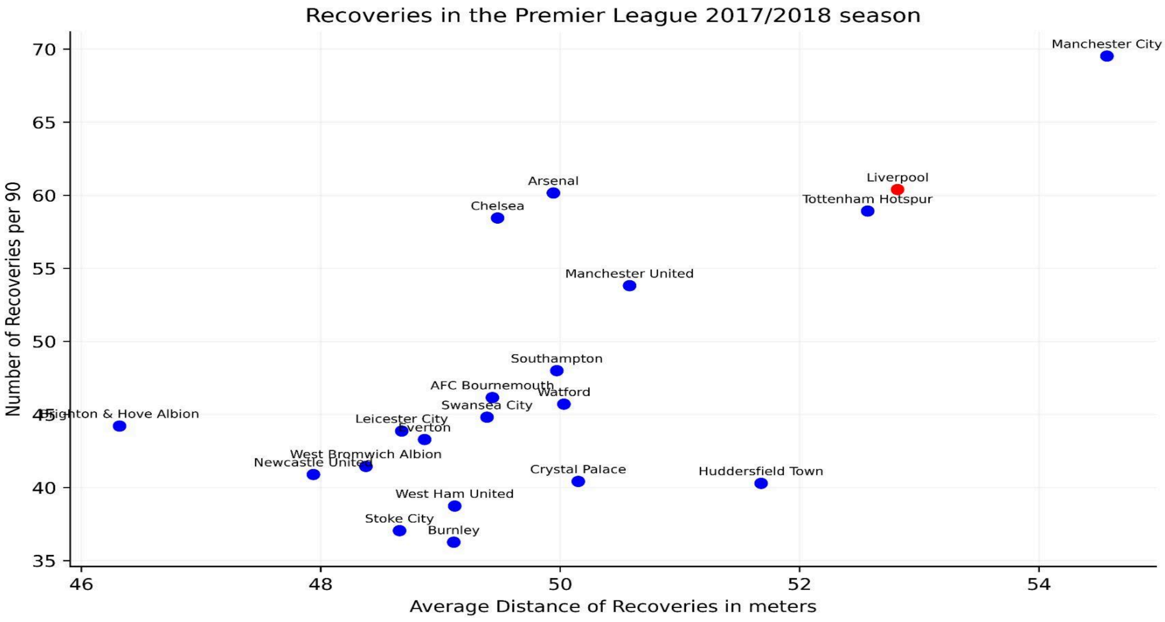
Technical Report - Denis Dervishi

Introduction

Analyzing the 2017/2018 season it was clear that Liverpool adopted the ‘gegenpress’ methodology imposed by Jurgen Klopp. An intense and relentless high press to provoke turnovers and scoring opportunities. They pressed more efficiently and much higher up the pitch resulting in recoveries in better positions. This can be shown through the numbers in the Recovery_Threat_Model.ipnyb (at the end of the Jupyter Notebook) where I calculated the average distance of a recovery (how high up the pitch the recovery was made on average) in the Premier League 2017/2018 season per team and also the number of total recoveries as seen below.

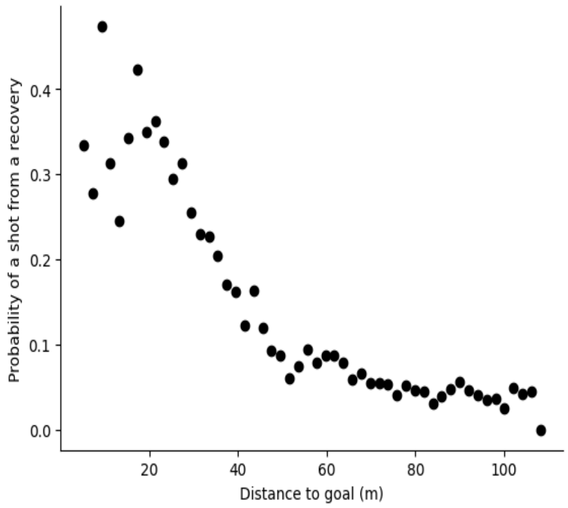


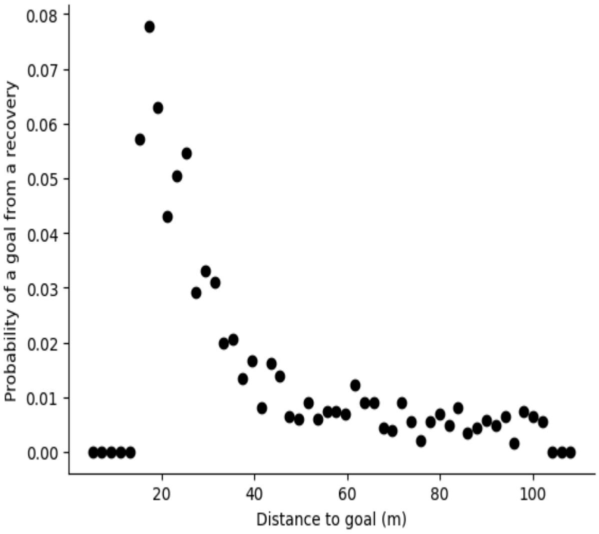
Liverpool is the second-best team in both metrics just behind Manchester City who won the league. It is evident that recoveries are important in football and they should be further explored. This is one of the main reasons that a Recovery Threat Model was created which tried to capture the effectiveness of this style of play. This model isolated all recoveries and tracked if in the future the result of that recovery resulted in a shot or a goal. A primary reason was to calculate the threat of a recovery and see which players' recoveries lead to most shots or goals per 90 minutes.

Preparing the data and building the model

To make the model I created the Recovery_Threat_Model.ipnyb file and first imported all the necessary libraries. Then I loaded into a data frame the Wyscout event data for the Premier League 2017/2018 season. I removed all the events from the data frame that were not important for recovery-based possession chains. I isolated all possession chains and added if they ended in a shot or not and if they ended in a goal or not. I removed all possession chains that had a set piece included to isolate only recovery-based possession chains. I added a column to show the length of the specific possession chain which will be a feature of one of the models and removed all possession chains that do not have at least 3 connected events.

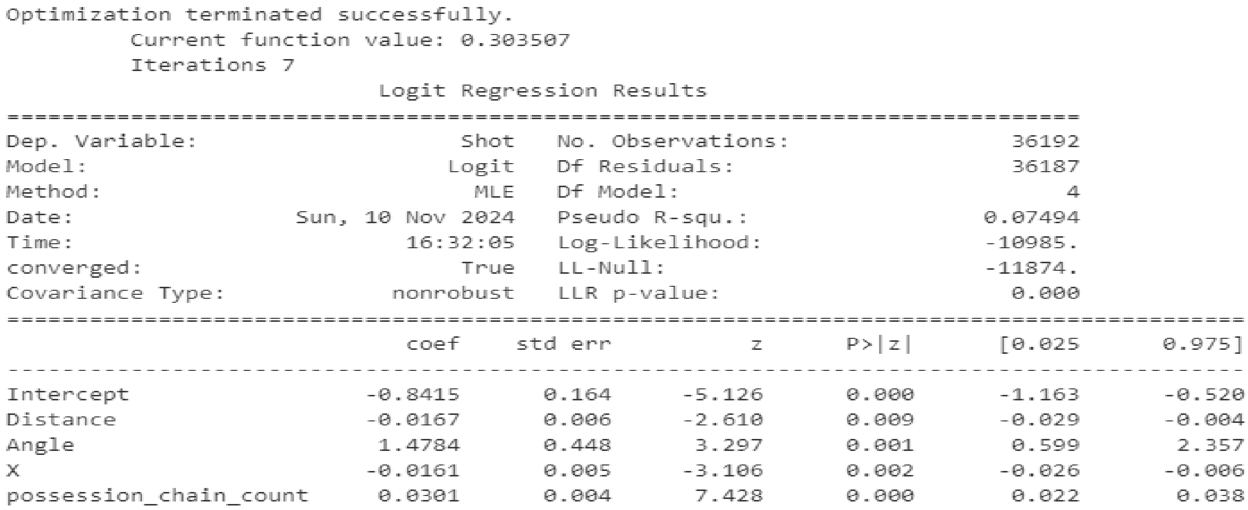
The main features of this model will be Angle, Distance, X and possession chain length. I add them to the data frame with their corresponding formulas. I made a scatter plot to see if there is some relation between Distance and probability of a shot from recovery and Distance and probability of a goal from recovery. The plots show there could be some correlation between the two. The plots are shown below.





The Logistic Regression Models

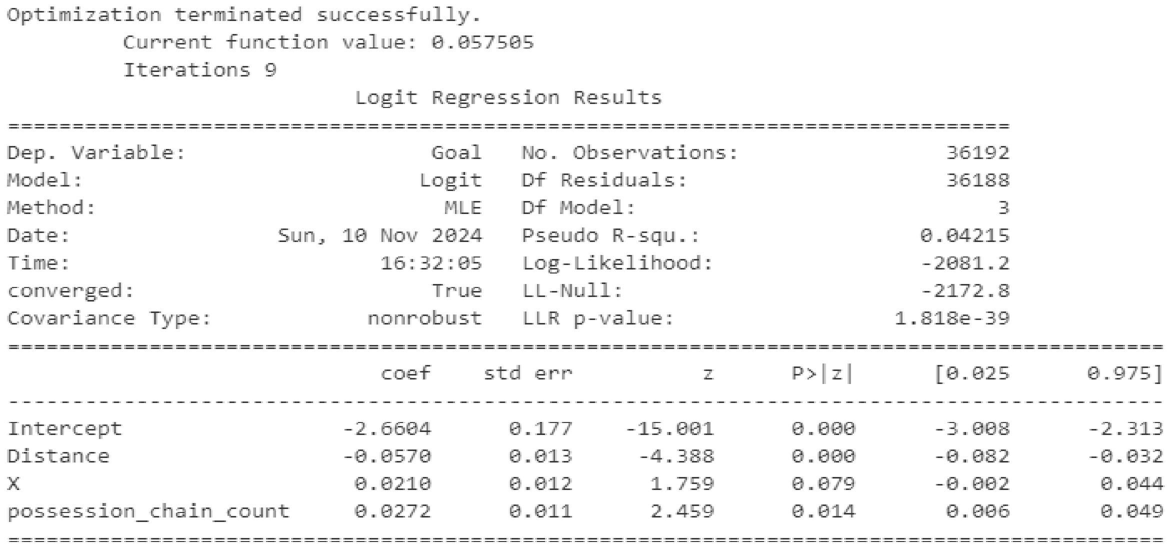
Furthermore, it is time to build the logistic regression model. The logistic regression model was used because of its interpretability. The first model is the recovery to shot model and its features are as follows: distance, angle, X and possession chain length. The first model's summary is shown below.



```
Optimization terminated successfully.
         Current function value: 0.303507
         Iterations 7
                           Logit Regression Results                           
==============================================================================
Dep. Variable:                   Shot   No. Observations:                36192
Model:                          Logit   Df Residuals:                    36187
Method:                           MLE   Df Model:                            4
Date:                Sun, 10 Nov 2024   Pseudo R-squ.:                 0.07494
Time:                        16:32:05   Log-Likelihood:                -10985.
converged:                       True   LL-Null:                       -11874.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 -0.8415      0.164     -5.126      0.000      -1.163      -0.520
Distance                  -0.0167      0.006     -2.610      0.009      -0.029      -0.004
Angle                      1.4784      0.448      3.297      0.001       0.599       2.357
X                         -0.0161      0.005     -3.106      0.002      -0.026      -0.006
possession_chain_count     0.0301      0.004      7.428      0.000       0.022       0.038
==========================================================================================
```

The first model’s p-values are all lower than 0,05 which makes it a very solid model so I continue with it. I do not change any of the features.

The second model is the recovery to goal model and its features are as follows: distance, X and possession chain length. Angle was removed from this model since it proved to correlate too much with other features. The second model’s summary is shown below.



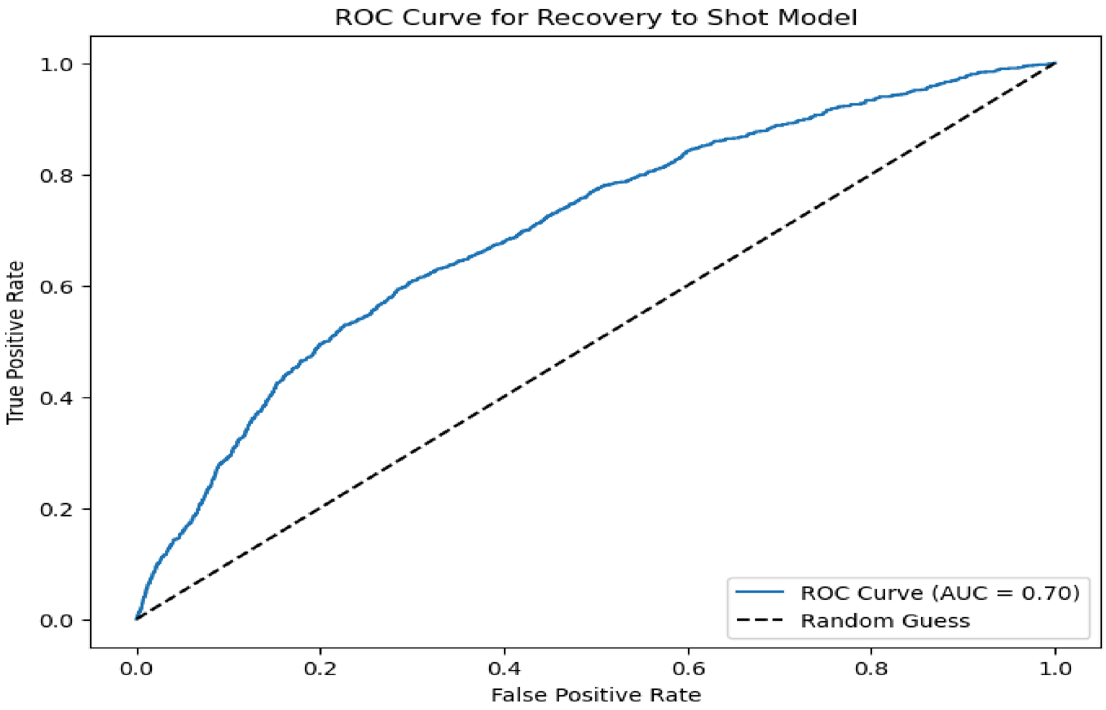
```
Optimization terminated successfully.
         Current function value: 0.057505
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                   Goal   No. Observations:                36192
Model:                          Logit   Df Residuals:                    36188
Method:                           MLE   Df Model:                            3
Date:                Sun, 10 Nov 2024   Pseudo R-squ.:                 0.04215
Time:                        16:32:05   Log-Likelihood:                -2081.2
converged:                       True   LL-Null:                       -2172.8
Covariance Type:            nonrobust   LLR p-value:                 1.818e-39
===========================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                  -2.6604      0.177    -15.001      0.000      -3.008      -2.313
Distance                   -0.0570      0.013     -4.388      0.000      -0.082      -0.032
X                           0.0210      0.012      1.759      0.079      -0.002       0.044
possession_chain_count      0.0272      0.011      2.459      0.014       0.006       0.049
===========================================================================================
```

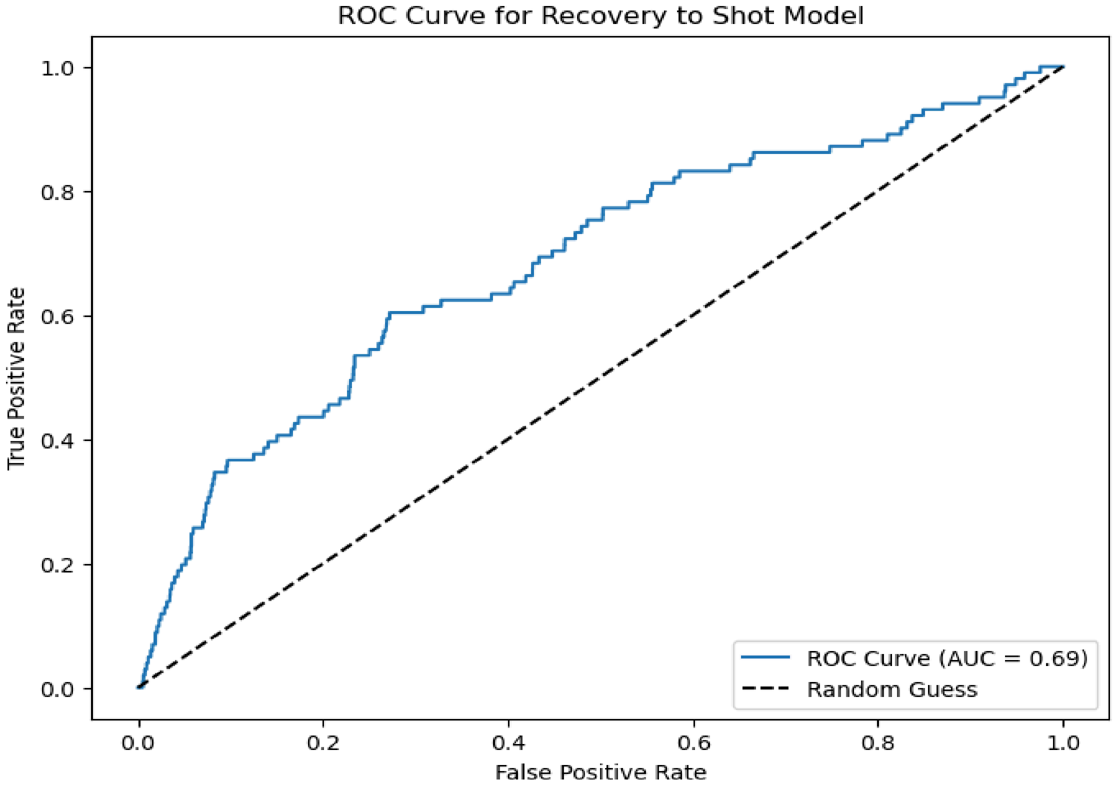
This model has all p-values of the features except X lower than 0,05 so it is a good model and I will continue with it.

The next step is to add player data from the Wyscout regarding their names and minutes played in the 2017/2018 season. Then it is time for normalizing the data to find out the best Premier League players according to both metrics. I eliminated all players who have not played more than 900 minutes. I plotted the data and found that Liverpool’s players performed exceptionally well at both of these metrics once again showing the importance of recoveries in Liverpool’s game plan. As two of those players (Philippe Coutinho and Emre Can) left Liverpool that season when writing my scout report and analyzing the rest of the leagues I will try to find suitable replacements based on their performance in these two models. I run the OtherLeaguesRecoveries.py file to get sorted_goals.json and sorted_shots.json files which are used to showcase the best players performing in these two models in La Liga, Bundesliga, Serie A and Ligue 1. I conclude that Julian Brandt and Diego Demme are suitable replacements. Both are performing extremely well in both models making them solid replacements for Philippe Coutinho and Emre Can.

ROC Scores and the Conclusion

To analyze the model’s predictability ROC Score needs to be measured. At the end of the Recovery_Threat_Model.ipnyb file, I calculated the ROC scores for both models. Down below ROC Curves are shown for both models.





In conclusion, both models show decent performance with ROC scores around 0.69-0.70, suggesting they can make fairly accurate predictions. However, with the current features, there's still plenty of room to improve. By adding more relevant features, I believe the model's predictability can be significantly enhanced, leading to better scouting suggestions. To do this, I'll need more data and will likely need to explore other leagues to fine-tune the model further and fully unlock its potential.