Data-Driven Risk Assessment for Team Performance Improvement:

A Comprehensive Analysis and Machine Learning Approach

231265
Celine Wu
January 26th 2024
Applied Data Science and Artificial Intelligence

Breda University
OF APPLIED SCIENCES

Index

1 Introduction

The client, NAC, came to us with the business problem of improving team performance and overall team success. When it comes to improving as a team the players are extremely important. My approach to the client's business problem is to use historical data, in particular the occurrence of red cards and yellow cards to develop a risk assessment method. The models evaluate and categorise players into distinct risk groups, high risk, risky, and low risk. Understanding a player's behaviour on the field can potentially impact the team's performance. By implementing a data-driven approach, the team can get actionable insight and create strategic decisions to improve the team as a whole.

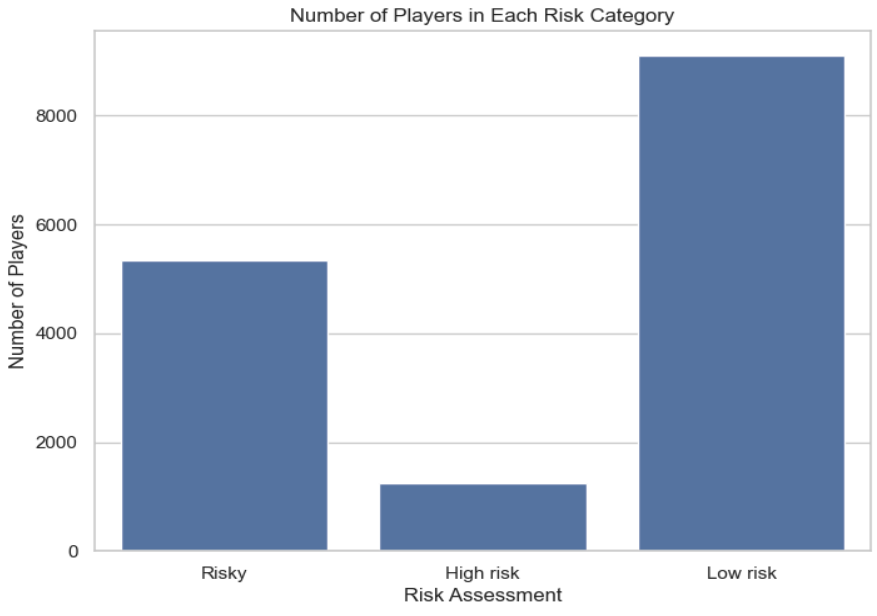
2 Exploratory Data Analysis

The dataset, provided by Breda University of Applied Sciences, contains 16,535 records and 115 features. This dataset contains various different feature types, such as numerical and categorical. A couple of examples of numerical features would be the red cards and yellow cards; the categorical features include team and positions.

I started exploring my dataset by identifying how many missing values there were. After identifying the column with the most missing values, I looked deeper into why there are so many missing values. Many columns had missing values due to the position of the players, such as the "Goal" column. Full-Back players are defensive players, so they typically will not have any goals, which, as a result, the column will contain missing values. Missing values in the "Goal" column and other columns alike were replaced with zero. The column "Contact_expires" was unique; I could not just replace the NaN values with zero. I was not sure how to deal with the missing data, so I decided to visualise it. By visualising the data, I saw that there were a couple of outliers. Based on that, I decided to use the median of "Contract_expires" to replace the missing values. This is because the median is less sensitive to outliers compared to the mean. While "Contract_expires" did contain a couple of outliers, it was not significant enough for me to drop them. The last column with a lot of missing data was the "Foot" column. I decided to replace the missing values in the foot column with the mode. This is because which foot a player uses does not impact their risk assessment. Lastly, rows with many missing values got dropped. I set the threshold to 115, meaning that all rows with at least 115 non-Nan values would be kept. My last step was creating a new column, which I named "Risk_assessment." This column was to be used as the target variable in the different models. As this column and many other columns

are categorical, I had to encode it. Machine learning algorithms require numerical inputs; by using a label encoder, I assigned a numerical input to each category.

While exploring the dataset, I used mean and median to find more information about the age of players. The average age per player is 25.18, and the median is 25. This means that about 50% of the players are younger than 25, and 50% are older than 25. To determine my "Risk_assessment" column, I defined the conditions for each category using the red cards and yellow cards column. To get some more insight into the data, I calculated the mean, median, and variance of "red cards" and "yellow cards". For red cards, the mean is 0.16, the median 0.0, and the variance 0.1. This means that, on average, a player has 0.16 red cards; a median of 0.0 means that at least 50% of the players have zero red cards, and a variance of 0.17 suggests some variability in the number of red cards among players. For the yellow cards, the mean is 3.07, the median is 2.0, and the variance is 6.79. The same logic for the red cards applies to the yellow cards. I ran a frequency count for the "Risk_assessment" column; it showed that 9099 players are low risk, 5340 players are considered risky, and lastly, 1238 are high risk. It is not surprising that only about 8% of the players are considered high risk given that the median for red cards was 0.0.



Having different visualisations does not only make the data easier for me to comprehend, but also for the client. For the initial Exploratory Data Analysis, I plotted the dataset to visualise which columns had missing data. After the imputations, I ran the code again to make sure I had no missing data. I used a boxplot to plot the distribution of the red cards and yellow cards. The plot from the red cards shows a red vertical line; this suggests that there is very little variability in the data, meaning that the middle 50% of the data have very similar values. The plot for the yellow cards suggests that there is a skewed distribution; the tail on the right side of the median is longer, which means that there are fewer but larger values on the right side. The left side is shorter, which means that most data points are concentrated there. Another chart I used was the bar chart; I used it to visualise the amount of high risk, risky, and low-risk players. By visualising this, I could see that there are very few high-risk players with the guidelines I set. I refined these conditions, resulting in stricter guidelines for player classification.

Before I started working on the models, I had to define the features. I created a correlation matrix and set the column “Risk_assessment” as the target. This calculates the correlation of all the other columns with the risk assessment column. I chose to use the top 8 columns with the highest correlation as my features. The correlations were pretty low, with the highest being around 0.13. This indicates a positive but weak correlation. A positive correlation suggests a positive linear relationship between the two variables. It is also important to keep in mind that correlation does not equal causation, especially when a correlation is that low. Although the correlations are all pretty weak, the models produced a relatively good accuracy, with the highest being 65%. With the accuracy being 65% with such a low correlation, I was hopeful that adding more feature variables later would improve the accuracy.

The exploratory data analysis I did on this dataset showed that there is a low frequency in red cards and a higher variability in yellow cards. This explains the small percentage of high-risk players. The median of the red cards was 0.0, meaning that a significant amount of players don't get red cards. Furthermore, I chose 8 feature variables with the strongest correlation, which gave me an accuracy of 65%. Using this information, I am hoping to be able to improve the accuracy by adjusting the conditions and creating stricter guidelines of what is considered high risk, risky, and low risk, as well as adding more features to see how it affects the accuracy.

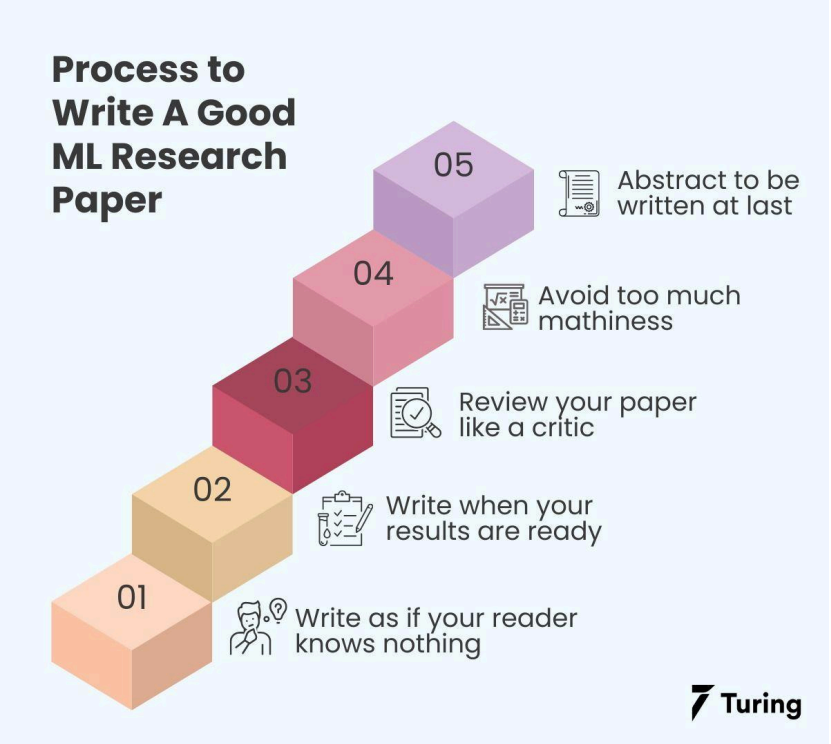


Figure 1. Process of writing a ML research paper (Source: https://www.turing.com/kb/how-to-write-research-paper-in-machine-learning-area)

1 Machine Learning

1.1 Method

The model that I concluded was the best fit for this task is the Random Forest model. Random Forest is an ensemble learning method, it creates multiple decision trees during the training process and merges the outcomes of all the decision trees into a single more accurate and stable prediction. After extensive exploration and hyperparameter tuning, my model had an accuracy of 69%, which admittedly is slightly lower, when compared to the Gradient Boosting Trees. While accuracy is important when picking the best model, there are many other factors to take into consideration. I performed a cross validation on both models and calculated the mean of the cross validation scores. When you solely look at the outcome of the cross validation, you can see that the scores are more consistent across different testing sets in the Random Forest model. This can also be seen with the mean of the cross validation scores, the Random Forest has a higher average than the Gradient Boosting Trees. Furthermore, Random Forests are a lot more computationally efficient, in my experience Gradient Boosting Trees take more than double the amount of time to run in comparison to Random Forests. So while accuracy is important it is not the sole determination for choosing the best model.

1.2 Model evaluation

The models I chose to evaluate are Random Forest and Gradient Boosting Trees, the reason for this is because these two were the highest performing models. I used 4 different metrics of evaluation; accuracy, confusion matrix, classification report, and cross validation. Accuracy can be a good indicator of the overall model performance and is easy to interpret. I specifically included a confusion matrix because my task is about risk assessment, which means

it is even more important to understand false positives and false negatives. False positives can lead to unnecessary caution and false negatives can pose risks, a confusion matrix helps with identifying these scenarios. A classification report includes, precision, recall, f1-score, and support. Precision is the ratio of correctly predicted positives divided by total predicted positives, the precision is high when the model makes less false positive predictions. Recall is the ratio of correctly predicted positives divided by all actual positives, the recall is high when the model successfully identifies a large portion of actual positive instances. The f1-score is the weighted average of precision and recall, if precision and recall is important the f1-score can be used to find a good balance. Support is the number of actual occurrences of the class in that dataset or column. Cross validation helps assess how well the model generalises unseen data, this helps to ensure that the model is not overfitted and can perform on new, unseen data. Based on the accuracy and classification report, Gradient Boosting Trees performed slightly better than the Random Forest model. The precision was slightly higher for all three classes, meaning that the Gradient Boosting Forest correctly predicted more true positives than the Random forest model. For the recall class zero and class two were higher in the Gradient Boosting Forest model, and class one was higher in the Random Forest model. Based on the result of the classification report, the Gradient Boosting Tree is a better model to use as the accuracy, precision, and recall is higher than the Random Forest. After performing the Cross validation on both models, I concluded that the Random Forest model was more reliable and stable. Random Forest was way more consistent, with the accuracy scores ranging from 0.64 to 0.69. Meanwhile the Gradient Boosting Tree, ranged from 0.09 to 0.69. While Gradient Boosting Trees did perform slightly better in the classification report, Random Forest performed way better in the cross validation.

1.3 Model improvement

I started optimising my model by experimenting with different features, I ended up using every column except the ones I used to determine the conditions for the risk assessment column. Initially I used grid search for both the Random Forest and the Gradient Boosting Tree models. However, due to the fact that it took over 4 hours to run the grid search for the Gradient Boosting Tree, I switched to using the random search approach as it is more computationally efficient. The random search provided me with a combination of hyperparameters from the specified search space. The hyperparameters I used are, “max_depth”, “min_samples_split”, “min_samples_leaf”, and “n_estimators”. I implemented the hyperparameters and the assigned values into their respective models. The accuracy of the Random Forest remained stable at 69% while the Gradient Boosting Tree saw a slight to 70%. While the output is not ideal, it is still important to add hyperparameters. Models with well tuned hyperparameters are more likely to perform well under different datasets and conditions not only that hyperparameters control the model’s complexity and tuning them can help with finding the right balance between underfitting and overfitting.

2 Ethical Considerations

The three elements for an ethical organisation includes, ethical company, ethical process and tools, and ethical people. For these three elements to relate to NAC, they would have to have a Compliance Team and HR Department for ethical companies, a development team that is accountable for embedding ethical practices for ethical process and tools, and lastly for ethical people they would have to have an employee training department. During our visit at the NAC stadium, we were informed that NAC is structured more like a family business rather than a professional one. There are no specific roles or job titles for the employees as everyone is like a close knit family. Moreover, we were informed that NAC lacks any explicit ethical guidelines, and it suggests a potential absence of a development team. An organisation without a development team may be more susceptible to data leaks, posing a risk of non-compliance with GDPR. The ethical problem I identified within NAC, is their lack of ethical guidelines. NAC could start by developing and adhering to clear policies related to data protection and privacy, this ensures compliance with GDPR and safeguarding the personal information of fans, employees, and other stakeholders.

3 Recommendations

My recommendation to improve team performance and overall team success, would be to utilise the developed risk assessment method. By implementing this, it would allow NAC to gain insight into individual player behaviour. Incorporating these insights into the player's development strategy can mitigate risks and could enhance team performance. Furthermore, NAC should also implement policies related to data protection and privacy to ensure compliance with GDPR.

4 References

OpenAI. ChatGPT. Source text previous academic year, adapted by ChatGPT and iterated upon by Celine WU. Prompt: 'Summarise this text'', (26-01-24).

Games

Media

Hotel

Facility

Built Environment

Logistics

Tourism

Leisure & Events

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility

Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

PHONE
+31 76 533 22 03
E-MAIL
communications@buas.nl
WEBSITE
www.BUas.nl

DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES