Marquette Basketball Analytics and Prediction Project

Team Members: Nathan Rusch, Mark Yoingo, Sienna Ifill, Allan Akkathara

Abstract

After Marquette's lackluster performance and completely avoidable loss against North Carolina State in the 2023 March NCAA March Madness tournament. We feel that staff on the Marquette basketball team, specifically trainers, need to be more familiar with the weak points of the Marquette Men's basketball team to avoid future losses. We plan to make a predictor that analyzes the previous games of the Marquette Men's basketball team and returns their weakest attributes as a whole i.e., shooting, rebounding, free throws, etc. This learning model will then allow them to highlight these weaknesses in practice and win more games. Not only will this benefit the Basketball team, but the whole school, as doing better in athletics will bring more students to apply to Marquette University and games that are predicted to be a win or a close game are more likely to bring in more revenue from ticket sales. Our model also successfully predicted a Marquette win in the upcoming game against Dayton on December 14, 2024. This tool can help the Marquette basketball team make informed decisions and improve their performance in future games.

Goals

Utilizing machine learning methods like logistic regression and clustering, we hope to analyze previous performance statistics provided by Marquette University regarding the Basketball Team to provide insight on what might be an area to target for better development of the Basketball team. We hope to not only analyze the performance of the team as a whole but also the players as individuals so that we can then extrapolate their information onto the team and allow us to personalize the training regimen for the individual players on the team. Overall we hope that our Model will allow staff on the Marquette basketball team to go further in the Marquette NCAA tournament in March.

Dataset

Our primary data source for this project is game reports from sports-reference.com, which offers a comprehensive collection of basketball statistics specific to Marquette University. Additionally, we plan to incorporate third-party game statistics and analytics to enhance the scope of our analysis. The dataset will be divided into a training set and a test set to support both individual player analyses and game statistic modeling.

While we have not yet implemented full data integration due to the specificity of our current dataset, we recognize the value of expanding our data to include similar statistics for other NCAA college teams. However, obtaining and integrating comparable data has proven challenging. If time and resources permit, we aim to pursue further data integration to improve the robustness of our models.

For now, the dataset we have assembled is sufficient for our current tasks. Below is a sample of the key parameters we will work on within our analysis.

Marquette

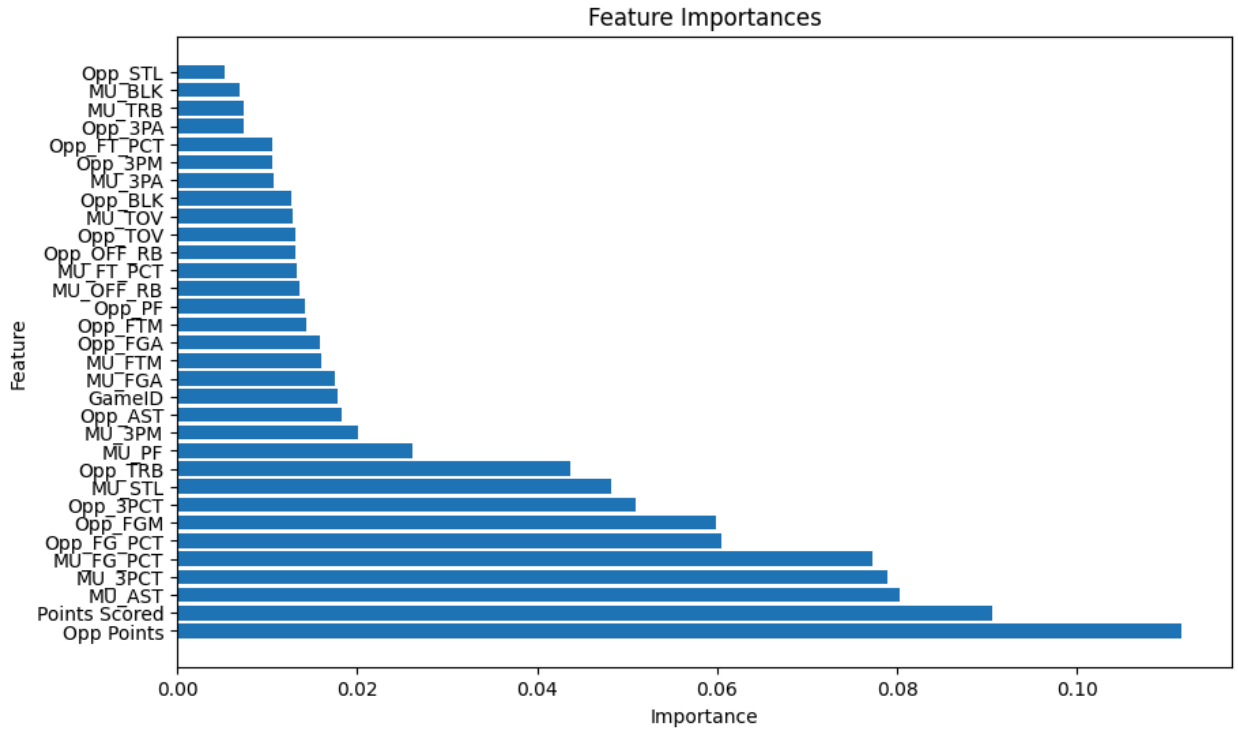
Tm -- Points

FG -- Field Goals
FGA -- Field Goal Attempts
FG% -- Field Goal Percentage
3P -- 3-Point Field Goals
3PA -- 3-Point Field Goal Attempts
3P% -- 3-Point Field Goal Percentage
FT -- Free Throws
FTA -- Free Throw Attempts
FT% -- Free Throw Percentage
ORB -- Offensive Rebounds
TRB -- Total Rebounds
AST -- Assists
STL -- Steals
BLK -- Blocks
TOV -- Turnovers
PF -- Personal Fouls

Opponent

Opp -- Opponent Points
Opp_FG -- Opponent Field Goals
Opp_FGA -- Opponent Field Goal Attempts
Opp_FG% -- Opponent Field Goal Percentage
Opp_3P -- Opponent 3-Point Field Goal Attempts
Opp_3PA -- Opponent 3-Point Field Goal Attempts
Opp_3P% -- Opponent 3-Point Field Goal Percentage
Opp_FT -- Opponent Free Throws
Opp_FTA -- Opponent Free Throw Attempts
Opp_FT% -- Opponent Free Throw Percentage
Opp_ORB -- Opponent Offensive Rebounds
Opp_TRB -- Opponent Total Rebounds
Opp_AST -- Opponent Assists
Opp_STL -- Opponent Steals
Opp_BLK -- Opponent Blocks
Opp_TOV -- Opponent Turnovers
Opp_PF -- Opponent Personal Fouls

Using the Random Forest Classifier, We ranked the above parameters:



Tools

Scikit - learn, pandas, vs code, matplotlib, seaborn, NumPy

Literature review

https://www.covers.com/ncaab/george-mason-vs-marquette-predictions-picks-friday-11-8-2024#Same-game_parlay - This site is one of many sites that describe predictions on Marquette basketball team vs other schools. This site gave us insight into how prediction models are used in the real world

https://sportsbook.draftkings.com/teams/basketball/mens-college-basketball/marquette-golden-eagles--odds - This site is an online betting site that users use to earn money which could also provide insight on how people use predictions to make decisions or how we could develop a prediction model that help users make informed decisions.

GitHub Link

Click here to view our GitHub repository.

Work we have completed

As of this point in time, we have found key discoveries within Marquette games using association rules, weaknesses, and strengths in the teams using clustering, and a prediction model that predicts game outcomes using logistic regression. We also have a preprocessed and cleaned dataset which we used for our models and findings.

Data Pre-Processing

Fortunately, sports-reference.com is very detailed and uploads very clean datasets, drastically reducing the stress of this project. However, the data still needed to be manipulated from its original form for our work. First, two columns needed to be removed because one was entirely blank and only for website formatting, and the other also originally appeared to mostly be filled with null values. Now it is clear that this row indicates whether Marquette is playing at home, away, or at a neutral site, which could have been useful for our data but we realized too late. However, since Marquette cannot decide to exclusively play at home anyway, we are doubtful about how helpful the data could have been. Next, columns were renamed from 'Unnamed: 0' through 'Unnamed: 6', 'School' through 'School.15', and 'Opponent' through 'Opponent.15' to names that would explain what the columns represent. For example, 'School.15' and 'Opponent.15' became 'MU_PF' and 'Opp_PF', which represent Marquette and their opponent's personal fouls respectively for each game.

Additionally, the data type for each row was just an object when initially scanned in. All the applicable rows were changed to integer or float objects so they could properly be manipulated. Furthermore, wins and losses were adjusted to appear as a zero or one respectively. Using these wins and losses, we divided the dataset into a data frame of Marquette's wins and Marquette's losses so we can examine what leads to a win and what leads to a loss. Further categorical breakdowns were made for one hot encoding in preparation for association rules.

Association Rules Report

We chose to begin with association rules to see if there were any connections between attributes and winning or losing that we would not instinctively think of. For example, Marquette and their opponents scoring above or below a certain threshold is expected. We want to know if other attributes we might not initially credit have a higher contribution to a win than we would expect. We used apriori and association rules on a one-hot encoded data frame of Marquette's wins and of Marquette's losses, called wins_dummies and loss_dummies respectively. For each data frame, we used a minimum support and a minimum confidence of 0.6. Sorting from highest to lowest lift, here were our results:

```
wins_assoc.head(8)
```

✓ 0.1s Python

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	represen
4	(Opp Points_61-80)	(Opp_FGM_21-30)	0.777778	0.777778	0.666667	0.857143	1.102041	
5	(Opp_FGM_21-30)	(Opp Points_61-80)	0.777778	0.777778	0.666667	0.857143	1.102041	
0	(Opp Points_61-80)	(MU_STL_6-10)	0.777778	0.740741	0.629630	0.809524	1.092857	
31	(Opp_FGM_21-30)	(Opp Points_61-80, Opp_BLK_<=5)	0.777778	0.740741	0.629630	0.809524	1.092857	
29	(Opp Points_61-80)	(Opp_BLK_<=5, Opp_FGM_21-30)	0.777778	0.740741	0.629630	0.809524	1.092857	
28	(Opp_BLK_<=5, Opp_FGM_21-30)	(Opp Points_61-80)	0.740741	0.777778	0.629630	0.850000	1.092857	
26	(Opp Points_61-80, Opp_BLK_<=5)	(Opp_FGM_21-30)	0.740741	0.777778	0.629630	0.850000	1.092857	
1	(MU_STL_6-10)	(Opp Points_61-80)	0.740741	0.777778	0.629630	0.850000	1.092857	

```
loss_assoc.head(8)
```

✓ 0.0s Python

```
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/mlxtend/frequent_
  cert_metric = np.where(certainty_denom == 0, 0, certainty_num / certainty_denom)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representat
1163	(MU_FTM_6-10, MU_FGM_21-30)	(MU_TRB_21-30, W/L, MU_BLK_<=5)	0.7	0.6	0.6	0.857143	1.428571	
386	(MU_FTM_6-10)	(MU_TOV_6-10, MU_FGM_21-30)	0.7	0.6	0.6	0.857143	1.428571	
777	(MU_FTM_6-10, W/L, MU_BLK_<=5)	(MU_TRB_21-30)	0.7	0.6	0.6	0.857143	1.428571	
1152	(MU_FTM_6-10, MU_FGM_21-30, W/L)	(MU_TRB_21-30, MU_BLK_<=5)	0.7	0.6	0.6	0.857143	1.428571	
1188	(MU_FTM_6-10, MU_FGM_21-30, W/L)	(MU_TOV_6-10, MU_BLK_<=5)	0.7	0.6	0.6	0.857143	1.428571	
636	(MU_FTM_6-10, MU_FGM_21-30, W/L)	(MU_TRB_21-30)	0.7	0.6	0.6	0.857143	1.428571	
373	(MU_FTM_6-10)	(MU_TRB_21-30, MU_FGM_21-30)	0.7	0.6	0.6	0.857143	1.428571	
675	(MU_FTM_6-10)	(MU_TOV_6-10, MU_FGM_21-30, W/L)	0.7	0.6	0.6	0.857143	1.428571	

In our wins_assoc, we can see that our wins are mostly determined by the points the opponent scores, as we might expect. Our opponents scoring in the range of 61-80 points appear in each of our top associations. What stands out is the appearance of the attribute 'MU_STL_6-10', which means that Marquette often steals 6-10 balls in games they win. This can give the coaches something to focus on when practicing.

In our loss_assoc, we can see that our losses are entirely dependent on our stats rather than our opponents. The first attribute that stands out to me is that Marquette only makes 6-10 free throws in each of our top associations. This is an incredibly low number and something that I have noticed attending and watching Marquette basketball games; if they start losing, they start missing almost every free throw. This is an even bigger problem because free throws are "free points," since there are no defenders to worry about while shooting. Marquette should try to recreate this feeling in practice by having drills where the starting players start down by several points and they have many free-throw attempts. Another way to recreate the stress of trying to make free throws while down would be between drills to randomly call out a player and they need to make a certain number or percentage of free throws, otherwise the whole team is required to do some punishment such as holding a plank for a minute.

Clustering Report: Strengths vs. Weaknesses

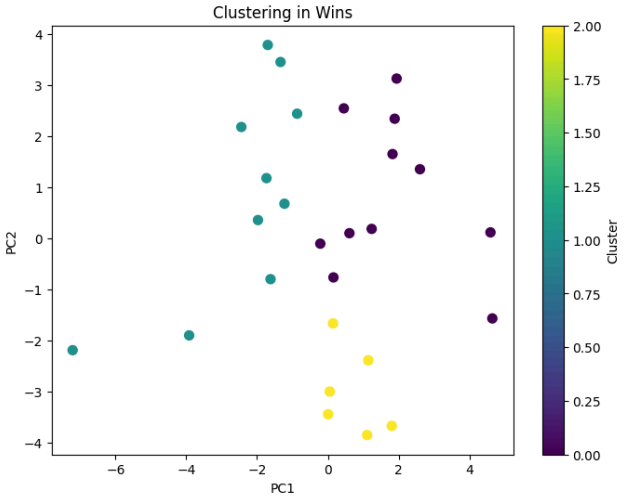
In our original statement, we sought to pinpoint the strengths and weaknesses of the Marquette men's basketball team. In doing so we hoped to use our findings to create a future game plan that would allow the men's basketball team to better succeed in future games and tournaments.

Method:

To satisfy the needs of this task, we decided that it would be best to utilize a clustering model to pinpoint the strengths and weaknesses of the Marquette basketball team. By utilizing the PCA method in the SKlearn library to cluster our data based on the strongest relationships within our dataset. Further, by using a PCA decomp function, we were able to get the specific features utilized to construct the principal components of the clustering graph and the degree to which they positively or negatively impacted the formation of the overall principal components.

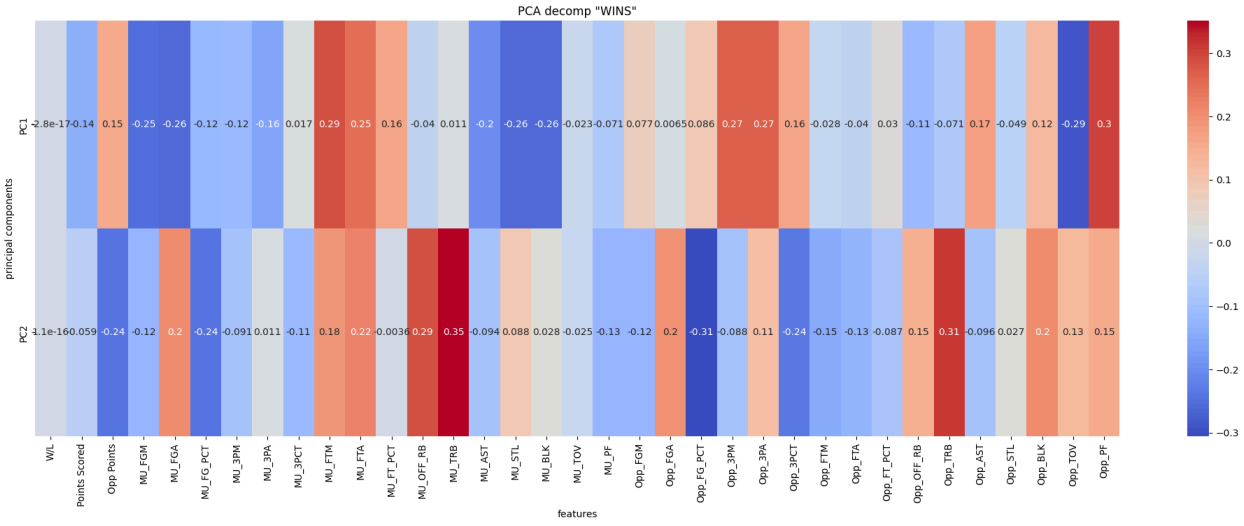
Upon execution of our original plan, there were several issues in the idea of our implementation. While clustering and principal component decomposition may give us hidden relationships within the data, it lumps games resulting in a win and games resulting in a loss together, causing issues with interpretation and the ability to conclude the data produced. To remedy this, we split our data frame into two separate frames accounting for the context of a Marquette win and a Marquette loss. By doing so we were able to contextualize the results of the PCA and better infer the strengths and weaknesses of the Marquette men's Basketball team.

Results:



Clustering/Decomp -> Win:

The clustering model generated by the principal components created upon running the SciKitlearn PCA function is included above. Added below is the Principal component decomposition for its respective clustering graph. Both graphs are in the context of a Marquette win.

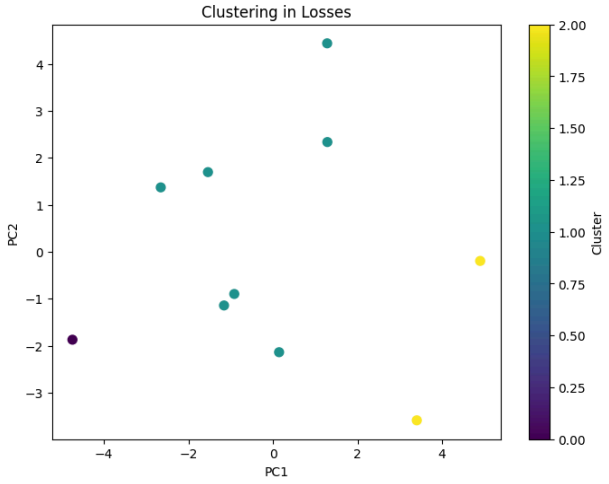


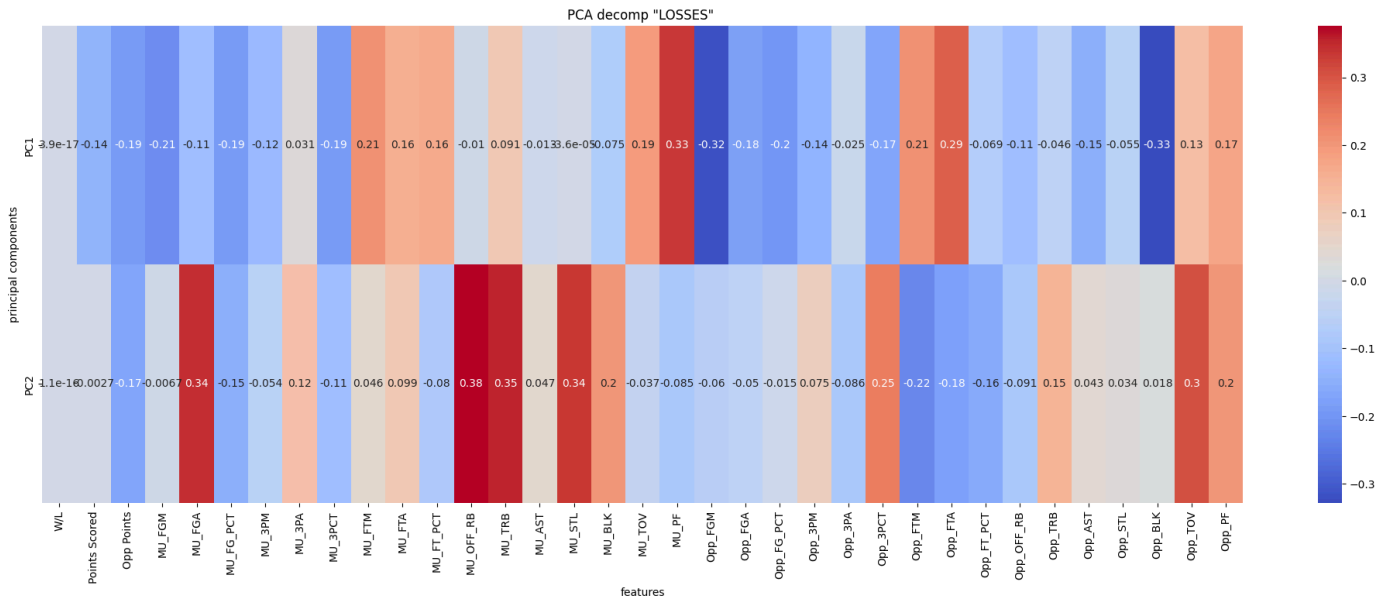
Through inspection of the positive and negatively influential features from the PCA decomposition graph, we were able to infer the strengths of the basketball team. Upon inspection of the principal component decomposition for either component, it is apparent that Marquette's defensive attributes are often the most positively influential and negatively influential features within the graph. Exhibited by a 0.35% positive utilization of Marquette “total rebounds” in PCA2 as well as a consistently positive utilization of the opponent's offensive features within PCA1. From these utilizations we can interpret that, in the context of a win, Marquette typically has a strong defense that allows them to shut down the offense of the

opponent. An entry that might exemplify this is a Marquette win with Marquette having high offensive stats and their opponent having low offensive stats. From this consistent pattern demonstrated by the decomposition, we can infer that Marquette's defense is one of the team's strengths and can be leveraged in future games.

Clustering/Decomp -> Loss:

Included to the right is the clustering model generated by Marquette losses and included below is the principal component decomposition for the clustering model provided.





Through inspection of the PCA decomp for Marquette losses, a very clear image is painted in regards to their weaknesses as a team. Within PCA2 the most influential components include the Marquette field goal attempts (shots attempted) and Marquette total rebounds. In PCA1 a

major negative contributor is opponent field goals made (shots made), and opponent blocks. From this, we can come to several conclusions. In the context of a loss, when examining statistics like field goal attempts juxtaposed to total rebounds and a lack of total field goals made, one can infer that in a losing game, Marquette basketball will take many shots that fail to go in and further fail to rebound these shots. Suggesting that Marquette's offensive strategy, specifically shooting and rebounding their shots, is one of their greatest weak points as a team.

Predicting Marquette Basketball Wins: A Logistic Regression Approach

We used a logistic regression model to predict the outcome of Marquette University basketball games. The model leverages past Marquette game data to predict whether Marquette will win or lose against a specific opponent.

Data Collection and Preparation:

We collected data that contained important marquette game stats from their previous games starting from 2021. We chose the cutoff range at 2021 since this is the year the current head coach Shaka Smart took his role as head coach. Team players change every year but the coach remains the same so this seems ideal. This gave us a dataset with around 108 observations to play with. We then pre-processed and cleaned this data to use for our models. We dropped all columns that provided very little to no effect on the outcome of a game like the dates, string values like name, etc. Each row in the dataset contained important game stats of Marquette like Field goals, three points made, free throws, etc, and the game stats of the opponent they versed on some specific day.

Feature Selection:

To refine the predictive power of the model, We used two feature selection techniques: a correlation matrix and a random forest classifier.

- The correlation matrix helped identify and remove features that were highly correlated with each other, as multicollinearity can negatively impact the performance of a logistic regression model.
- The random forest classifier provided an important ranking of features, enabling us to focus on the most impactful variables for predicting game outcomes.

After applying these techniques, the dataset was reduced to a smaller set of significant parameters. This reduced-parameter model outperformed the original model, both in terms of accuracy and interpretability. I didn't try much feature engineering since a lot of the features were already a combination of other features but if I had more time, I would

like to see if we could improve the accuracy and model score with better feature engineering.

Model Improvement:

Initially, the dataset lacked sufficient data for robust prediction because we only used games played in the current year 2023-24, which led to an underperforming model. To address this, I added the game data from 2021 to the CSV file. This improved the sample size, providing a better foundation for training the model.

The enhanced dataset, combined with the reduced-parameter model, yielded much better results than the earlier attempts with the full set of features or small sample sizes. The final logistic regression model demonstrated improved accuracy and stability, making it a reliable predictor of Marquette basketball game outcomes.

Results:

Looking at the initial model stats below when the dataset didn't have enough features, the accuracy was 0.875, 1 false positive, 50% of the predicted positive instances were positive, only 86% of the actual positive instances were correctly identified and class 1, the f1 score was 67%.

```
Accuracy: 0.875
Confusion Matrix:
 [[6 1]
 [0 1]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.86      0.92         7
           1       0.50      1.00      0.67         1

    accuracy                           0.88         8
   macro avg       0.75      0.93      0.79         8
weighted avg       0.94      0.88      0.89         8
```

The second model with more data and a reduced and less correlated model provided much better results overall. This model received an accuracy of 95%. It Incorrectly predicted 1 negative instance but other than that only recall for case 1 was slightly behind, with only 83% of the actual positive instances being correctly identified. The model got a good score on everything else.

```
Accuracy: 0.9523809523809523
Confusion Matrix:
 [[15  0]
 [ 1  5]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        15
           1       1.00      0.83      0.91         6

    accuracy                           0.95        21
   macro avg       0.97      0.92      0.94        21
weighted avg       0.96      0.95      0.95        21
```

Application: Predicting upcoming Marquette vs Dayton game (Dec 14, 2024)

To predict the upcoming game, I first calculated the average for Marquette game stats. This was not too difficult since I had the data and just needed to do the calculation. I faced some obstacles doing the same for Dayton. Luckily, Dayton had data on their team in the same format. I made a web scraping script using the beautifulSoups4 library in Python, extracted the data, and created a data frame with the means. I then appended Dayton's data onto Marquette's data and used this as input for the prediction.

```
[158]  print("Predicted Outcome 0 if marqutte will win and 1 if Dayton will win:", dayton_pred)
       ✓ 0.0s

...    Predicted Outcome 0 if marqutte will win and 1 if Dayton will win: [0]
```

The model predicted Marquette to win! The output being 0 might be slightly confusing but we set 0 as our win condition in the data and therefore 0 means win and 1 is loss.

Future goals

In the future, we aim to explore the implementation of a perceptron model or a neural network to predict game outcomes. Predicting the result of basketball games is inherently challenging due to their unpredictable nature, as numerous variables can influence the outcome. While logistic regression is a powerful tool for binary classification, it may have limitations in capturing the nonlinear relationships in complex sports scenarios. Neural networks are particularly well-suited for this task because of their ability to learn complex patterns and relationships in data.

Additionally, we plan to investigate model ensemble techniques, better data integration, and combining multiple models to enhance predictive accuracy. This combined approach could provide insights into future game outcomes based on the current team's strengths and weaknesses. It may also help identify areas where the team could improve to gain a competitive edge.

Once we establish reliable models, our goal is to develop a graphical user interface (GUI) or application to implement our research findings. Such a tool could be beneficial to the Marquette team, offering actionable insights and strategic support.

Conclusion

Marquette's men's basketball team is overall a very strong team, exhibited by the number of entries within the winning clustering model and the lack of entries within the losing clustering model. However, this lack of entries allows us to better understand the mechanics of the team that is causing them to fail. Revisiting the previous analysis, it's apparent that Marquette's greatest weakness is their ability to consistently make shots and rebound any missed shots, resulting in a turnover. We suggest that the team focus on their defensive plays as that has proven successful according to our analysis, and to focus on shooting and rebounding in practice.

References, Links, and Citations

Articles:

Building My First Machine Learning Model | NBA Prediction Algorithm | by Alexander Fayad | Towards Data Science

bing.com/ck/a?!&&p=5e9d2fc82bd423ba548721828cf654a2b704ddcce1a73bb272a6fd901352e537JmltdHM9MTczMTk3NDQwMA&ptn=3&ver=2&hsh=4&fclid=0db72137-cc7f-6069-00f3-317bcd016134&psq=sports+analytics+machine+learning+model&u=a1aHR0cHM6Ly93d3cuY291cnNlcmEub3JnL2xlYXJuL21hY2hpbmUtbGVhcm5pbmctc3BvcnRzLWFuYWx5dGljcz9tc29ja2lkPTBkYjcyMTM3Y2M3ZjYwNjkwMGYzMzE3YmNkMDE2MTM0&ntb=1

Learner Reviews & Feedback for Introduction to Machine Learning in Sports Analytics Course | Coursera

GitHub Link: https://github.com/SiennaIfill/COSC_5610_final (in case the first one doesn't work)

Useful Links:

2024-25 Men's Basketball Cumulative Statistics

-

Marquette Golden Eagles 2024-25 Men's College Basketball Stats - ESPN
-
2023-24 Marquette Golden Eagles Men's Roster and Stats | College Basketball at Sports-Reference.com
-
2023-24 Marquette Golden Eagles Men's Gamelogs | College Basketball at Sports-Reference.com

Provided above are various data frames regarding the personal statistics and game statistics of the Marquette Men's basketball team.

Contributions Appendix:
Data Pre-Processing: Sienna Ifill
Association Rules Report: Sienna Ifill
Strengths vs Weaknesses Clustering Model: Nathan Rusch
Logistic regression for game prediction: Allan Akkathara
Problem introduction, Future improvement, Tools, dataset: Mark Yoingco