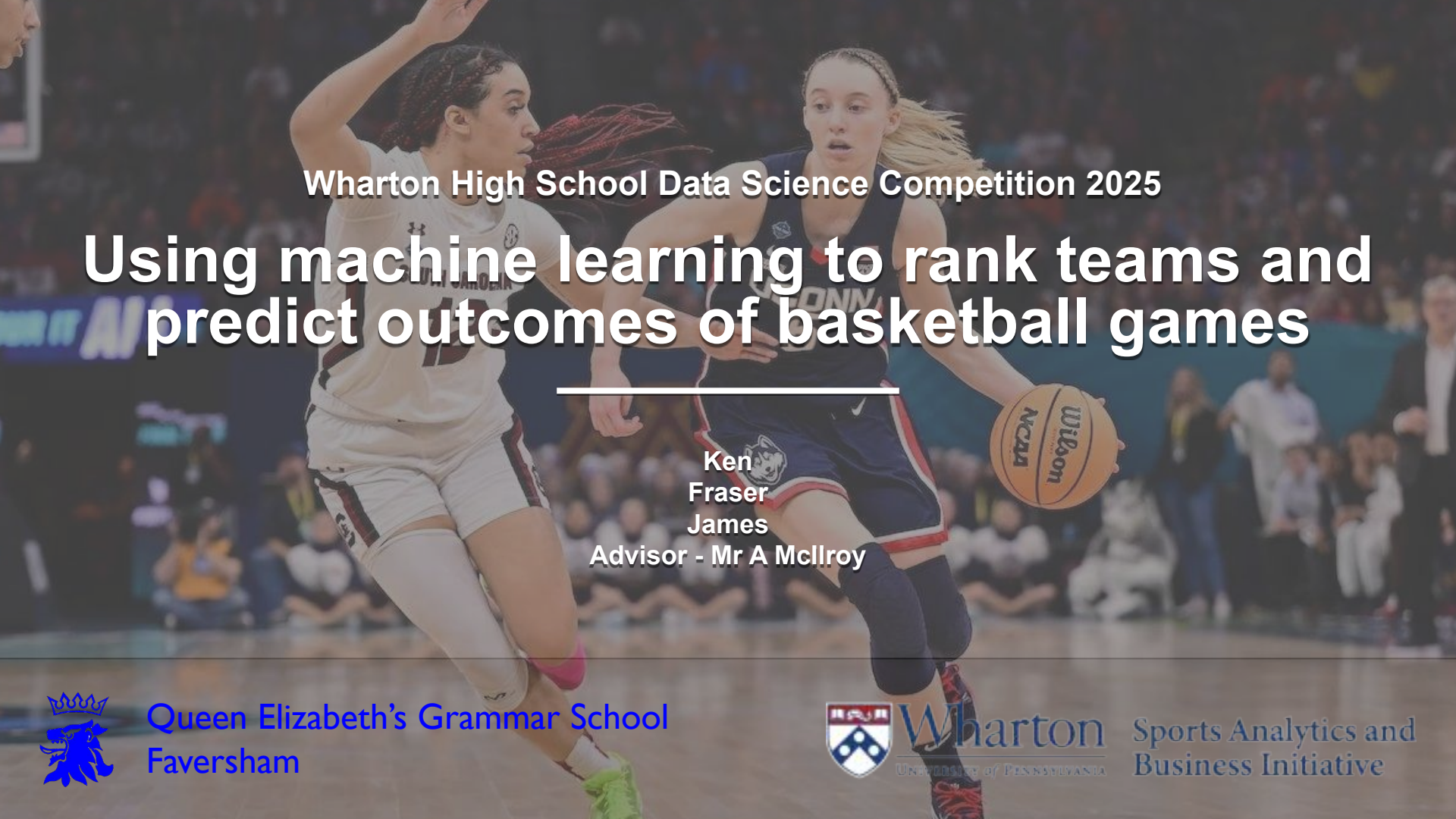


Wharton High School Data Science Competition 2025

Using machine learning to rank teams and predict outcomes of basketball games

Ken
Fraser
James
Advisor - Mr A McIlroy

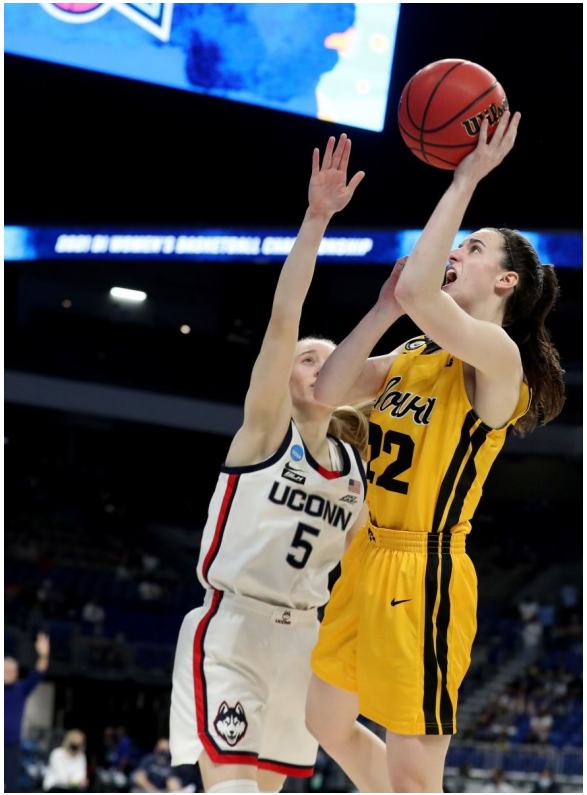
Queen Elizabeth's Grammar School
Faversham

Wharton
University of Pennsylvania
Sports Analytics and Business Initiative

Table of Contents

1. Introduction and Background Research

- Previously, Stanford won the 2021 Women's NCAA title, beating Arizona 54-53.
- Haley Jones was named MVP.
- Analysts predict NCAA basketball outcomes using statistical models, machine learning, expert polls, and tools like KenPom for team efficiency.
- Bracket simulations also help, but the unpredictable nature of the tournament makes predictions difficult.



Sources: ESPN, collegeinsider.com, analytics8.com

Research Question - Can Machine Learning effectively be used in sports?

- Binary classification problem (two outcomes/labels)
  - In basketball, games can only result in a win or loss due to overtime rules.
- And thus we decided that a machine learning classification algorithm can be employed
  - These algorithms also produce, quantitative results which can be interpreted as probabilities.

Definition:
Classification Algorithms try to predict the correct label of given input data.

Data Cleaning

- Aggregate match data
- Non-D1 team data was excluded

Data Transformation

- Machine Learning models perform better with strong linear correlations
- Weaker correlations can be transformed using polynomial and log functions

Correlation Analysis

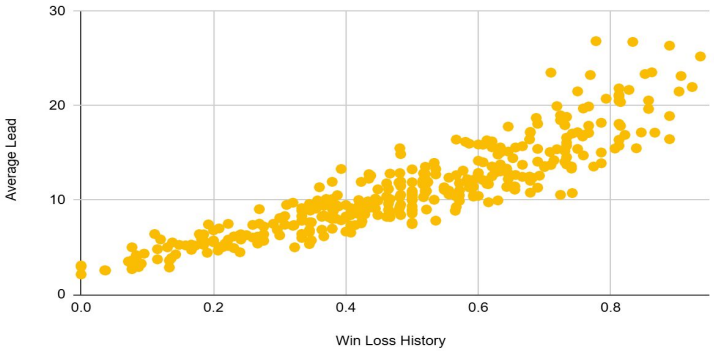
- Analyse and evaluate the relationship between variables.
- Spearman's rank correlation coefficient
- Select the most correlated variables to use as inputs in our models.

Data Transformation

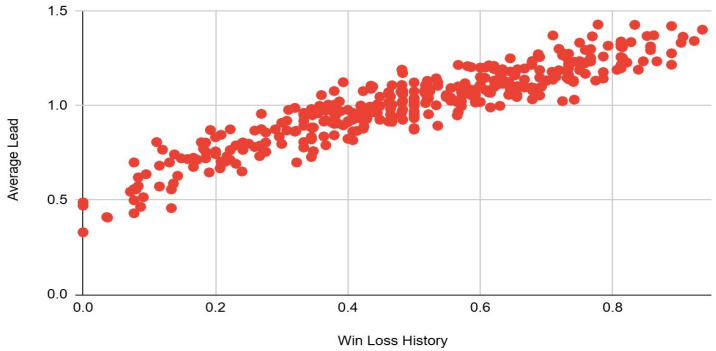
- Machine Learning models perform better with strong linear correlations
- Weaker correlations can be transformed using polynomial and log functions

Example, Average Lead:

Average Lead (untransformed)

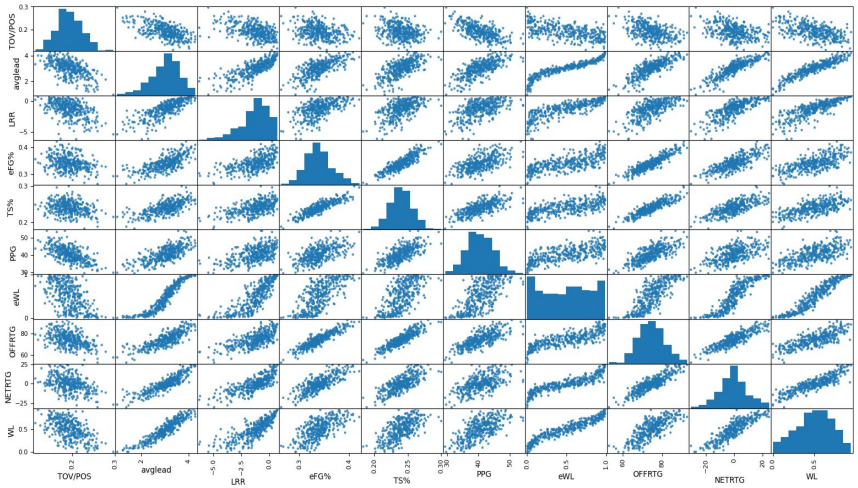


Average Lead (log, cubed)



Correlation Analysis

Scatter matrix created in Python



- Analyse and evaluate the relationship between variables.
- Spearman's rank correlation coefficient
- Select the most correlated variables to use as inputs in our models.

Use of LLMs in developing algorithms - ChatGPT

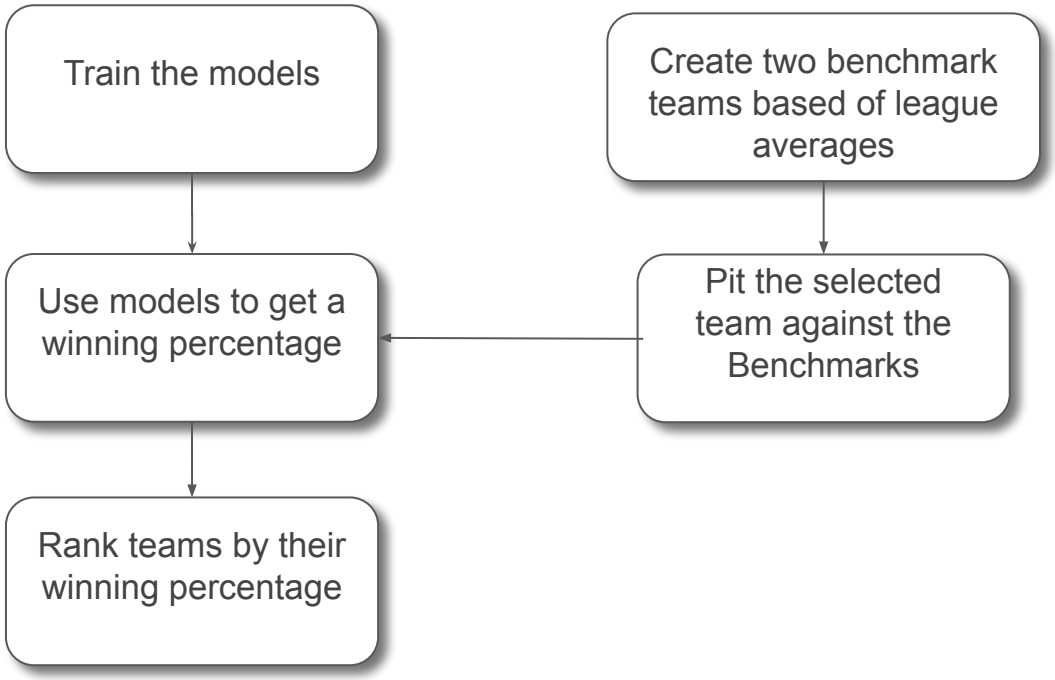
- Team had no prior experience/knowledge of machine learning
- Basic Python knowledge
- No experience using python for machine learning
- We researched different models to understand their concepts
- Used Chat GPT to develop the code for the models.

Evaluation

- If we were better equipped with the knowledge, we would have created the models ourselves
- Chat GPT allowed us to create and evaluate multiple models in the small time frame

- Many sports use ELO systems (point allocation)
- Benchmark teams
  - Average of the league
  - Average of the region

Ranking System



Classification Models

Random Forests

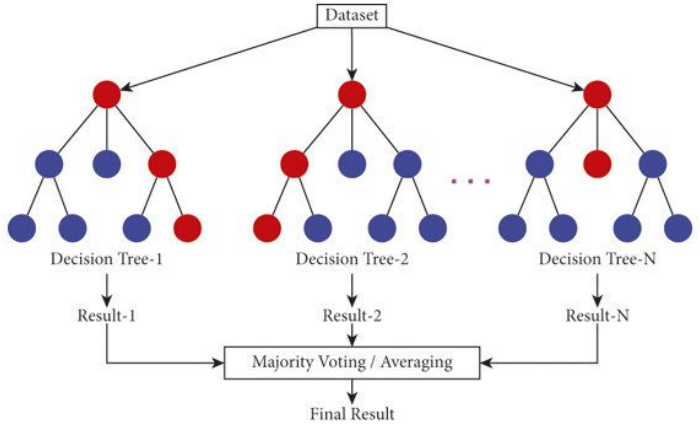
- Combines multiple independent decision trees

Pros

- Usually very accurate
- Reduces overfitting

Cons

- Resource Intensive



Source: researchgate.net

Classification Models

XGBoost (**Ex**treme **G**radient **Boost**ing)

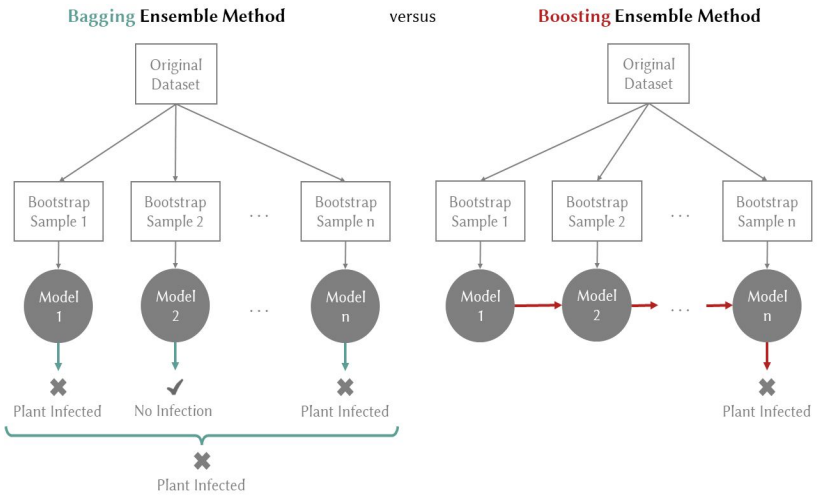
- Builds multiple decision trees (sequential)
    - New trees correct old trees

Pros

- Can reduce overfitting
- Fast training with large datasets

Cons

- Resource Intensive

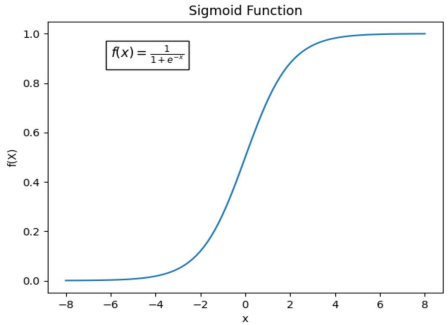


Source: towardsdatascience.com

Classification Models

Logistic Regression

- Estimates the probability of a label using logistic (sigmoid) function



Source: medium.com

Pros

- Can reduce overfitting
- Resource Inexpensive

Cons

- Performs poorly with nonlinear relationships between features

Classification Models

Naive Bayes

- Estimates the probability of a label using Bayes' Theorem
- Conditional probability

Pros

- Efficient

Cons

- More effective with text data
- Zero frequency problem
- Naive Assumption
    - Assumes conditions are independent

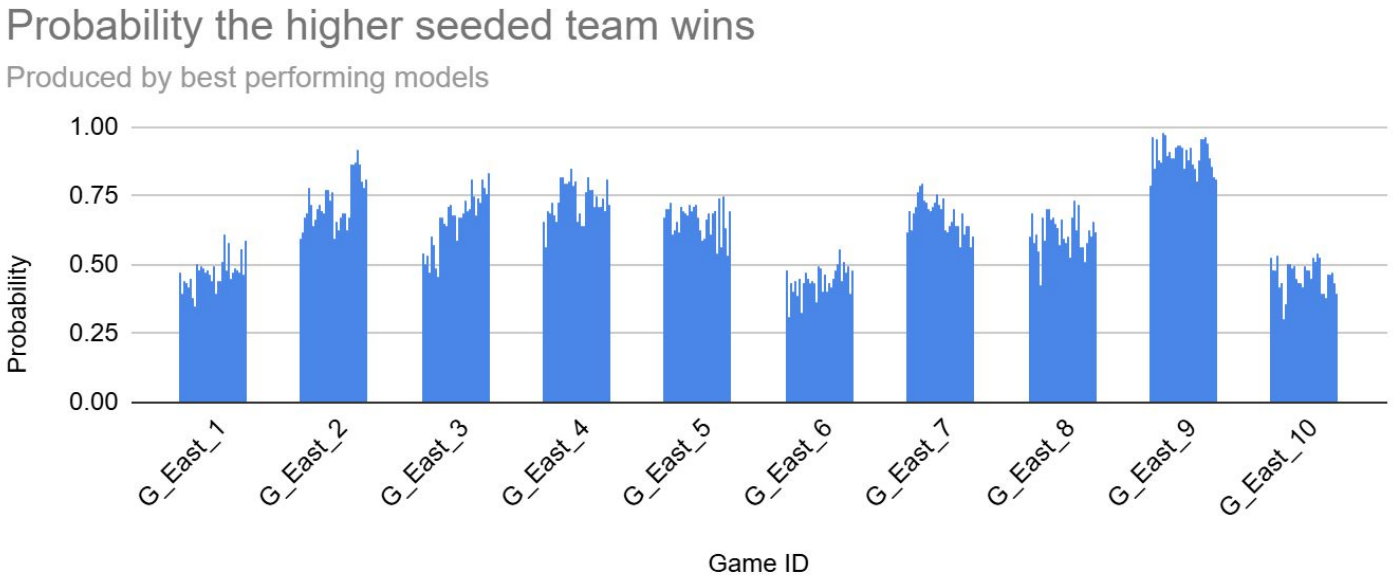
Definition:
The Zero Frequency Problem:
If no conditions are present, it causes the probability to be zero.

Model Evaluation

Model Type	Accuracy	AUC	F1
XGboost	0.7677	0.8499	0.8137
Logistic Regression	0.7768	0.8572	0.8181
Naive Bayes	0.7609	0.8414	0.7926
Random Forest	0.7732	0.8515	0.8166

We decided to discontinue the use of results from all Naive Bayes models as a result of its relatively lower metrics and the zero frequency problem.

Playoff Probabilities



Results

	SOUTH		WEST		NORTH	
Rank	Team	Win Loss History	Team	Win Loss History	Team	Win Loss History
1	Louisville Cardinals	0.8621	BYU Cougars	0.8889	South Carolina Gamecocks	0.9355
2	Iowa State Cyclones	0.8125	Stanford Cardinal	0.9032	Jackson State Lady Tigers	0.7692
3	Toledo Rockets	0.8667	Baylor Bears	0.8125	Stephen F Austin Ladyjacks	0.8571
4	Ohio State Buckeyes	0.7931	Nebraska Cornhuskers	0.7500	Lsu Tigers	0.8148
5	Iowa Hawkeyes	0.7667	Texas Longhorns	0.8125	Ucf Knights	0.8889
6	IU Indianapolis Jaguars	0.8571	South Dakota State Jackrabbits	0.7097	Belmont Bruins	0.7586
7	Virginia Tech Hokies	0.7188	Gonzaga Bulldogs	0.8125	Florida Gulf Coast Eagles	0.9231
8	Dayton Flyers	0.8462	South Dakota Coyotes	0.8276	Tennessee Lady Volunteers	0.7419
9	Michigan Wolverines	0.8148	UNLV Lady Rebels	0.8125	Ole Miss Rebels	0.7241
10	Notre Dame Fighting Irish	0.7333	Arizona Wildcats	0.7308	Mercer Bears	0.8125
11	Missouri State Lady Bears	0.7667	Creighton Bluejays	0.6897	Troy Trojans	0.7586
12	Indiana Hoosiers	0.7333	New Mexico Lobos	0.7333	Georgia Lady Bulldogs	0.6786
13	Depaul Blue Demon	0.6875	Utah Utes	0.6452	Middle Tennessee Blue Raiders	0.7667
14	Kentucky Wildcats	0.6333	Oklahoma Sooners	0.7500	Georgia Tech Yellow Jackets	0.6774
15	Murray State Racers	0.6786	Colorado Buffaloes	0.7333	South Florida Bulls	0.7742
16	Cleveland State Vikings	0.7083	Oregon Ducks	0.6207	Arkansas Razorbacks	0.5806

- Unweighted average of probabilities
- Final ranks did not reflect win loss history

Can Machine Learning effectively be used in sports?

Answering the research question created at the beginning of our methodology

Conclusion: Machine learning method is effective with some drawbacks

Pros:

- High model accuracy
  - Mean accuracy - 0.7727
  - Peak accuracy - 0.8047

Cons:

- Sports is affected by other factors
  - Luck, sentiment, condition etc
- Limited dataset

With the right resources, experience and knowledge, machine learning is an effective tool for sports

- Our model is tailored to this league
- Insights are specific to this league

Discussion

Through SHAP values, we can look at metrics to focus on through coaching

- Expected Win Loss
- Average Lead
- Net Rating
- Turnovers
- Free throw scoring