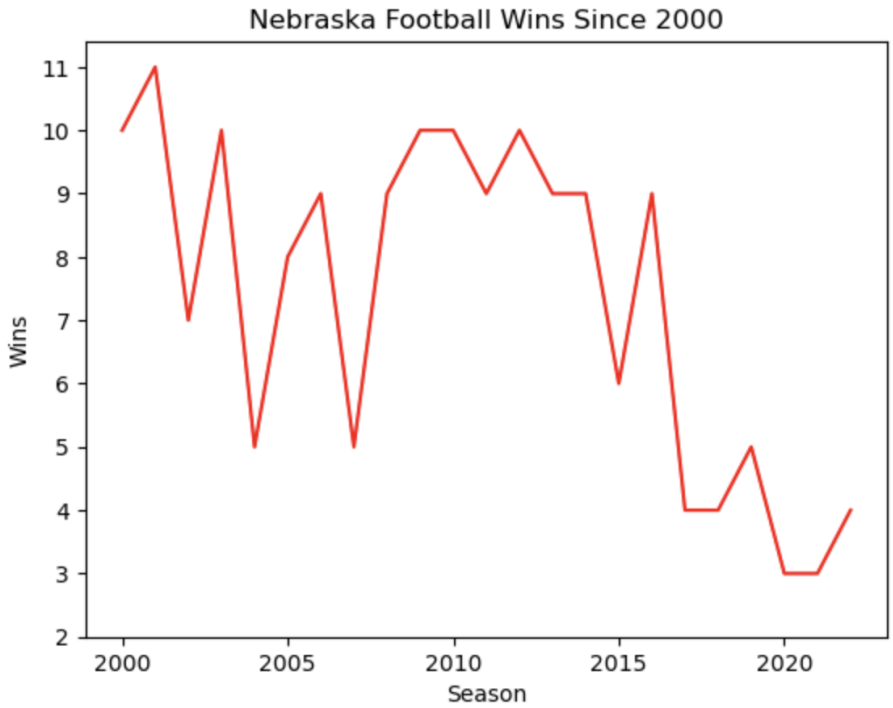
**Using Football Game Data to Build Better Practices**

The data used in this report includes box score data for every Nebraska Cornhuskers football game since September of 1962. This includes data such as date, time and weather during the game as well as rushing, passing, turnovers, penalties for both teams. It carries almost all of this information for every game up until present day. This data could be very useful for the coaching staff or players in order to discover the most effective way to develop a more successful program. Another use of this data has been the investigation of the correlation between variables. For example, the correlation between rushing yards and scoring is higher than that of passing yards and scoring.

In the last few years, Nebraska's football team has not achieved as much success as it has in the 2000's as seen in the plot below.



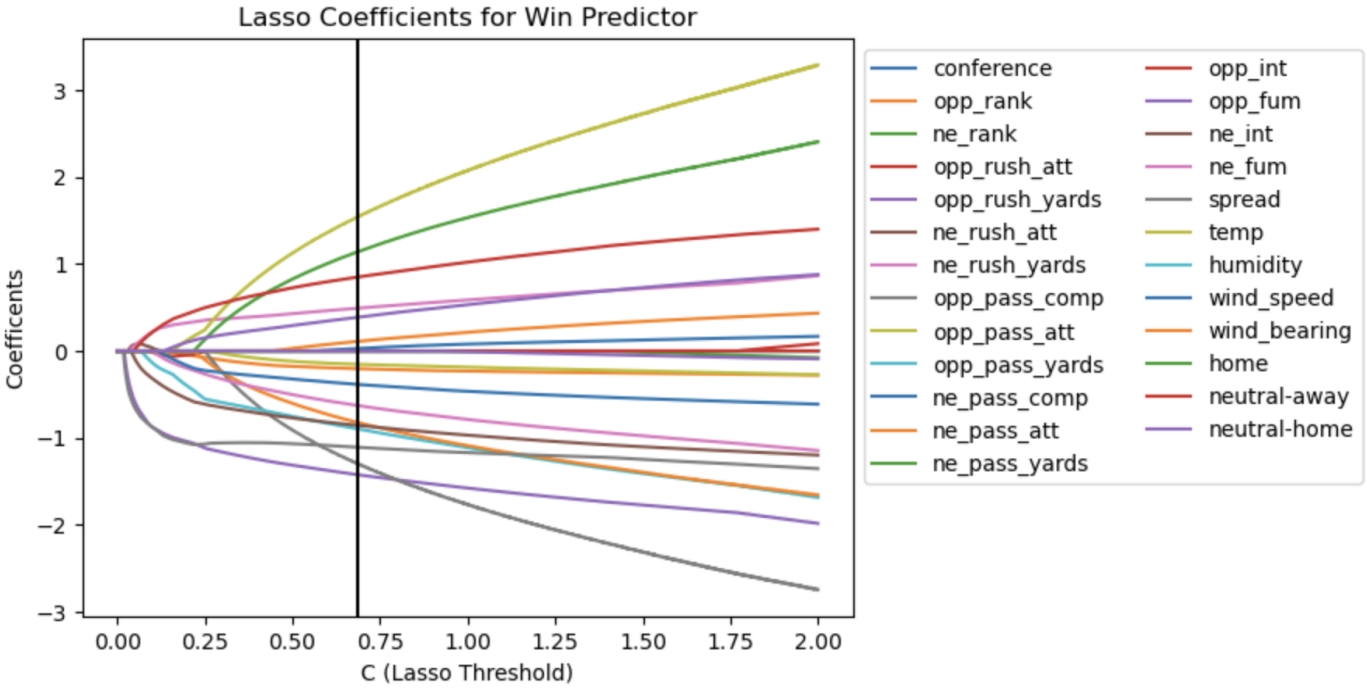

The above graph shows the program's wins throughout the seasons 2000 - 2022. Note that in 2020 Nebraska only played 8 games due to the Covid 19 pandemic.

As we can see, before 2017 the team could quite consistently achieve well above 5 win seasons. Since 2017, the highest they have reached is 5 in 2019 , with a drop to 3 in 2020 and 2021. In order to return to their prior status as a program, they need to identify the areas in which they need to improve in order to win more games. Due to the nature of this task, I chose to fit a lasso model to predict if the team would win or not using games from the year 2000 until present. I did this because a lasso model will go through feature selection to show which of the variables is most important in deciding a win. Through looping through several models with various C thresholds on training data, I found the threshold with the largest AUC on the test data. At this threshold we can see that the variables not having strictly to do with how the teams performed have little influence on the outcome of the game.

| humidity | wind_speed | wind_bearing | home | neutral-away | neutral-home |
|---|---|---|---|---|---|
| 0.0 | -0.384770 | 0.108819 | 0.000000 | 0.0 | 0.000000 |
| 0.0 | -0.385089 | 0.109276 | 0.000000 | 0.0 | 0.000000 |
| 0.0 | -0.385046 | 0.109584 | 0.000000 | 0.0 | 0.000000 |
| 0.0 | -0.385742 | 0.110149 | 0.000000 | 0.0 | 0.000000 |
| 0.0 | -0.385894 | 0.110377 | 0.000000 | 0.0 | 0.000000 |

Lasso models use a parameter to limit the coefficients in the model. Higher coefficients designate higher importance of the variable in the prediction. Coefficients of 0 show that the lasso model does not include these variables in the model at this threshold.

It is good that we can get rid of these variables because unfortunately, looking at the coefficient plots, it is quite difficult to understand due to the amount of variables. With all the variables included the coefficient plot looks like the jumbled mess below.




As stated above, higher absolute value coefficients designate variables as important predictors in the models. The black vertical line signifies the threshold which we found with the lowest AUC.

In order to solve this problem, I separated the variables into Nebraska's statistics and their opponents statistics. These can largely be thought of as offensive (their statistics) and defensive (opponents statistics), which can help us isolate what specifically the team needs to improve. I created two logistic classifier models with these separated explanatory variables to find their accuracy in predicting a win. To find their accuracy I simply created confusion matrices for each

model. Using a model consisting only of the Nebraska variables yielded the following confusion matrix.

| **Nebraska Statistics** | Predicted Loss | Predicted Win |
|---|---|---|
| Lost | 83 | 37 |
| Win | 27 | 134 |

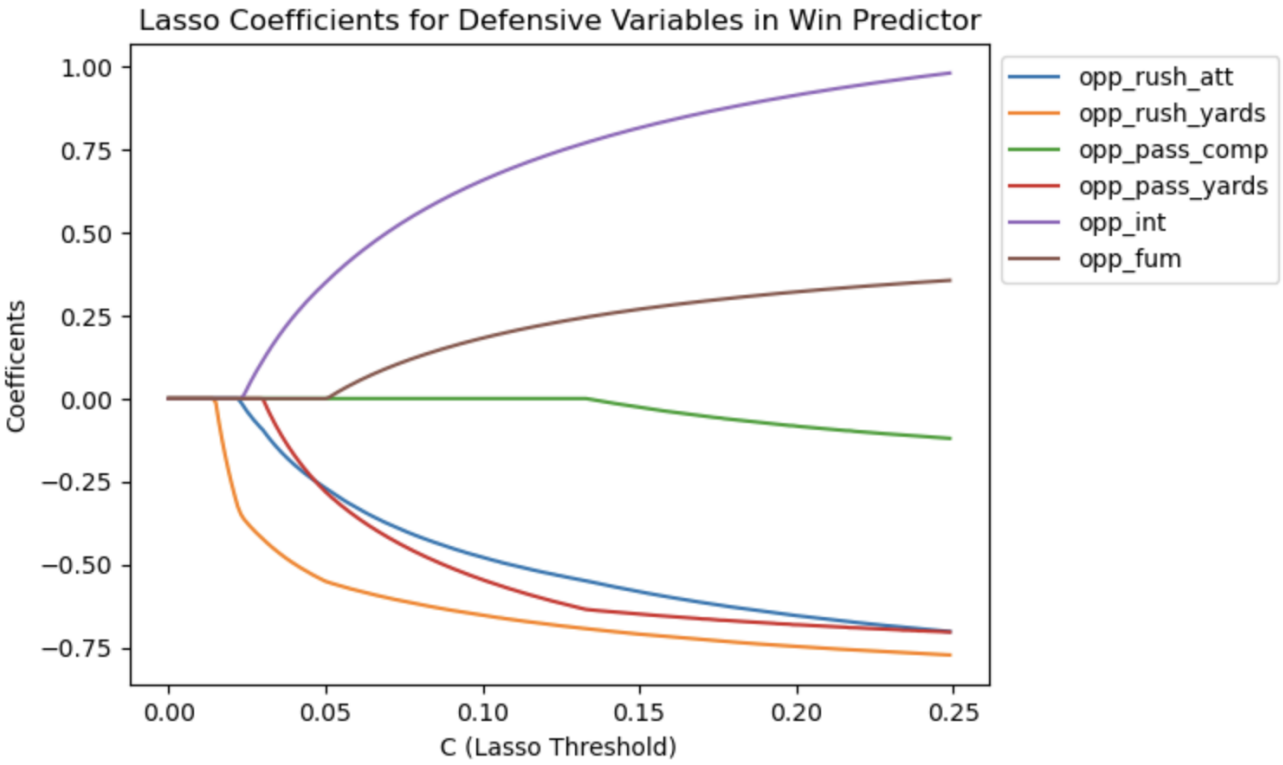
Confusion matrix of Logistic Regression model consisting of variables 'ne_rush_att', 'ne_rush_yards', 'ne_pass_comp', 'ne_pass_yards', 'ne_int' and 'ne_fum'.

Similarly, using a model only consisting of opponent variables, the following confusion matrix was created.

| **Opponent Statistics** | Predicted Loss | Predicted Win |
|---|---|---|
| Lost | 97 | 23 |
| Win | 20 | 141 |

Confusion matrix of Logistic Regression model consisting of variables 'opp_rush_att', 'opp_rush_yards', 'opp_pass_comp', 'opp_pass_yards', 'opp_int' and 'opp_fum'.

The models resulted in accuracies of 0.758 and 0.819 respectively, meaning that the model fit using opponent data is a more accurate predictor of a win. Now that I knew that these variables were more significant, I recreated the plot from before but with this more limited number of variables. The resulting graph is much easier to interpret and gives us a good idea of the importance of each variable in each game.




Again, higher absolute values of coefficients designate higher importance in our prediction model.

From this model we can see that the opponents rushing yards and interceptions thrown are the greatest predictors of this win model. What does this mean in actuality though, how can we use this information to win more games. Through this we have learned that the most important factors for Nebraska to win is to force more interceptions and minimize rush yards. Coaching staff can use this information to center practices around pressuring the backfield because this would help them accomplish those two goals and hopefully set them on the track to winning more games.

# Bibliography


Source for data set:

Data Set Name: Nebraska Football Box Scores 1962-2022,

Author: CVIAXMIWNPTR

Updated 5 Months Ago

https://www.kaggle.com/datasets/cviaxmiwnptr/nebraska-boxscores-19622019

Source for other data investigation:

https://www.kaggle.com/code/kerneler/starter-nebraska-boxscores-1962-2019-8e5841f7-1