**Prem Footy Predictor: The Final Report**

**Introduction:**

The Prem Footy Predictor trains on past data to predict the outcomes of Premier League football matches. Given a series of statistics and some insightful metrics built on them, the Prem Footy Predictor will return either an ‘H’, ‘D’, or ‘A’ – home victory, draw, away victory. Unlike many other sports, football has not seen a renowned match predictor (bar [Paul the Octopus](http://my.fakingnews.firstpost.com/files/2014/06/Paul.jpg)), and there are many potential reasons for this: the lack of statistics recorded and the unavailability of those that are; the free-flowing nature of the game without much in-match coach interference; the vulnerability to individual players and random flashes of brilliance. Despite all this, we sought to see if we couldn’t just build a good predictor.

**Data Collection & Analysis:**

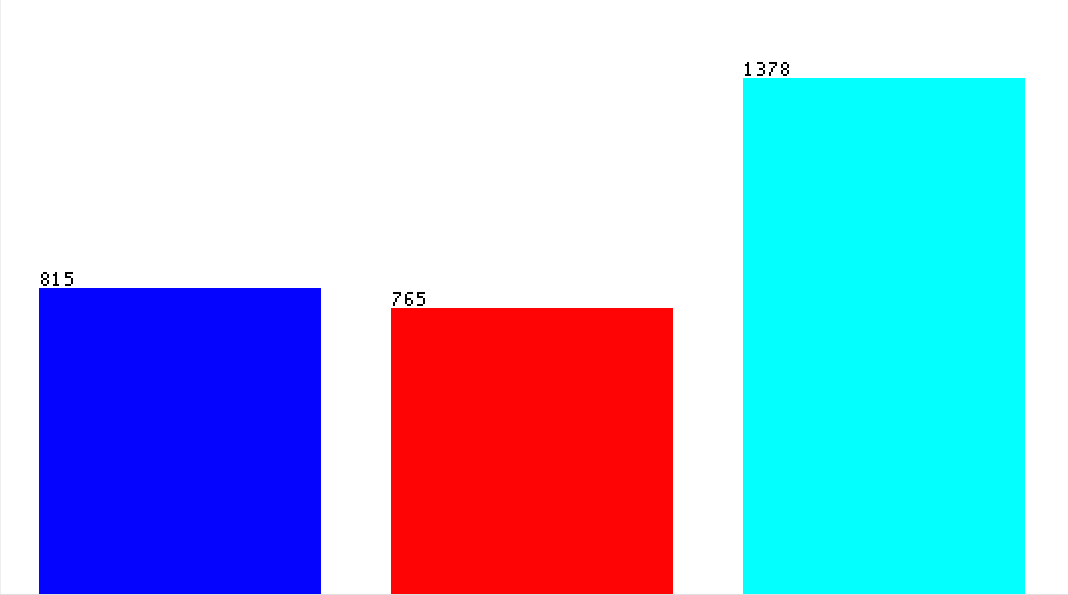
We went about creating 1 dataset with statistics for every match played in 8 years in the Premier League. The raw statistics available to us were:

*{Score, #Shots (on & off target), Possession, #Corners, #Offsides, #YC, #RC}*

We made sure to store these statistics, and then additionally create another dataset of more-insightful metrics based off of this data. Some of the metrics included:

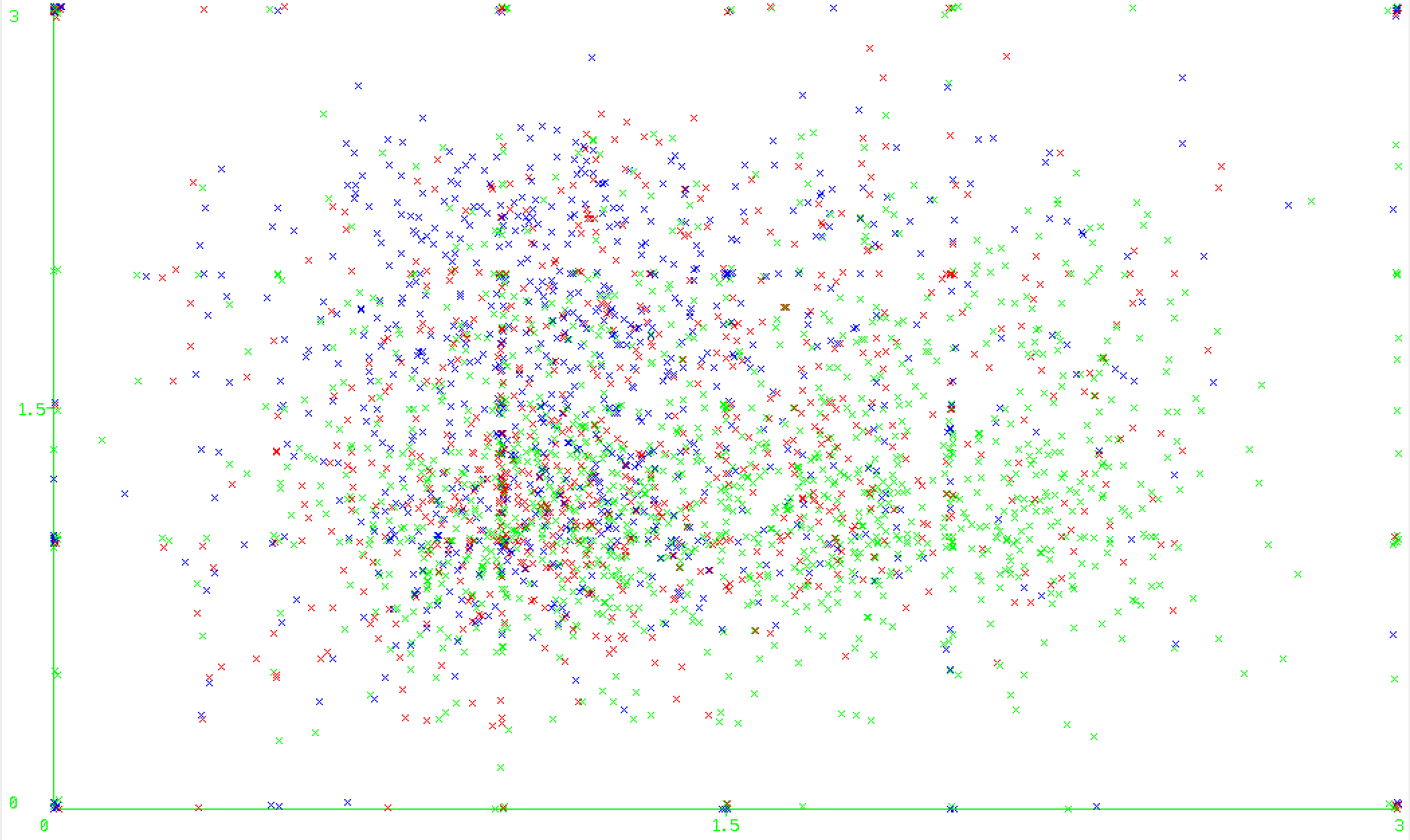
{,,,,}

Refer to *Appendix A* for a full-list of the statistics.

****Our data consisted of ~3,000 match samples with outcomes distributed as shown in Figure 1 below. The distribution of the outcomes with respect to the statistics is shown in its entirety in *Appendix B*.

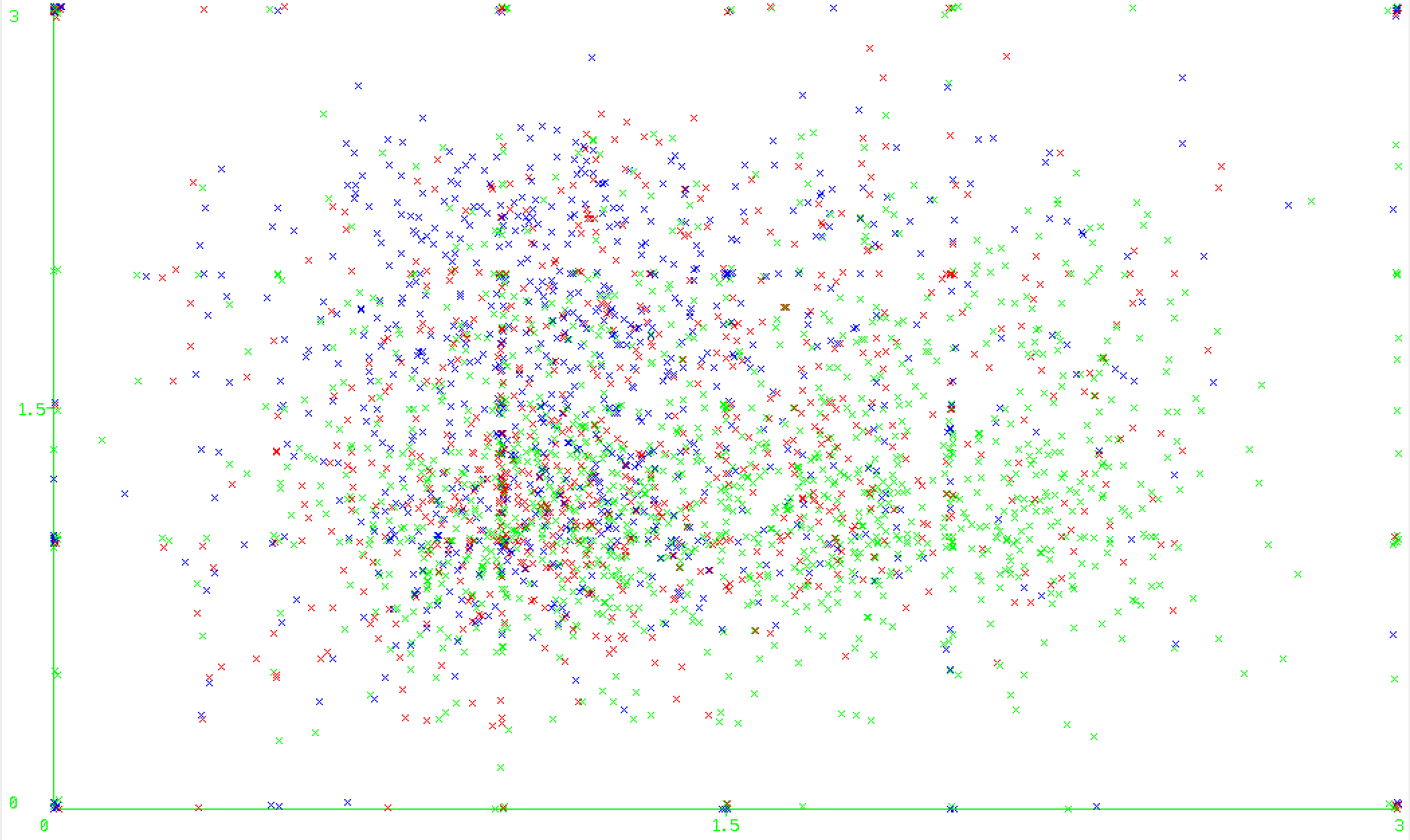
**Figure 1:** Dataset Outcome Distribution

{Blue, Red, Teal} = {Away, Draw, Home}

**Figure 2:** home\_team\_pointspergame(X) vs. away\_teampointspergame(Y)

{Blue, Red, Green} = {Away, Draw, Home}

The most significant conclusion of our data collection was the strong evidence we discovered in support of the home-field advantage. Figure 2 and Figure 3 below both reflect a motif of weaker teams at home fighting for draws and stronger teams at home consistently winning matches against lower-ranked opponents. The contrasts between the performances of strong teams at home and strong teams away as well as the contrasts between the performances of weak teams at home and weak teams away give strong evidence for the homefield advantage.

****

**Figure 3:** hometeam\_leaguestanding (X) vs. awayteam\_leaguestanding (Y)

**Training and Testing:**

While evaluating algorithms, we used 10-fold cross-validation on the entire sample-space of each dataset we constructed. Our outcomes with most algorithms were approximately the same across both datasets, with each reporting accuracy between 45% and 53% on the latter insightful metric dataset, and between 38-48 on the raw data dataset.

Most significantly:

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Bayes Net | 48.783% |
| Complement Naïve Bayes | 47.9378% |
| Naïve Bayes | 49.1886% |
| Logistic | 51.4875% |
| Multilayer Perceptron | 46.146% |
| RBFNetwork | 49.3239% |
| SimpleLogistic | 52.5693% |
| IBK | 45.1427% |
| BFTree | 51.3861% |
| J48 | 46.7884% |
| NBTree | 50.169% |
| RandomForest | 50.6085% |

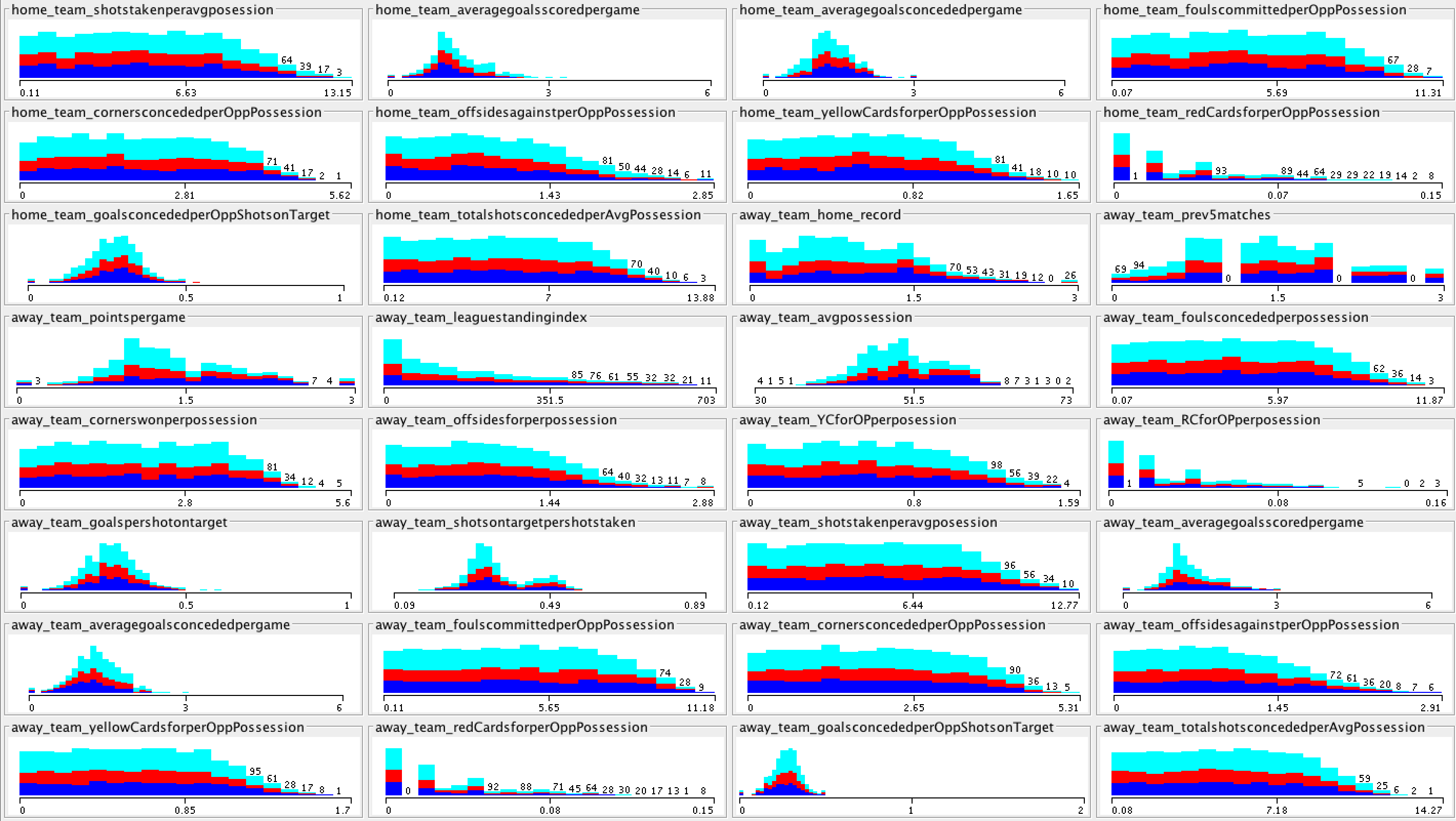
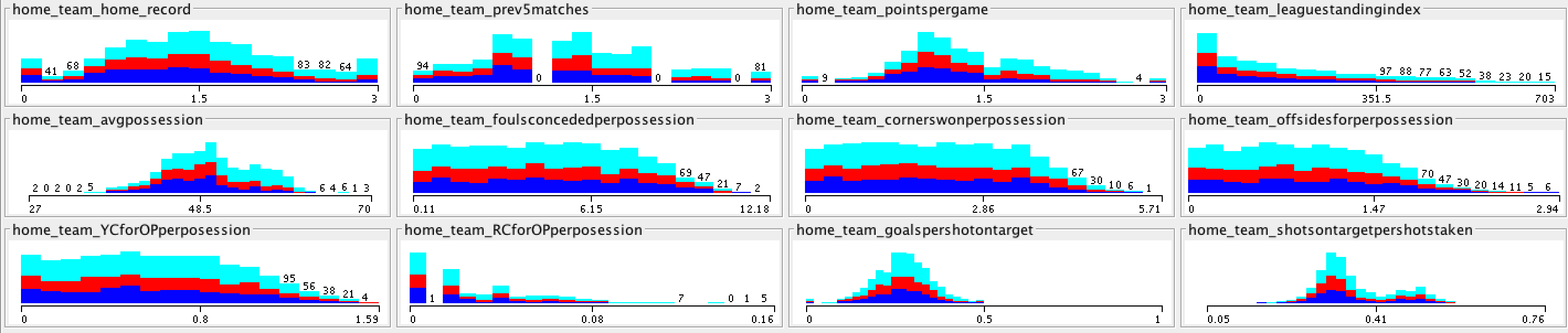
The relatively poor accuracy of the k-Nearest-Neighbors function (IBK) and the successes of SimpleLogistc and Logistic functions may be symptomatic of a broader truth, which may be that statistics combine to determine the outcomes of matches, and that they must be analyzed simultaneously with respect to one another. The logistic functions can model a large range of relationships between variables, and so its success alludes to that. The k-Nearest-Neighbors function does not offer any more data-interpretation other than comparing it to nodes with similar statistics, and in a game as susceptible to moments of inspiration, and flashes of brilliance, as football, there is no telling what sort of outliers the nearest neighbor(s) could be.

**Conclusion & Moving Forward:**

Working with a limited dataset in a field ultimately governed by chance, we expected to face a lot of difficulty, and really sought to explore rather than solve. At Prem Footy Predictor’s best, it increases the accuracy rate from the 1/3rd of a random guess to the ½ of a calculated prediction. To further build this out and see higher accuracy rates, we would need to incorporate in more statistics and there are 2 fronts of doing so: you can add more team-statistics such as passes completed, chances created, possession in area of a field etc… and you can incorporate player based statistics. Unfortunately neither of these is readily available, especially without paying a ghastly subscription fee. All in all the Prem Footy Predictor has achieved a significant improvement on match prediction, and has revealed some interesting relationships between variables as touched upon above.

*Appendix A: Insightful Metrics*

home\_team\_home\_record: continuous.  
home\_team\_prev5matches: continuous.  
home\_team\_pointspergame: continuous.  
home\_team\_leaguestandingindex: continuous.  
home\_team\_avgpossession: continuous.  
home\_team\_foulsconcededperpossession: continuous.  
home\_team\_cornerswonperpossession: continuous.  
home\_team\_offsidesforperpossession: continuous.  
home\_team\_YCforOPperposession: continuous.  
home\_team\_RCforOPperposession: continuous.  
home\_team\_goalspershotontarget: continuous.  
home\_team\_shotsontargetpershotstaken: continuous.  
home\_team\_shotstakenperavgposession: continuous.  
home\_team\_averagegoalsscoredpergame: continuous.  
home\_team\_averagegoalsconcededpergame: continuous.  
home\_team\_foulscommittedperOppPossession: continuous.  
home\_team\_cornersconcededperOppPossession: continuous.  
home\_team\_offsidesagainstperOppPossession: continuous.  
home\_team\_yellowCardsforperOppPossession: continuous.  
home\_team\_redCardsforperOppPossession: continuous.  
home\_team\_goalsconcededperOppShotsonTarget: continuous.  
home\_team\_totalshotsconcededperAvgPossession: continuous.  
away\_team\_home\_record: continuous.  
away\_team\_prev5matches: continuous.  
away\_team\_pointspergame: continuous.  
away\_team\_leaguestandingindex: continuous.  
away\_team\_avgpossession: continuous.  
away\_team\_foulsconcededperpossession: continuous.  
away\_team\_cornerswonperpossession: continuous.  
away\_team\_offsidesforperpossession: continuous.  
away\_team\_YCforOPperposession: continuous.  
away\_team\_RCforOPperposession: continuous.  
away\_team\_goalspershotontarget: continuous.  
away\_team\_shotsontargetpershotstaken: continuous.  
away\_team\_shotstakenperavgposession: continuous.  
away\_team\_averagegoalsscoredpergame: continuous.  
away\_team\_averagegoalsconcededpergame: continuous.  
away\_team\_foulscommittedperOppPossession: continuous.  
away\_team\_cornersconcededperOppPossession: continuous.  
away\_team\_offsidesagainstperOppPossession: continuous.  
away\_team\_yellowCardsforperOppPossession: continuous.  
away\_team\_redCardsforperOppPossession: continuous.  
away\_team\_goalsconcededperOppShotsonTarget: continuous.  
away\_team\_totalshotsconcededperAvgPossession: continuous.

*Appendix B: Outcomes for features*