Predicting Women's NCAA Lacrosse Game Outcomes with Linear Support Vector Machines

Oliver Reidmiller

Introduction

This poster investigates how Support Vector Machines, an advanced geometric-based machine learning tool, excel at predicting game outcomes with accuracy, correctly predicting 3 out of 4 game outcomes. Through my analysis, I reveal valuable insights obtained by applying SVMs to real-time statistics from the 2023 Women's NCAA DI-DIII Lacrosse season. By considering both offensive and defensive metrics, I demonstrate how SVMs classify game results as wins or losses.

Objective

To create a Support Vector Machine model capable of accurately classifying a given game as a win or loss using statistics about the team up until that point in the season. By leveraging real-time statistics from the 2023 season, I intend to train the SVM model to analyze both offensive and defensive metrics and make informed predictions about game outcomes. Through this analysis, I aim to demonstrate the effectiveness of SVMs in providing valuable insights for strategic decision-making in collegiate lacrosse.

Data

Table 01

Summary of Statistics (N=3,579)

Variable	Mean	St. Dev	Min	Max
Average Shots	27.4	5.9	5.3	46.5
Average Goals	12	3.4	1.3	22
Win	0.5	0.5	0	1
Average Faceoff Percentage	0.5	0.1	0.2	0.9
Average Shot Percentage	0.4	0.1	0.2	0.7
Record	0.5	0.3	0	1
Games Played	2.5	4.8	0	22
Average Offensive Efficiency	0.3	0.1	0.02	0.6
Average Defensive Efficiency	0.3	0.1	0.03	0.6
Opponent Average Shots	27.3	5.9	5.3	44
Opponent Average Goals	11.8	3.5	1	23
Opponent Average Faceoffs	0.5	0.1	0.1	0.8
Opponnet Average Shot Percentage	0.4	0.1	0.1	0.7
Opponent Record	0.5	0.3	0	1
Opponent Games Played	2.5	4.8	0	21
Opponent Offensive Effciency	0.3	0.1	0	0.6
Opponent Defensive Efficiency	0.3	0.1	0.04	0.6

Methods

Table 02

Key Formulas

$$Offensive\ Efficiency = \frac{Goals}{Possessions}$$

$$Defensive\ Efficiency = \frac{Goals\ Allowed}{Opponent\ Possessions}$$

$$Shot\ Percentage = \frac{Shots}{Goals}$$

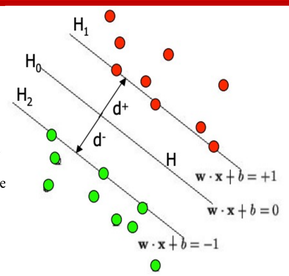
Figure 01

Support Vector Machines Explained

Main Goal

Find the best boundary that separates different groups of data points

1. **Optimize Hyperplane:** SVM finds the best hyperplane to separate classes, maximizing the distance between the hyperplane and the closest data points.
2. **Maximize Margin**: SVM aims to increase the margin, which is the distance between the decision boundary (hyperplane) and the nearest data points from each class. This enhances generalization and classifier robustness, improving performance on new data.



Input: Support vectors, or data points closest to the hyperplane
Output: Weights for each feature, whose linear combination predicts the value of y

Figure 02. R. Berwick, Village Idiot. An Idiot's guide to Support vector machines (SVMs). https://web.mit.edu/6.034/wwwbob/svm.pdf.

Results

Figure 02

Correlation Coefficient Heat Map

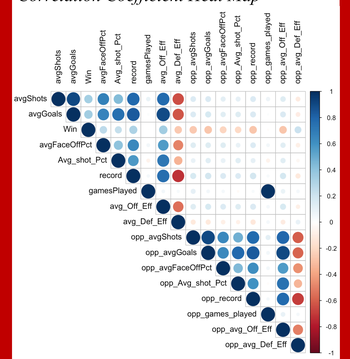
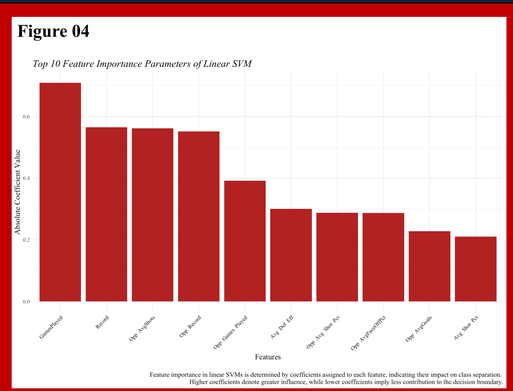


Figure 03

Figure 04

Top 10 Feature Importance Parameters of Linear SVM

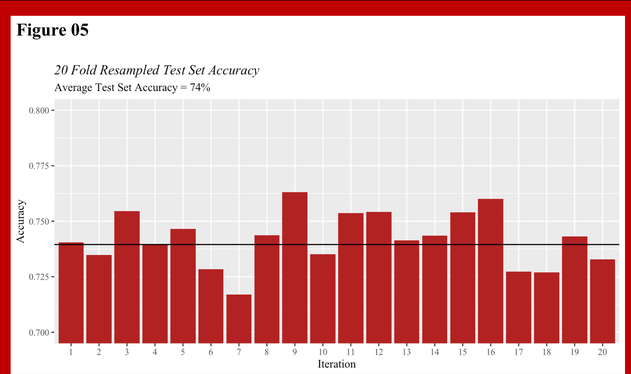


Feature importance in linear SVMs is determined by coefficients assigned to each feature, indicating their impact on class separation. Higher coefficients denote greater influence, while lower coefficients imply less contribution to the decision boundary.

Figure 05

20 Fold Resampled Test Set Accuracy

Average Test Set Accuracy = 74%



Conclusions

After creating a 10-fold cross-validated linear SVM model, the resampled test set accuracy using the 2023 season data is approximately 74%, equivalent to correctly predicting around 3 out of 4 games. While this accuracy is satisfactory, it falls short of expectations. This discrepancy can be attributed to the limited amount of data available throughout the season, which primarily depends on the number of games played. Consequently, certain features, such as the number of games played, emerge as highly important factors in the predictive model.

However, when using data from the end of the season to predict games during the season, the model demonstrates a significantly improved resampled test-set accuracy of 88%. This suggests that as more data points are generated throughout the season, the model's accuracy increases due to reduced standard error and variance.

Looking ahead, there are several avenues for improving the model's performance. These include the creation of new variables, the addition of more diverse variables to the dataset, and the incorporation of team performance data from previous seasons. By considering factors such as team consistency across seasons, the model can make more accurate predictions, particularly in early-season games where historical data may play a crucial role in forecasting outcomes.

References

LacrosseReferencePro. 2024. https://pro.lacrossereference.com/

R. Berwick, Village Idiot. (n.d.) An Idiot's guide to Support vector machines (SVMs). [Image]. Retrieved from https://web.mit.edu/6.034/wwwbob/svm.pdf.

Dickinson College Dept. of Economics