Predicting the Results of AFL Matches

Shaun Sewell

N9509623

Group 28

I. Introduction

In recent years there has been a dramatic increase in the volume of statistical information generated and gathered by various sporting leagues around the world, with the AFL being no exception. Where there is an increase in available information there is often a corresponding demand to use and profit from any insights that can be gleaned by the use of big data and machine learning methods. Within the AFL community, the key useful outcomes that can be predicted from data analysis is the performance of teams and outcomes of matches.

This report seeks to test the predictability of AFL matches by using an approach first applied to NBA matches by Aryan & Sharafat (2012). In doing so the intent is to show whether approaches that successfully predict winners in one sport can be directly applied to another sport.

II. Related Works

This project is based primarily on the paper, 'A Novel Approach to Predicting the Results of NBA Matches' by Aryan & Sharafat (2012). Their methodology predicted the statistical output of each team for every match played in order to predict the outcome; this differed from previously applied models in that it considered the predicted the performance of the opposing team in a match, not just the team in question (Aryan & Sharafat 2012:1). Overall, their model was effective and more comprehensively considered the expected result of a match, rather than merely a single team's ability to win a match. One potential result of using multiple predictions (as opposed to a single win-lose prediction of other models) is that the compounding error might decrease the overall efficacy of the prediction model. However, Aryan & Sharafat (2012) found that their error margin was in line with other single-outcome models with the added benefit of considering both teams in a match.

III. Data

The data used in this project was gathered from the AFL statistics website footywire.com pertaining to matches played from 2012 to 2018. For each game a set of 25 statistical features were gathered (see Table 1). There were additional statistic features available for some years, as the statistics gathering became more complex, however only the 25 features available for *all* seasons were used. The scraper also gathered results of the matches and the home team in each match to determine if the home game advantage was a factor in match outcomes.

Table 1: Statistical Features

Disposals	Scoring Shots
Kicks	Goal Assists
Handballs	Inside 50s
Marks	Rebound 50s
Tackles	Contested Possessions
Hitouts	Uncontested Possessions
Clearances	Effective Disposals
Clangers	Disposal Efficiency
Frees For	Contested Marks
Frees Against	Marks Inside 50
Goals Kicked	One Percenters
Behinds Kicked	Bounces
Rushed Behinds	

IV. Model Overview

The model proposed by Aryan & Sharafat (2012; Figure 1) is comprised of three components:

- Statistic Prediction;
- Feature formation; and
- Result prediction.

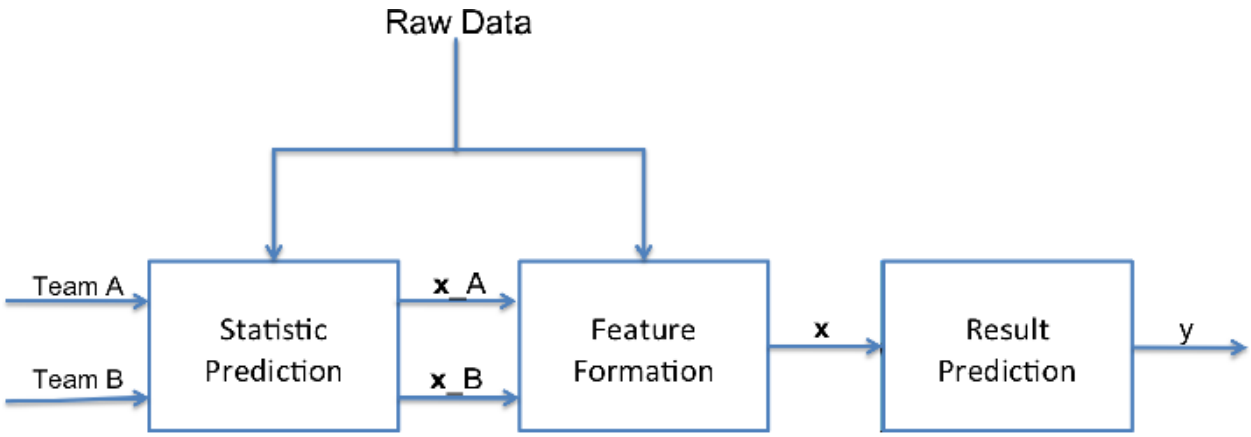

Figure 1: Model Overview (Aryan & Sharafat 2012).

Statistics prediction involves the predicting of the output of a team based on previous performances. Aryan & Sharafat's (2012) model involves four different predictive models:

1. Running average: a running average of the statistical features of the team in previous games.

2. Exponentially decaying average: similar to the running average except the value of every game in the average is weighted by a factor of $0 < \alpha < 1$ using the formula below.
$$g_{n+1} = \frac{1-\alpha}{1-\alpha^n}\sum_{i=1}^{n} \alpha^{n-i} g_i ,$$
where n is the number of games already played and g_i is the statistical features of the ith game.
3. Home and away average: same as the running average except it keeps tally of home and away games.
4. Cluster based prediction: predicting the outcome based upon the type of opposing team. To do this first PCA is performed in order t reduce the pool of features to only those of the most significance. Then k-means cluster is performed using various values of k. And finally using the cluster assignments a running average of each teams performance against every cluster is recorded.

Feature formation then takes the individual team statistics prediction and combines the two teams involved in any match. By doing this the feature vector will contain information about both teams playing a match as opposed to just a single team. The implementation for this is to simply take the difference of the two team predictions, i.e:
$$x = x_A - x_B$$
In the result component an output of $y = 1$ would represent a win for team A and $y = 0$ would be the inverse result.

Result prediction takes the feature vector formed in the previous step and uses it as input for the three machine learning algorithms assessed: linear regression, logistic regression and support vector machines (SVM).

V. Implementation & Results

Using the gathered data, the statistical predictions and feature formations were precomputed in order to reduce the computational demand required. By doing this, the feature vectors become independent of the source season data allowing for the creation of larger training and testing sets. Training and testing were conducted using the hold-out cross validation method with a training set containing 70% of the examples. The errors reported are the averages of 100 randomized iterations of this method.

The forward search feature selection algorithm was used to select the features that minimized the result error from each of the machine learning algorithms. Furthermore, the algorithms

were configured using the built-in Matlab functions which optimized the parameters. By doing this, a more consistent result can be generated across the various statistic predictions.

The results can be broken down into four sections each relating to the different statistic prediction methods.

Running average: The results of the running average are shown in Table 2. From the results we see that SVM is the better predictor by a tiny margin from logistic regression. With both algorithms being consistent from training to testing. Whereas linear regression performs well on the training set but based on its results from the testing set appears to overfit.

Home & Away average: The results from the home and away average (See Table 2) follow the same pattern as the running average with SVM being better overall but linear regression showing clear evidence of overfitting. The overall error across the classifiers is higher than for running average which seems to indicate some issues with the statistic prediction. The major issue is likely caused by the fact that quite a few teams share a home ground. This leads to the wrong average being used in many cases throughout the season.

Figure 2: Results from Running average & Home and Away average.

	Running Average		**Home & Away Average**	
Classifier	Training Error	Testing Error	Training Error	Testing Error
Linear Regression	0.3096	0.3323	0.3388	0.3715
Logistic Regression	0.3153	0.3222	0.3822	0.3895
SVM	0.3149	0.3211	0.3624	0.3724

Decaying average: The decaying average was predicted using a range of values for α. The results are shown in Figure 2. As can be seen in the results the closer α gets to 1 the closer the error gets to the running average error. This indicates that past match results are just as important as the most recent results. Also, the same overfitting of the linear regression is seen as α approaches 1. The few results around α=0.4 show the SVM classifier with a dramatically worse error than any other classifier this is likely due to the predicted features becoming less separable.

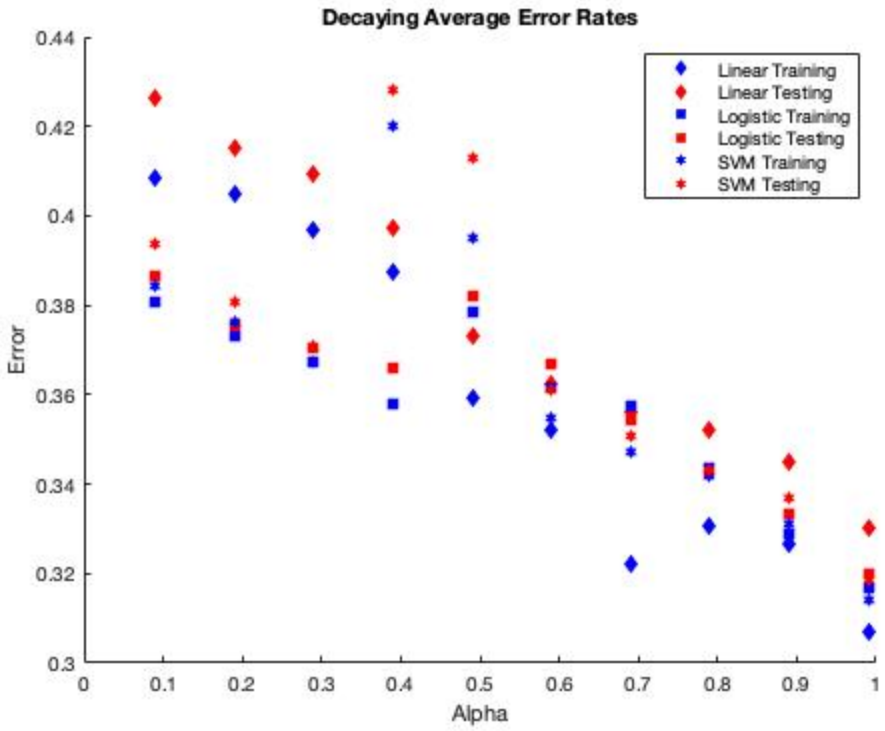

Figure 3: Error as a function of α.

Cluster-based average: For the cluster-based average, a range of clusters and PCA components were used during the statistic predictions. The results are shown in Figures 3-5. A trend emerges where the error rates decrease as the number of clusters increase. This indicates that the clustering of teams increases the accuracy of the statistic predictions.

Additionally, we see that as the number of components used is increased the variability of the errors decrease. This is likely due to the features becoming more separable as available information is increased. Overall logistic regression performs the best overall when using this prediction method with ten clusters.

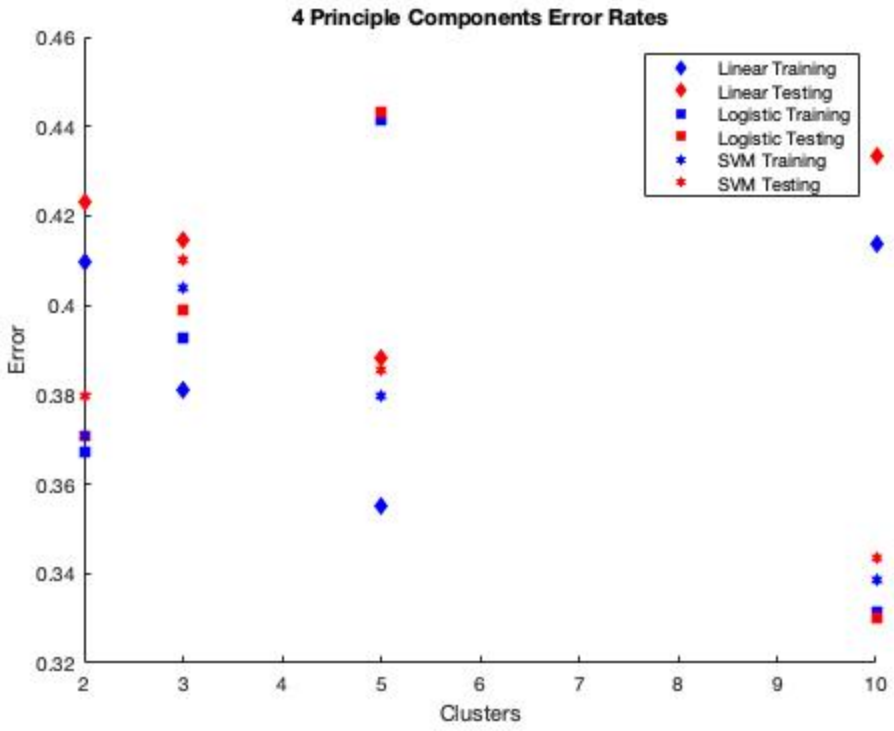

Figure 4: Four Principle Components.

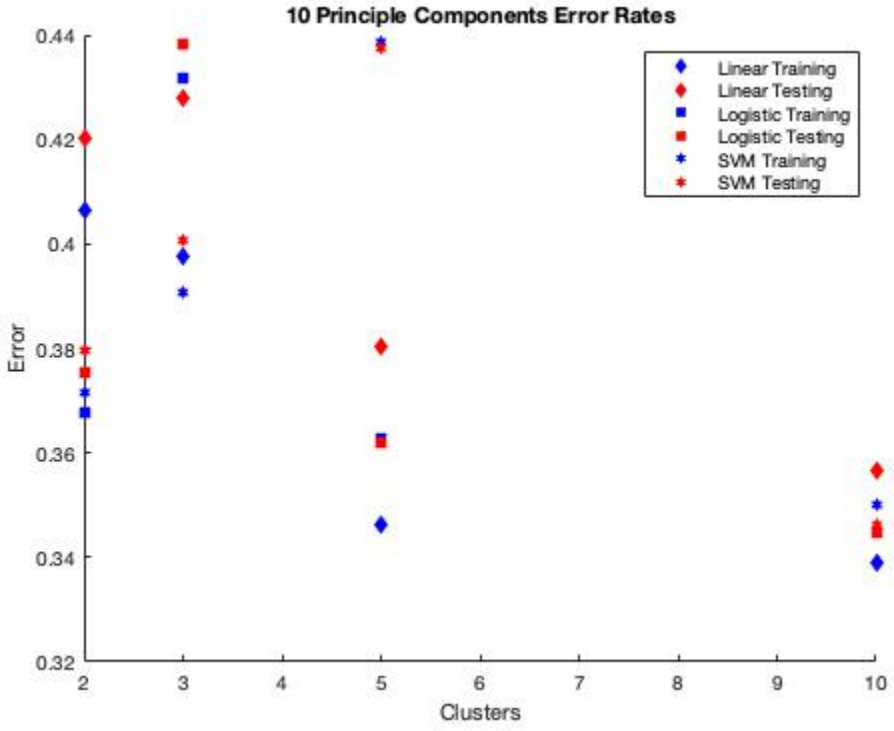

Figure 5: Ten Principle Components.

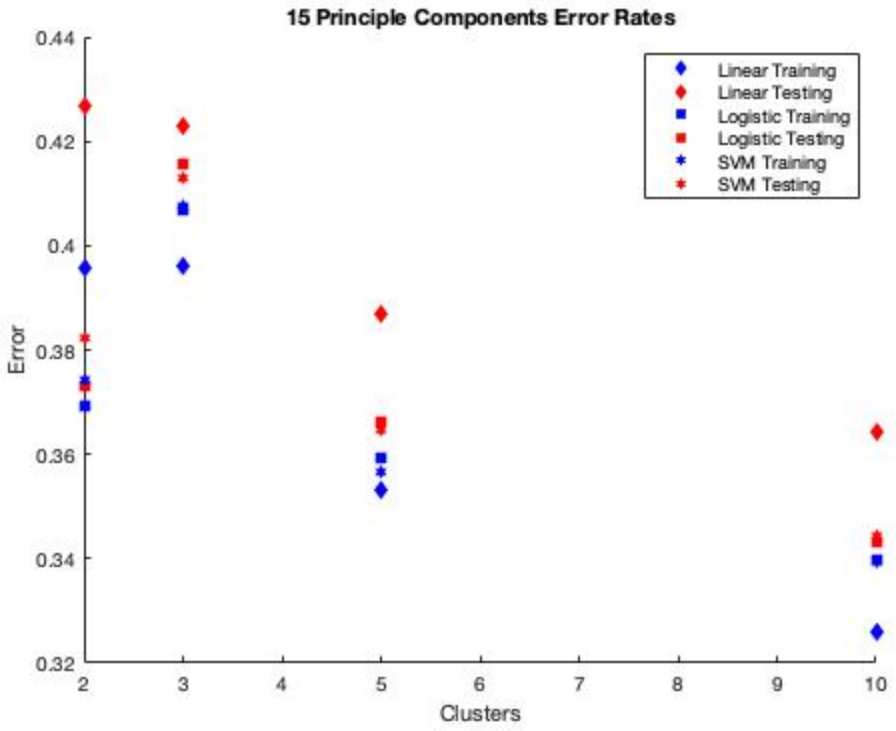

Figure 6: Fifteen Principle Components.

VI. Conclusion

In conclusion, the best classifier to predict the outcomes of AFL matches was the SVM using a running average, with a logistic regression a close second. The linear regression overfit the training set consistently as, as such, was the least accurate classifier. The source method and application to NBA games by Aryan & Sharafat (2012) yielded the opposite results – with linear regression being the most accurate classifier and SVM being the least accurate. This opposite result when applied to AFL statistics is likely because the statistical features, though fewer, are more separable. A more accurate statistics predictor based upon the individual players in each game may yield more accurate results, taking this model one step further.

VII. Reference

Aryan, O. and A. R. Sharafat 2012 *A Novel Approach to Predicting the Results of the NBA Matches*. Available online from http://cs229.stanford.edu/proj2014/Omid%20Aryan,%20Ali%20Reza%20Sharafat,%20A%20Novel%20Approach%20to%20Predicting%20the%20Results%20of%20NBA%20Matches.pdf.