

## Executive Summary

In the game of soccer, there are countless different ways a possession can turn out. We must:

- "Create a network for the ball passing between players, where each player is a node and each pass constitutes a link between players. Use your passing network to identify network patterns, such as dyadic and triadic configurations and team formations. Also consider other structural indicators and network properties across the games. You should explore multiple scales such as, but not limited to, micro (pairwise) to macro (all players) when looking at interactions, and time such as short (minute-to-minute) to long (entire game or entire season)."
- "Identify performance indicators that reflect successful teamwork (in addition to points or wins) such as diversity in the types of plays, coordination among players or distribution of contributions. You also may consider other team level processes, such as adaptability, flexibility, tempo, or flow. It may be important to clarify whether strategies are universally effective or dependent on opponents' counter-strategies. Use the performance indicators and team level processes that you have identified to create a model that captures structural, configurational, and dynamical aspects of teamwork."
- "Use the insights gained from your teamwork model to inform the coach about what kinds of structural strategies have been effective for the Huskies. Advise the coach on what changes the network analysis indicates that they should make next season to improve team success."
- "Your analysis of the Huskies has allowed you to consider group dynamics in a controlled setting of a team sport. Understanding the complex set of factors that make some groups perform better than others is critical for how societies develop and innovate. As our societies increasingly solve problems involving teams, can you generalize your findings to say something about how to design more effective teams? What other aspects of teamwork would need to be captured to develop generalized models of team performance?"

[1]

## Background/Theory

Being able to work as a team is crucial in everyday life whether it be in the workplace, in the classroom, on the sports field, etc. Recently, there have been many new advances in predictive modeling technology, and we were able to use these to our advantage when looking at a very tangible example of teamwork: soccer. There are many factors that can influence the outcome of a soccer match, but one of the most prominent is teamwork. We sought out to analyze the passing patterns that give our team the greatest chance at success.

## Assumptions and Justification

- First of all, we are classifying a shot to be a win. We believe it's more efficient to do it that way rather than finding when a goal was actually scored. We agreed if a shot got off that didn't result in a goal, that doesn't have to mean it was a bad shot selection, as it could have simply been a bad shot by the shooter or a good save by the goalkeeper.
- Secondly, we are using the outcome of each possession instead of individual events. We chose to do this because many events may have a positive effect on the game without causing an instant shot. Turnovers are almost exclusively negative so any event that ends a possession without a shot is undesirable. However, there are passes made that improve the team's position, but don't result in a shot. A pass from a contested possession to an open man, for example. Another example is a combination teammates who may have good chemistry that causes their interactions to have increased rates of shots.
- Thirdly, in many parts of our analysis we need a binary outcome, i.e win or loss. In this case a tie is not an option, so we are giving the win to the team with greater number of passes. The reasoning behind this is that we are assuming that more passes means the team had more possession. Generally if a team has significantly more possession, they were in control of the game while their opponents were on the defense.
- Lastly, home field advantage is not a factor for the purpose of our project. We did not group wins or losses by whether the huskies were home or away, so we are not focused on that aspect. We are assuming that the huskies and the opponent play at the same level whether playing at home or traveling on the road. We are also assuming ideal weather for each game.

## Methodology and Results

- The first thing we did is find the mean field position for each player on the Huskies starting squad.
- Next we created a matrix showing the number of passes made by a player that went to each individual teammates.
- The previous two steps were combined graphically for visualization purposes:

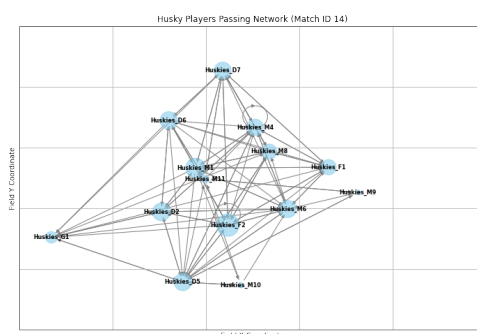


Figure 1: Husky Passing Network

- Our next step was to create a key factors matrix with the following arguments: Match ID, Own Score, Opponent Score, Total Passes, Total Shots, Opponent Pass, Opponent Shot, Shot differential and Goal Differential.

MatchID	OwnScore	OpponentScore	TotalPasses	OppPasses	TotalShots	OppShots	pass_diff	shot_diff	goal_diff
1	1	0	369	197	8	10	172	-2	1
2	1	1	180	416	7	18	-236	-11	0
3	0	2	324	471	7	18	-147	-11	-2
4	0	3	354	345	9	15	9	-6	-3
5	0	4	382	373	6	12	9	-6	-4

Figure 2: Key Factors Matrix

- To test the advantage of passing the ball more we created a scatterplot with passes on x-axis and goals scored on the y-axis Huskies (Blue) Opponent (Red):

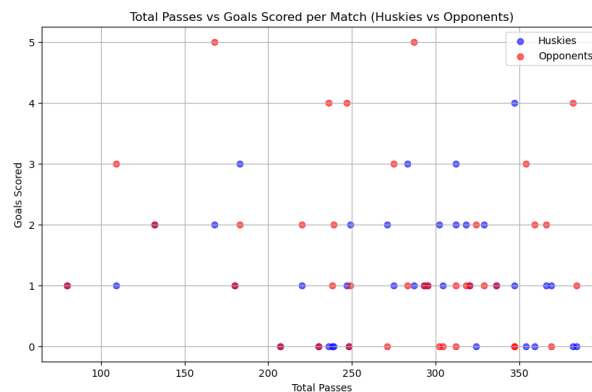


Figure 3: Husky Passes vs Goals Scored

- As you can see from Figure 3, the Huskies tended to score more when they passed more. Secondly, notice that the opposition scored less when the Huskies pass the ball more, as this allowed the Huskies to keep possession. The more the Huskies controlled possession, the more shot opportunities they were able to get, while also forcing the opposition to defend.
- Next, using the average position of each player, we used a k-Means clustering algorithm to find the formation the team played in a given match. We then compared to find which formations were most effective. We also used this find opposing formations that the Huskies struggled with.
- To find the formations used, we applied a k-Means clustering algorithm. K-Means starts by choosing k number of starting points, in this case we were looking for the quantity of: defenders, midfielders, and forwards. Thus the appropriate value is  $k = 3$ . These points are chosen randomly. Then the following process is repeated until every player stayed in the same cluster:
- After the points are chosen, every player was assigned to the closest point. These are the clusters. The points move to the center of their cluster and the process is repeated.
- We then used the formations of each match and appended them to their corresponding match. This allowed us to look at win-loss record based on formation used. (Note that these formations may not be the intended formations, however, it is what the team ended

up playing. This is because we took the average position for each player throughout the game. For example, they could have been trying to play a 4-3-3, and maybe they were getting heavily pressured and accidentally dropped back to a 4-4-2.)

Husky_Formation	MatchID
4-3-3	5
4-4-2	6
5-3-2	1
5-3-3	1

Figure 4: Formations won with

Husky_Formation	MatchID
3-5-2	2
4-2-4	1
4-3-3	2
4-4-2	7
4-5-1	1
5-4-1	2

Figure 5: Formations lost with

- As you can see from Figures 4 and 5, the Huskies tend to win a lot using the 4-3-3 and the 4-4-2 formations. These are typically the two most common formations in soccer. You can also see that they lost a lot when using the 4-4-2, but they only lost 2 games using the 4-3-3. This makes their 4-3-3 formation record 5 wins and 2 losses, and their 4-4-2 record 6 wins and 7 losses. It would be smart for the Huskies to play a 4-3-3 formation next season because they have the best record while using it.

Opponent_Formation	MatchID
3-5-2	1
4-3-3	1
4-4-2	7
4-5-1	2
4-5-2	1
5-3-2	1

Figure 6: Formations won against

	MatchID
Opponent_Formation	
4-3-3	1
4-4-2	5
4-5-1	3
4-5-2	1
5-3-2	1
5-4-1	4

Figure 7: Formations lost against

- Figures 6 and 7 give somewhat inconclusive data, because while the Huskies lost 5 times to the 4-4-2, they also won 7 times against it, so they aren't horrible at defending it. They are, however, not good at playing against the 5-4-1 as they had 0 wins and 4 losses, so maybe they could work on defending against a defensive play style such as that in practice.
- Our next action was to find the amount triatic and diatic passes made involving each combination of players. Triatic meaning passes made between three players. While diatic means passes made back and forth between players.

```
(Huskies_M1, Huskies_M3)    312
(Huskies_F2, Huskies_M1)    280
(Huskies_D1, Huskies_D3)    211
(Huskies_D1, Huskies_M1)    198
(Huskies_D5, Huskies_M1)    177
(Huskies_D5, Huskies_F2)    160
(Huskies_D1, Huskies_M3)    153
(Huskies_D4, Huskies_M1)    151
(Huskies_D1, Huskies_G1)    150
(Huskies_D3, Huskies_M1)    143
Name: count, dtype: int64
```

Figure 8: Diatic Passes

```
(Huskies_D3, Huskies_D1, Huskies_D4)    22
(Huskies_D1, Huskies_D2, Huskies_D3)    21
(Huskies_D4, Huskies_D1, Huskies_D3)    19
(Huskies_D3, Huskies_G1, Huskies_D1)    17
(Huskies_D1, Huskies_G1, Huskies_D3)    17
(Huskies_M3, Huskies_M1, Huskies_F2)    17
(Huskies_M1, Huskies_F2, Huskies_D5)    17
(Huskies_D5, Huskies_F2, Huskies_M1)    16
(Huskies_M4, Huskies_F2, Huskies_M1)    15
(Huskies_D1, Huskies_D3, Huskies_G1)    15
Name: count, dtype: int64
```

Figure 9: Triatic Passes

- We also created a scatter plot with with passes on x-axis and shots on y-axis:

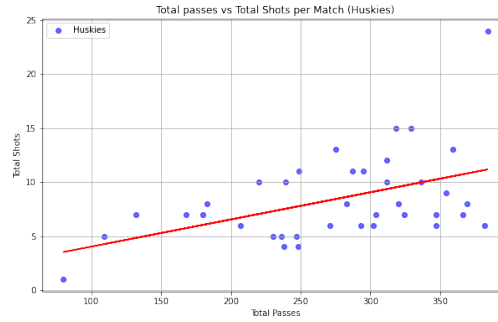


Figure 10: Husky Passes vs Shots

- We noticed here that there looked to be a correlation between passes and shots (which we expected) so we threw a regression line on there to better visualize the data. The equation for this line is  $y = 0.0251455x + 1.5159595$ . This shows that assuming you start with 1.5 passes before the game even starts, the data says it takes about 40 passes to get a shot ( $0.025 * 40 = 1$ ). As we learned in high school statistics, correlation does not equal causation, but this was an interesting bit of information that we found useful.
- We also found how many passes occurred in each ten minute portion of the game and expressed via a bar graph:

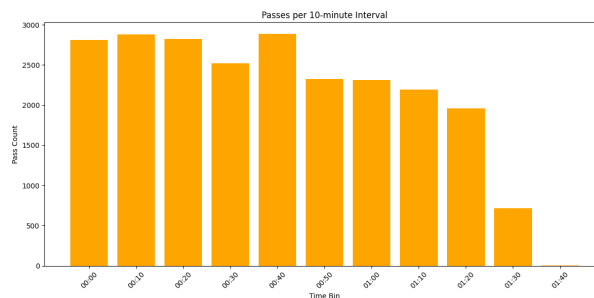


Figure 11: Passes per 10 minutes

- Figure 11 is just an example from one game, however most games followed a pattern similar to what is seen here. As you can see there is a decrease in passes as the game progresses, with a substantial decrease in the last ten minutes. There are two likely explanations to this:
- Firstly, The Huskies are poorly conditioned, and as the match progresses they get tired and start making less passes.
- Secondly, towards the end of a match one team will likely be trying to defend the lead and will be slowing down the pace of play, and as a result making less frequent passes. On the other hand one team will be trying to get back in the game, forcing them to make desperate passes that are likely to be inaccurate in the hopes of getting another shot. If the pass attempt fails for any reason it is not counted.
- To prepare for further analysis, we categorized events into possessions and made a matrix. Contains: Event count, possession duration, positioning, distance covered, number of passes, and the target variable: if the possession's outcome was a shot

- Before starting the analysis we needed to prepare the data. To accomplish this we started with a Principal Component Analysis (PCA). PCA projects data into subsets that maximize variance, this breaks the data into new features that have a lower dimension. These new features maintain the important information and are easier for machine learning algorithms to use.
- We also applied a standard scalar, which normalizes the data using the following formula:

$$y_i = \frac{x_i - \mu_x}{\sigma_x}$$

- This transforms the mean to  $\mu_x = 0$ , and  $\sigma_x = 1$ . Bringing features to the same magnitude.
- After we performed the preprocessing on the data we ran a logistic regression model to predict the shots. The logistic regression used the logistic sigmoid:

$$\theta = \frac{1}{1 + e^{-f}}$$

- The function then classifies data by whether it is above the axis on the logistic sigmoid.
- $\theta$ ,  $X_i$  (Variables), and  $Y$ (Shots) are then optimized using gradient descent based on the following pairwise function:

$$L(\theta, X_i, y_i) = \begin{cases} -\ln(\sigma * f) & \text{if } y = 1 \\ -\ln(1 - \sigma * f) & \text{if } y = 0 \end{cases}$$

- The model was able to reach 97% accuracy. However, the recall was ill-defined. The reasoning being this being that the dataset was very imbalanced, and the model just guessed that every possession did not end in a shot, as only 3% of possessions resulted in a shot. To overcome this, we created a Random Forest Classifier.
- Random Forest Classifiers make decisions by making N number of decision trees. Each iteration will randomly omit and repeat various data points. Then by taking the average of the results it makes its classification, in our case either a shot or not a shot. Each individual decision tree is a collection of if-then statements. An example node in this case would be if the number of passes in a possession was greater than or equal 5. Once there is no more ways to divide the data, the terminal node is reached and the tree makes its prediction.
- The Random Forest Classifier performed with the same 97% accuracy as the Logistic Regression, however the precision was 74% while the Logistic Regression was under 50%. Also it provided a defined value for recall of 77%. Overall the Random Forest Classifier provided real predictions with a high accuracy. The Random Forest also provided feature importance that can be used to find which events are most useful in deciding the outcome of a possession:

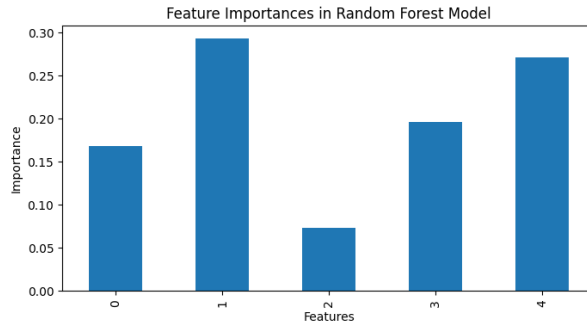


Figure 12: Feature Importance

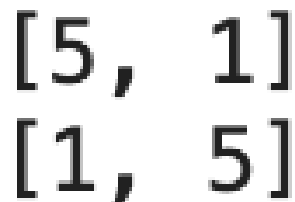
- As shown by figure 12, the most important features for determining the outcome of a possession are numbers one and four. From this we can infer that the average x coordinates and the number of passes have the greatest influence on the outcome of a possession. This makes sense as the x coordinates reflected what part of the field the ball is in, so high x values would mean the Huskies are in a better field position to take a shot. Additionally, the greater number of passes made in a possession, the more pressure is applied to the opposing defense. As there is increased pressure applied, there is a greater chance that there will be a breakthrough resulting in a shot.
- Pezalli score: We used the Pezalli score as a way to measure how efficient the Huskies were with each game.[2]

$$Pezalli = \frac{OwnGoals \times OppShots}{OwnShots \times OppGoals}$$

- The higher the Pezalli score, the better the huskies performed. This ratio increases as the huskies become more efficient at scoring goals and as they make their opponents more inefficient at scoring goals.
- We created another model to predict the wins based on how well a team spreads out their passes within the game. For this we ran a Logistic regression based on these 5 variables for the huskies:
  - *p\_passes\_x*: mean of passes by player.
    - Coefficient: 0.777409
    - Interpretation: This coefficient shows a strong relationship between the average number of passes a player makes and winning. This suggests that players who are more involved in passing tend to contribute significantly to their team's success, so getting the entire team involved will help the team win more games.
  - *p\_passes\_y*: variance of passes by player.
    - Coefficient: 0.010342
    - Interpretation: The small coefficient for the variance in passes by player indicates that a slight increase in this variance slightly increases the chances of winning. So for the huskies they probably have a few key players they get the ball to (or maybe their forwards) that tend to score the goals which lead to games won.
  - *z\_passes\_x*: mean of passes by zone.
    - Coefficient: -1.233196



- Interpretation: The negative coefficient here indicates that higher average passes within specific zones are associated with a lower probability of winning. This shows that as you spread the ball all over the field you tend to get less productive. Instead the huskies need to focus on passing the ball to key areas to get high quality shots.
- *z\_passes\_y*: variance of passes by zone.
  - Coefficient: -0.197159
  - Interpretation: Similar to '*z\_passes\_x*', a higher variance in the number of passes by zone negatively affects the likelihood of winning. We interpret this as passing to meaningless areas of the field lead to worse results resulting in reducing the team's overall effectiveness. As we saw in the random forest model that the location was very important on seeing the outcome of the play so we think there should be more emphasis on getting the ball to key areas.
- *total\_passes*: total passes in the game.
  - Coefficient: -0.078626
  - Interpretation: A negative coefficient for total passes suggests that simply accumulating a higher number of passes is not key to winning games. More important than a high number of passes is quality passes to get good shots at the net.
- *total\_shots*: total shots in the game.
  - Coefficient: -0.152623
  - Interpretation: This small positive coefficient indicates that an increase in total shots slightly increases the chances of winning. This one is a little weird for us because we feel like this should be the opposite. We interpret this similar to number of passes as the number of shots does not directly correlate to winning, more importantly we need to have high quality scoring chances and shots.

$$\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$$

Figure 13: Confusion Matrix

- Our model predicted correctly 83% of the games it was tested on with 5 true positives and 5 true negatives.
  - Soccer Strategy Implications:
    - \* Balanced Passing: While passing is crucial, effective use of the ball in terms of progressing towards the opponent's goal and creating meaningful scoring opportunities is more important than mere passing volume.
    - \* Effective Zone Utilization: Overloading certain zones with passes can be less effective if not strategically planned. Teams should focus on dynamic movements and varying their play to break down defenses.
    - \* Player Roles: Variation in player passing rates might be beneficial to a point, suggesting different roles within the team from play makers to finishers.

## Letter to Huskies coach

Dear Huskies' coach,

In response to your requests we have compiled a report of our analysis. In this letter we will summarize our findings in a easy to understand format:

We found the the Huskies tend to perform best with a 4-3-3 formation; the record was 5 wins and 2 losses. This was the only formation with a significant sample that yielded a winning record. This is supported by the Random Forest Classifier as it found that keeping the ball in the offensive side of the field worked well for the team. We also found that the team struggles playing against the 5-4-1. The Husky offense should be better prepared in future seasons against heavy defensive formations.

We also suggest that the team should play the Tiki-Taka play style. We believe the Tiki-Taka play style will improve their record because it emphasizes making frequent passes to keep the opponent moving, eventually exhausting the defense leading to offensive breakthroughs. This does not mean making a ton of meaningless passes as our logistic model shows that is not directly correlated, we want to put an emphasis on making a lot of quality passes to create a good run at the net. Going along with this we suggest that you highly emphasis getting the ball to a good area to get quality shots off instead of taking as many shots as possible because quantity alone does not have a good correlation for winning games.

Finally, we found some evidence that the team struggled with fatigue this season. Often the passing rates of the Huskies steadily decreased as the game drew on. There are a variety of factors that influence this such as the score, however, it also suggests that the players are under-conditioned. The solution is to simply better condition the Huskies. This is a low-risk change to implement because in soccer, you can never be in too good of shape.

Sincerely, Mater, Joshua, Jason, and Ryan

## References

- [1] [www.mathmodels.org](http://www.mathmodels.org/Problems/2020/ICM-D/index.html). "Problems." Accessed April 19, 2024. <https://www.mathmodels.org/Problems/2020/ICM-D/index.html>.
- [2] Flueck, Alexander J. 2005. *ECE 100* [online]. Chicago: Illinois Institute of Technology, Electrical and Computer Engineering Department, 2005 [cited 30 August 2005]. Available from World Wide Web: (<http://www.ece.iit.edu/~flueck/ece100>).