Holden Svacha
Group #4

How Teams Win in the Big Ten Football Conference

Introduction

College Football is a very large culture in the U.S., with thousands of people packing into stadiums to watch the physical sport every year, with millions watching behind a TV. As a student at a Big Ten school I also avidly watch the sport and cheer for green and white, and I wonder why the red and white team that we shall not name sees much more success than us. It's easy to think, "Well they're just better than us", but it's hard to find the exact correlation between the team and its success. With so many variables to keep track of there's no wonder why football has such a large need for data scientists and analysts like us.

Every MSU student has one thing on their mind, "Why can't we be as good as we were a couple years ago?". It's a good question, considering that MSU has the 5th highest winning percentage in the Big Ten against 11 other teams, at a win to loss ratio of .601. If that were true this season, we would've gone close to 7-6 rather than 4-8, so what gives? With teams like Ohio State having a consistently good team every year, to the point where they are the only team in the Big Ten to not have a season with 9 or more losses in its existence, what makes teams successful in the Big Ten?

Our question that we wanted to answer in our research was, "What makes teams win more games in the Big Ten?". To answer this we had to break it into three key questions, which are: "Does a Good Offense Always Win Games?", "Does Defense Win Games?", and "Does Special Teams Really Matter?". These questions are important since football is broken up into three different qualities which are offense, defense, and special teams. If we were able to find out which qualities are more important than others, we could make a prediction model and find out what really makes teams win games.

Methodology

For our research, we used the "College Football Team Stats Seasons 2013 to 2023", and aggregated the data over 11 years including only teams that are in the Big Ten as of 2023. We're including only Big Ten teams since that's the conference that we are focused on, and we aggregated the data to have more to work with as it would cause issues if we had data from only one year for each of the teams.

Holden Svacha
Group #4

Finding our sub-variables

When we started our research we knew we needed to split offense, defense, and special teams into smaller sub-variables such as total yards, interceptions, and punt return yards. To find which variables we would choose we decided it would be best to find which variables have the highest correlation to win ratio, the highest correlated variables being the ones that have the most impact on the win ratio or change it the most. To do this, we decided we would create a correlation matrix, a data method that is used to find how correlated a variable is to another given variable, which in this case would be the win ratio. At this point it is important to note that we created the win ratio variable as it was not included in the data, and the reason we are using the ratio is because not every team has the same number of games played.

Pre-Regression Analysis

We knew that we wanted to look at the sub-variables that we chose before we started looking into regression plots and other prediction models, so we decided we'd use bar charts highlighting each teams success in a given variable. This way, it would be easier to identify any mistakes we made later on as it would be easy to spot if one teams regression data didn't align with what we saw from our data previously.

Regression Data

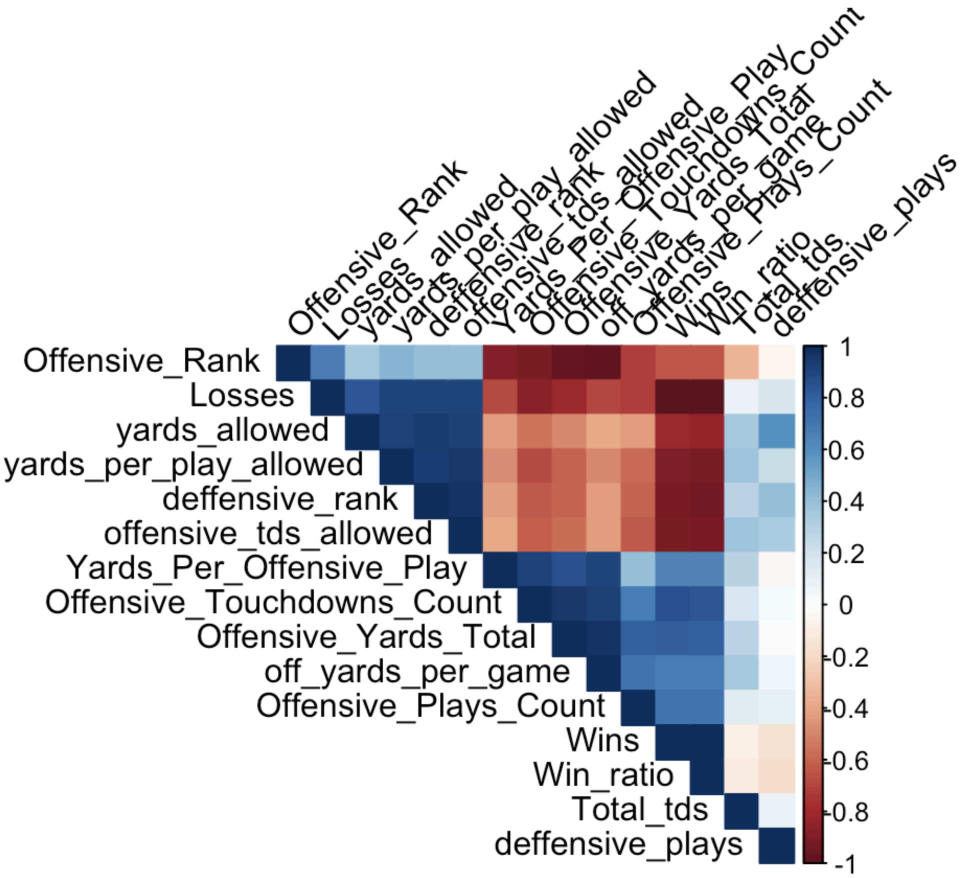
This is where we thought that everything would come together, when we find out which variables are the most correlated with win ratio, and we can start making predictive models to decide what really goes behind a good team. The plan was that we would choose our variables that we see fit and see how much they impact the win ratio given every one unit increase of such variables, this way we could use our own knowledge of football to determine which variables truly have the most impact on win ratio. Once we get our chosen variables, we can then make a "perfect" model which will show the predicted win ratio for each team given how well they do in our chosen variables, and we'll know if our model is correct if the win ratio's roughly match with the true win ratios.

Holden Svacha
Group #4

Results

Using a Correlation Matrix

For our correlation matrix we used the cor_matrix command which is a part of the corrplot package in R-Studio, which proved to be sufficient in finding our variables that had the highest correlation. How the correlation matrix works is it looks at the data from each year and compares each variable to each other, if a variable increases when another variable increases, it will have a positive correlation, but if the variable decreases when it increases, it will have a negative correlation. The further away a correlation value is from zero, the higher correlated the given variable is to another variable, with positive meaning it increases the variable and negative meaning it decreases it. With that out of the way let’s take a look at one of our correlation matrices, we’ll use our offense matrix and go from there.



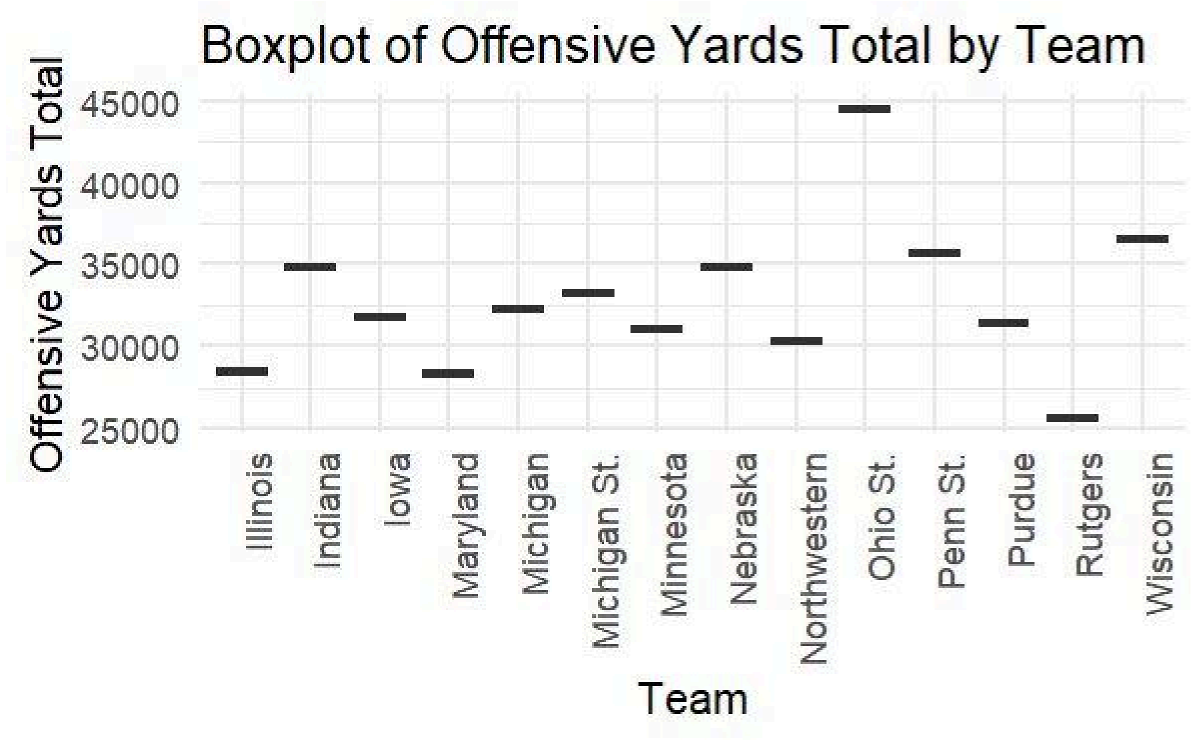
Holden Svacha
Group #4

As we can see from our matrix, correlation varies with which variables you are comparing to one another. To find out which variables correlate to win ratio, we'd want to look at the third to last row in this matrix and see which columns are highly correlated to our third to last row. In this matrix, we notice that Win_ratio has a high positive correlation to Yards_Per_Offensive_Play, which means that the more the Yards_Per_Offensive_Play variable increases, the more the Win_ratio variable increases. This is a variable we would end up using in our Pre-Regression analysis as well as our Regression Data, as it could be an important variable in the question we are trying to answer.

We would then repeat this step for a defensive and special teams matrix until we have the variables that we believe are most significantly correlated with Win_ratio.

Creating Bar Charts of our Correlated Variables

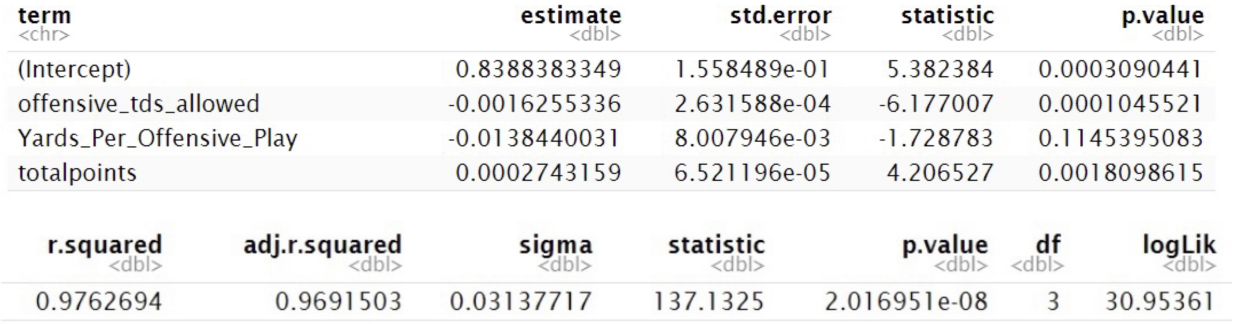
As I mentioned before, once we got our chosen variables from our correlation matrix, we wanted to further analyze them using bar charts. We chose bar charts since it would allow us to see each team's given statistic in the variable we were trying to analyze. The way we will create this bar chart is with the ggplot package, creating a ggplot and using the geom_boxplot function to create a bar chart that is grouped by Team. In this example we're going to use the Yards_Per_Offensive_Play variable since it's the same variable we used for the last example. Upon finishing the code for our bar chart, this is what we come up with.



From this we can see that there’s a lot of variety with each team in accordance to the Yards_Per_Offensive_Play variable, with Northwestern having the lowest at 33.22 and Ohio State having the highest at 47.06. At this point it is worth noting that our data is aggregated over 11 years which means that Northwestern gets an average of around 3.2 yards per offensive play in any given year while Ohio State gets around 4.3 yards per offensive play on average in any given year. With this much variety in the variable and its data matching up with what we’ve found in our background analysis, it would make a perfect variable for our next step, which is to create predictive tables of our win ratio in accordance to a given variable.

Prediction Model using Regression

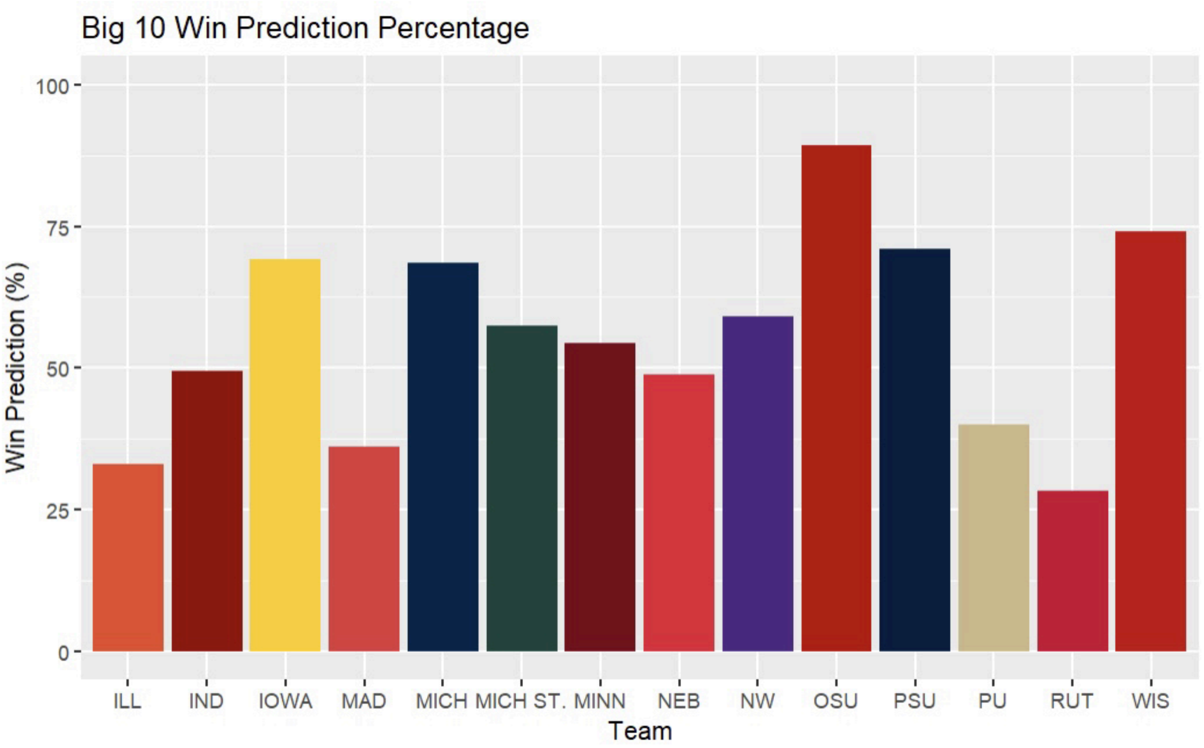
Now that we’ve chosen our variables, we can begin creating prediction models to see which variables have the most impact on Win_ratio. During this process we use the lm() function which is a part of the ‘stats’ package in R-Studio to create different regression models with variables that we chose from our previous steps. An example of our “perfect” model is below, which includes the main values we look at during this step which are “estimate” and “r.squared”. Estimate is how much your chosen variable increases for every one unit increase in a different variable, for example we can see in our “perfect” model that Win_ratio begins at 0.8388 and decreases by 0.0016 for every one unit increase in offensive_tds_allowed and increases by 0.00027 for every one unit increase in totalpoints. This makes sense as the opponent scoring touchdowns will result in more losses while points scored increases your chances of winning in football. R.squared is the other variable we looked at which is in simple terms how “fit” how model is, or how trustworthy it is. We wanted our model to have a high r.squared value, that way we knew it was as accurate as it could be.



```
term                          estimate     std.error    statistic        p.value
<chr>                            <dbl>         <dbl>        <dbl>          <dbl>
(Intercept)               0.8388383349  1.558489e-01     5.382384   0.0003090441
offensive_tds_allowed    -0.0016255336  2.631588e-04    -6.177007   0.0001045521
Yards_Per_Offensive_Play -0.0138440031  8.007946e-03    -1.728783   0.1145395083
totalpoints               0.0002743159  6.521196e-05     4.206527   0.0018098615

  r.squared   adj.r.squared        sigma    statistic         p.value    df      logLik
      <dbl>           <dbl>        <dbl>        <dbl>           <dbl> <dbl>       <dbl>
  0.9762694       0.9691503   0.03137717     137.1325    2.016951e-08     3    30.95361
```

Now that we have our perfect model, we can make new dataframes for each team which include the variables given in our perfect model. We then use the “predict” function which is a part of the “margins” package in R-studio, to find a confidence interval of Win_ratio for each team. A confidence interval is an interval including a middle point and two endpoints, the middle point being the best fit value with the end points being opposite ends of the best fit value determining that the model is some percentage confident that the true sample mean lies between the endpoints, the percentage in our case would be 95%. Once we get our confidence intervals for each team, we can make a beautiful prediction bar chart including all our teams and their predicted win ratio using the variables from our “perfect” model.



Conclusion

Summary

We were able to learn a lot from our research on what variables influence how successful teams are, with the main points being that offense had the largest impact on how successful a team is with special teams variables being nearly negligible. This does make sense as offense is the most direct way of scoring points in football, while special teams sees the field significantly less than offense or defense.

Critique of Methodology

Looking back at our methodology, while our final results proved to be sufficient, some of our previous questions could’ve used a bit more coverage. I believe it would’ve been beneficial had we made regression models of offense, defense, and special teams before we made our “perfect model”. That way we could see how each variable affects the win ratio compared to the others, and effectively tell whether offense or defense mattered more. Were we to do this again I think we would make regression models of offense, defense, and special teams; then take the two variables with the highest “r.squared” and furthest away from zero “estimate” values to use in our “perfect” model, that way variables from each category would be used in our “perfect” model, further answering our question.

Limitations/Reliability

One of the limitations we had while working with our data was we couldn’t use the offensive rank or defensive rank variables since the teams were originally ranked with all teams in the NCAA I division, rather than only Big Ten teams. Were we able to use those variables I imagine our “perfect” model would’ve had a much higher r.squared value and overall more trustworthy.

Future Research

If future research is ever done on this same topic, I think something that would possibly be interesting would be to create new variables for offense, defense, and special teams of their impact toward winning the game based on their time on the field.

Holden Svacha
Group #4

References

Knoop, T. (2022, May 26). *Ranking Big Ten teams by all-time winning percentage*. WolverinesWire. Retrieved April 20, 2024. https://wolverineswire.usatoday.com/lists/big-ten-football-teams-all-time-winning-percentage/

NCAA.com. (2024, April 19). *NCAA College Football FBS stats*. Retrieved April 20, 2024. https://www.ncaa.com/stats/football/fbs

College Football Team stats Seasons 2013 to 2023. (2024, March 31). Kaggle. Dataset Created by Jeff Gallini. Retrieved April 20, 2024. https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019/data