Appendix to the Project detailing the work

Step 0. All the code and datafiles involved are made available in the following Github repo : https://github.com/bosemessi/SBODN. The notebooks can be found in the Codes directory, while the datafiles are in the Data directory of this repo. All of the coding is done in Python.

Step 1. Extract all the event data from the WC2018 games and combine into a single file. This is done in the "DownloadWC2018DataFromStatsbomb.ipynb" notebook.

Step 2. Extract the game minutes played by each player and combine. It's done one game at a time : by noting down whether the player started or subbed on, and whether he got subbed off or redcarded or played the full game. This is done in the https://github.com/bosemessi/SBODN/blob/main/Codes/EDA_Minutes.ipynb notebook.

Step 3. To get all the pass and carry related metrics, this notebook is used : https://github.com/bosemessi/SBODN/blob/main/Codes/EDA_Passing%26Carrying.ipynb. We use the following definition for progressive passes/carries : if the ball moves closer to the goal by 25 % as compared to where it was before, or it gets into the box, the pass/carry is progressive.

Step 4. To get xT and xT Facilitated, this notebook is used : https://github.com/bosemessi/SBODN/blob/main/Codes/EDA_xT.ipynb. It takes all successful passes and carries and assigns a “threat value” based on the Expected Threat model developed by Karun Singh (https://karun.in/blog/expected-threat.html). xT Facilitated is defined as the xT of the subsequent move. This captures the effects where a player can invite pressure on himself and release a teammate who is free.

Step 5. To get XG Chain and XG Buildup, this notebook is used : https://github.com/bosemessi/SBODN/blob/main/Codes/EDA_XGCHAIN.ipynb. It looks at all possession sequences of a team that contains atleast one open play shot, and assigns the xG value of the shot to the participants of the possession sequence.

Step 6. To get all defensive metrics, this notebook is used : https://github.com/bosemessi/SBODN/blob/main/Codes/EDA_Possession_Defence.ipynb. It iterates over each game and extracts all relevant defensive metrics + dribbles + turnovers. It collects pressures and successful pressures from the possession sequences as well.

Step 7. To build an Expected Pass model and use that to evaluate Pass Completion Above Expected, this notebook is used : https://github.com/bosemessi/SBODN/blob/main/Codes/PassCompletionModel.ipynb. It used Scikit-Learn library of Python to train Classifier models on the data to identify pass completion probability. Random Forest and XGBOOST models have been tried out. XGBOOST with hyperparameter grid search gives the best result.

Step 8. Combine all the different metrics and calculate the percentiles for the Center backs here : https://github.com/bosemessi/SBODN/blob/main/Codes/EDA_Percentiles.ipynb. A minute cutoff of 180 minutes is used, as well as a minimum pass cutoff of 100 open play passes. Defensive metrics are possession-adjusted to mitigate the possession-heavy style effects of certain teams. xT and xT Facilitated are adjusted per 100 passes to mitigate the effect of high-pass volume players who could rack up a lot of xT just by making short forward passes. XG Chain and XG Buildup are adjusted per 10 open play shots of the team to mitigate the effect of different teams being different volume shooters.

Step 9. All the different plots : scatters, bar charts and progressive pass/carry maps are generated here : https://github.com/bosemessi/SBODN/blob/main/Codes/Plotting.ipynb.