DSC 630 Term

Project Milestone 4

Zack DeNoto, Chris Goodwin, and Kenny Waite

Question/Hypothesis

“Can we predict the number of wins a college basketball team will have in a given year?”

Why?

How?

The Data

We will use Andrew Sundberg's *College Basketball Dataset* found on Kaggle.

This dataset contains statistics about every NCAA Division I college basketball team from 2015 - 2020.

The Data

24 columns (Both numeric and categoricals)

Range from standard basketball statistics like turnover percentage and three point shooting percentage, to more complex analytic values like power rating and adjusted tempo

Exploratory Data Analysis

Initial analysis to identify relevant data points

Leverage relevant data points in an attempt to build an accurate prediction model for the number of wins

Our analysis will focus on years 2015-2019

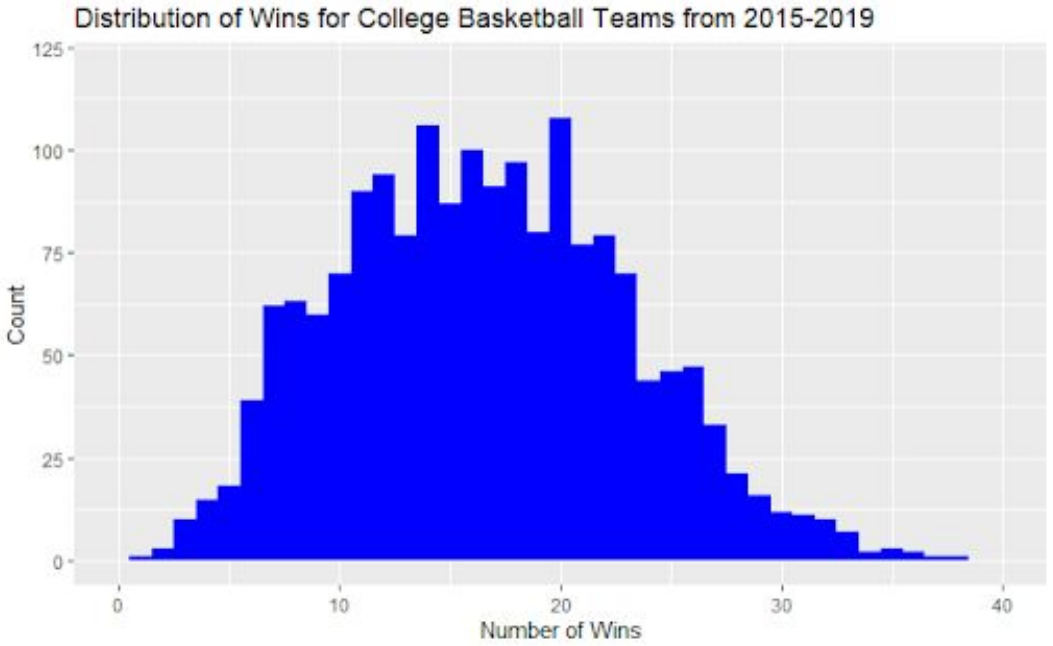
Exclude 2020 from our model due to the COVID-19 pandemic

Exploratory Data Analysis - Target Variable

Target Variable - **Wins**

Symmetric with a slight positive skew

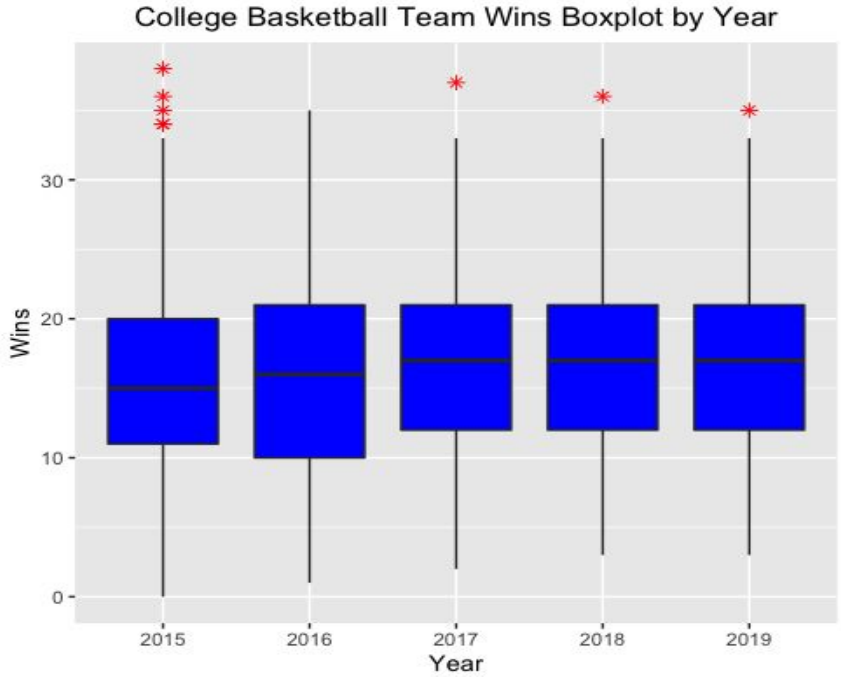
No data transformations needed



Exploratory Data Analysis - Number of Wins

Reviewed outliers and historical win data

Identified these totals as accurate and necessary for inclusion



Exploratory Data Analysis - Feature Selection

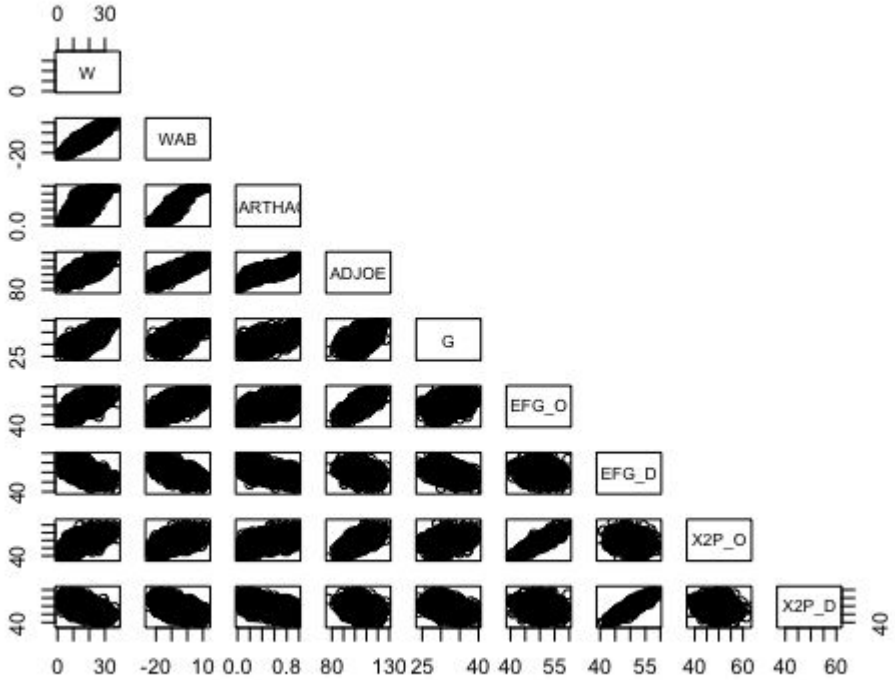
Identified variables with a correlation greater than 0.50 or less than -0.50

These variables will be the focus of our modeling

	Variable	Cor
1	WAB	0.90502921
2	BARTHAG	0.81451207
3	ADJOE	0.75453166
4	G	0.70883800
5	ADJDE	-0.69075290
6	EFG_O	0.61783930
7	EFG_D	-0.60914371
8	X2P_O	0.58580590
9	X2P_D	-0.52955814

Exploratory Data Analysis - Feature Selection

Correlation matrix shows strength and direction of relationships



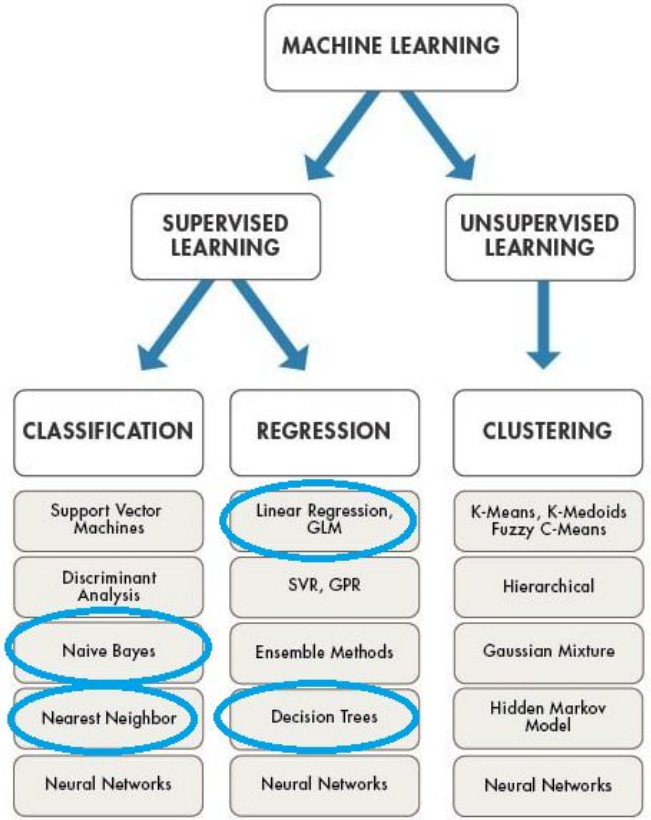
Modeling

K-Nearest Neighbor (KNN)

Naive Bayes

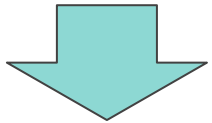
Random Forest

Ridge Regression



Preliminary Modeling Results

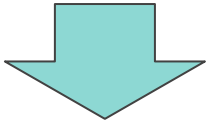
Model	Full Dataset	Unranked	Ranked	2015	2016	2017	2018	2019	Average Accuracy per Model
KNN Accuracy	8.0%	7.3%	7.6%	4.9%	4.1%	4.1%	7.7%	4.8%	6.2%
Naives Bayes Accuracy	12.1%	13.6%	8.8%	11.4%	10.6%	10.6%	11.4%	8.1%	10.8%
Random Forest Accuracy	11.0%	13.3%	8.8%	6.5%	8.1%	8.1%	11.0%	6.0%	9.4%
Ridge Regression Accuracy	0.1%	0.1%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	0.1%
Average Accuracy per Group	7.8%	8.6%	6.3%	5.9%	5.7%	5.7%	7.5%	4.7%	6.6%



Model	Full Dataset	Unranked	Ranked	2015	2016	2017	2018	2019	Average Accuracy per Model
KNN Accuracy	10.0%	11.1%	8.0%	11.0%	13.8%	13.8%	13.4%	8.1%	10.0%
Naives Bayes Accuracy	11.5%	12.6%	12.6%	11.0%	15.4%	15.4%	16.7%	8.9%	12.2%
Random Forest Accuracy	15.0%	12.8%	13.0%	12.2%	13.4%	13.4%	15.0%	12.9%	13.2%
Ridge Regression Accuracy	0.2%	0.1%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	0.1%
Average Accuracy per Group	9.1%	9.1%	8.4%	8.6%	10.7%	10.7%	11.3%	7.5%	8.9%

Adding Ensemble Techniques

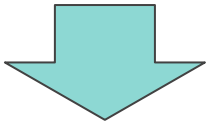
10 Estimators	Unranked	Full Dataset	Ranked	2015	2016	2017	2018	2019	Total
K-NN Test accuracy score	8.1%	5.9%	11.8%	8.5%	3.3%	3.8%	6.1%	6.6%	6.7%
Naives Bayes Accuracy	14.7%	12.8%	9.8%	10.8%	10.8%	9.9%	15.5%	8.9%	11.6%
Random Forest Test accuracy score	14.3%	14.7%	15.2%	10.3%	11.7%	15.0%	14.1%	15.0%	13.8%
Ride Regression Test accuracy score	0.1%	0.2%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.1%
AdaBoost	11.5%	9.3%	12.3%	8.9%	8.9%	14.1%	8.5%	8.9%	10.3%
GradientBoost	14.8%	13.3%	15.2%	10.3%	12.7%	14.6%	10.3%	11.7%	12.9%
GradientRegression	10.7%	11.1%	13.2%	13.1%	16.9%	14.6%	12.7%	11.3%	12.9%
Total	10.6%	9.3%	10.7%	8.2%	7.9%	9.5%	9.1%	8.5%	9.2%



200 Estimators	Unranked	Full Dataset	Ranked	2015	2016	2017	2018	2019	Total
K-NN Test accuracy score	7.9%	7.2%	8.8%	10.3%	3.8%	5.2%	6.6%	7.0%	7.1%
Naives Bayes Accuracy	13.5%	12.9%	8.8%	12.2%	14.1%	8.9%	13.6%	8.9%	11.6%
Random Forest Test accuracy score	16.2%	15.4%	14.7%	12.7%	11.7%	11.7%	17.8%	12.2%	14.1%
Ride Regression Test accuracy score	0.4%	0.0%	0.0%	0.9%	0.0%	0.0%	0.0%	0.0%	0.2%
AdaBoost	9.5%	9.9%	14.7%	9.9%	9.9%	8.5%	6.1%	13.6%	10.3%
GradientBoost	15.1%	13.8%	14.2%	17.4%	12.7%	8.5%	11.3%	17.4%	13.8%
GradientRegression	27.9%	27.6%	22.1%	21.1%	27.2%	23.0%	21.6%	21.1%	24.0%
Total	12.9%	9.9%	10.2%	10.6%	8.7%	7.1%	9.2%	9.9%	9.8%

Adding Ensemble Techniques Continued

10 Estimators	Unranked	Full Dataset	Ranked	2015	2016	2017	2018	2019	Total
K-NN Test accuracy score	13.7%	11.6%	11.3%	8.0%	9.4%	10.8%	14.6%	10.8%	11.3%
Naives Bayes Accuracy	12.8%	13.1%	10.8%	9.9%	17.4%	14.6%	15.5%	11.3%	13.1%
Random Forest Test accuracy score	16.8%	15.6%	16.2%	10.8%	16.9%	15.0%	13.6%	16.0%	15.1%
Ride Regression Test accuracy score	0.0%	0.1%	0.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.1%
AdaBoost	10.0%	11.5%	12.7%	8.5%	7.0%	8.5%	8.9%	8.5%	9.4%
GradientBoost	15.5%	15.7%	13.7%	14.6%	14.1%	13.6%	15.5%	9.9%	14.1%
GradientRegression	9.2%	10.4%	12.7%	10.8%	14.6%	14.6%	15.5%	13.6%	12.7%
Total	11.1%	11.3%	10.8%	8.7%	10.8%	10.4%	11.3%	9.4%	10.5%



100 Estimators	Unranked	Full Dataset	Ranked	2015	2016	2017	2018	2019	Total
K-NN Test accuracy score	13.4%	11.7%	5.4%	7.0%	8.0%	7.0%	12.2%	10.8%	9.4%
Naives Bayes Accuracy	14.0%	11.3%	9.8%	11.7%	12.7%	8.0%	13.6%	12.7%	11.7%
Random Forest Test accuracy score	15.0%	15.8%	15.7%	12.7%	13.6%	8.5%	17.4%	11.7%	13.8%
Ride Regression Test accuracy score	0.1%	0.1%	0.0%	0.9%	0.0%	0.0%	0.0%	0.0%	0.1%
AdaBoost	12.0%	10.0%	10.8%	10.8%	6.6%	11.3%	10.3%	9.9%	10.2%
GradientBoost	15.6%	13.4%	13.2%	9.9%	16.0%	10.3%	13.1%	11.7%	12.9%
GradientRegression	20.0%	21.1%	16.7%	19.2%	21.6%	16.0%	15.0%	19.2%	18.6%
Total	12.9%	10.4%	9.2%	8.8%	9.5%	7.5%	11.1%	9.5%	9.8%

Results

- Grouping Impact
- Ensemble Impact
- Best and worst accuracies

Accuracy within 0 games off	32.39%
Accuracy within 1 game off	59.51%
Accuracy within 2 games off	89.79%
Accuracy within 3 games off	98.01%
Accuracy within 4 games off	100.00%

Conclusion and What's Next

- Exploratory data analysis broke down the data
- Initial modeling accuracies improved when we only included the variables with high correlations
- Boosting models improved overall accuracy substantially
- Where do we go from here?