Cade Kilbride

Executive Report

Project Description

This project addresses the challenge of evaluating what factors lead to the success of a Division 1 NCAA Men's Lacrosse Team. By taking all of the tracked metrics, and all of the game scores, I developed multiple machine learning algorithms to perform data analysis and further understand the impact that different parts of the game have on a team's success. The models I have created are not necessarily perfect predictors of outright winners, but they offer valuable insight that shows which statistics are more intertwined with winning.

Lacrosse is one of the fastest growing sports in the world, and with this rise in popularity comes changes within the game itself. Especially recently, the leaders of the sports are trying to make the game as entertaining as possible for the viewer. This has led to some major rule changes that have had significant impact on the way the game is played. One of these rule changes was the addition of a shot clock beginning in 2019. The addition of the shot clock meant teams would have to play a quicker brand of lacrosse, which ultimately leads to more scoring. This rule was also tweaked just two years later to make the game even more fast paced. In addition, in this same year, the substitution box was shortened to increase the amount of transition opportunities, which in turn, also leads to higher scores. Another major rule change that has greatly impacted the game was the removal of the "motorcycle" grip from faceoffs. This made the skill gap within faceoffs less prevalent, and since after every goal there is a faceoff to compete for possession, this evened the possession battle of lacrosse games. They are still making tweaks to these rules, and others, every year.

All of these changes have led to closer games, more entertainment for the viewer, and major changes in strategy. An example of this is the change in how a team stops the other team from clearing the ball (also called riding). Since 2020, there has been a major increase in what is called the 10 Man Ride. The 10 Man Ride is a more aggressive style of playing lacrosse, and has taken over college lacrosse. Many teams have implemented this as their only style of riding, and every NCAA D1 team has played against a 10 Man Ride at some point this year. This has led to a much higher emphasis on clearing and riding in lacrosse, and this project aims to expose more subtle statistics like these.

Methodology

The data used in this project comes directly from the NCAA site, and comes from the years 2017-2024. The data before 2017 has lots of missing values, and therefore is not reliable information, and the 2025 data is not complete. I collected the scores of every game within this range, and collected each season average statistic for each team leading up to that date (Assists per Game, Caused Turnovers per Game, Average Clearing Pct, Average Faceoff Pct, Ground Balls per Game, Average Man Down Defense Pct, Average Man Up Offense Pct, Average Opponent Clearing Percentage, Goals per Game, Saves per Game, Goals Allowed per Game, Average Margin, Turnovers per Game, and Shot Pct). I used web scraping (Beautiful Soup) to aid me, and used a python script to clean and merge this data. For each game, there was all of the statistics listed above for each team playing, gathered from that season and leading up to the date before the game.

These statistics were helpful, but there needed to be a metric to show how hard a team's schedule has been. Without this metric, if a team with a bad schedule played a team with a hard schedule, then this would hurt the training. I decided to use the NCAA Lacrosse Ratings Percentage Index (RPI) as this metric. For a team, 25% of RPI is that team's record, 50% is their opponents' cumulative record, and 25%

is their opponents' opponents' cumulative record. While this is not a perfect metric to show the strength of schedule, it gives an objective metric to show how good a team is. I calculated this for each team in each game on that date, so the result of the game has no effect on their RPI before that game.

I split the data for everything into before 2020 and after 2020. This was due to a couple of factors. First, one of the goals of this project is to look at how rule changes affect each part of the game. The two major rule changes were the addition of the shot clock, and the changing of faceoffs. The shot clock was added in 2019, and tweaked in 2021, while faceoffs were changed in 2020. In order to see the effect of both of these factors, I decided to split before 2020. The pre 2020 data is the seasons 2017-2020, and the post 2020 data is the seasons 2020-2024. In addition, opponent clear pct was not tracked until the 2020 season, so this made training the models much easier.

This project approaches the success of a team as a classification problem. Since the datasheet is built off of the game scores, we can look at whether this team won as our target variable. I used Logistic Regression (with and without regularization), Random Forest, and XGBoost to create my models in order to see the results. These all offer their own advantages and disadvantages. Logistic Regression offers a simple way to look at the linear relationship between winning and everything else. A Random Forest can handle non-linear relationships, and interaction a little bit better than a Logistic Regression model. XGBoost is the most complex model, and handles non-linear data the best. I ran all of these models with differential features (Team 1 Stat - Team 2 Stat). For each of these models, I also created a learning curve graph to show if the model was overfitting or underfitting.

After running these models, I also did some manual feature engineering to try and improve performance, specifically on the logistic regression. I chose to only run the logistic regression with these new features because the random forest and the XGBoost already take into account the complex interactions. The first major feature engineering I did was look at how a team's statistic matches up with the opposing team's adversarial statistic, for example Team 1 Offense vs Team 2 Defense. I did this for most of the statistics, and re-trained the logistic regression on these new features. Next, I looked at rolling averages over three games. This allowed the data to show momentum throughout a season. I also re-trained the logistic regression model with this new data.

Results and Discussion

Raw Data

Table 1

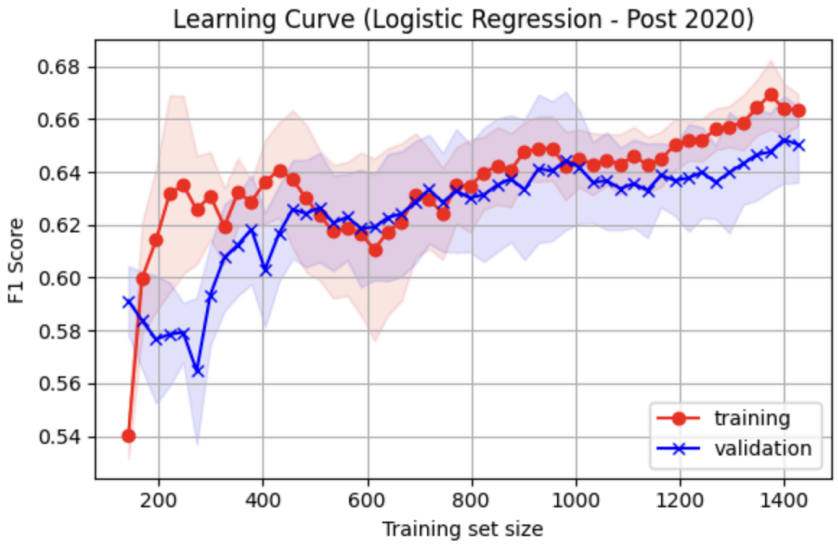
Model	Pre 2020 Accuracy	Pre 2020 F1 Score	Post 2020 Accuracy	Post 2020 F1 Score
Logistic Regression	0.706	0.550	0.728	0.645
Logistic Regression L1 Regularization	0.690 (c=0.1)	0.531 (c=0.1)	0.739 (c=0.01)	0.674 (c=0.01)
Logistic Regression L2 Regularization	0.702 (c=0.1)	0.547 (c=0.1)	0.723 (c=0.01)	0.682 (c=0.01)
Random Forest	0.687	0.554	0.723	0.635
XGBoost	0.665	0.534	0.700	0.640

Table 2

Model	Dataset	Pre 2020 Accuracy	Pre 2020 F1 Score	Post 2020 Accuracy	Post 2020 F1 Score
Logistic Regression	Adversarial Data Only	0.661	0.465	0.711	0.617
Logistic Regression	Adversarial and Original Data	0.665 (c=0.1)	0.487 (c=0.1)	0.717 (c=0.01)	0.635 (c=0.01)
Logistic Regression	Rolling Data Only	0.649	0.463	0.678	0.605
Logistic Regression	Rolling and Original Data	0.702	0.558	0.720	0.630
Logistic Regression with L1 Regularization	Rolling and Original Data	0.706 (c=0.5)	0.566 (c=0.5)	0.731 (c=0.05)	0.657 (c=0.05)
Logistic Regression with L2 Regularization	Rolling and Original Data	0.690 (c=0.01)	0.558 (c=0.01)	0.714 (c=0.01)	0.648 (c=0.01)

Model Performance

After running the original logistic regression, I created a learning curve plot (Figure 1 below) to see if our model was overfitting or underfitting. Both the pre 2020 and the post 2020 graph looked similar, and they appear to neither be overfitting or underfitting. I believe they are not underfitting because the models were getting accuracies of 70%, which is just under the average win rate for favorites. I knew they weren't overfitting because the training and validation was close together.



I used f1-score, and accuracy to determine the performance of each model. For the logistic regression, I compared the weights of the model parameters to determine the importance of each factor. For the random forest, I used feature importance. For the XGBoost, I used mostly gain to determine the importance.

We can see that our logistic regression, specifically with L1 regularization, on our raw dataset actually had the best performance out of all of our models. This shows that the relationship within our data appears to be linear. Our L1 regularization model also performed better than no regularization or L2 regularization, showing that there may only be a small number of features that are important. That being said, the difference in performance in these models was small, and in order to fully understand which is best, we would need more data.

Feature Importance

For the basic logistic regression (averaged over all of the different models), the most impactful factors pre 2020 were RPI (average weight of 0.35), Turnovers per Game (-0.239), Assists per Game (0.221), Saves per Game (-0.202), and Faceoff Percent (1.89). For the post 2020, these were RPI (0.365), Turnovers per Game (-0.246), Average Margin (0.141), Shot Percentage (0.124), and Goals Allowed per Game (-0.112).

In the Random Forest, the pre 2020 features with the biggest importance were RPI (0.107), Average Margin (0.097), Assists per Game (0.076), Turnovers per Game (0.073), and Goals per Game (0.070). In the post 2020, these were RPI (0.140), Average Margin (0.104), Goals per Game (0.078), Turnovers per Game (0.072), and Ground Balls per Game (0.063).

For the XGBoost, we can look at the gain to see how important a feature is. For pre 2020, the five most important features were Average Margin (1.81), RPI (1.77), Saves per Game (1.22), Shot Percentage (1.40), and Assists per Game (1.04). For post 2020, these were RPI (3.967), Average Margin (2.74), Saves per Game (1.39), Turnovers per Game (1.31), and Faceoff Percent (1.26).

Using this raw data, we can see the importance of multiple different statistics in determining a team's success. Firstly, RPI is consistently either the highest, or second highest in importance. This makes sense, since RPI is a basic metric of how good a team is. The next most prevalent metric is Turnovers per Game, which makes an appearance in all of the top five most important factors except for one. Turnovers also have a negative correlation in the logistic regression, which makes sense since more turnovers leads to a worse game. Average Margin is also widely prevalent, as well as Faceoff Percent.

One of the more surprising results was the negative correlation that Saves per Game had, because you might assume that more saves leads to a better result. When you dive deeper into the way this statistic works, you will find that more saves per game means more shots on goal per game, which will lead to more goals being scored.

Another surprising result is the lack of Clearing Percent or Opponent Clearing Percent, especially in the post 2020 data. This shows that these statistics may not be as important as other features with a higher importance.

We can also use these top 5 metrics to show the difference from pre 2020 to after the rule changes in 2020. We can see that RPI is more involved in determining a team's success post 2020. In all three of these models, the importance metric for RPI increased after 2020. We can also see that Assists per Game appears in the top 4 twice in the pre 2020, and none after 2020.

Both of these results probably stem from the rule changes. Both of the rule changes mentioned in the introduction will increase the speed of play. This means worse teams can hold on to the ball longer, which would decrease possessions and allow for more "randomness" to occur. For the decrease in

importance for the Assists per Game, a faster pace of play doesn't allow teams to wait for the perfect shot. In lacrosse, assisted goals are generally considered to be better shots, but since teams can't hold the ball anymore, they are forced to take more unassisted shots.

Although our models performed worse after the feature engineering, these can also offer us valuable insight into our feature importance. For the adversarial data, the shooting percentage data was much more valuable than in the logistic regressions, which shows us the importance of having a solid goalie. The adversarial data didn't offer much difference between pre 2020 and post 2020.

For the rolling averages models, there weren't any features that consistently stuck out, showing that momentum may not be that important for individual statistics.

Conclusion

This project demonstrates how data-driven machine learning models can give valuable insights into data analytics. We can see how using different models can lead to understanding how different statistics interact with each other. In this project, logistic regression performed the best, showing these interactions may be more linear.

We can also see the important factors in winning a lacrosse game. The importance of turnovers show that teams should focus on ball security, and teams that do better at faceoffs tend to win more. In today's game, many coaches say that clearing and riding will win a game, but from this data, we can see that this may not be the case.

Importantly, this project also shows subtleties that are important to understanding how rule changes can affect college lacrosse games. The increase of importance in RPI might suggest that there are less upsets due to these rule changes, which is actually the opposite of what the rule changes were meant to do. The decrease of the importance of assists per game might mean individual efforts have become largely more important in today's game.

In future work, exploring more interaction between these features could give valuable insight that this project's models could not give. In addition, using a wider range of data that is not currently tracked (time of possession, player performance, etc.) could give us more important statistics than we currently have.