Data Mining in Sports

Implementation and analysis of machine learning prediction models to find potential influential factors in results of matches in the Australian Football League

Bachelor Thesis

Submitted at the
IMC Fachhochschule Krems
(University of Applied Sciences)

Bachelor Programme Informatics

by

Thomas Gallagher

for the award of academic degree
Bachelor of Science in Engineering (BSc)

under the supervision of
Dr. Deepak Dhungana

Submitted on 02.02.2023

Declaration of honour

I declare on my word of honour that I have written this Bachelor Thesis on my own and that I have not used any sources or resources other than stated and that I have marked those passages and/or ideas that were either verbally or textually extracted from sources. This also applies to drawings, sketches, graphic representations as well as to sources from the internet. The Bachelor Thesis has not been submitted in this or similar form for assessment at any other domestic or foreign post-secondary educational institution and has not been published elsewhere. The present Bachelor Thesis complies with the version submitted electronically.

Thomas Gallagher
02.02.2023

Abstract

Despite the large amounts of data produced by the sporting industry every year, there has been relatively little crossover between academic data science and professional sports. On the other hand, Data Mining techniques such as Machine Learning, Neural Networks and Association Methods, have seen rapid increases in their complexity and use over a wide variety of fields and disciplines. This paper aims to address this issue of a lack of data science applications in sports, with regards to the Australian Football League (AFL), by applying Data Mining techniques to improve upon already utilised data analysis techniques present in modern professional sports. In this paper statistics and results from the previous 10 AFL seasons will be assessed to identify key features and possible trends and use them to create more explainable white box prediction models. This will help not only sports professionals and data scientists but also casual viewers to understand the finer statistical details behind sports.

Keywords: Data Analytics, Machine Learning, Sports Prediction, Neural Network

Table of Contents

List of Tables

1
Introduction

This paper will apply Data Mining techniques on Australian Rules Football data in an attempt to identify any trends that may occur in individual seasons and then further apply this information into a Neural Network to create a prediction model for subsequent games.

The field of Data Mining has seen a rapid increase as we enter the digital age dominated digital information also commonly known as data. While many believe data mining to simply be the extraction of data it is actually a much broader topic. Data extraction is only the first step in data mining, the goal of data mining is to extract 'previously unknown' and unseen patterns from this data [1], much of which would be impossible to uncover without the help of computer systems. Data Mining is most commonly applied in commercial business. There it can be used to 'help identify and predict individual and aggregate behaviour' [2], or more commonly to predict how customers will consume/buy products and their buying habits so that they can more efficiently market their products in specific situations. Data Mining in sports however is much less advanced, despite massive amounts of data being produced by the sporting industry, it is mainly kept behind closed doors, with much of it focusing on player and team performance.

That's not to say that there have been no recent attempts at data mining in sports but much of this has been in, classification [3], action recognition [4] and image recognition [5]. While more have aimed to create prediction models [6][7][8], none of these models aim to identify trends within the sports apply these as part of their prediction model. In Australian Football, one of the most data rich sports [9], data is used in many ways from real time analysis of players to modelling games based on Geographical Positioning System (GPS) data and also in prediction models. An issue with many prediction models is the lack of transparency in their decision making, a feature of this is the idea of black box and white box models. Black box models are more accurate than white box model but much more difficult to interpret to see how they came to certain decisions. In this paper a combination of models will be used to attempt to overcome this lack of interpretability. These models will be used to extract meaningful interpretable data on trends in Australian Football and then feed this information into a stronger prediction model. Teams and fans alike will be able to see what has the greatest effects on their performances and how to overcome these factors.

As this paper is discussing Australian Football a basic background should be given to understand some of the terms and structure of the game, as well as to understand the structure of the data set.

1. On match day a team is made up of **22 players**, 18 on the field and 4 reserve players

2. The aim is to score more points than the opponent by kicking the ball through either the **goals**, worth 6 points, or **behinds**, worth 1 point.

3. Each team plays **22 games** per season, each season occurs between March and October in the Australian winter.

4. The finals are contested between the top 8 teams from the regular season, the finals series lasts for **4 weeks**

5. There are currently **18 teams** in the competition.

6. 10 teams are based in Victoria. New South Wales, Queensland, South Australia and Western Australia each have 2 teams in the competition.

The final stage of the practical work will be to input the analysed data into a Neural Network to create a prediction model. Neural Networks are computing systems which aim to 'mimic the way that biological neurons signal to one another'[10]. Neural Networks have a multitude of uses in a variety of fields including medical diagnosis by image classification, targeted marketing and financial predictions through behaviour data analysis and financial data processing, natural language processing and time series analysis. Neural Networks in sports...

Why will this be useful

Deeper analysis of trends and features

Uncover new features

Improved prediction

2

Objectives and Research Questions

2.0.1 Research Questions

The area of focus of the paper is the Implementation and analysis of machine learning models to find potential influential factors in results of matches in the Australian Football League (AFL). From here I have defined the following research questions which will be explored:

> Can data mining be used to find, key features and patterns and explain results in professional sports?
>
> Can these extracted trends and features be used to create an explainable prediction model?
>
> How can the results be utilised in the sporting industry?
>
> If the model can reliably predict outcomes, how many rounds will the model need before it becomes reliable?

2.0.2 Objectives

The first section of this paper will be the exploration of the data set with the aim of identifying trends and important features within the data. Important features are features that have a greater impact on the final result of a prediction model. Trends will be defined as features or feature patterns that occur in multiple seasons. Both known and unknown features and trends will be extracted and from the data set. Known features will be defined as factors that are currently used in analysis of sports games, which stem from less scientific analysis of data. These known features are generally used in predictions of games by both domain experts and casual observers and are widely accepted to be true, but there has been little research to provide evidence to the claims. The unknown features and trends will be discovered during the data exploration and will be represented by the important features, the factors which have the greatest effect on the outcome of the model, extracted during the feature engineering process.

In sports analysis there are many factors that are said to have an impact on results by domain experts and casual viewers alike, however there has been little effort to back up these claims using data science techniques. The idea behind exploring known trends or features is to analyse whether there is any truth to these claims in, it is also useful to take domain knowledge from experts, even if it second hand as in this case. The main points that will be explored from this side will be the effects of weather conditions, length between games and travel on future results. Both are suggested to negatively impact a team's future performance. In the case of wet weather, it is believed that it can negatively affect a team's performance in future weeks, as the game becomes more contested thus requiring a greater physical output. While with travel it is suggested that spending long times in planes both before and after games can affect a teams' recovery and preparations.

While trends can be extremely useful for prediction they can also be difficult to interpret and extract from data, which another reason why the expert analysis is being explored. It gives a starting point from which the data can be analysed. The further extraction of trends will be entirely theoretical and based solely on interpretation of the data set. This will be very important for the project as it will hopefully enable unseeable trends to be found, rather than just proving or disproving already voiced opinions. These trends would be much more valuable in the industry as they would give professionals a completely new insight into their analytics.

The second key part will be the implementation of a prediction model. Prediction models are used regularly in sports. However as there are so many unknown factors in these prediction models it is very difficult to explain their predictions and how they came to these decisions. The aim is to create a prediction model based on the extracted trends and important features, that is more explainable than other models. Assuming that the prediction model does reliably work, one more key piece of information will be extracted, the final research question, what number of rounds/games does the model need before it can reliably predict the outcomes of games? This can be useful not only for this model but for other prediction models for the AFL, as it can give a greater understanding of when they will become reliable during a season. There are many algorithms and techniques that can be used in Data Mining and sports, this paper will mainly utilise the following:

> Linear Regression: Models the relationship between an output (dependent variable) and one or more input values (independent variables)
>
> Decision Trees: A tree like predictive models that shows where certain decisions were made via nodes.
>
> Random Forest: A collection of decision trees, that uses a random subset of the data in each iteration.
>
> Long Term Short Term Neural Networks (LSTM): A type Neural Network that introduces the concept of memory and feedback connections allowing it to process data sequences.

The dataset was taken from Kaggle (https://www.kaggle.com/datasets/stoney71/aflstats), it is a comprehensive dataset, containing statistics from every AFL game between 2012 and 2021. It is comprised of 3 tables, Games (containing game data) has 12 columns, Players (containing player information) has 7 columns, and Stats (containing individual player statistics for every game) has 31 columns. Overall there are over 90000 data points across the datasets.

However there are some values that will be sourced from other websites to enhance the dataset, Fantasy Score and Supercoach Scores (player ratings after a game) will be taken from footywire.com. Travel time data will be manually updated. Days between games will be calculated via a python script in pre-processing.

The project will be completed using python as the main programming language. There will be numerous libraries used during the project, but they will be identified during implementation, at this stage it is impossible to know every library that will be used.

3

Scientific Method

3.1 Description of the Scientific Method

3.1.1 Phase 1: Feature Engineering

The first phase of the paper is the Data Mining on the entire dataset using Regression and Random Forest models to extract important features and trends from the dataset. This is central to answering the first research question, can data mining be used to find, key features and patterns and explain results in professional sports? These models will be built on the entire featured dataset in multiple independent tests utilising different sections on the data. The main focus will be on the games table as it will be used in the second model, the experimentation on the other tables will be done to extract importance values for individual player data and team selection. Each player will have a performance score calculated for each game based on table 1. The importance values will be a standardised value between 1 and 10. It will be calculated using a players previous 5 games with most recent games having higher weights (1, 0.8, 0.6, 0.4, 0.2) and diving it by 3 (the sum of the weights) and added to the games table as “Home Team Player Importance” and “Away Team Player Importance”

Once the values have been defined the model can be built and trained. For testing a new table will be created building on the games table. It will contain a combination of the features: Disposals, Kicks, Handballs, Tackles, Free Kicks, Hit Outs, Contested Possessions, Player Importance, Marks, Contested Marks, Inside 50s, Rebounds, Clangers, Team Name, Travel Distance and Break Between Games for both teams as well as Rainfall. This will try to predict which team won. 20 percent of the data will be held out for testing purposes.

Stat	Value +/-
Kick	4
Handball	2
Mark	3
Goal	6
Behind	1
Hitout	2
Tackles	3
Rebounds	1
Inside 50s	2
Clearances	3
Clangers	-3
Free Against	-3
Contested Marks	3
One Percenters	3
Goal Assist	2

Table 3.1: Players count for each stat is multiplied by the value.

Evaluation in this phase is the feature importance metrics. Random Forests will be used for finding feature importance as they are naturally capable of discerning important features. Ideally 5 features would be identified that consistently are found to have the greatest impact on predicting the result of a model and the 5 features that consistently have the least impact on the result. To realise this outcome a random forest model containing 100 trees and the 5 metrics that were found most and least important on average will be taken as the fitting metrics.

3.1.2 Phase 2: Predictive Model

The second phase has to do with the building of a prediction model, using the knowledge of important features gained from the previous phase. The prediction model will be created using a LTSM Neural Network. The data acquisition and pre-processing will be much shorter as the majority had been done previously in the first phase of the project. Again a table will be built that is similar to the table used in the random forest model. However, the 5 least important features will be removed (and the corresponding feature for the opposite team). The scores of the home and away team will be included in the dataset as well as the round number and year. The dataset will then be separated by year and ordered by the round. Because the model will be predicting future events the training and test data split will not be static. The initial tuning phase will hold out only the final round, making the model theoretically as powerful as possible. But in different iterations more or less rounds will be held out, this is crucial to see if there is a certain point where the models begin to see an increase in their predictive powers. Evaluation on the models The main metric I will use to evaluate my model is the Mean Average Percentage Error (MAPE), which is “the proportion of the average absolute difference between projected and true values divided by the true value” [11], and the Weighted Mean Average Percentage Error (WMAPE) will be used if the dataset is too small to calculate the MAPE properly. Previous prediction models utilising machine learning have been able to reach a best accuracy of 68.1 percent and average accuracy of 65.1 percent [12]. Knowing that models are capable of reaching these accuracies, the target accuracy of the prediction model will be greater than those values, ideally above 70

4

Theoretical Background

4.0.1 Data Mining

Although data mining is not a big subject in sports it is a major influence in many other industries. Data Mining has a very broad use case in both Supervised and Unsupervised learning. Supervised learning is a subcategory of machine learning that deals with data where the input and output is known, as is the case with the data in this paper, it aims to map input data to resulting output data. Supervised learning can be further divided into two categories Classification, where the output is a class or category and Regression where the output is continuous.

There are many state of the art data mining techniques in both Classification and Regression. There have been some Classification Methods already mentioned in the paper, Decision Trees and Neural. Other widely used algorithms include K Nearest Neighbour (KNN), Support Vector Machines (SVM) and Bayesian Methods [13]. The K Nearest Neighbours attempts to classify a data point based on the class of its Nearest Neighbour points, a SVM is a method that aims to define boundaries to separate a space into classes. Bayesian Methods are more complex it is based upon a method that “combines prior information about a population parameter with new evidence from information contained in a sample to guide the statistical inference process” [14], the basis of the inference is gained through an application of Bayes Theorem [14]. The main application of these algorithms is to forecast how a certain input will act based on their characteristics, in the finance industry this can be deciding how likely a customer is to default on their payments based on their previous behaviours, in the medical industry these can be used to predict how likely a patient is to have a certain medical condition.

Regression techniques differ from classification techniques as they aim to forecast a continuous variable. Common techniques include Linear Regression, Multi Linear Regression and Polynomial Regression. Linear and Multi Linear regression are both similar in that they aim to predict a variable Y from a predictor or set of predictors in the case of multi linear regression. Polynomial Regression is used when the resulting Y variable is on a non-linear scale, and a straight line of best fit cannot be used to describe the outcome as is the case in linear and multi linear regression. Regression techniques are used in various fields in the medical field they can be used to predict the likelihood of a patient to survive a certain medical condition [13] or for predicting the value of houses based on the features of its geographical location.

Clustering techniques differ from the previous two as they are unsupervised techniques, meaning that only input variables are given and the algorithms must find hidden patterns within the data that connects the points. Clustering techniques generally fall into two categories, distance based and density based. Distance based methods look at how data points fit into a general area around them, while density based methods are more fluid and look at how data points sequentially connected based on density and separated based on low density. Distance based methods include K Means and Hierarchical clustering. The K Means algorithm an arbitrary number of clusters must be predefined, the data set is then divided into the predefined number of clusters by a similarity metric and the centre of each cluster is shifted until it is in an optimal position. In hierarchical clustering there are two approaches, Divisive and Agglomerative. Divisive is a top down approach, every point starts in one cluster which is divided through each iteration. Agglomerative is the opposite, every point starts as its own class which are then merged with similar classes through each iteration. These methods struggle with arbitrary shapes, which is where density clustering is useful. Density Clustering as the name suggests considers the density of data points rather than a spherical area used as the distance based methods use. Clustering methods are used in industries to identify patterns in data, it can be used to separate different stages of diseases based on provided images [13] in the medical industry, or in much more advanced fields such as document recognition.

4.0.2 Data Mining in Sports

Although there are many state of the art use cases of data mining, they have limited use in the field of sports. As discussed in “Sports Data Mining” by Robert P. Schumaker et al. [15]. This text looks at which data should be collected to properly performing data mining in sports and “how to best make use of it”. It also discusses the issues faced by those entering data mining in sports not only in the initial phases but also the final phases of data mining. One of these difficulties is the relationships between sports organisations and their data, many were unwilling to embrace data mining techniques at the time of the publication of the book. They defined 5 levels of relationships between organisations and the data they produced. Shown below in table 2, taken from Chapter 1, page 2 from the book.

Level	Relationship
One	No relationship
Two	Human domain experts make predictions using instinct and gut feeling
Three	Human domain experts make predictions using historical data
Four	Use of statistics in the decision-making process
Five	Use of data mining in the decision-making process

Table 4.1: Hierarchy of Sport and Sport Data Relationships.

Schumaker claims in the text that at the time of publishing many organisations resided in level 3 or 4 of the hierarchy. The issue is that there is currently no straightforward way to identify if organisations are using data mining and how they are using it. Another issue highlighted in the text is the inconsistencies and misinterpretations of statistical sports data. As is pointed out in the text many long standing statistics are misleading in modern contexts where we now have a better grasp on these methods and much greater available computing abilities. However, these methods still remain prominent and in use today, as such the book seeks to define new methods to ensure proper algorithms and applications of data analysis and data mining are used. In the book the Data-Information-Knowledge-Wisdom Hierarchy (DIKW) is used to separate techniques and use cases into each of the categories, to identify the best methods and algorithms to be used on each use case.

5

Related Works

5.1 Related Work

Moving into the field of academic papers there are only a handful of resources that explore similar issues to this paper. The most common areas of focus are American Football, Basketball and Baseball.

Carson Leung and Kyle Joseph, explore a prediction model for American Football in their paper, “Sports Data Mining: Predicting results for the college football” [6]. They employ an interesting dropout technique for their prediction model, in which they do not explore the results of the two teams that will be predicted, instead focusing on “a set of teams that are the most similar to each of the competing teams” [6], [page 716]. Similarly, they identify statistics on which the final model will be based, but only 4 instead of the larger set as in the model in this paper.

On the more analytical side is the paper “Sports analytics— Evaluation of basketball players and team” by Vangelis Sarlis and corresponding author Christos Tjortjis. This paper looks at and “evaluates the existing performance analytics used in Europe and NBA (in USA) basketball” [7]. They identify two metric types, Player and Team, and four criteria groups; Key Performance Indicators (KPIs), Defensive criteria, Offensive Criteria and Overall Performance Criteria, as well as outlining a Comparison Matrix and Data Mining Techniques used in sports analysis. They also explore if they can optimise existing performance analysis metrics that are used in basketball.

An overall review of the current techniques used in sports prediction (as of 2013) is given in the paper “A review of Data Mining Techniques for Result Prediction in Sports”, M. Haghighat et al., [16]. 6 Methods are analysed in the resource of which two, Artificial Neural Networks and Decision Trees are used in this paper. Bayesian Method, Logistic Regression, Support Vector Machines and Fuzzy Methods are the other methods that are analysed. Since nearly 10 years have passed since this resource was published, there has been many improvements to the models that are used in this paper, specifically in the field of Neural Networks.

Very little comprehensive work has been completed regarding Data Science and Australian Football. In 2008 McCabe and Travathan briefly explored the usability of neural networks in predicting the outcomes of sports in their conference paper “Artificial Intelligence in Sports Prediction” [12]. Here they reached an average accuracy of 65 percent using 11 performance metrics. Other studies conducted on AFL data include “Using meta-regression data mining to improve predictions of performance based on heart rate dynamics for Australian football” [17], which focuses on individual player performance as opposed to team performance prediction, but does follow a similar scientific method, utilising Random Forest algorithms to extract key features from the dataset, which is more complex featuring GPS data on top of more traditional performance measures. Also utilising GPS data is “The effect of team formation on defensive performance in Australian [18], which analysed a team’s defensive positioning and the effect it had on defensive performance. These both used similar techniques but the dataset differed from what will be used in this paper.

6

Summary

6.1 Summary

Overall this paper intends to expand the understanding and use of Data Mining in Professional Sports, specifically the Australian Football League. It aims to combine and build upon methods already being used in sports data mining to explore the following research questions:

> Can data mining be used to find key features and patterns and explain results in professional sports?
>
> Can these extracted trends and features be used to create an explainable prediction model?
>
> How can the results be utilised in the sporting industry?
>
> If the model can reliably predict outcomes, how many rounds will the model need before it becomes reliable?

In the paper Decision Trees, Random Forests, Linear Regression Models and LSTM Neural Networks will be utilised to extract important features and trends in the Australian Football League and use them to implement an explainable prediction model.

A two phase approach will be taken to create the final model. The first being feature engineering and trend analysis, using the Decision Tree and Random Forest techniques and Linear Regression. This will be done iteratively until the 5 most important and least important features are identified. The second phase will be the building of the prediction model, utilising the results from the first phase, with the aim of achieving an average accuracy greater than 65.1 percent.

Bibliography

[1] TechGuy, “Differences between data mining, machine learning and deep learning,” 2021. [Accessed June 12. 2022].

[2] E. P. D. P. C. Apte, B. Liu and P. Smyth, “Business applications of data mining,” *Communications of the ACM*, vol. 45, no. 8, pp. 49–53, 2002.

[3] H. Weiwei, “Classification of sport actions using principal component analysis and random forest based on three-dimensional data,” *Displays*, vol. 72, no. 102135, pp. 1–9, 2022.

[4] Y. Men, “Intelligent sports prediction analysis system based on improved gaussian fuzzy algorithm,” *Alexandria Engineering Journal*, vol. 61, no. 61, pp. 5351–5359, 2021.

[5] M. I.-K. Kristina Host, “An overview of human action recognition in sports based on computer vision,” *Heliyon*, vol. 8, no. 6, pp. 1–25, 2022.

[6] K. W. J. Carson K. Leung, “Sports data mining: predicting results for the college football games,” *Procedia Computer Science*, vol. 35, no. 35, pp. 710–719, 2014.

[7] C. T. Vangelis Sarlis, “Sports analytics— evaluation of basketball players and team performance,” *Information Systems*, vol. 93, no. 101562, pp. 1–19, 2020.

[8] Qiyun Zhang, Xuyun Zhang, Hongsheng Hu, Cauzhong Li, Yinping Lin, Rui Mae, “Sports match prediction model for training and exercise using attention-based lstm network,” *Digital Communications and Networks*, 2020. https://doi.org/10.1016/j.dcan.2021.08.008.

[9] VICE Sports, “Analytics in the afl - the most data rich sport on earth,” 2016. [Accessed June 13. 2022].

[10] IBM, “What are neural networks?.” [Accessed March 17. 2023.

[11] V. Lendave, “A guide to different evaluation metrics for time series forecasting models,” 2021. [Accessed June 13. 2022].

[12] J. T. Alan McCabe, “Artificial intelligence in sports prediction,” in *Fifth International Conference on Information Technology: New Generations (itng 2008)*, IEEE, 1996. 10.1109/ITNG.2008.203.

[13] S.Yamini, Dr.V.Khanaa, Dr.Krishna Mohantha, “A state of the art review on various data mining techniques,” *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 5, no. 3, pp. 2802–2808, 2016.

[14] E. Gregerson, “Bayesian analysis,” 2011. [Accessed June 14. 2022].

[15] H. C. Robert P. Schumaker, Osama K. Solieman, *SPORTS DATA MINING*. Springer Link, 2010.

[16] N. N. Maral Haghighat, Hamid Rastegari, “A review of data mining techniques for result prediction in sports,” *Advances in Computer Science: an International Journal,*, vol. 2, no. 5, pp. 7–12, 2013.

[17] D. J. R. A. S. D. J. C. Herbert F. Jelinek, Andrei Kelarev, “Using meta-regression data mining to improve predictions ofperformance based on heart rate dynamics for australian football,” *Applied Soft Computing*, vol. 14, pp. 81–87, 2014.

[18] L. B. D. B. D. Mitchell F. Aarons, Christopher M. Young, “The effect of team formation on defensive performance in australian football,” *Journal of Science and Medicine in Sport*, vol. 25, pp. 178–182, 2022.