PREDICTING BROWNLOW VOTES FROM GAME PERFORMANCE

Project Group 15

COMP20008 (Elements of data processing)

Elements of data processing Assignment 2 report
(AFL dataset)

Report contents

Research question

Within the Australian Football League the 'leading individual award' called the Brownlow medal (image 1) is awarded to the 'Fairest and best player' (AFL 2022). The decision making process behind the point system is as follows:

Points	1	2	3
Player	3rd best player	2nd best player	Best player

For these points how 'fair' and 'best' a player performs is determined by the field umpires at the end of each match and is not necessarily determined by a certain statistic such as number of goals.

The goal of this project is to find which game characteristics are most correlated with Brownlow votes to ultimately try and predict how many votes a player could receive based off their game performance. Determining which characteristics are important will become a valuable tool for not only for players who wish to increase their chances of winning the medal but also viewers of the sport giving them a better idea of what makes a player 'best'.

Image 1: The Brownlow medal (ABC 2022)
https://www.abc.net.au/news/2022-09-10/afl-moves-brownlow-medal-ceremony-due-to-queen-funeral/101426210

Dataset

For this project one dataset 'AFL-2022-totals' from afltabels.com was used. This dataset was given in the form of a CSV file containing a number of different statistics surrounding the sport including:

- Games played
- Kicks
- Goals
- Brownlow votes

for AFL players in 2022. In total 25 different statistics were provided all in integer form. Unfortunately, the dataset chose to omit 0's from the dataset instead leaving these values blank making it difficult to decide whether a value was missing or was in fact a 0.

Audience

This report is aimed at football players and viewers giving them a better idea of how to either improve their performance and likelihood of winning a medal, or giving viewers a better idea of what makes a player 'best' according to a field umpire.

Pre-processing (Wrangling techniques)

Since the dataset was already in a useful format (CSV), it required no modification before pre-processing although values such as player and team were of no use so were removed.

The first issue was missing values in the dataset with the features 'Hit outs' and 'Brownlow votes' missing the most. For these two features most of the missing values were most likely zeros since most players will not score 'Hit outs' (as this is reserved for the 2 'ruckman' per team) and only 3 players per game can score any Brownlow votes. An alternative method would be to make each missing value the mean of that feature however this does not make sense in this case as a player who has not earned a Brownlow point would now be considered as possibly being a 'fair' and 'best' player. The rest of the missing values also appear on players who score much lower in other features making it very likely that these values should be zero.

Another problem in this dataset is that each player appears to have played in a different number of games. To fix this issue each player's stats were divided by the number of games played to find an average for each statistic per game. Finally checks were made to ensure that each value was:

- A positive number (Since none of these stats could be negative or a letter, the presence of which would indicate an error in input)
- Less than 1000 (Values over 1000 would not be possible and again would most likely be a misinput)
- A possible value (The maximum possible Brownlow points per game is 3, more would not be possible. The maximum number of goals could not exceed number of kicks + free kicks)

No values were found to be outside of these ranges.

Analysis

Methods (Feature selection, Evaluation methods, Performance metrics)

This AFL dataset contains information for each player and player's game statistics. For example kicks (KI), marks (MK), handballs (HB), disposals (DI), goals (GL), hit outs (HO), tackles (TK), and contested possessions (CP), etc.

Initially, we cleaned and pre-processed our data to make sure it was uniform, complete, and consistent. After that, we conducted a correlation analysis to investigate how Brownlow votes were related to various game characteristics. Considering continuous data in our dataset, we figured out that we might need to apply techniques like linear regression and Pearson correlation to better analyze our data.

The correlation analysis (Table 1) showed a strong association(choose the correlation coefficients greater than 0.5) between Brownlow votes and a selection of variables: handballs (HB), disposals (DI), inside 50s (IF), clearances (CL), and contested possessions (CP). Therefore, these five variables were identified as principal predictors for our model.

Table 1: Pearson correlation coefficients of game characteristics with Brownlow votes

Variable	Handballs (HB)	Disposals (DI)	Inside 50s (IF)	Clearances (CL)	contested possessions (CP)
Pearson correlation (in 3 decimal places)	0.551	0.584	0.602	0.568	0.636

We created a scatter plot matrix (Figure 1) using seaborn library to obtain whether these features have a linear relationship with our target variable. Pairplot mainly shows the relationship between two variables. From the figure, we found that HB and CP have strong association with the other four explanatory variables, this is the problem of collinearity that makes the interpretation of regression much more difficult. From the last column that displays the relationship between the explanatory variables and response variable, which illustrates the relationships didn't show a good linear relationship and the transformation of variables seems to be necessary.

Before creating the linear regression model, we need to split our dataset into training and test sets for enhancing the generalizability of the model. We randomly split the dataset in a 7:3 ratio, 70% of data used to train the machine learning model, and 30% of data used to evaluate the performance of the model.

To account for potential non-linear relationships between the independant and the dependent variable, we applied a Polynomial transformation on the independent variables. Polynomial regression made the order of feature variables higher to improve the interpretability of the model so that it can better fit the non-linear relation.

Since our dataset after preprocessing still included plenty of zero values, we employed bootstrap sampling to handle imbalanced data and manufacture training and test sets. We chose 500 samples randomly from training data and repeated 500 times for each sample, then we reported its mean scores from the corresponding test sets to estimate performance results.

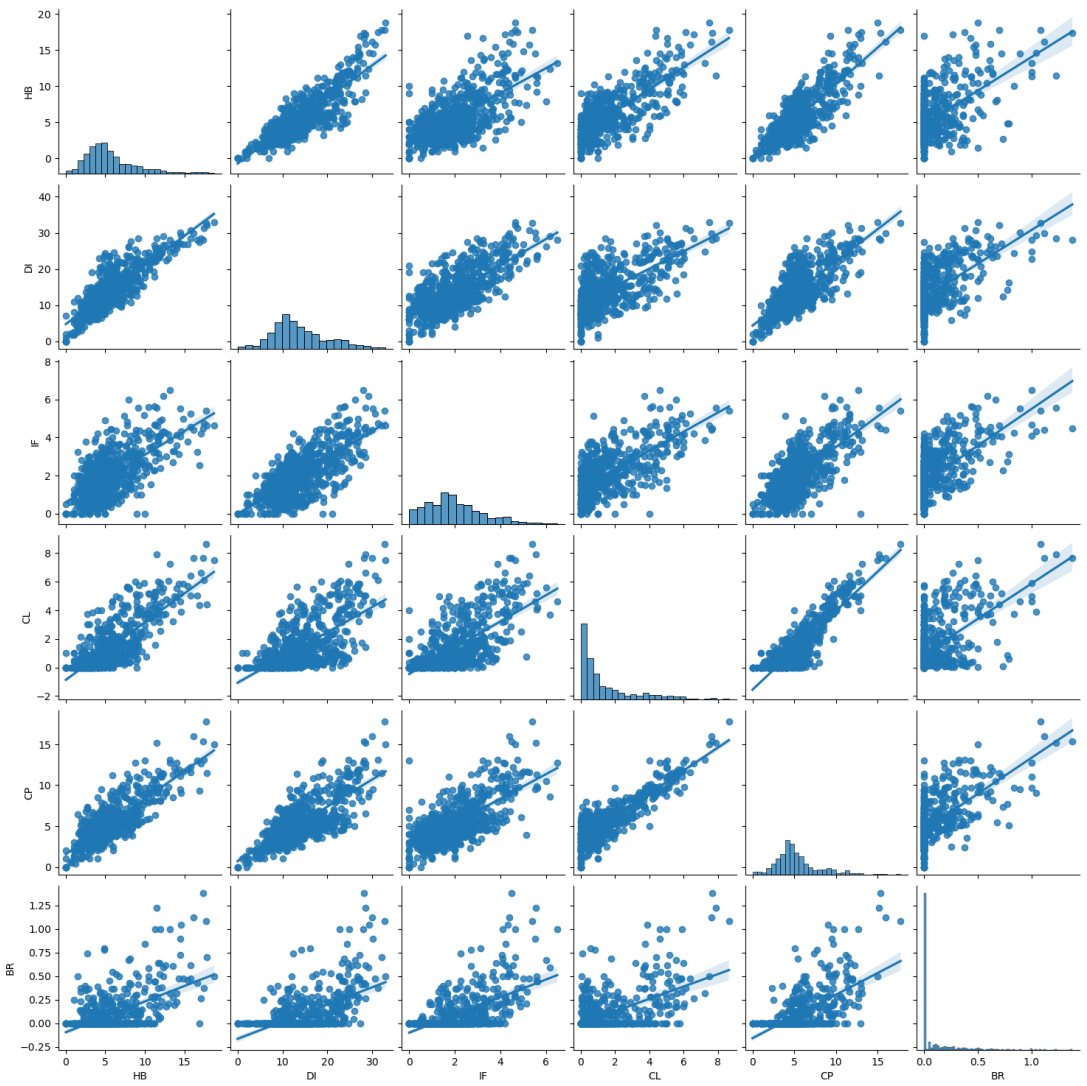


Figure 1: Pairplot of five primary features against Brownlow votes

Interpretation of results

Modelling

Firstly, we fit a multiple regression model with five relevant game characteristics as independent variables to better understand whether there is a significant influence to the Brownlow votes. And also using a Polynomial Regression with these features transformed to the second degree. We calculated the mean square error(MSE) and R^2 of training and test sets separately to see the model performance. Furthermore, note that there exists collinearity between the explanatory variables(see Appendix), we created another Polynomial Regression model with CP as the only explanatory variable since it has the highest correlation with the target variable. Subsequently, compared with the same results calculated by Bootstrap validation metric, which evaluated the reliability of the generalization ability of the model.

(In 3 decimal places)	Training R^2	Test R^2	Test MSE	Bootstrap(k=500) R^2	Bootstrap(k=500) MSE
Model 1	0.683	0.438	0.013	0.563	0.015
Model 2 (with CP only)	0.489	0.400	0.014	0.442	0.020

Table 2: Regression Model performance scores(MSE & R^2)

Discussion

We will interpret the model and discuss the performance scores. From Table 2 above, low MSE values for test set(0.013) and Bootstrap sample(0.015) indicate that our model shows a good performance. R^2 also increased after doing Bootstrapping which means a better fit of the model and a better performance, suggesting that there exists causality between these five game characteristics and our target variable, and this learning model could predict the votes in the future based on these characteristics. However, when we drawed the residual plot(see Appendix), we observed that the residuals didn't randomly distribute around 0 and it likely displayed some kind of pattern, which revealed that the assumptions for linear regression were not satisfied and this is contradictory with the results we obtained. In addition, the boxplot of Brownlow votes showed that it has too many outliers that lead to a negative effect on the generalization ability and accuracy of our model. Therefore, the valuable discoveries tend to suggest that the linear regression model may not be the most appropriate learning model for predicting the number of Brownlow votes a player would receive.

Evaluation(Improvement)

Since the significant proportion of outliers affected the model prediction, we attempted to change the supervised learning method with the K-nearest neighbor classifier that treats the Brownlow votes as a discrete variable. Given that the dataset after cleaning represents the average statistics for each player each game, which means the number of the Brownlow votes each game is either 0 or 1 for each player(according to our dataset). Hence, we applied Mutual Information to deal with non-linear relationships and implemented the discretization(Binning) to the target variable and into 2 bins. We obtained the correlation between variables as well. Furthermore, we built the K-NN model that included the same features to predict the Brownlow votes. Similarly, we did training-test sets split, and we chose the hyperparameter k with 3. Then we evaluated the model performance using accuracy metric and found that the value is quite high(0.995) which is almost 1. Recall that accuracy metric is a misleading indicator of performance on imbalanced problems, then alternative metrics such as precision or recall may be appropriate to evaluate. Finally, we used the Bootstrap sampling approach(500 samples from the training set) to get a range of performance scores(Table 3). In this case, we consider that the precision may be more appropriate than recall since we need to avoid the mistaken identification. Low precision rate is likely to indicate that the model couldn't predict accurately, possibly because of the imbalanced data problem and there is no obvious discrimination between labels. Confusion matrix could also perform the imbalanced problem(see Appendix).

Table 3: Classification Model Performance metrics values for Bootstrap samples

	Mean Accuracy	Mean Precision	Mean Recall
Bootstrapping (In 3 decimal places)	0.973	0.339	0.197

Limitations and future improvements

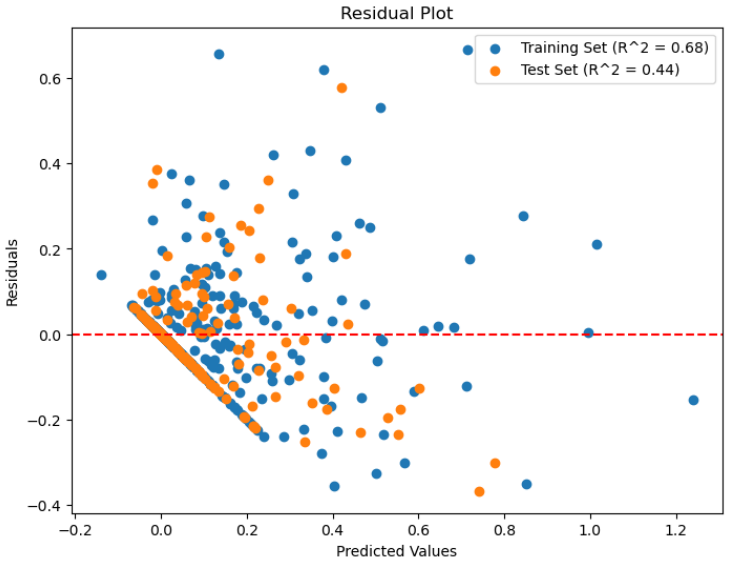
There's still problems with our model. Firstly, we have not processed discrete values. Due to the problem of data scale, if we remove all discrete values, the data scale of this model will be reduced by more than 60 percent, which may cause this data to lose the significance of analysis. Voting is the result of multiple votes, which is categorical data. This dataset could not obtain the data of separate votes. In addition, objective factors of voting also had a lot of influence on the data, such as the popularity of the team or players and the opponent faced by the team members, which would affect their brownlow votes. Hence, in the future, if we can get more relevant data such as team popularity that are factors that affect people's vote, we can then provide a more precise prediction. Secondly, during the process of discretizing the data into multiple bins, we didn't create 3 or more bins(the maximum votes is 3 each game though the dataset didn't appear that) since this is more difficult to analyze its performance which is beyond our ability. Finally, based on the categorical attribute of votes, other models such as logistic regression model may be more suitable.

Appendix

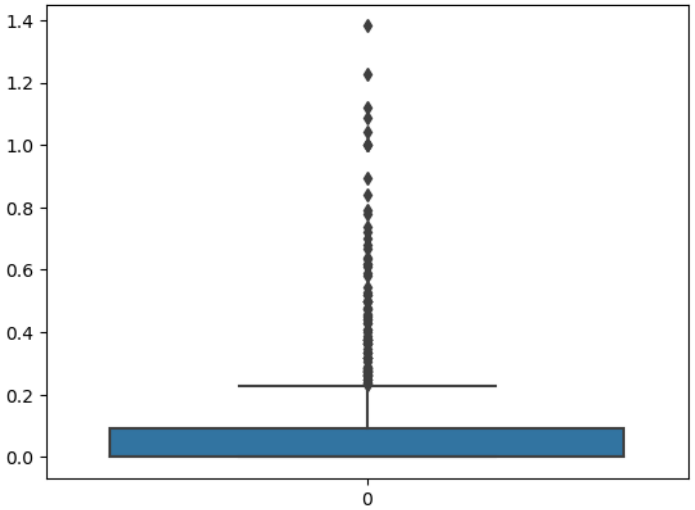
Correlation between explanatory variables

	HB	DI	IF	CL	CP	BR
HB	1.000000	0.859024	0.656623	0.783693	0.826426	0.550714
DI	0.859024	1.000000	0.725310	0.651273	0.768870	0.583512
IF	0.656623	0.725310	1.000000	0.678611	0.696912	0.601772
CL	0.783693	0.651273	0.678611	1.000000	0.876429	0.568212
CP	0.826426	0.768870	0.696912	0.876429	1.000000	0.636283
BR	0.550714	0.583512	0.601772	0.568212	0.636283	1.000000

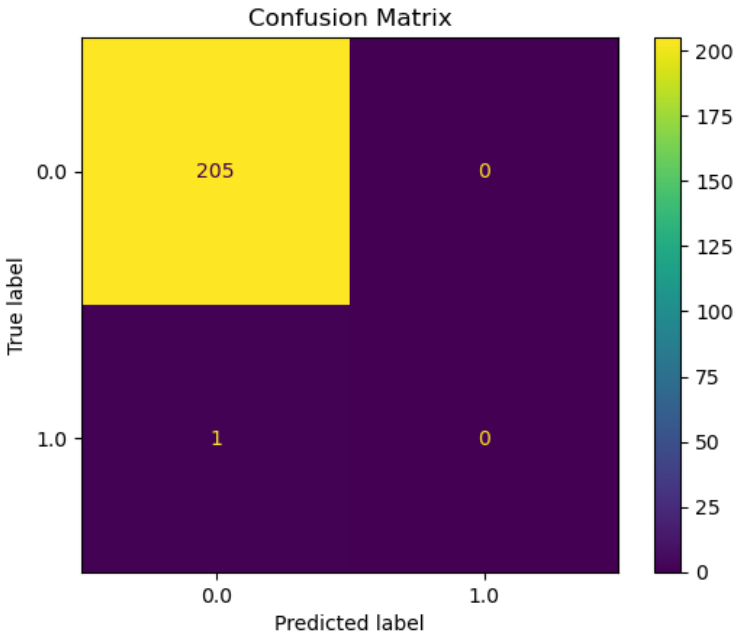
Residual plot



Boxplot of Brownlow votes



Confusion matrix



Resources

Brownlow medal history – AFL (2022) https://www.afl.com.au/brownlow-medal/history