Project Check-In 3

Model Selection and Preliminary Results Report

In this project, we are aiming to build a machine learning model that can accurately predict the winners of the Champions League tournament for the upcoming 2024 season. The initial step involved selecting a suitable dataset that contained relevant information about previous Champions League matches. We had originally chosen a different dataset for exploration, however, after evaluating multiple other options, we settled on a different dataset that provided more specific, detailed, and spelt-out statistics for each team's performance, including matches played, wins, draws, losses, goals scored, goals conceded, and more.

Data Cleaning and Feature Selection/Engineering

Before proceeding with model training and development, we performed essential data cleaning tasks to ensure the integrity and quality of our dataset. We checked for missing values and duplicates, as well as addressed any inconsistencies that could potentially affect the model's performance. To handle negative goal values and ensure data consistency, we implemented a custom function that imputed zero values for negative goals scored or conceded.

```
# Function to handle any possible inconsistencies in our dataset
def fix_inconsistencies(data):
    # Check for and impute any negative goal values with a 0
    data['goals_scored'] = data['goals_scored'].where(data['goals_scored'] >= 0, 0)
    data['goals_conceded'] = data['goals_conceded'].where(data['goals_conceded'] >= 0, 0)

    # Ensure that matches played = wins + draws + losses
    data = data[data['wins'] + data['draws'] + data['losts'] == data['match_played']]

    # Check that Goal Diff = Goals For - Goals Against
    data = data[data['goals_scored'] - data['goals_conceded'] == data['gd']]

    return data

cleaned_df = fix_inconsistencies(data.copy()) # Using a copy of the original data set
cleaned_df
```

Feature engineering played a crucial role in our analysis. We renamed certain features to improve clarity and dropped unnecessary columns, such as the 'year' attribute, as they we deemed it somewhat irrelevant for the models's training and main task of prediction. Additionally, we encoded the 'team' column to facilitate better handling of categorical data.

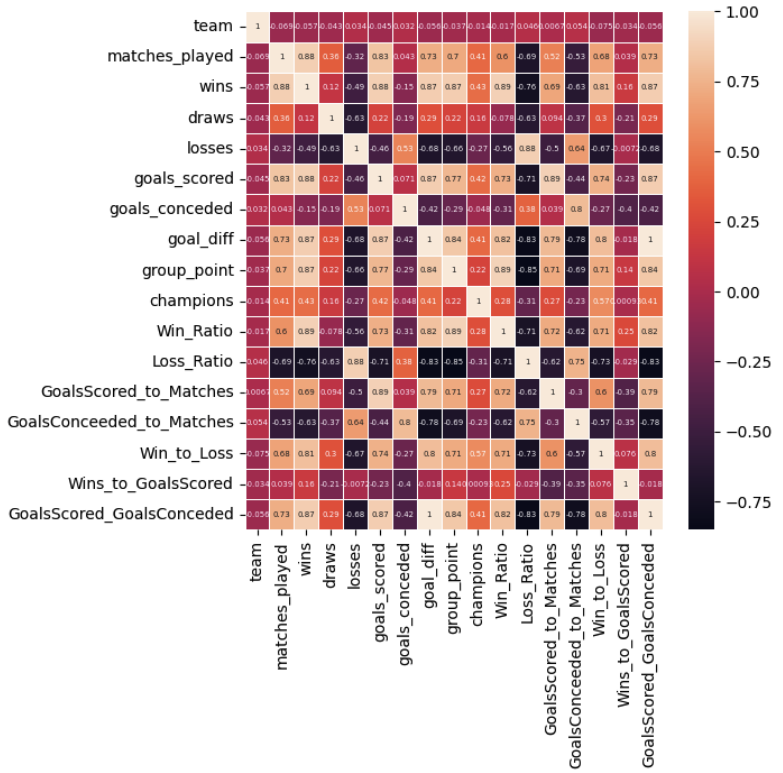
One of the key aspects of our feature engineering process was the creation of new ratio features. We are still working on evaluating and choosing the best new features to use for this model. However, these new ratios were calculated based on existing attributes, such as the ratio of wins to matches played, losses to matches played, goals scored to matches played, and goals conceded to matches played. We also derived features like

'Win_to_Loss', *'Wins_to_GoalsScored'*, and '*GoalsScored_GoalsConceded'*, which could potentially capture important patterns and enhance the model's predictive capabilities.

Model Selection and Correlation Analysis

To identify the most promising features for our model, we conducted a correlation analysis. This involved calculating, visualizing, and plotting a correlation matrix for the features, which greatly helped to reveal the strength and direction (positive or negative) of relationships between the features and the target variable ('*champions'*). We employed the use of the Seaborn heatmap function to visualize this correlation matrix. From the displayed graphic, we could easily identify features that exhibited strong positive or negative correlations with the target variable. Features that show high correlations with the target variable are considered valuable for inclusion in the model, as they could contribute significantly to accurate predictions. Conversely, features with weak or negligible correlations were candidates for elimination from our dataset since they might not add much predictive power.

Furthermore, the correlation analysis helped us detect and address multicollinearity, a indicating where the dataset features/variables are highly correlated with each other. This could possibly introduce redundancy, instability, and inaccuracy in the model, so we aimed to either remove or combine highly correlated features to improve model performance.



Moving forward, we plan to perform Principal Component analyses on the available features in addition to the correlation matrix to further optimize the best features. We will utilize the insights we gain from these processes to select the most informative features and construct our machine learning model. We plan to use a random forest algorithm to train this model with different sets of selected features, while recording important metrics like accuracy, precision, and recall on each iteration. We will then go ahead with the combination of features that produce the best performance metrics.

Through the sequential process of feature selection, model training, and evaluation, we aim to develop a robust and reasonably accurate model that can predict the winners of the 2024 Champions League tournament.