ACL milestone 1 Report

Team 19

October 2025

Overview

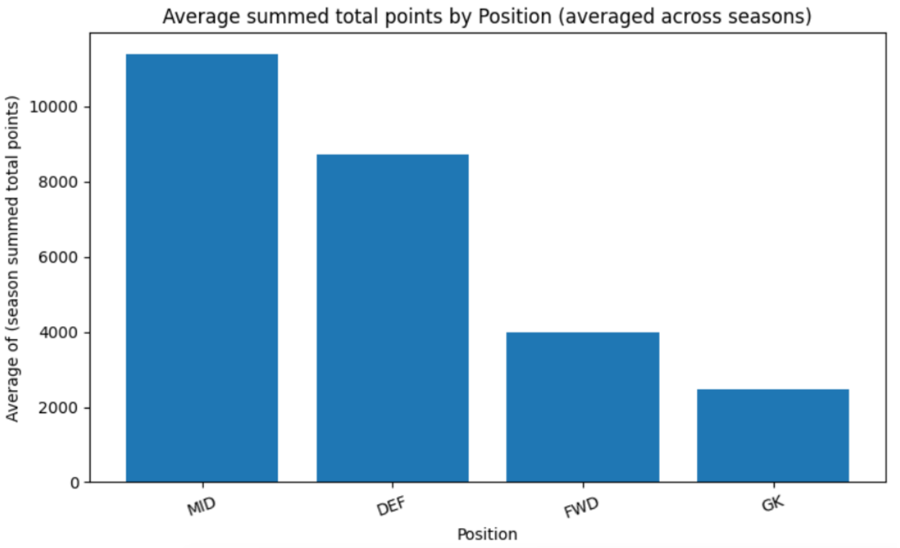
An analytical text report answering and visualizing the data engineering questions. The report also includes the features used in the predictive model and why each feature was selected.

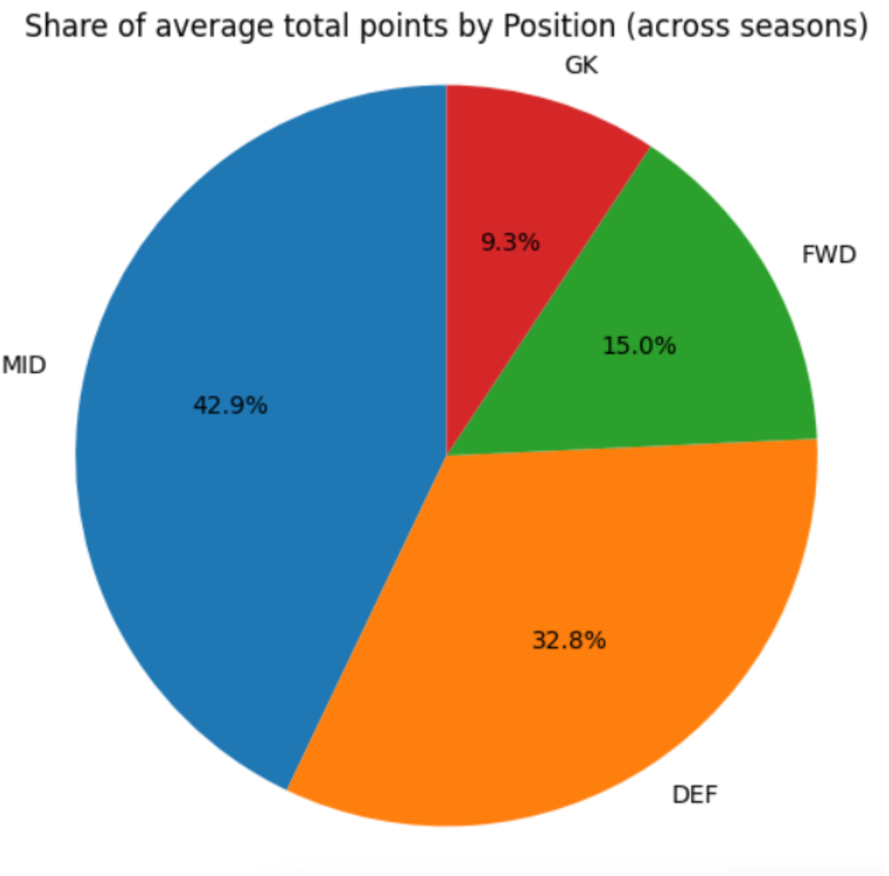
Data Engineering Questions

a. Across the seasons, which player positions (e.g., goalkeeper, defender, midfielder, etc.) score the largest sum of total points on average?

Description:
To determine which positions contribute the most fantasy value, we first summed total points by position for each season. We then averaged those per-season sums across all seasons to obtain the expected seasonal contribution per position. Visualization used a bar chart (for clear comparison of average season totals) complemented by a percentage table/pie to illustrate each position's share. This approach reveals the relative aggregate contributions of goalkeepers, defenders, midfielders, and forwards while controlling for season-to-season variability. Limitations include that this aggregate measure does not capture per-player efficiency (points per 90 minutes) or cost-effectiveness; follow-up analyses are recommended to explore those dimensions.

Visualization screenshots:





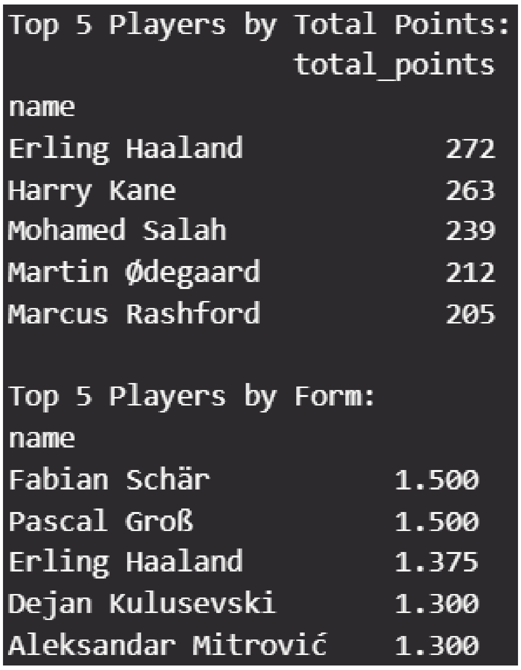
Analysis of Average Total Points by Position (Across Seasons):
The visualizations reveal that midfielders (MID) consistently contribute the largest share of total points across Premier League seasons, averaging about 43% of all points, followed by defenders (DEF) at 33%, forwards (FWD) at 15%, and goalkeepers (GK) at 9%. This pattern indicates that midfielders play the most influential role in Fantasy Premier League scoring, reflecting their involvement in both attacking and defensive actions, goals, assists, and bonus points. The bar chart reinforces this dominance, showing midfielders leading by a wide margin in total seasonal points. These insights help fantasy managers prioritize investing in high-performing midfielders to maximize returns across gameweeks.

---

b. Using the form feature, how did the performance of the top five players evolve across gameweeks during the 2022–23 FPL season? Are the top players in form the same top players with the highest total points?

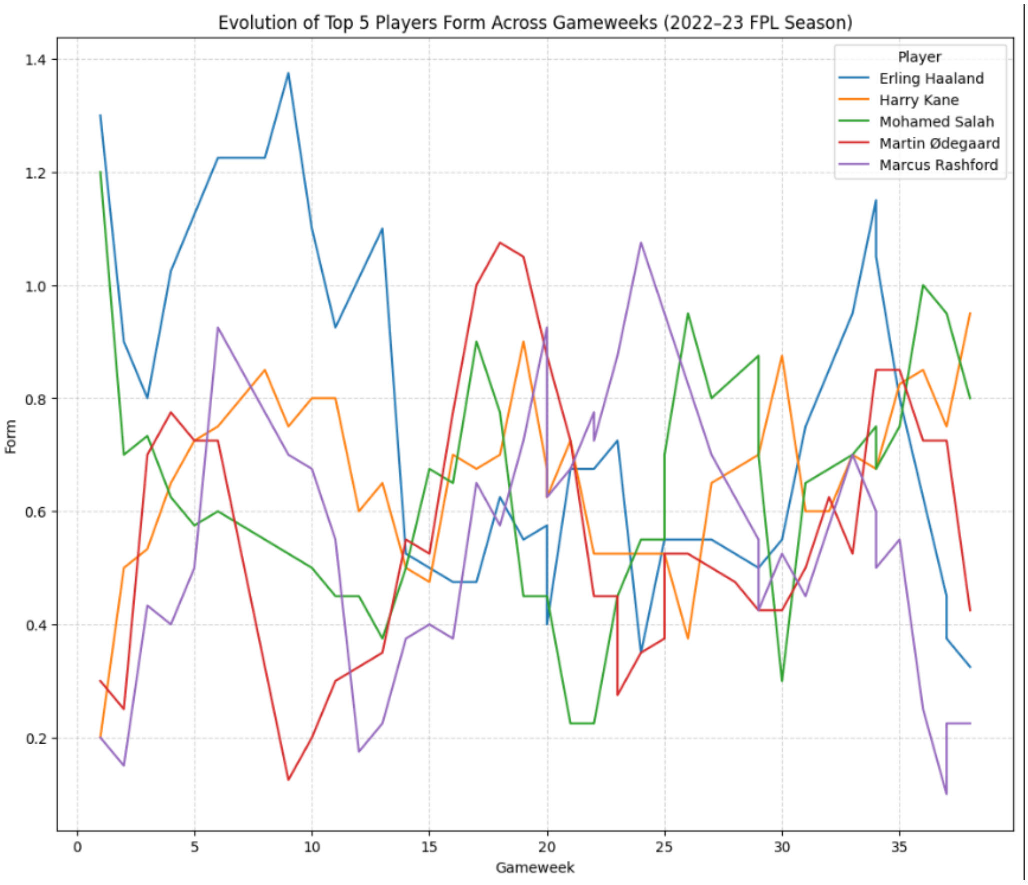
Description:
To examine how top performers' form evolved and whether form aligns with season-long scoring, we filtered the dataset to the 2022–23 season and computed each player's rolling form (average total points over the previous four gameweeks, divided by 10). We identified the season's top five players by total-points and plotted their form across gameweeks to visualize week-by-week momentum and dry spells. Separately, we computed the top five players by peak form (their maximum form value) and compared the two lists to check overlap. The visualization (line chart of form vs GW for the top-5-by-total) reveals temporal patterns, bursts of high form, injuries or benchings as sudden drops, and whether top scorers sustained consistent form or relied on isolated high-scoring weeks. Comparing peak-form leaders with total-points leaders tests whether short-term hot streaks translate to season-long output: often they do not entirely overlap, some players reach very high short-term form but lack season-long volume, while top scorers may show steadier, moderate form across many weeks. In your report, note this distinction and the implication for fantasy managers: form is valuable for short-term transfer/lineup decisions, while total-points better reflects cumulative value across the season. Limitations: form depends on the chosen 4-week window and excludes popularity/context features; consider complementing this analysis with minutes-played, injuries, and value (cost) to assess reliability of form signals.

Visualization screenshots:



```
Top 5 Players by Total Points:
                total_points
name
Erling Haaland           272
Harry Kane               263
Mohamed Salah            239
Martin Ødegaard          212
Marcus Rashford          205

Top 5 Players by Form:
name
Fabian Schär          1.500
Pascal Groß           1.500
Erling Haaland        1.375
Dejan Kulusevski      1.300
Aleksandar Mitrović   1.300
```



Analysis of Top Players' Form Evolution (2022–23 FPL Season):
The chart shows how the form of the top five players by total points Haaland, Kane, Salah, Ødegaard, and Rashford changed across the 2022-23 season. Haaland's strong early peaks reflect his explosive start, while Kane and Salah maintained steady consistency. Rashford and odegaard both had a mid-season surge and total points players reveals limited overlap. Short-term performers like Schär and Groß reached high form briefly but lacked season-long consistency. This indicates that form captures short-term momentum, whereas total points reflects sustained performance over the season.

Model Evaluation Metrics

To evaluate the regression model's performance, several standard metrics were used: **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and the **Coefficient of Determination** (R^2). Each provides unique insights into the model's accuracy and reliability.

1. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

The MAE measures the average magnitude of errors between the predicted ($\hat{y}_i$) and actual (y_i) values, without considering the direction of the errors. It provides an intuitive measure of average prediction error in the same units as the target variable. Lower MAE values indicate higher model accuracy.

2. Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

The MSE penalizes larger errors more heavily by squaring them, making it sensitive to outliers. While it provides a sense of overall error magnitude, its squared units make direct interpretation less intuitive compared to MAE.

3. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

The RMSE is the square root of MSE, returning the error to the original scale of the target variable. It emphasizes large errors but remains interpretable in the same units as the predictions. A smaller RMSE signifies better model performance and stability.

4. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

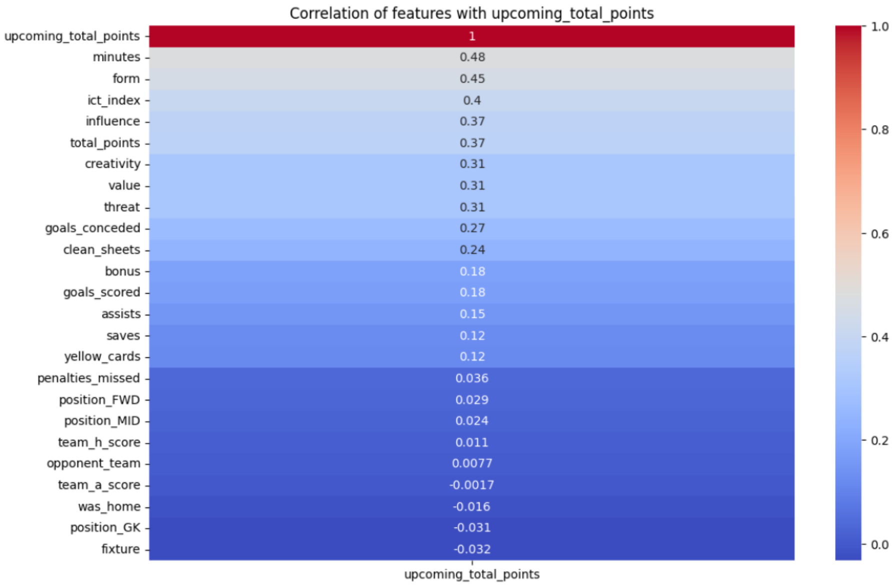
The R^2 score measures how well the model explains the variance in the target variable. It ranges from 0 to 1, where higher values indicate a better fit. An R^2 of 1 represents perfect predictions, while values near 0 suggest the model fails to capture underlying patterns in the data.

Feature Selection

Four different feature sets were chosen to train the model on and then each model's result was collected and compared with the rest.

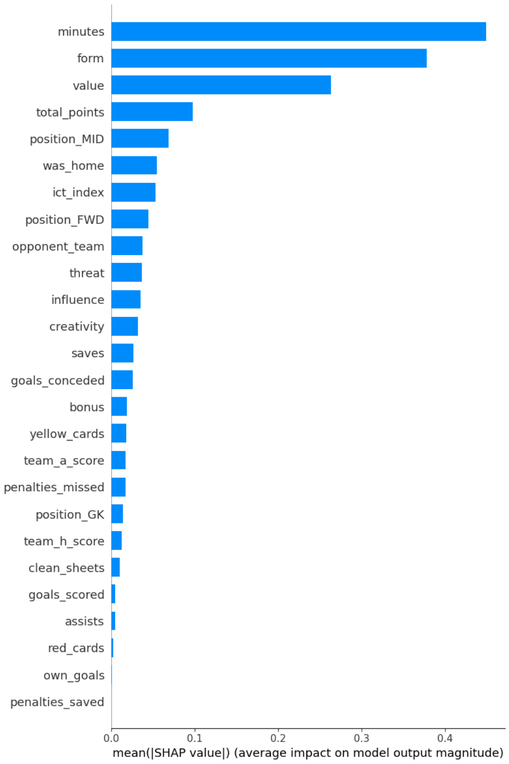
Importance:

The correlation matrix shows how strongly numerical features in the dataset are related to each other. It helps identify patterns of dependence, for example, which variables move together or inversely, guiding feature selection and reducing multicollinearity before model training. **Correlation Matrix:**



Importance:
This matrix shows how each feature changes the upcoming total points.

Shap on all features:

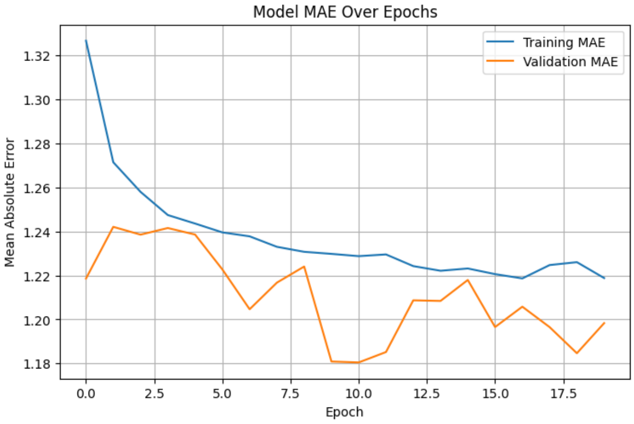


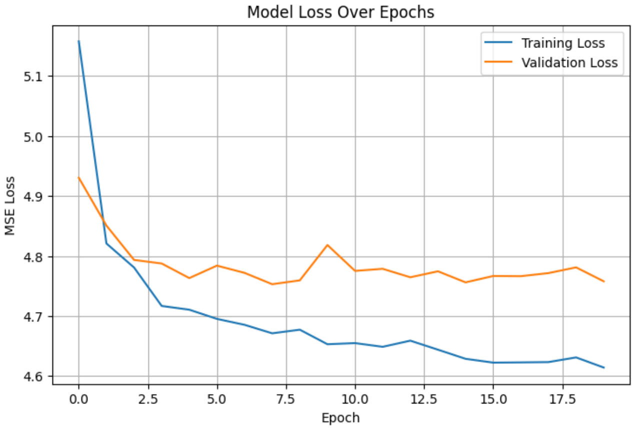
Importance:
This shows how each feature changes the upcoming total points measured by the model that was trained on the baseline feature set.

Baseline Feature Set: Has all features
Numerical:(minutes, form, ict-index, influence, total-points, creativity, value, threat, goals-conceded, clean-sheets, bonus, goals-scored, assists, saves, yellow-cards, penalties-missed, own-goals, penalties-saved, team-h-score, opponent-team, team-a-score, red-cards, fixture, element)
Categorical:(position, was home)

took 20 epochs
metrics: MAE: 1.1955, MSE: 4.6054, RMSE: 2.1460, R2: 0.2877

Screenshots:

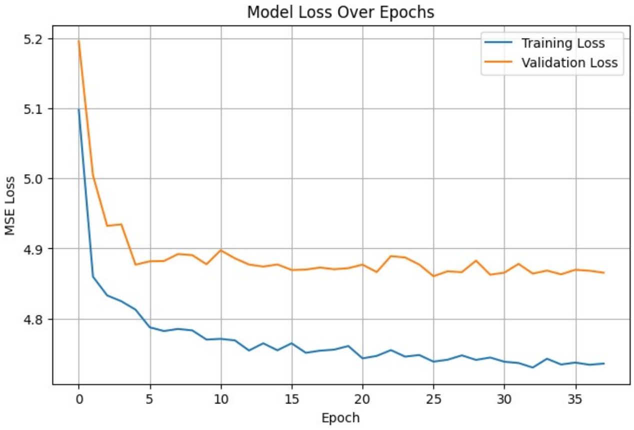


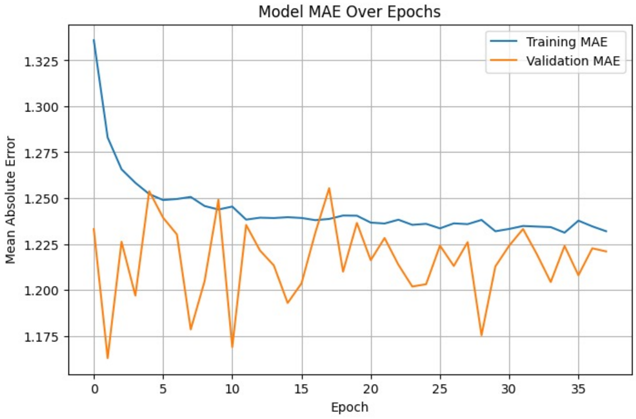


Why These Features Were Chosen And Results:
These features were chosen to be the baseline features so it consisted of all columns in the dataset that could be used to predict the upcoming total points. Only discarded arbitrary columns such as team names, season , GW, element, fixture and the likes.

1st Feature Set: Best 5 features in SHAP
Numerical:(minutes, form, total points)
Categorical:(position, was home)
took 38 epochs
metrics: MAE: 1.1996, MSE: 4.6817, RMSE: 2.1637, R2: 0.2760

Screenshots:



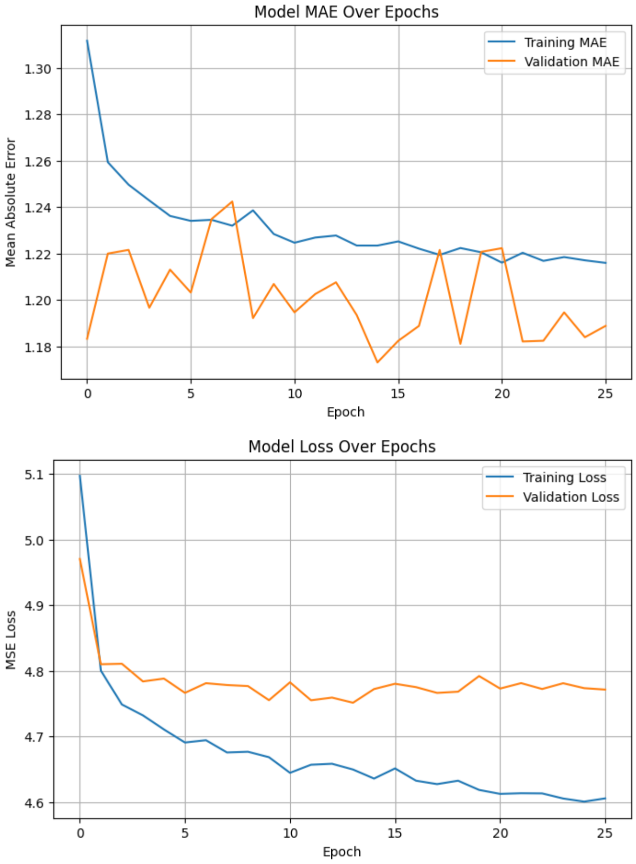


Why These Features Were Chosen And Results:

We chose these features as these were the most influencial features regarding predicting the total upcoming points according to the global SHAP graph.

2nd Feature Set: Removed features with least SHAP value (penalties saved, own goals, red card)
Numerical:(assists, bonus, clean sheets, creativity, goals-conceded, goals-scored, ict-index, influence, minutes, penalties-missed, saves, team-a-score, team-h-score, threat, total-points, value, yellow-cards, opponent-team, form)
Categorical:(position, was home)
took 26 epochs
metrics: MAE: 1.1744, MSE: 4.6039, RMSE: 2.1457, R2: 0.2880

Screenshots:

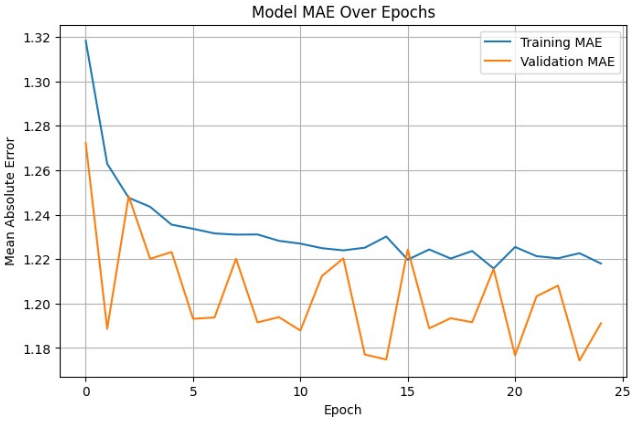


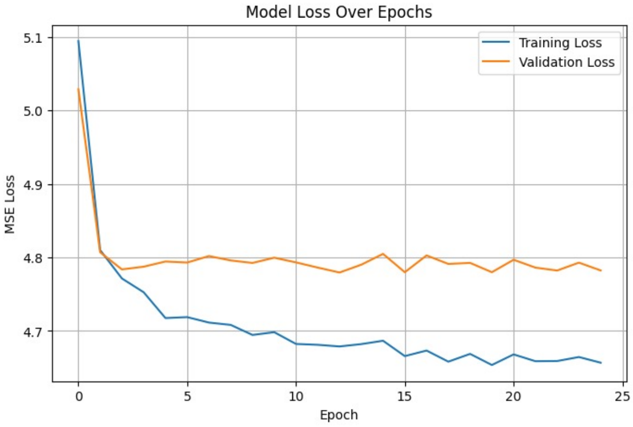
Why These Features Were Chosen And Results:

We went with a different approach in this feature set to compare which approach returned better results. In this approach we removed the least influencial features according to the global SHAP.

3rd Feature Set: Best 10 features in SHAP
Numerical:(minutes, total-points, form, ict-index, influence, creativity, value, threat, goals-conceded, goals-scored)
took 25 epochs
metrics: MAE: 1.1965, MSE: 4.6193, RMSE: 2.1493, R2: 0.2856

Screenshots:





Why These Features Were Chosen And Results:

We increased the number of features to use from the SHAP to the best 10 and measured the performance.
