and the application of Hawk-Eye in tennis brought sports analytics to every household.[2] Putting into quantitative terms, the global sports analytics industry is expected to grow 21.3% annually from 2021 to 2028.[3] As one of the most popular sports in the world, soccer received a tremendous amount of attention from data scientists and analysts. Top clubs in the English Premier League have 15 analysts on board to help improve performance of the team using data.[4] Ultimately, soccer analytics focus on winning, and this involves breaking down complex spatial and temporal data from continuous movement made by 22 players on the pitch. Summary statistics such as number of shots made on target are not useful because they fail to present the decisive moments in the game. The result of a 90-minute long game can be directly

[1] Hitesh Kashyap, "A Primer on Sports Analytics: A New Dimension of Sports," Analytics India Magazine, August 5, 2021, https://analyticsindiamag.com/a-primer-on-sports-analytics-a-new-dimension-of-sports/#:%7E:text=Some%20of%20the%20organizations%20started,of%20new%20experience%20to%20sports.&text=These%20companies%20and%20many%20more%20are%20exploring%20the%20applications%20of%20sports%20analytics.

[2] "Forget Big Data, Computer Vision Is the next Moneyball," Plainsight, April 1, 2021, https://plainsight.ai/forget-big-data-computer-vision-is-the-next-moneyball/.

[3] "Global Sports Analytics Market Size, Share &amp; Trends Analysis Report by Component (Software, Service), by Analysis Type (on-Field, off-Field), by Sports (Football, Cricket, Basketball, Baseball), and Segment Forecasts, 2021-2028," Research and Markets - Market Research Reports - Welcome, April 2021, https://www.researchandmarkets.com/reports/5415591/global-sports-analytics-market-size-share-and.

[4] Justin Harper, "Data Experts Are Becoming Football's Best Signings," BBC News (BBC, March 5, 2021), https://www.bbc.com/news/business-56164159.

related to the events that happened in a 10 minute turnaround, and identifying that vital moment becomes the key. Researchers are focusing on predicting game results by using granular data collected from either video or text commentary. This paper also explores a similar approach by utilizing event-based data collected from soccer matches.

2-Related Research

Event driven analysis has shown to be very effective in breakdown key elements of soccer games by researchers and analysts. Given the unstructured and complex nature of soccer matches, every researcher needs to answer the following two questions: first, how to define and categorize an "event" in a continuous game? Second, how to determine the value and correlations between events that led to the game result? The following two studies provided useful yet different methods that guided this research.

In order to identify events with high impact, a straightforward approach would be to identify the key differences in game action between winning and losing teams. Oberstone looked into 24 pitch actions data for all 20 clubs in the English Premier League during the 2007-2008 season.[5] He started with a simple regression analysis to select an independent and statistically significant set of pitch actions in order to predict the overall points earned after 38 games in the season. Acknowledging the issue of multicollinearity, he performed backward elimination to drill down to 6 significant pitch actions in the end. For example, he identified that teams can get additional 4 points if they can increase their percentage of goals scored outside the box by 5%. His additional research looked at competitiveness by dividing the club performances into three buckets: top 4 clubs that yield consistent performance across multiple seasons, middle 12 clubs, and bottom 4 clubs that were at risk of relegation. By comparing performances between the three buckets, he identified another 13 factors that showed significant differences through analysis of variance. This research offered insights about key event types that are likely to help predicting game results. However, this research lacks spatial and temporal considerations of those statistics. Furthermore, this research assumes independence between pitch actions, thus ignoring the combined impact of key event actions during the game.

Min, Kim et al. introduced a more complex framework for results prediction by incorporating a rule-based reasoner and Bayesian networks together.[6] The rule-based reasoning part focused on domain knowledge outside of the game, such as defensive rating and strength of different formations. Bayesian networks were used to measure the flow of the games based on its strategy and outcome, incorporating uncertainty of the game by introducing probabilities from each decision. This study

[5] Joel Oberstone, "Differentiating the Top English Premier League Football Clubs from the ...," researchgate, January 2009, https://www.researchgate.net/publication/46554872_Differentiating_the_Top_English_Premier_League_Football_Clubs_from_the_Rest_of_the_Pack_Identifying_the_Keys_to_Success.

[6] Byungho Min et al., "A Compound Framework for Sports Results Prediction: A Football Case Study," Knowledge-Based Systems 21, no. 7 (March 21, 2008): pp. 551-562, https://doi.org/10.1016/j.knosys.2008.03.016.

broke down a game into ten time frames, and used the combined model to derive the outcome of a strategy and team fatigue level from one frame to another. This allowed them to connect flows during the game when predicting the outcome. In order to evaluate this framework, this study used the World Cup data in 2002 to simulate games and generated a ranking against the actual result as well as historic predictor data as a benchmark. This framework was able to correctly predict 6 out of the actual top 8 countries from the competition, higher than the benchmark of 5. T-tests were also performed to confirm that the framework is significantly better than the benchmark predictor. Contrary to others, this study utilized Bayesian Network to represent the continuous movement of the game but it also presented several drawbacks. First, strategies are not objective and do not reflect the events on the pitch directly. For example, substituting in a defensive player might imply to improve the team's defense for some coaches, but it could also switch the formation to signal more attacks. Therefore, miss-interpretation of strategy could lead to erroneous prediction of the game. Second, this study evaluated the model with less than 20 games, which might fail to capture various styles and soccer events compared to an entire season of matches.

3-Data Description

For this project, a dataset was obtained from Kaggle containing football (soccer) game statistics and events.[7] This dataset covers 9,074 games with 941,009 events collected from 5 European football leagues for 6 seasons (2012 to 2017). Game statistics contain summary data from basic team information such as name, home location, country and final score. It also includes betting oddings for the home team to win, draw or lose the game. Event data was extracted from text commentary to describe every action in the game, including when (the time it occurred), where (position on the pitch), who (team/player who initiated the event), how (type of the event), and what (detailed action of the event). Event data contains both the original text commentary and extracted event data columns. Only the event data columns were used for this research.

4-Exploratory Data Analysis

The raw data is stored in two different datasets: game and event. Game data consists of 10,112 rows and 18 columns, with each row representing a soccer match uniquely identified by the field “id_odsp”. Event data consists of 941,009 rows and 22 columns, with each row representing an event during the game (a shot, for example), uniquely identified by “id_event”. These two datasets can be associated with the key “id_odsp”. Appendix 1 contains column names in this dataset.

There are some missing data features. For the game dataset, these odd features contain more than 90% NA values: odd_over, odd_under, odd_bts, and odd_bts_n. They were dropped for this project. For the event data set, some features are also NA due to

[7] Alin Secareanu, “Football Events,” Kaggle, January 25, 2017, https://www.kaggle.com/datasets/secareanualin/football-events.

the nature of the event itself. For example, shot events include shot result as a feature, but it will be left empty for passing or substitution events. In addition, the number of events varies between games, therefore feature space might be sparse for games that lack event activities.

5-Data Preparation

Utilizing home score (fthg) and away score (ftag) features in the game dataset, the output class was constructed as a binary value for this project. When the home team won the game (home score greater than away score), class value is 1. Otherwise, class value is 0 for draw and lost games.

Based on summary statistics of the raw data, the average number of events per soccer match is 104 and about 75% of the games have less than 115 events. Therefore, only the first 100 events per game are considered for this project. Games without any events will be dropped due to a lack of features.

The raw datasets used numeric integers to represent the ID of different values in the feature. For example, the event_type feature contains 11 unique values, where 1 indicates shot, 2 indicates corner and so on. These features were converted into groups of binary values for the input feature space. After transforming the original data using Python, the resulting dataset contained 3114 features in total that were either numeric or binary values.

Using the ratio of 20%, 70%, and 10% to split between development, validation, and final test datasets, there were 1836 cases (games) identified for development, 6427 cases for cross-validation, and 919 for the final test.

6-Baseline Performance

Baseline performance was evaluated using the entire development set with 1836 cases. Simple logistic regression with default setting was chosen as the baseline model with 5-fold cross-validation. Model achieved an accuracy of 65.69% and a kappa statistic of 0.2817. Appendix 2 contains a summary of the baseline model.

7-Error Analysis

For the error analysis, 1836 development cases were first split into two sets with 999 and 837 records separately. The first set was used as development data to train the model, and the second set was used to conduct error analysis. Using WEKA, simple logistic regression was chosen as the model.

Using the training set, the baseline model correctly classifies 53.25% of the instances with a Kappa statistic of 0.0569 using 5-fold cross validation. After supplying the test set, the performance became 51.61% as accuracy and 0.0226 as the kappa statistics.

Features with high absolute weights were collected and recorded in Appendix 3.

From the test set result, false positive and false negative games were reviewed in order to determine the root causes of errors. By comparing those games with highly weighted features, here are some instances that offered several insights:

1. Game zRp4p5Xc/: Predicted 1, Actual 0

This was a soccer game in England back in 2017 between West Ham (Home) and

Arsenal (Away). The final score was the away team winning by 5 to 1, which indicates that the away team won pretty easily. This result can be reflected by the odds, which are 5.9 for the home team win, 4.41 for draw, and 1.65 for away team win. Given the weights for those odds, the combined result from odds features will contribute to prediction of 0 (not win for the home team). However, for game event features, some of the key events with high feature weights were overlooked. For this game, the home team produced some good stats (key pass, shot on target) at 93 minute, which is too late to change the game. Features like those pulled the result to be predicted as a win, which is why this game was falsely classified.

2. Game fgbLxs2T/: Predicted 0, Actual 1

This was a game in the Spanish league in 2013 between Valencia and Atletico Madrid with a total score of 2 to 0 as the home team scored a victory. The odds are pretty even as 2.65, 3.55, and 3, which means that the game event features will produce greater influence to the prediction instead. In this game, the game produced 7 yellow cards and 1 red card, with the home team getting 3 yellow cards and 1 red card. These bookings led to the model predicting that the home team is more likely to lose. Other important features for the home team were not picked up by the model.

Most of the errors were contributed by the collective impact from game events features, and this is because the feature space focuses too much on events and only considers those independently. Aggregating event data seems to be a logical solution to fix this problem. The proposed solution is to aggregate the event features into groups of ten. For example, instead of having 10 shot-related binary features between event 1 to event 10, those data will be merged together into one feature that reflects the performance during a period of game time.

The new feature space contained 318 features, significantly reduced from the previous volume (3114). After running a 5-fold cross validation with the same model and settings, The new model produced an accuracy of 65.69% and a kappa statistic of 0.2837 with the entire set of development data. The accuracy stayed the same with the overall baseline performance, but achieved a slightly better kappa statistic than the baseline (0.2817). This improvement was due to better performance on predicting winning games. Appendix 4 includes the summary of this new model.

The improvement from error analysis is not significant. However, the new model used a much smaller feature space by aggregating granular events into clusters. This could be helpful to reduce computation effort during model training and validation.

8-Tuning

This project selected Logistic Regression as the model for tuning instead of Simple logistic regression. In WEKA's logistic regression model, ridge value was selected as the tuning parameter, with default value of 10^-8. Ridge estimator is used for regularization in order to simplify the model. Changing the ridge estimator will affect model coefficients, therefore adding or reducing regularization.

To create a baseline performance with this new model for tuning, the development dataset with 1836 cases was used to train a Logistic model through WEKA. The baseline performance with default setting had an accuracy of 60.40% and the kappa statistic is 0.1953. Appendix 5 includes a summary of this baseline model.

The cross validation dataset contains 6427 cases, and was divided into 5 folds of train and test datasets for tuning. In addition, this tuning selected the following values for the ridge parameter: 10^-8, 10^-10, 10^-5, 10^1.

In stage 1, the optimal setting for the ridge parameter was selected on each fold. Appendix 6 contains the performance result per fold. In stage 3, test set performances using the optimal setting were compared with the same sets using the default setting for each fold. Appendix 7 contains the test set result with both settings. Although ridge value setting as 10^1 seemed to be optimal across each fold, test set performance didn't show a significant difference.

A significant test was performed between performances with optimal setting and default setting. The P-value between baseline (default setting) and new (optimal setting) is 0.9679, indicating that the change is insignificant.

Based on the result from tuning, the ridge parameter should not be optimized and default value (10^-8) was still the best setting for the model.

9-Final Evaluation

The final evaluation of the logistic regression model for this project was based on the final test set with 919 cases (games). The model was first trained with the merged data set of development and validation data, containing 8263 cases (games). There is no overlap between data sets and the final test set was not used and examined during any part of the model development process. Using logistic regression with the ridge parameter at 10^-8, the model correctly classified 584 out of the 919 cases (63.55%) and reached a kappa statistic of 0.2723. This was slightly worse than the baseline performance of 65.69% as accuracy and 0.2817 as kappa statistics.

10-Learning and Discussion

This research focused on breaking down summary statistics into lower levels in order to better understand the driving factors of football games. This project started from data transformation by merging two datasets into one, and setting up sparse binary feature space for nominal data. Results from the error analysis supported further data manipulation by setting up 10 groups of event data as features instead of having independent features for each event. This significantly reduced the feature space from 3114 features to 318 while achieving a similar baseline performance. Tuning focused on the ridge parameter of the logistic regression model, and set the optimal value of 10^-8. Final evaluation showed that a simple logistic regression model with lower-level statistical data of previous matches can support accurate predictions of future, unseen games.

An interesting finding is that granular event data from the game failed to offer real information to the prediction model. This could be due to two reasons: first, unnecessary event features introduced more noise than insight into the model; Second, event features were treated independently and failed to capture the “flow” of the game. Combining independent events into groups was a relatively easy approach, but an optimal solution should involve a probabilistic approach between event features as introduced by one of the research papers discussed in section 2.

A big change for this project was switching from multi-class classification (win, draw, lose) during the research proposal to binary classification (win, not win). This was due to the model's poor performance in predicting drawed games. This makes sense since features that indicate likely scoring activities for both teams will be considered by the model, and it is generally harder to find the middle section between two sides. Future research can focus on predicting draw games, and discover hidden insights that highlight the back and forth between two teams.

A limitation of this study is the lack of external game data involved in predictions. Although specific game event data are crucial, they were an outcome of the hidden strategy and analysis before the game. Incorporating external game data such as team fitness and high-level strategies from the coach can further improve the performance of game predictions.

References

“Forget Big Data, Computer Vision Is the next Moneyball.” Plainsight, April 1, 2021. https://plainsight.ai/forget-big-data-computer-vision-is-the-next-moneyball/.

“Global Sports Analytics Market Size, Share &amp; Trends Analysis Report by Component (Software, Service), by Analysis Type (on-Field, off-Field), by Sports (Football, Cricket, Basketball, Baseball), and Segment Forecasts, 2021-2028.” Research and Markets - Market Research Reports - Welcome, April 2021. https://www.researchandmarkets.com/reports/5415591/global-sports-analytics-market-size-share-and.

Harper, Justin. “Data Experts Are Becoming Football's Best Signings.” BBC News. BBC, March 5, 2021. https://www.bbc.com/news/business-56164159.

Kashyap, Hitesh. “A Primer on Sports Analytics: A New Dimension of Sports.” Analytics India Magazine, August 5, 2021. https://analyticsindiamag.com/a-primer-on-sports-analytics-a-new-dimension-of-sports/#:%7E:text=Some%20of%20the%20organizations%20started,of%20new%20experience%20to%20sports.&amp;text=These%20companies%20and%20many%20more%20are%20exploring%20the%20applications%20of%20sports%20analytics.

Min, Byungho, Jinhyuck Kim, Chongyoun Choe, Hyeonsang Eom, and R.I. (Bob) McKay. “A Compound Framework for Sports Results Prediction: A Football Case Study.” Knowledge-Based Systems 21, no. 7 (March 21, 2008): 551–62.

https://doi.org/10.1016/j.knosys.2008.03.016.

Oberstone, Joel. “Differentiating the Top English Premier League Football Clubs from the ...” researchgate, January 2009. https://www.researchgate.net/publication/46554872_Differentiating_the_Top_English_Premier_League_Football_Clubs_from_the_Rest_of_the_Pack_Identifying_the_Keys_to_Success.

Secareanu, Alin. “Football Events.” Kaggle, January 25, 2017. https://www.kaggle.com/datasets/secareanualin/football-events.

Appendix

Appendix 1.

Legend: Primary Key, *Secondary Key*

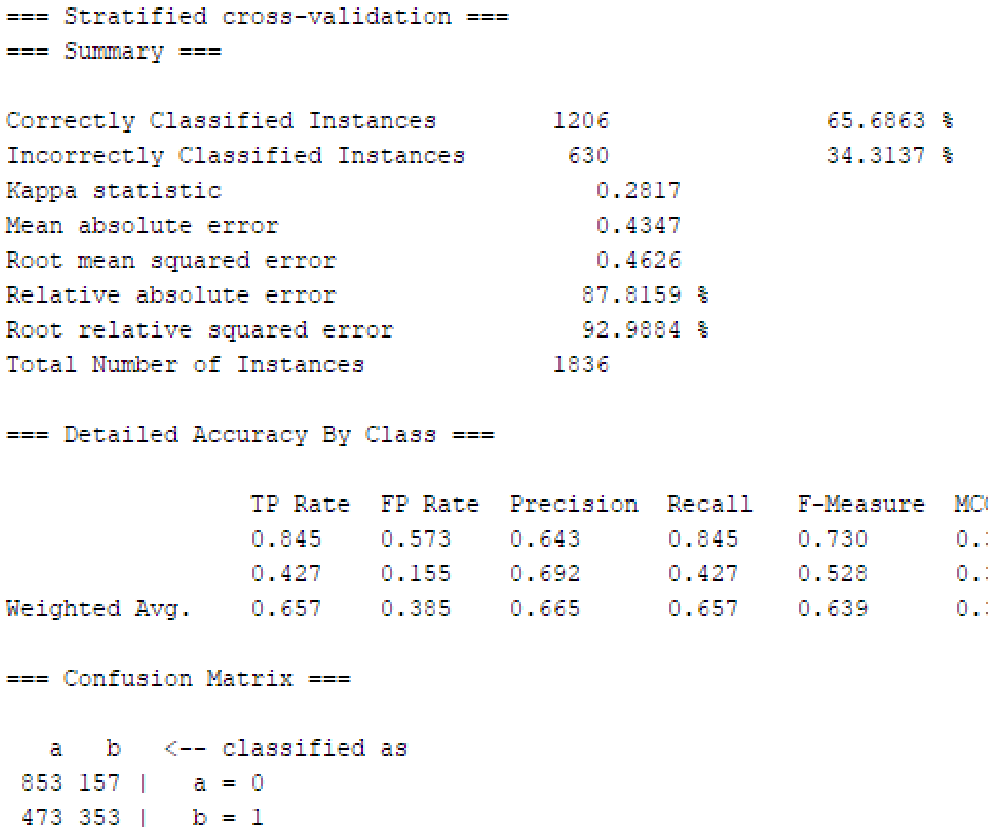
Game Data
id_odsp: game ID
link_odsp: game link
adv_stats: if contain events
date: date of the game
league: club league
season: year played
country: nation of league
ht: home team name
at: away team name
fthg: full-time home goals
ftag: full time away goals
odd_h: home win odds
odd_d: draw odds
odd_a: away win odds
odd_over: other odds

odd_under: other odds
odd_bts: other odds
odd_bts_n: other odds

Event Data
id_odsp: game ID
id_event: event ID
Sort_order: event order
Time: minute of game
Text: commentary
Event_type: primary event ID
Event_type2: secondary ID
Side: Home or away
Event_team: team name
Opponent: opponent name
Player: primary player
Player2: secondary player
Player_in: substitution in
Player_out: substitution out
Shot_place: shot placement
Shot_outcome: outcomes
Is_goal: resulted in goal
Location: pitch location
Bodypart: shot body part
Assist_method: assist type
Situation: game situation

Fast_break: fast break or not

Appendix 2.



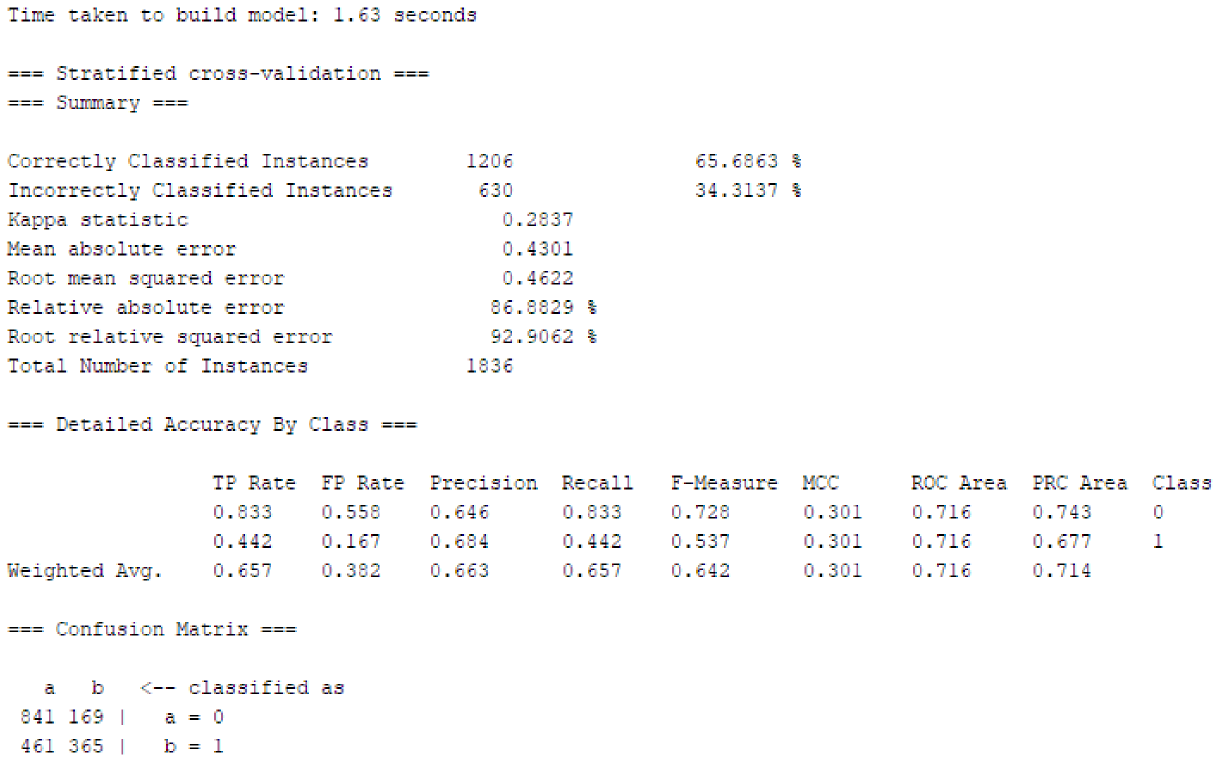
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1206               65.6863 %
Incorrectly Classified Instances       630               34.3137 %
Kappa statistic                          0.2817
Mean absolute error                      0.4347
Root mean squared error                  0.4626
Relative absolute error                 87.8159 %
Root relative squared error             92.9884 %
Total Number of Instances             1836

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MC
                 0.845    0.573    0.643      0.845    0.730      0.
                 0.427    0.155    0.692      0.427    0.528      0.
Weighted Avg.    0.657    0.385    0.665      0.657    0.639      0.

=== Confusion Matrix ===

   a   b   <-- classified as
 853 157 |   a = 0
 473 353 |   b = 1
```

Appendix 3.

Feature	**Weight**
odd_h	4.3954
odd_d	-9.5022
odd_a	3.9051
spain_country_binary=1	-9.5355
1-PenaltyConceded_EventType_binary=1	18.4556
1-SetPiece_Situation_binary=1	-33.3953

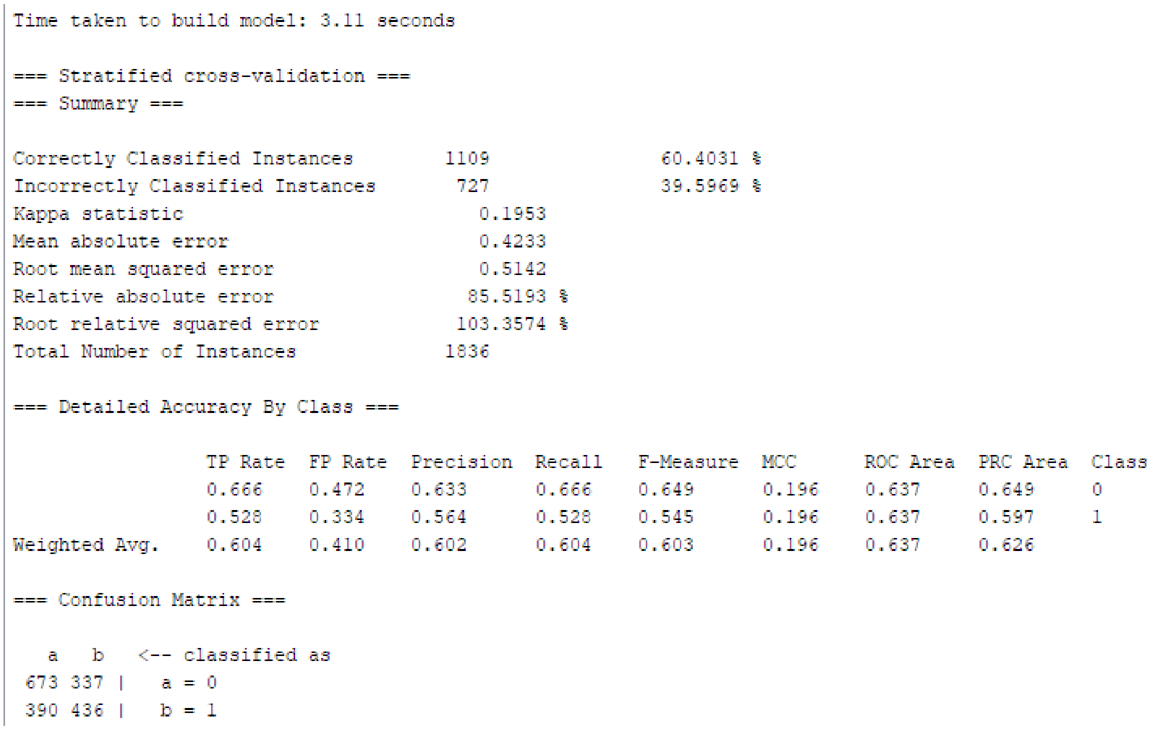
4-FreeKick_Situation_binary=1	28.6666
8-HitThePost_ShotOutcome_binary=1	-14.8898
12-PenaltyConceded_EventType_binary=1	22.8582
38-PenaltyConceded_EventType_binary=1	44.3982
51-fast_break=1	42.3161
62-OwnGoal_EventType2_binary=1	78.642
82-SendingOff_EventType2_binary=1	34.1281

Appendix 4.



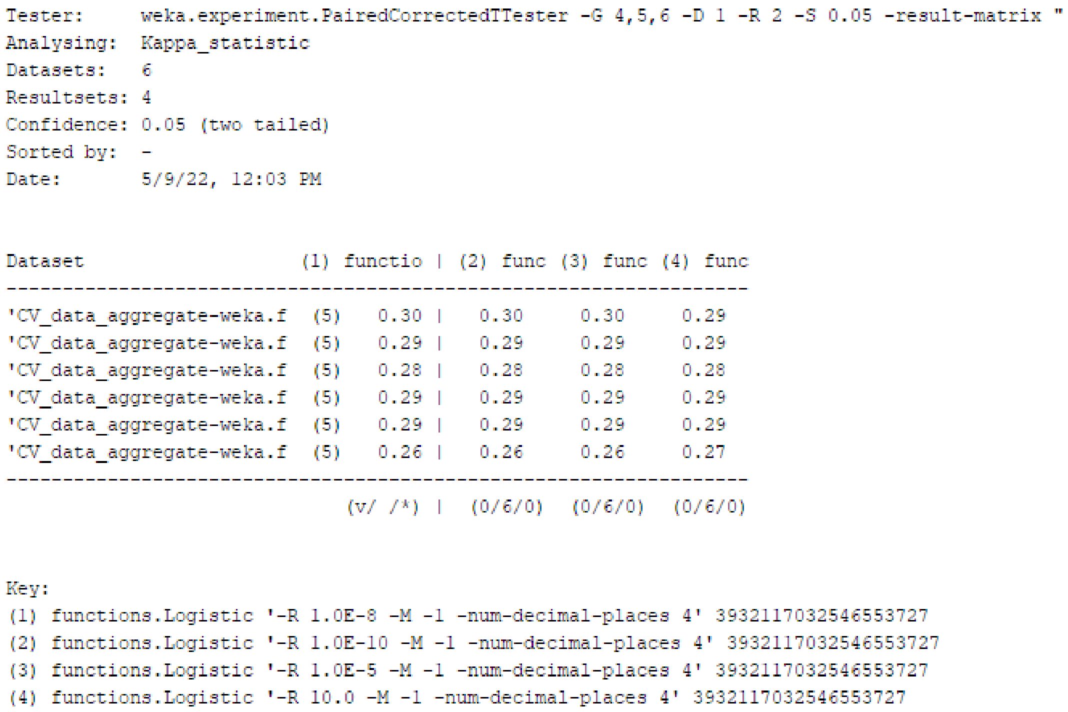
```
Time taken to build model: 1.63 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1206               65.6863 %
Incorrectly Classified Instances       630               34.3137 %
Kappa statistic                          0.2837
Mean absolute error                      0.4301
Root mean squared error                  0.4622
Relative absolute error                 86.8829 %
Root relative squared error             92.9062 %
Total Number of Instances             1836

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.558    0.646      0.833    0.728      0.301    0.716     0.743     0
                 0.442    0.167    0.684      0.442    0.537      0.301    0.716     0.677     1
Weighted Avg.    0.657    0.382    0.663      0.657    0.642      0.301    0.716     0.714

=== Confusion Matrix ===

   a   b   <-- classified as
 841 169 |   a = 0
 461 365 |   b = 1
```

Appendix 5.



```
Time taken to build model: 3.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1109               60.4031 %
Incorrectly Classified Instances       727               39.5969 %
Kappa statistic                          0.1953
Mean absolute error                      0.4233
Root mean squared error                  0.5142
Relative absolute error                 85.5193 %
Root relative squared error            103.3574 %
Total Number of Instances             1836

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.666    0.472    0.633      0.666    0.649      0.196    0.637     0.649     0
                 0.528    0.334    0.564      0.528    0.545      0.196    0.637     0.597     1
Weighted Avg.    0.604    0.410    0.602      0.604    0.603      0.196    0.637     0.626

=== Confusion Matrix ===

   a   b   <-- classified as
 673 337 |   a = 0
 390 436 |   b = 1
```

Appendix 6.



```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "
Analysing:  Kappa_statistic
Datasets:   6
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       5/9/22, 12:03 PM


Dataset                   (1) functio | (2) func (3) func (4) func
-----------------------------------------------------------------
'CV_data_aggregate-weka.f  (5)   0.30 |   0.30     0.30     0.29
'CV_data_aggregate-weka.f  (5)   0.29 |   0.29     0.29     0.29
'CV_data_aggregate-weka.f  (5)   0.28 |   0.28     0.28     0.28
'CV_data_aggregate-weka.f  (5)   0.29 |   0.29     0.29     0.29
'CV_data_aggregate-weka.f  (5)   0.29 |   0.29     0.29     0.29
'CV_data_aggregate-weka.f  (5)   0.26 |   0.26     0.26     0.27
-----------------------------------------------------------------
                              (v/ /*) |  (0/6/0)  (0/6/0)  (0/6/0)


Key:
(1) functions.Logistic '-R 1.0E-8 -M -1 -num-decimal-places 4' 3932117032546553727
(2) functions.Logistic '-R 1.0E-10 -M -1 -num-decimal-places 4' 3932117032546553727
(3) functions.Logistic '-R 1.0E-5 -M -1 -num-decimal-places 4' 3932117032546553727
(4) functions.Logistic '-R 10.0 -M -1 -num-decimal-places 4' 3932117032546553727
```

Appendix 7.

	Ridge Value						
Fold	10^-8	10^-10	10^-5	10^1	Optimal Setting	Test Set on Optimal	Test Set on Default
All	0.3	0.3	0.3	0.29			
1	0.29	0.29	0.29	0.29	10^1	0.2542	0.2451
2	0.28	0.28	0.28	0.28	10^1	0.2677	0.2656
3	0.29	0.29	0.29	0.29	10^1	0.2677	0.2754
4	0.29	0.29	0.29	0.29	10^1	0.2572	0.2605
5	0.26	0.26	0.26	0.27	10^1	0.3333	0.3329