Theo Lee, Jordan Lee-Kook, Yiqi Song

Project Debrief

1. **Effort Breakdown:**

   Overall, the project workload was divided very evenly across the three team members. The major milestones of the project were spearheaded by each of the team members with continued support and review from the others. Jordan initially took the lead on the idea generation and refinement, as well as gathering and cleaning the data for analysis. Theo took the lead on writing the code for the bulk of the model itself, bringing up key issues and developments to the team. Yiqi led the development of the presentation slides, laying out the overall structure and developing most of the content for it. The team dynamics were very positive with no major hurdles.

2. **Challenges and Surprises:**

   Initially, our plan was to conduct analyses of all players across the English Premier League. We quickly realized that this would be difficult to perform within the project timeframe since player metrics vary across positions. Here, we made our first major change to limit the scope to just analyze one specific position. Our second challenge was that our data did not have metrics specific to any one position, just team and game statistics. We were having trouble isolating and quantifying different players' contributions to each of the metrics. However, we saw that some of the statistics (goals saved, goals scored, etc.) were heavily influenced by goalkeeper efforts, and thus analyzing this metric would give an accurate portrayal of their skill level.

3. **Learnings and Improvements:**

   Our project utilized a synthetic dataset to stress-test AI applications. By using LLMs to automate data cleaning and troubleshooting, we significantly accelerated our workflow. The experience highlighted how modern startups deploy these technologies to maximize output and challenge traditional distribution models. Through this process, we learned that synthetic data still requires rigorous validation to handle edge cases effectively. Future iterations would benefit from refining our initial data generation parameters to reduce downstream noise.

4. **Course Contributions:**

   A lot of what we built for this project came directly from what we learned in CPSC 1710. The overall structure of how we handled the data, cleaned it up, and normalized the features followed the exact workflow we practiced in class. When we realized that predicting raw stats was too noisy, we drew on the course lessons about feature engineering and model limitations, which pushed us toward using K Means to uncover natural patterns instead of forcing a supervised model that did not fit the problem. The way we validated the clusters, visualized the tiers, and later built the transition matrix all came from the class discussions on how to reason about models over time and how to make sure a model’s output is actually meaningful.

At the same time, the soccer side of the project was very much our own. The idea of turning these clusters into goalkeeper archetypes, thinking through how teams might use GSAA, and creating a “Moneyball” style scouting tool came from our personal experience watching and analyzing the sport. We used AI in a pretty targeted way. It helped us write the short narrative scout notes and polish some of the language in our slides, but everything technical was built by us. All of the data work, modeling logic, metric design, and code came from the team. In the end, the project feels like a mix of what the course taught us and what we brought in from our own interests, with AI playing a supportive role.