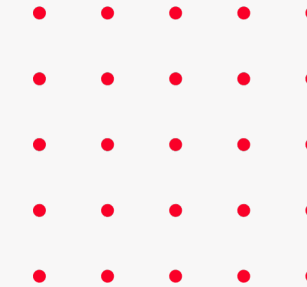




T01-G15

WNBA Predictions Playoff

Daniel Samuel Edim - 202400301 - 20%
Luis Vieira Relvas - 202108661 - 40%
Rodrigo Campos Rodrigues - 202108847 - 40%



Domain Description

WNBA Structure

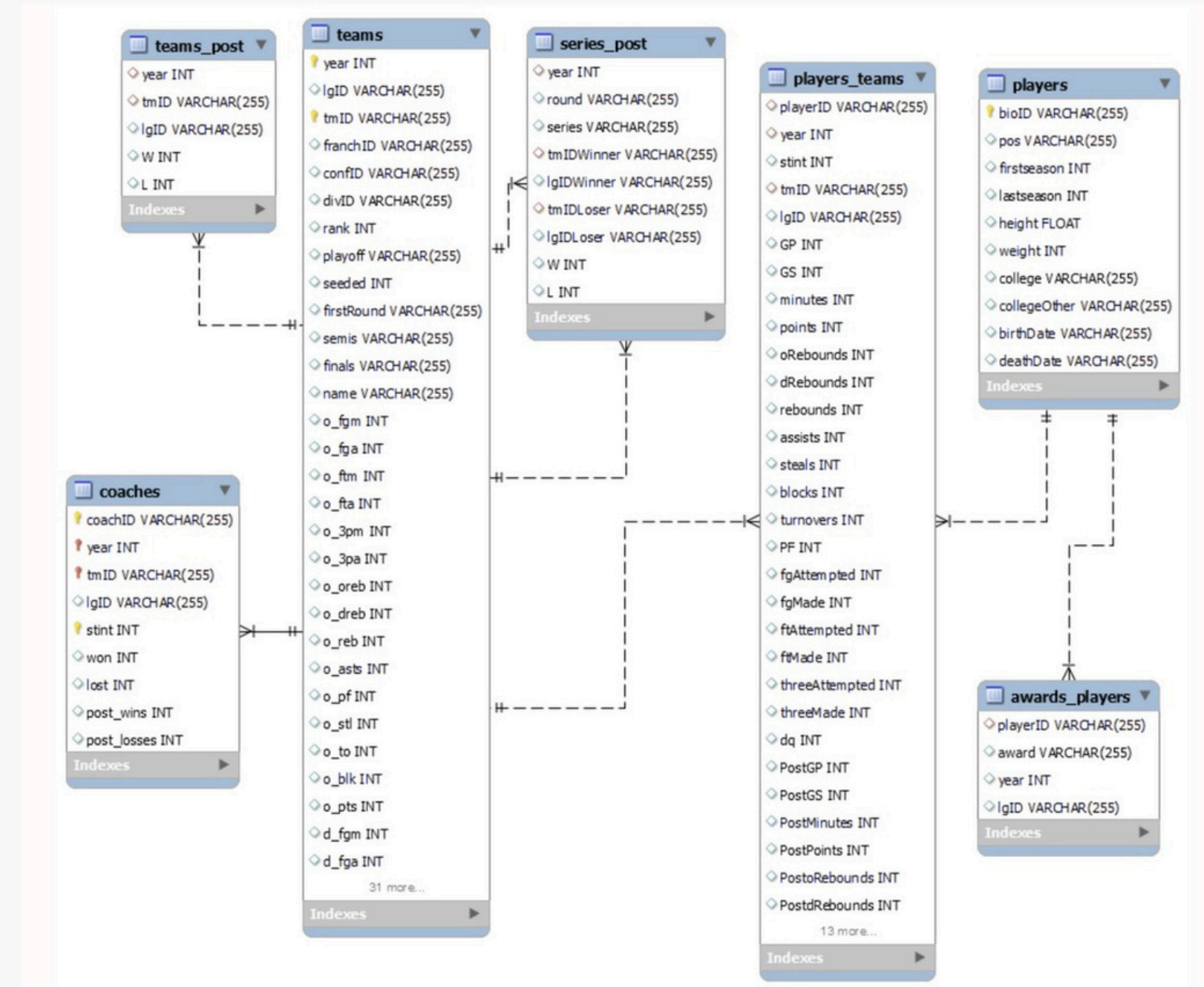
- Every season is distributed in 2 phases:
 - Regular phase where the best teams go to Playoffs phase
 - where the top eight teams will compete to decide the WNBA champion

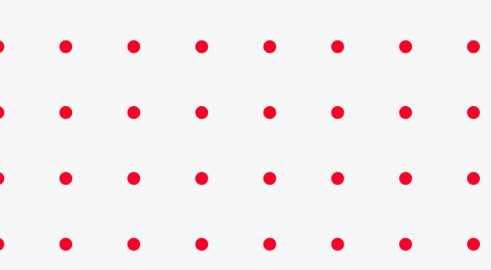
Dataset

- 10 years of data
- Teams - 142 entries (20 teams)
- Players - 893 entries (893 players)
- Coaches - 162 entries (57 coaches)
- Awards_Players - 95 entries (12 awards)

Project aim

- Using machine learning knowledge, predict which teams will qualify for the playoffs in a specific year.





Business Understanding

Analysis of Requirements with End User

- To meet with the requirements of the end users we should be capable of retrieving the teams that went to the playoff so that they can make predictions for the next years.
- Also they want to know which information from the datasets, have the most influence in the final predictions.

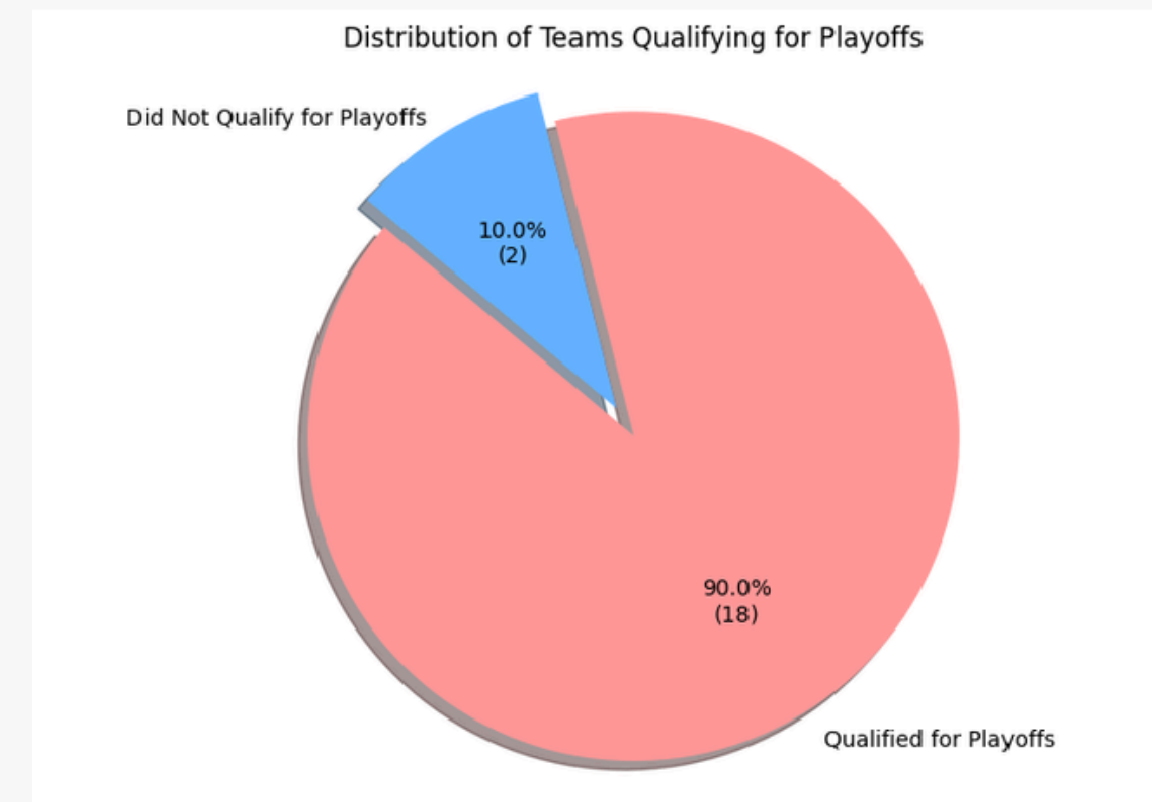
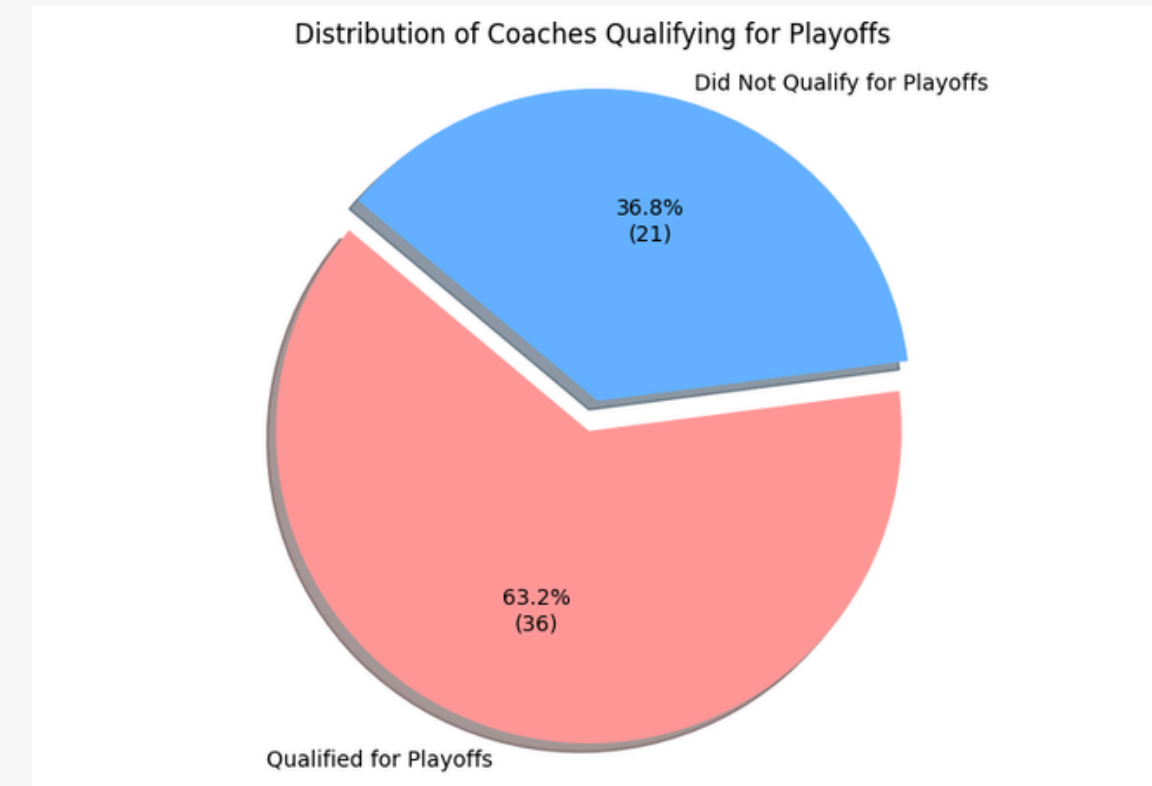
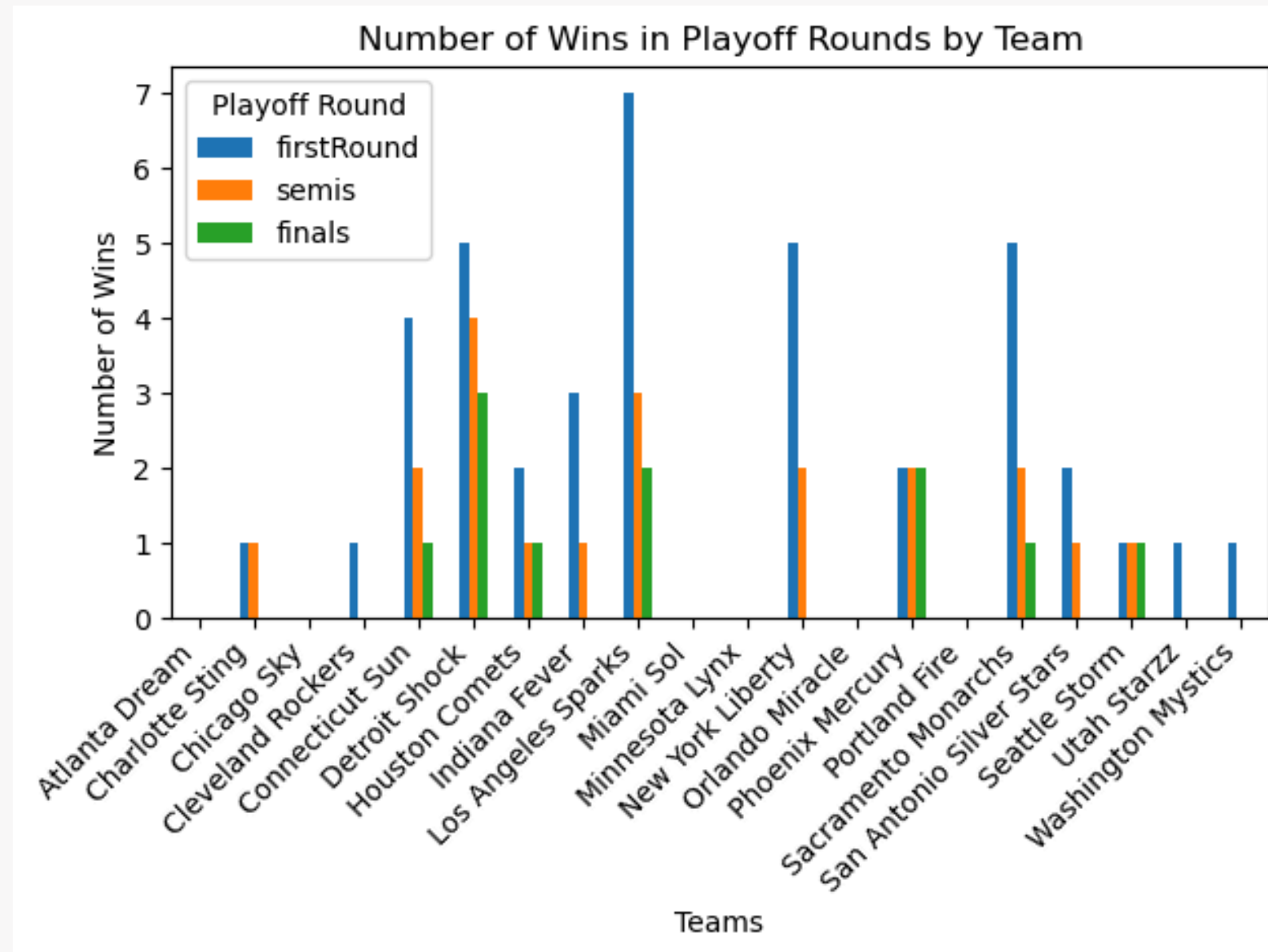
Business Goals

- Develop a system to accurately identify the teams that are most likely to qualify for the playoffs based on past seasons.
- Provide stakeholders, the decisions that were taken during the whole process.

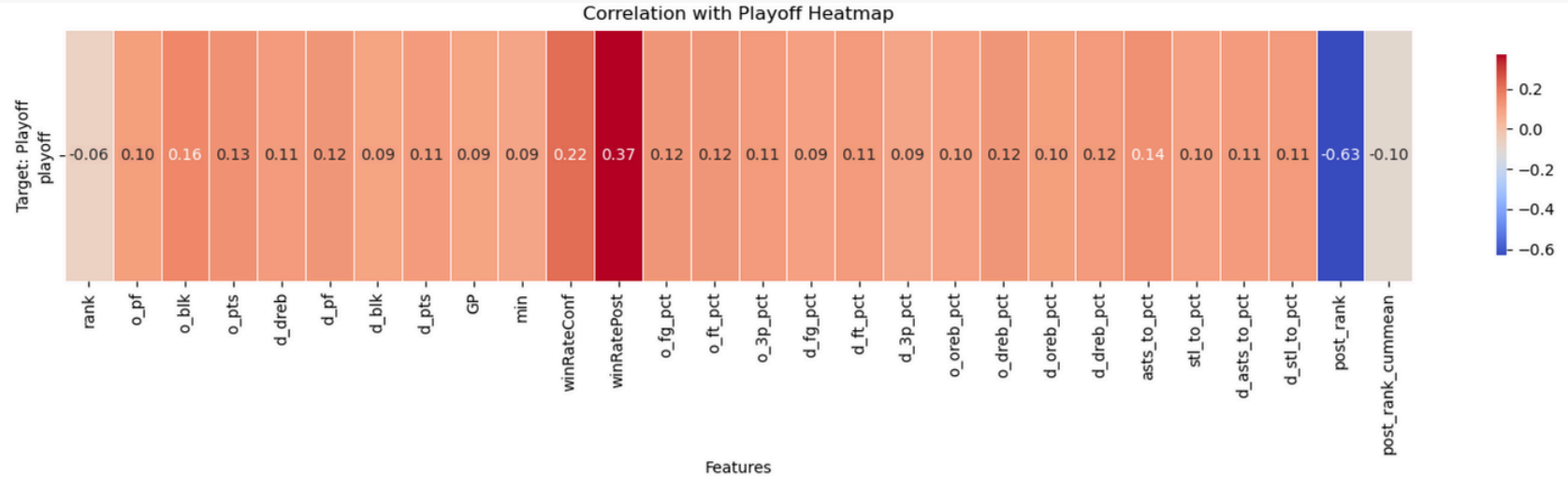
Business Goals to Data Mining Goals

- Build a model to predict which teams qualified for the playoff in a specific year, only having access to information about the previous seasons.
- Obtain at least an accuracy of 60 %.

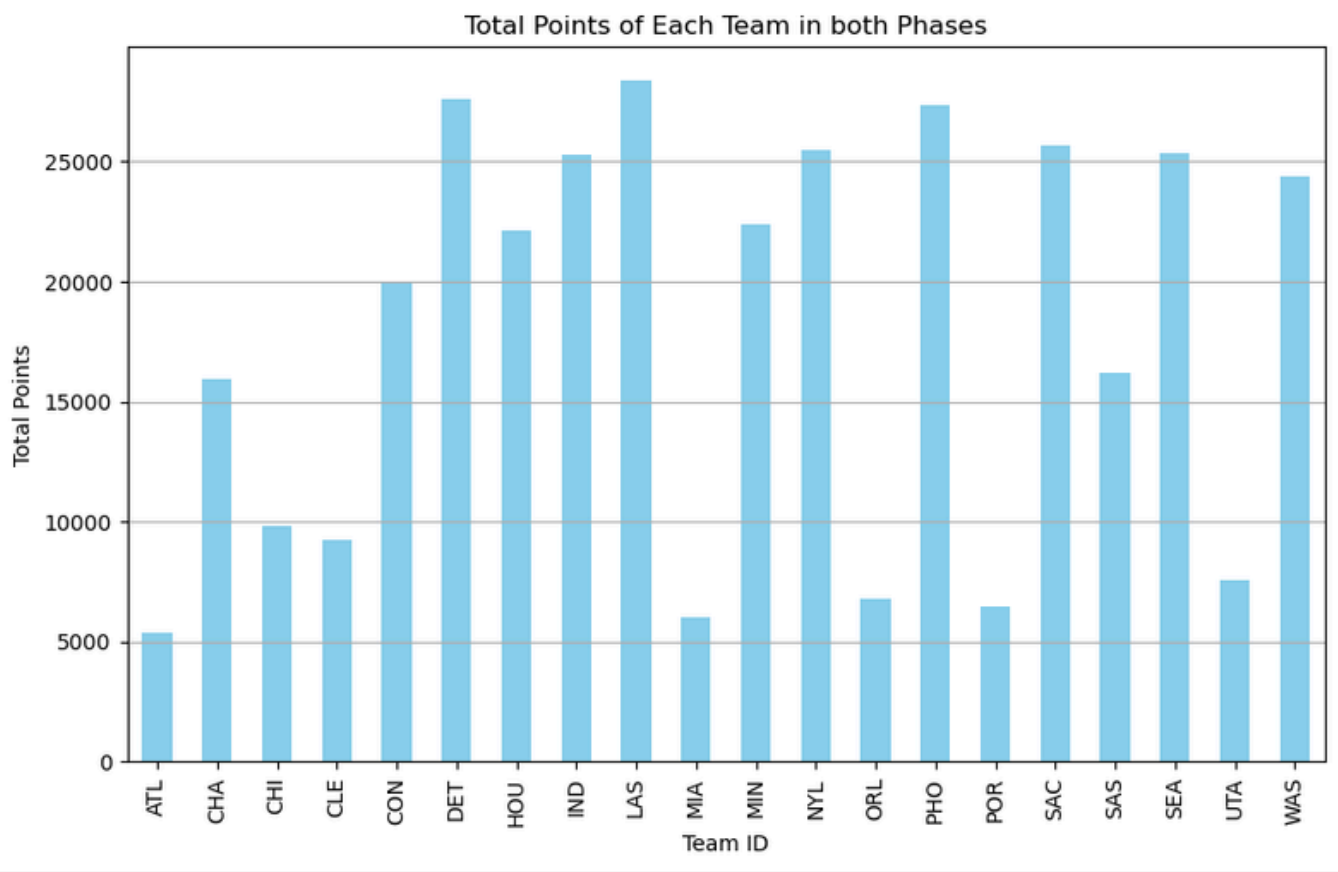
Exploratory Data Analysis



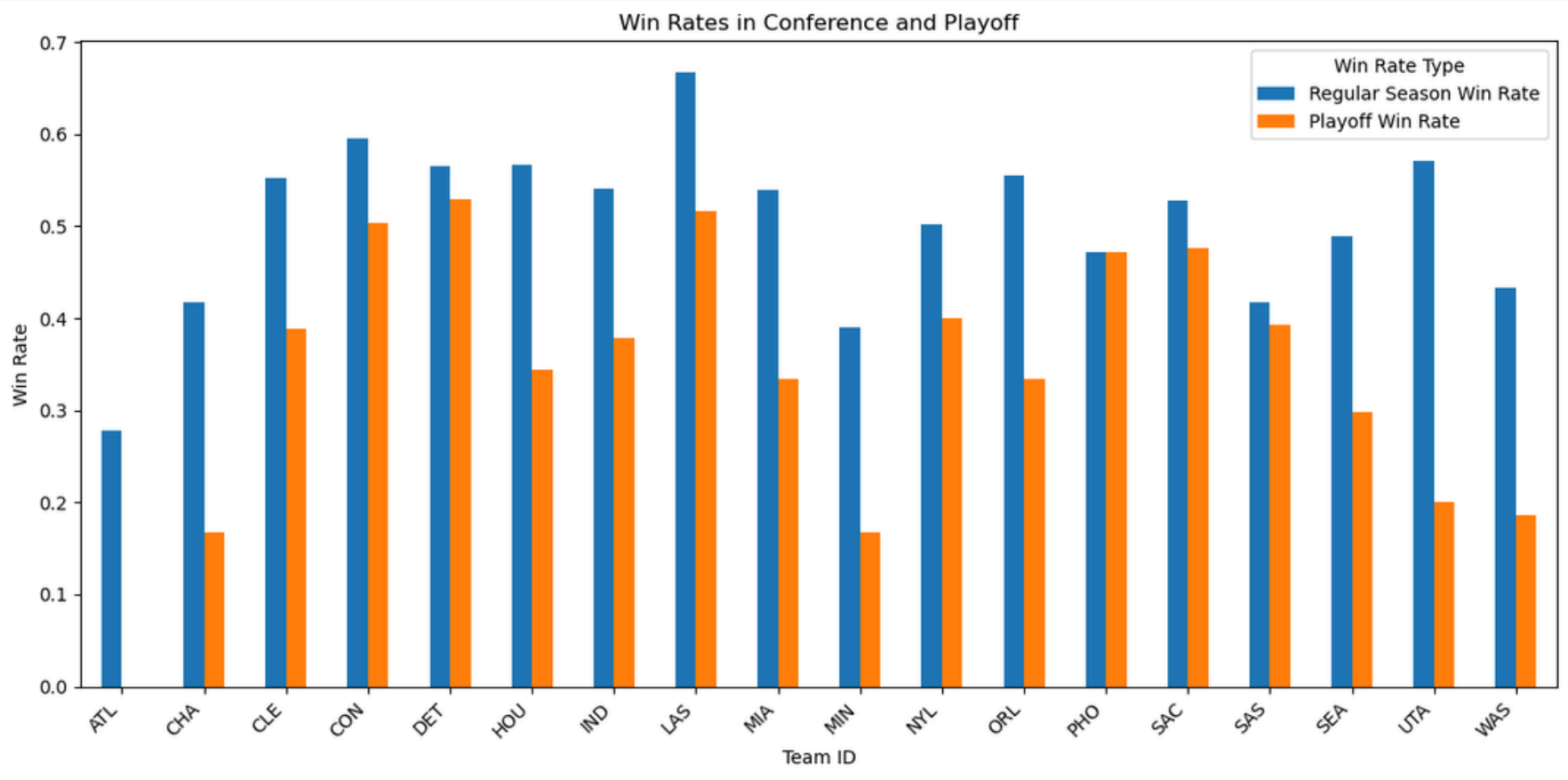
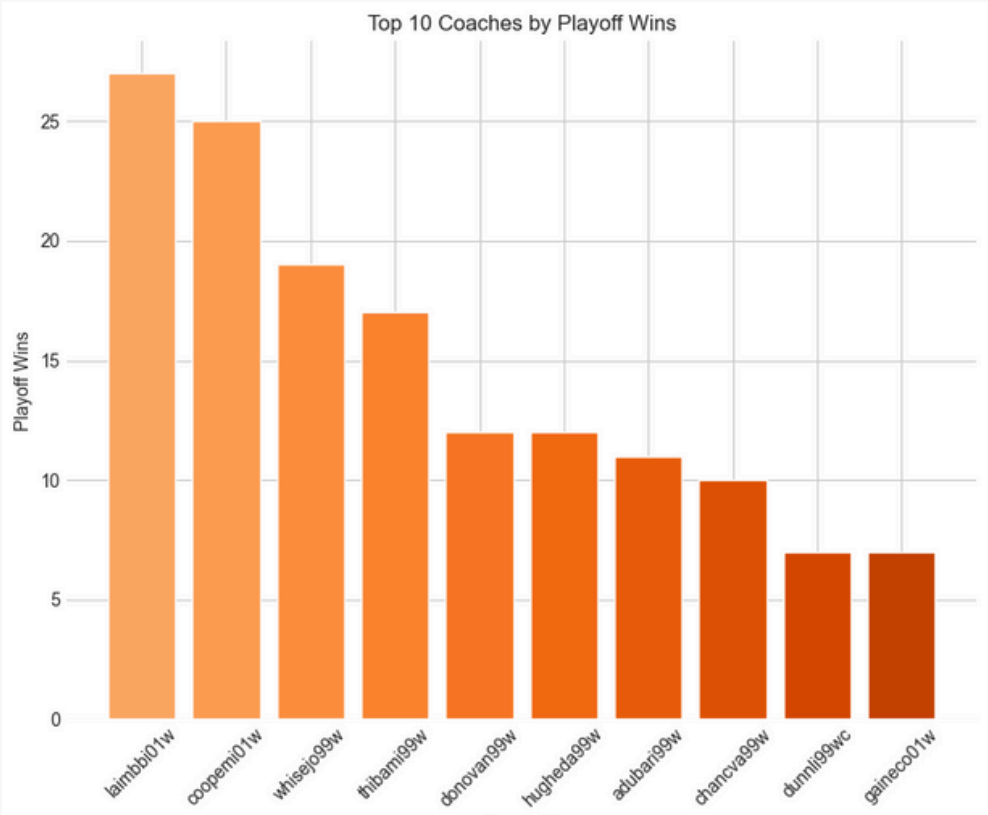
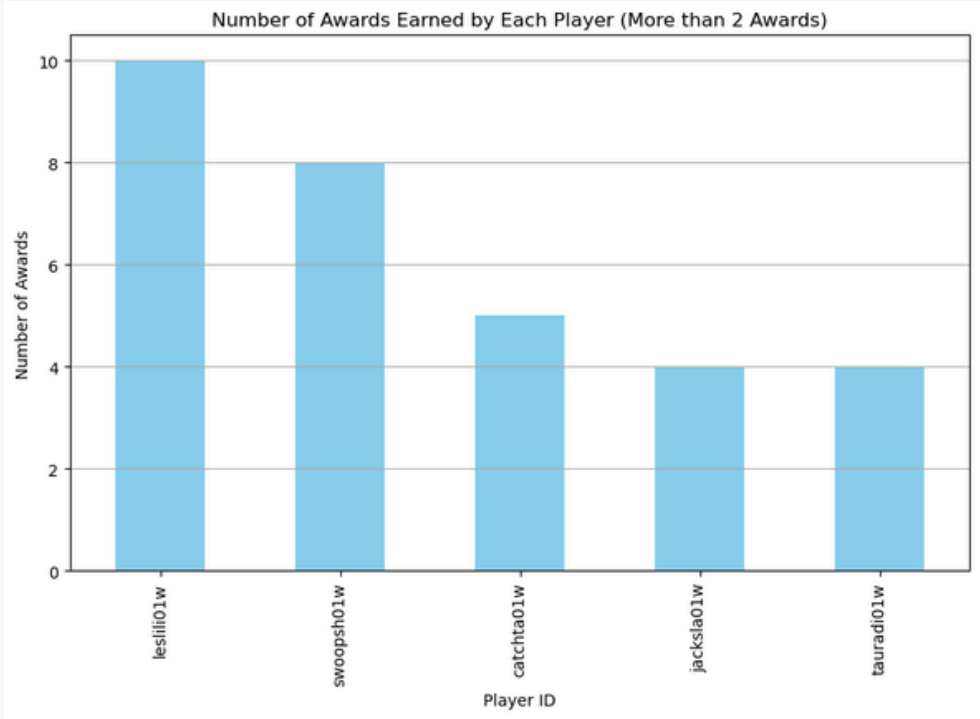
Exploratory Data Analysis



- Correlation matrix that shows the percentage of correlation between the columns from the teams_csv and the target variable (Playoff)



Exploratory Data Analysis





Problem Definition

Problem

- Every season consists of two different phases:
 - Regular phase: All teams compete in a balanced schedule where each team faces every other team multiple times. The goal of this phase is for teams to guarantee as many wins as possible, as their overall performance determines if they went to the playoff or not.
 - Playoff phase: To define the matches, a raffle is done, and the top eight teams will compete in a series of elimination rounds to define the WNBA champion.

Process Model

- The process model chosen was the CRISP-DM (Cross-Industry Standard Process for Data Mining).
 - Helped us in planning and executing the required tasks.

Objectives

- Solve a classification problem where the column to predict has binary values.
- Predict which teams will qualify for the playoffs in the next season.
- Evaluate the predictions from the model using metrics like accuracy, f1-score ...



Data Preparation

Pre-Processing data

- Before diving deep into the data we decided to replace categorical data with numerical values. To accomplish this, we used Label Encoder. We applied this transformations in all datasets that contained at least one of this columns: **ConfID** and **tmID**. In addition, we used the same encoder on our target variable, the **Playoff** column.

Feature Selection

- **Players dataset :**

- Removed all the players that are not assigned to any team.
- Removed the “firstseason” and “lastseason” columns.

- **Players Teams dataset:**

- Removed the “lgID” column.
- Removed all cases where the player hasn't played a single game or minute.

- **Coaches dataset:**

- Removed the “lgID” colum.
- Removed all cases where the coach has no win and loss.

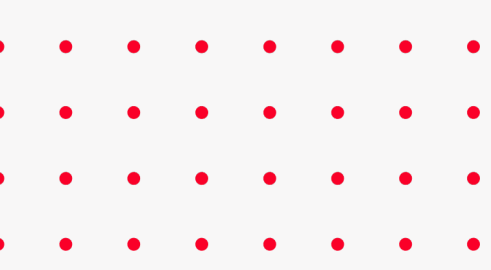
- **Teams dataset:**

- Removed the “lgID”, “divID”, “seeded”, “opptmORB”, “opptmDRB”, “opptmTRB”, “homeW”, “homeL”, “awayW”, “awayL”, “confW”, “confL”, “attend”, “arena”, “name”, “franchID”.

Data Preparation

- In every dataset that contained the *lgld* we decided to remove it because it always has the same value.
- In the datasets where we dropped other columns that are different from: *firstSeason*, *lastSeason*, *birthDate*, and *lgID*, before we dropped those columns we generated percentages derived from that in order to have a consistent and clearer dataset.
- Example :
 - To generate the Win/Loss Rate of each team for regular and playoffs phases, we selected the columns 'won' and 'lost', and 'W' and 'L', respectively, to create a new feature called Win/Loss Rate Regular and Win/Lost Rate Playoff.





Data Preparation

Outlier detection

- To check the outliers, we used the Z-Score. The ones that were relevant where in the weight, and we used a Regressor model to generate new values for it. There was also a outlier in the height although in the feature selection, the value is removed due to not having played any game.

Feature Engineering

To achieve better results, we used accumulative sums and means because they are useful to discover patterns over time. Also, we have generated new features:

- **Coaches dataset:**

Win Rate Conference, Win Rate Post Season.

- **Teams dataset :**

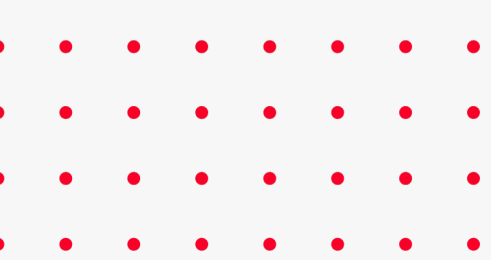
Win Rate Conference, Win Rate Post Season, Percentages for statistics, Post Season Rank.

- **Ratings:**

Squad

Team

Coach



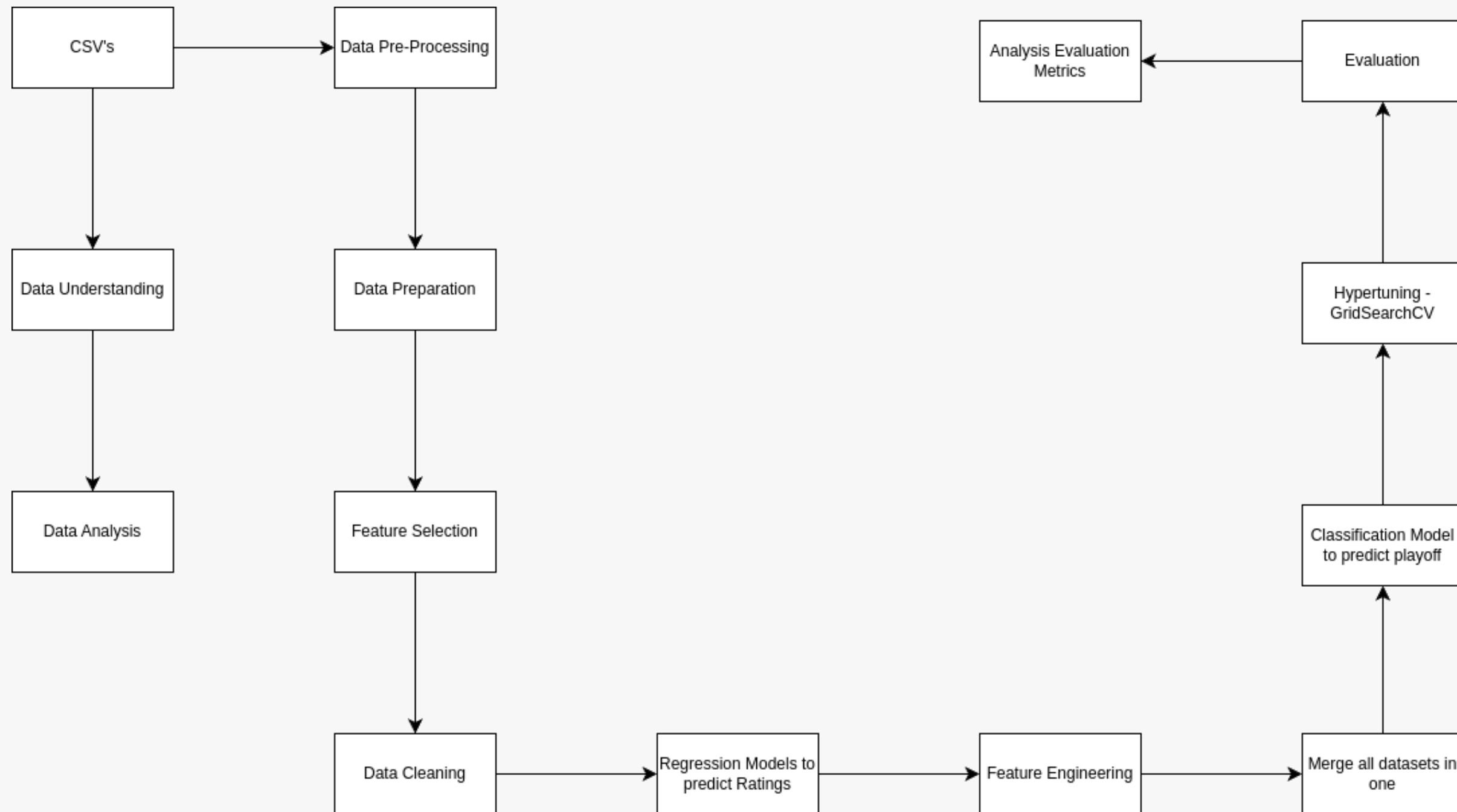
Data Preparation

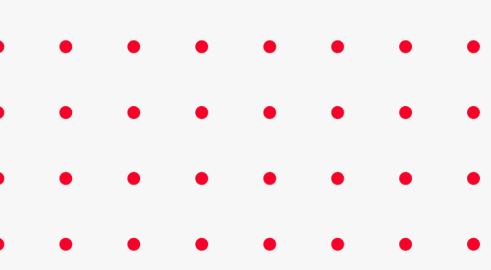
Inconsistencies

Very few inconsistencies were found in the dataset. For instance, in the Awards Players Dataset, the same award it written in two different ways.

playerID	award	year
lawsoka01w	Kim Perrot Sportsmanship	10
mcconsu01w	Kim Perrot Sportsmanship Award	1

Experimental Setup

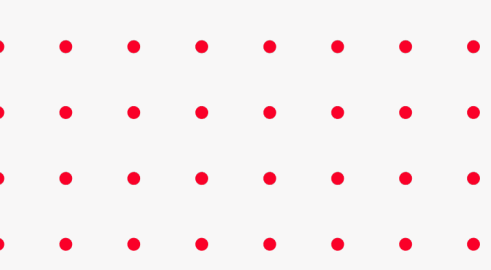




Experimental Setup

Team Rating, Coach Rating and Player Rating

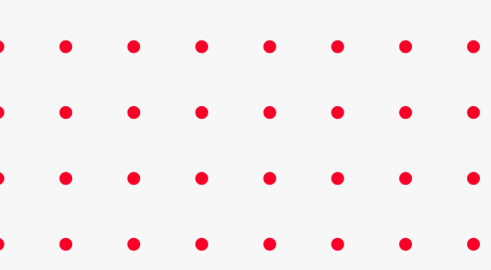
- To predict teams' success in the WNBA playoffs, we calculate three key ratings:
 - **Team Rating:** Reflects the overall performance and strength of the team, based on historical data and key metrics.
 - **Coach Rating:** Evaluates the impact of coaching team's success.
 - **Player Rating:** Measures individual player performance, factoring in key stats and contributions.
 - **Machine Learning Models:** These ratings are generated through some of the machine learning models we used: *Random Forest, Gradient Boosting, Linear Regression, Support Vector Regression, Decision Tree*
- In the calculation of all these ratings, we separate the features from the target variable, allowing us to focus on variables that contribute to our goal.
- In the predictions instead of using the Player Rating, we used the Squad Rating that is the mean of the top 10 players that belong to that team.



Experimental Setup

Accumulative Sum/Mean Sliding Windows

- As explained before, the Accumulative Sum and Mean Sliding Window processes have been essential helping us understand how well or badly a team has performed over multiple years.
- These techniques track the performances of the players, coaches, and teams, not just in a single season, but across a rolling window of seasons, giving us a more dynamic view of their overall path.
- By using a sliding window, we can continuously assess a team's performance based on a specific number of recent seasons, making it easier to spot improvements or declines over time. For instance, instead of simply looking at a team's performance in one year, we can evaluate their average performance over the past 2 years, allowing a more realistic approach when predicting whether a team will make it to the playoffs, as it reflects both recent form and historical consistency.
- This approach brings a lot of advantages:
 - Capturing trends over time;
 - Smoothing Outliers and Inconsistencies;
 - Long-Term Insights



Experimental Setup

Metrics

- Hyperparameter Tuning:
 - Plays a critical role in optimizing the performance and generalizability of a model.
 - Helps balance the trade-off between underfitting and overfitting to achieve better results.
- Binary Classification
 - Enables predictions to address problems with two possible outcomes (e.g., yes/no, true/false).

- Accuracy

Proportion of correct predictions compared to the total predictions

- Precision

Proportion of True Positives compared to all instances the model predicted as positive

- Recall

Proportion of True Positives compared to all actual positive instances

- F1-Score

Harmonic mean of precision and recall, balancing both metrics to provide a single performance measure

- ROC-AUC

Area under the ROC curve, measuring the model's ability to distinguish between positive and negative classes across all possible thresholds.



Results

- To check all of our experimental setups we decided to predict the playoff using the following models:
 - **K-Nearest Neighbors:**
 - **Description:** This algorithm predicts a team's playoff based on the performance of its nearest neighbors in feature space.
 - **Strength:** Simple and effective for smaller datasets.
 - **Challenge:** Can be sensitive to irrelevant features.
 - **DecisionTree:**
 - **Description:** This tree-like structure splits data based on feature values to make predictions.
 - **Strength:** Easy to interpret and understand.
 - **Challenge:** Insensitive to local optimum.
 - **Random Forest:**
 - **Description:** An ensemble model that combines multiple decision trees to improve accuracy and reduce overfitting.
 - **Strength:** Robust to overfitting and handles large datasets well.
 - **Challenge:** It can be computationally intensive with a large number of trees.



Results


- **Support Vector Classifier**

- **Description:** A model that separates classes using a hyperplane in a high-dimensional space.
- **Strength:** Works well for linear problems.
- **Challenge:** Requires careful tuning of hyperparameters.

- **Logistic Regression**

- **Description:** A statistical model that predicts probabilities and classifies data into binary categories.
- **Strength:** Simple, efficient, and interpretable.
- **Challenge:** Assumes linear relationships between features and the target.

- **Gradient Boosting**

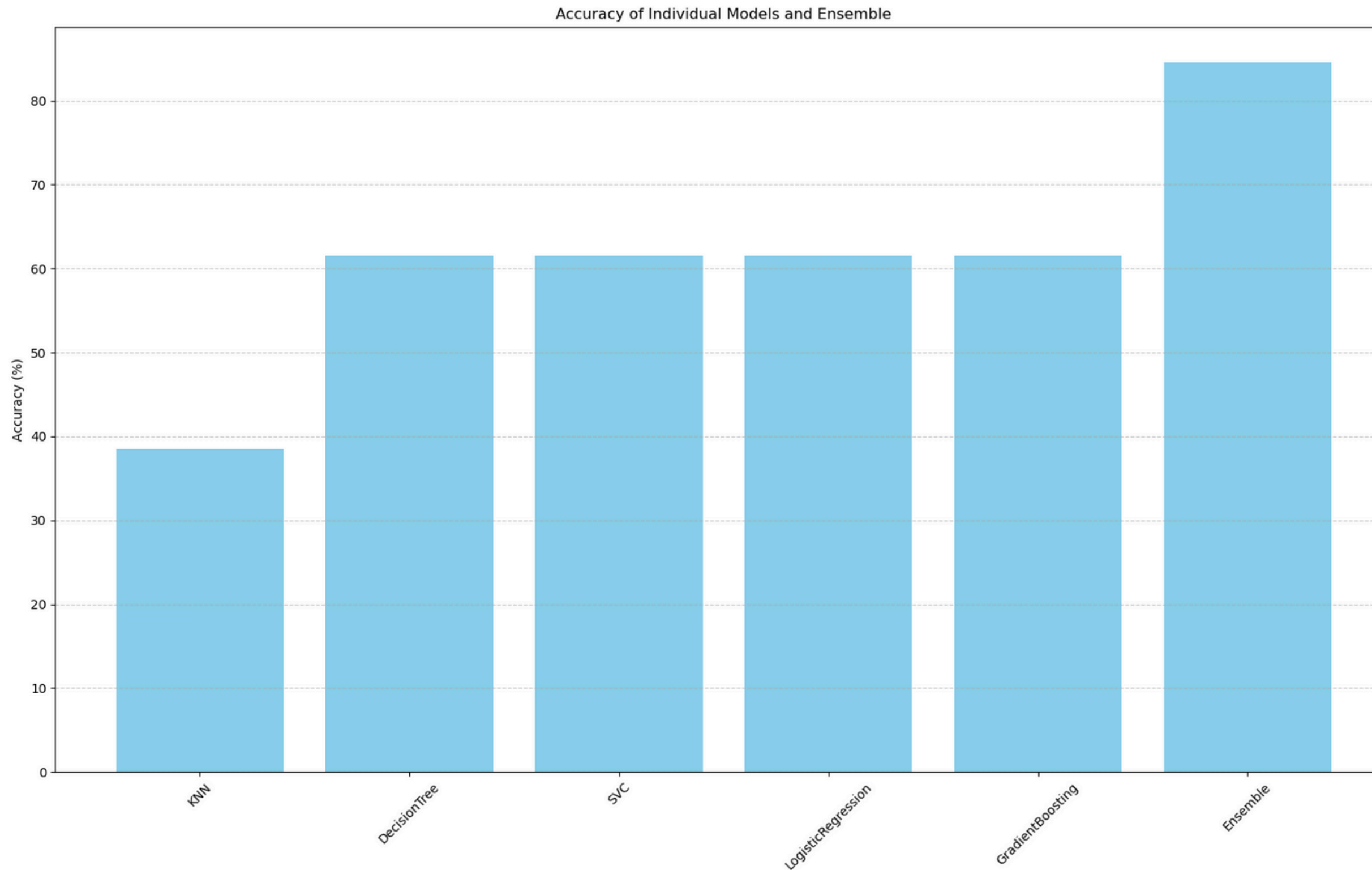
- **Description:** An ensemble method that builds models sequentially to correct errors from previous iterations.
 - **Strength:** High predictive power and flexibility in handling complex relationships.
 - **Challenge:** Computationally expensive, especially when dealing with large datasets.
- 



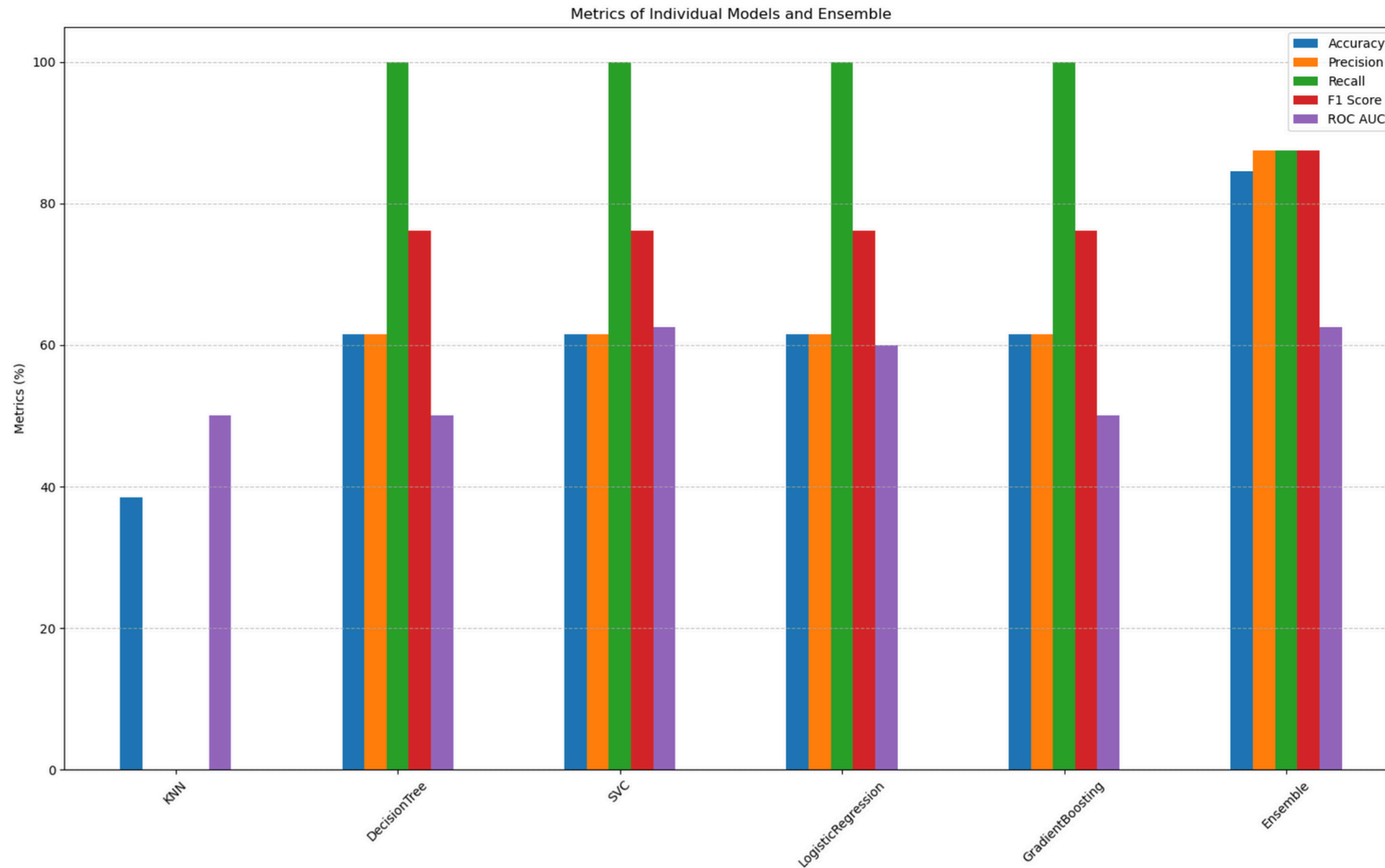
Results

- To improve the consistency and accuracy of our playoff predictions, we implemented an ensemble voting classifier. This approach combines the strengths of the multiple machine learning models, mentioned in the previous slides, upgrading their unique capabilities to create a more reliable prediction system.
- **Voting Mechanism:**
 - A soft voting strategy was used, combining the probabilistic predictions from each model, ensuring that the final prediction benefits from a balanced contribution among models.
- **Advantages of Ensemble:**
 - Increased Accuracy.
 - Robustness.
 - Scalability – this may be the most important aspect because in the future we will be predicting which teams went to the playoff in an undefined season only knowing the past years.

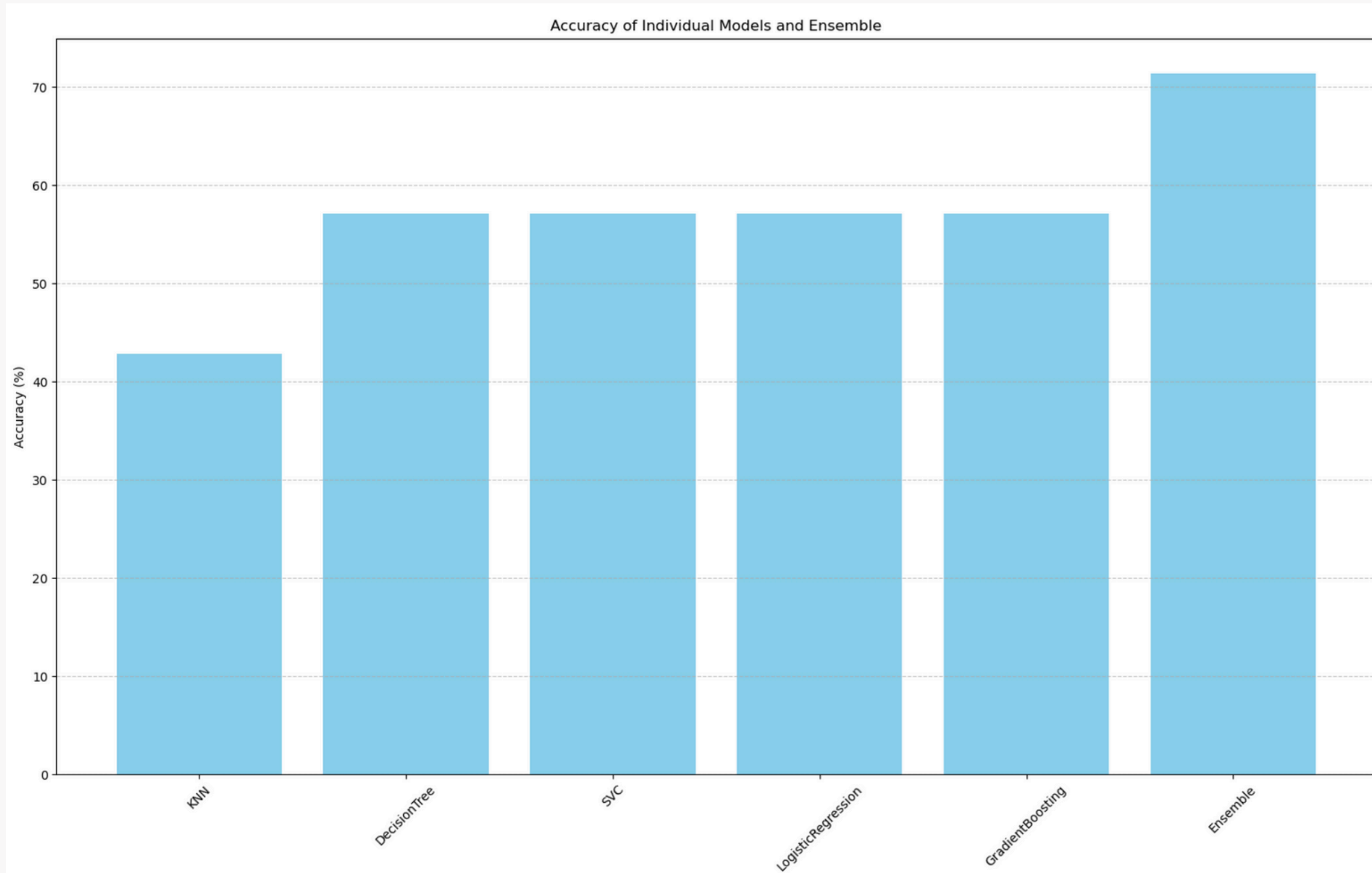
RESULTS SEASON 8



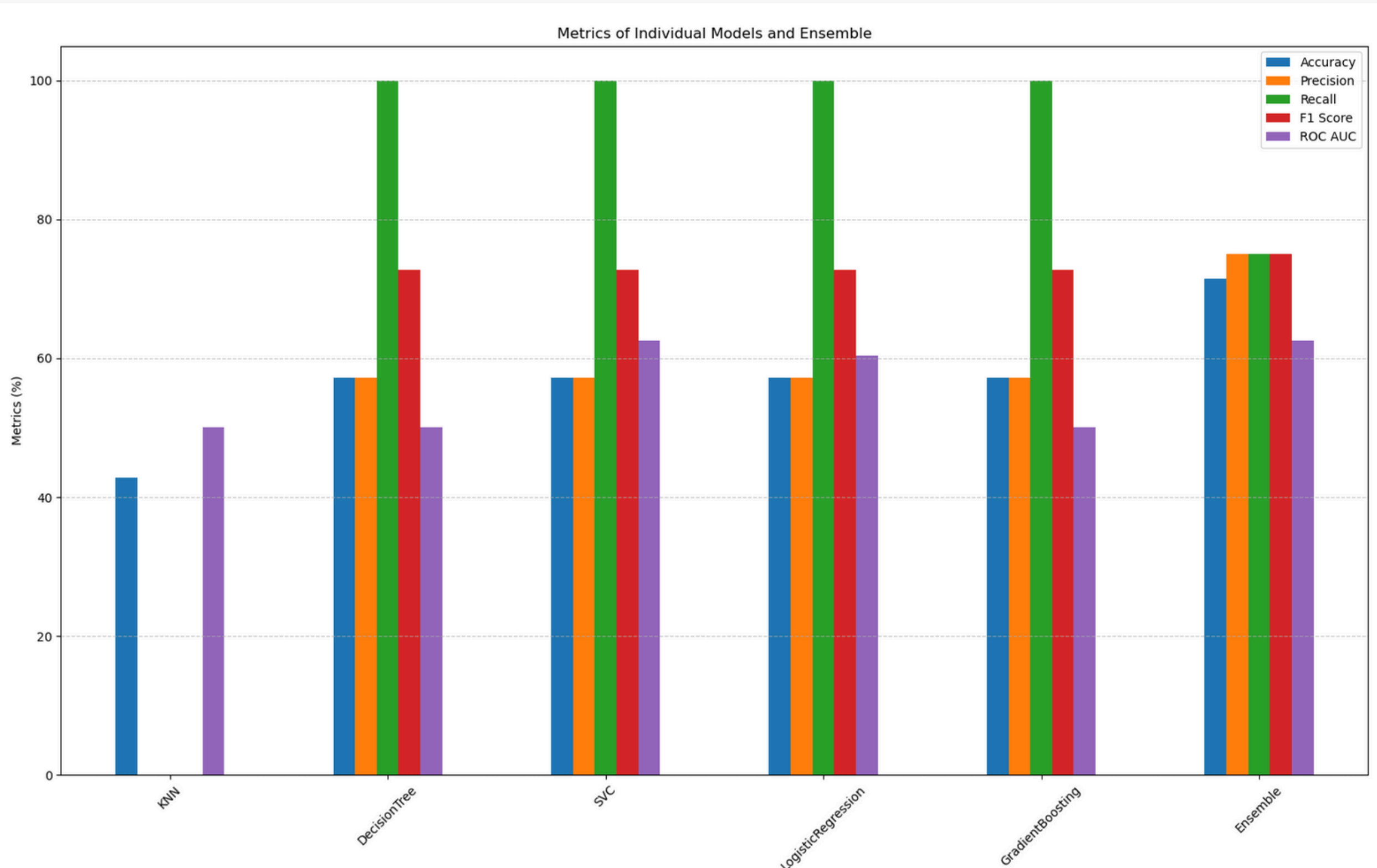
RESULTS SEASON 8



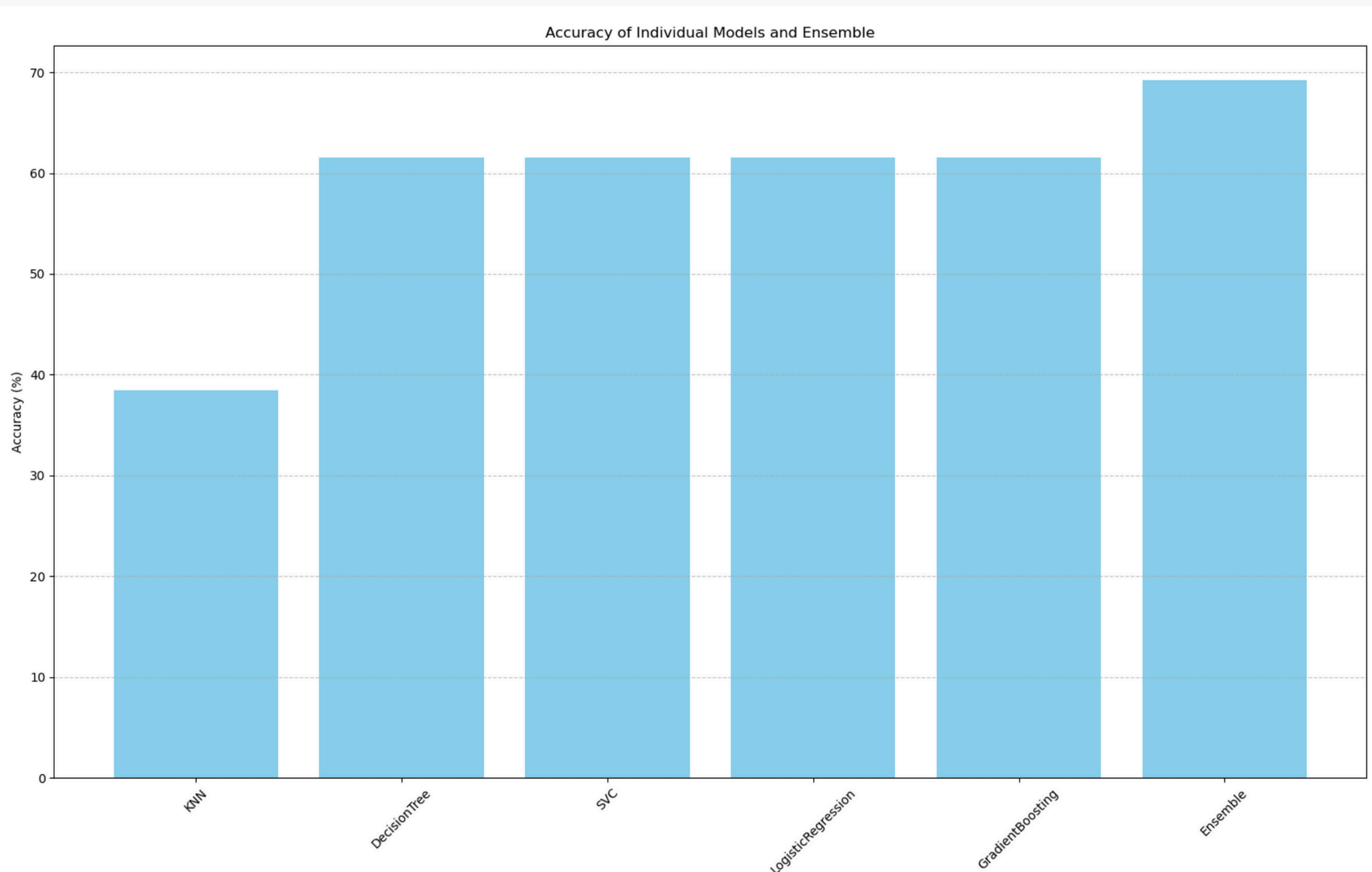
RESULTS SEASON 9



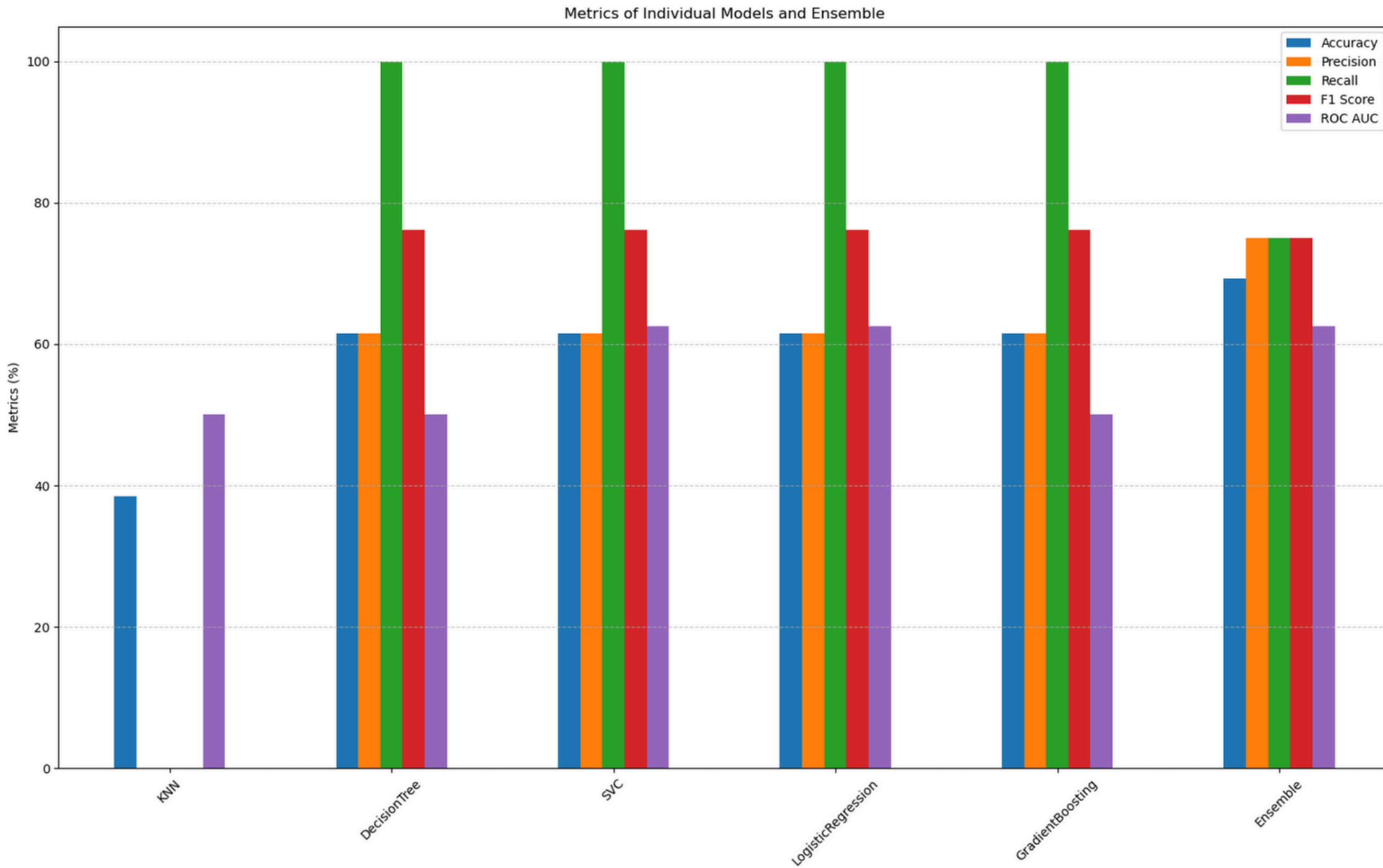
RESULTS SEASON 9

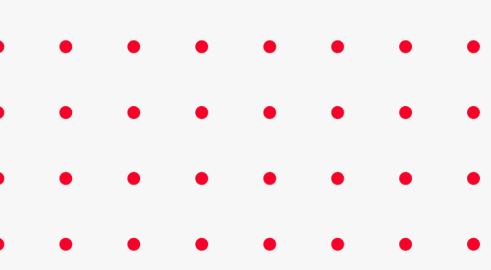


RESULTS SEASON 10



RESULTS SEASON 10





Conclusion

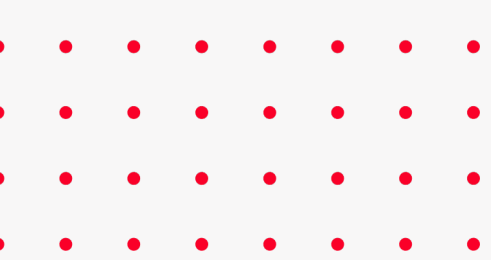
- In this project, we successfully combined data analysis, feature engineering, and machine learning techniques to predict which teams are the most likely to qualify for the WNBA playoffs.
- The use of accumulative sums and means to explore historical data provided us with a dynamic perspective on historical trends.
- Feature engineering helped calculate the statistics that have the most influence on our predictions.
- The implementation of the ensemble voting classifier was important in order to combine all the algorithms that we think that are the most appropriate to achieve our goal.
- Analyzed feature importance to highlight the most influential factors, ensuring usability and scalability for end users.
- Delivered a solution that not only predicts playoff teams with high consistency but also provides insights to support future decisions.



Kaggle Submission

- All submissions to Kaggle use the same algorithm, ensemble, to predict which teams went to the playoff.
- The submissions differ due to the algorithms used to predict the ratings for the team, squad, and coach. We decided to not change the prediction algorithm because the ensemble is a strong algorithm against overfitting and underfitting.
- In the first three submissions, we decided not to predict the squad ratings using a model, but instead use the mean of the top 10 players of each team considering the last seasons. In the last two submissions, we decided to, based on the previous ratings, use an algorithm to also predict it.
- The results predicted from the last submission, that were considered the best are:

	tmID	predicted_playoff
0	ATL	0
1	CHI	0
2	CON	1
3	IND	1
4	LAS	0
5	MIN	0
6	NYL	1
7	PHO	1
8	SAS	1
9	SEA	1
10	TUL	1
11	WAS	1



Annexes

Outliers Players

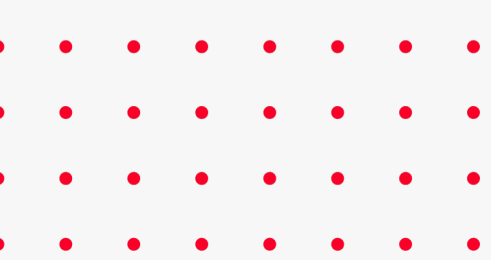
```
Outliers in Players:  
Outliers in weight: [0, 0, 0, 0, 0, 0, 253]  
Outliers in height: [9.0]
```

Outliers Coaches

```
Outliers in Coaches:  
Outliers in won: []  
Outliers in lost: []  
Outliers in post_wins: []  
Outliers in post_losses: []
```

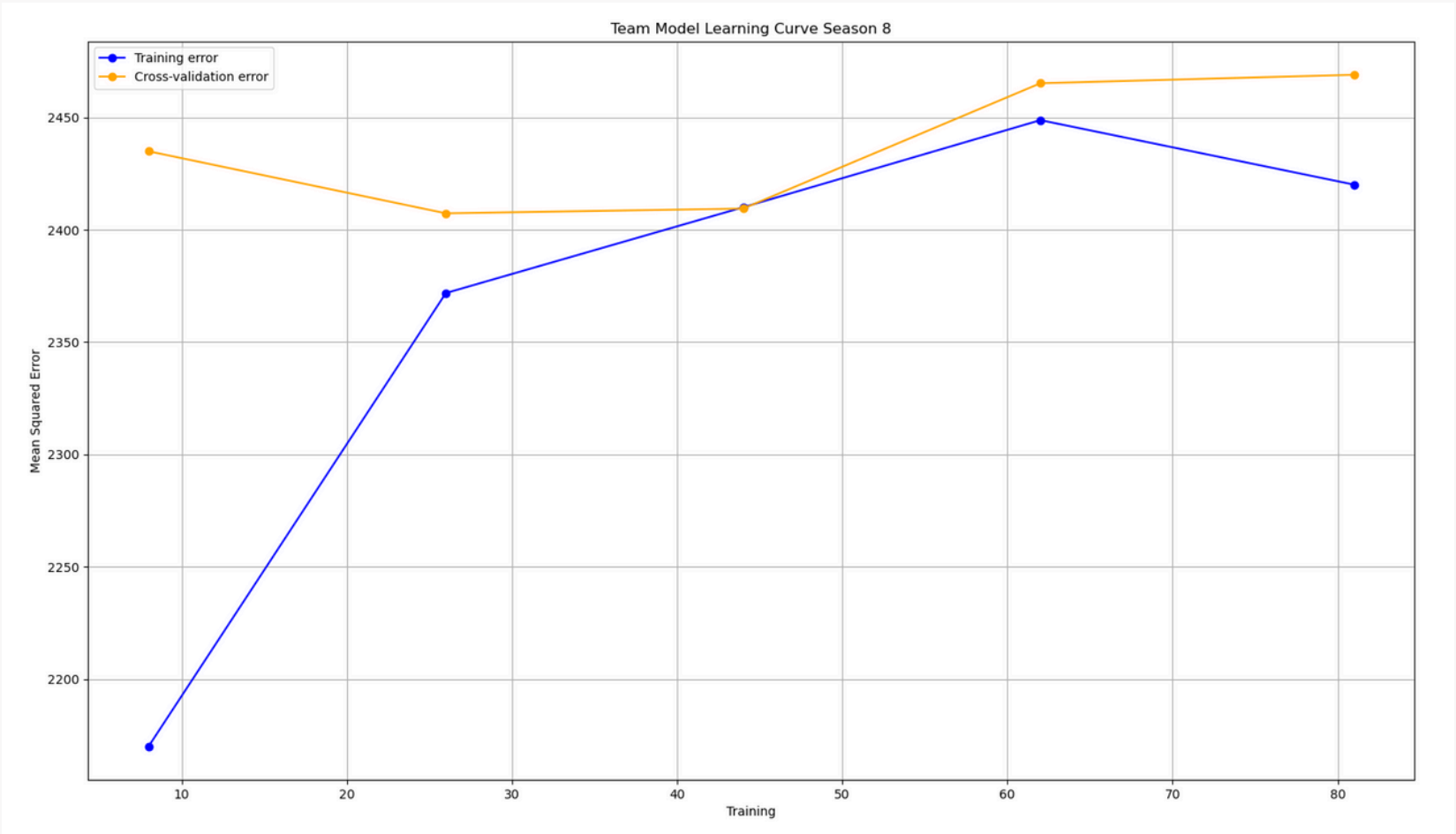
Outliers Teams

```
No discrepancies found for column o_fgm  
No discrepancies found for column o_fga  
No discrepancies found for column o_ftm  
No discrepancies found for column o_fta  
No discrepancies found for column o_3pm  
No discrepancies found for column o_3pa  
No discrepancies found for column o_oreb  
No discrepancies found for column o_dreb  
No discrepancies found for column o_reb  
No discrepancies found for column o_ast  
No discrepancies found for column o_pf  
No discrepancies found for column o_stl  
No discrepancies found for column o_to  
No discrepancies found for column o_blk  
No discrepancies found for column o_pts  
No discrepancies found for column d_fgm  
No discrepancies found for column d_fga  
No discrepancies found for column d_ftm  
No discrepancies found for column d_fta  
No discrepancies found for column d_3pm  
No discrepancies found for column d_3pa  
No discrepancies found for column GP
```

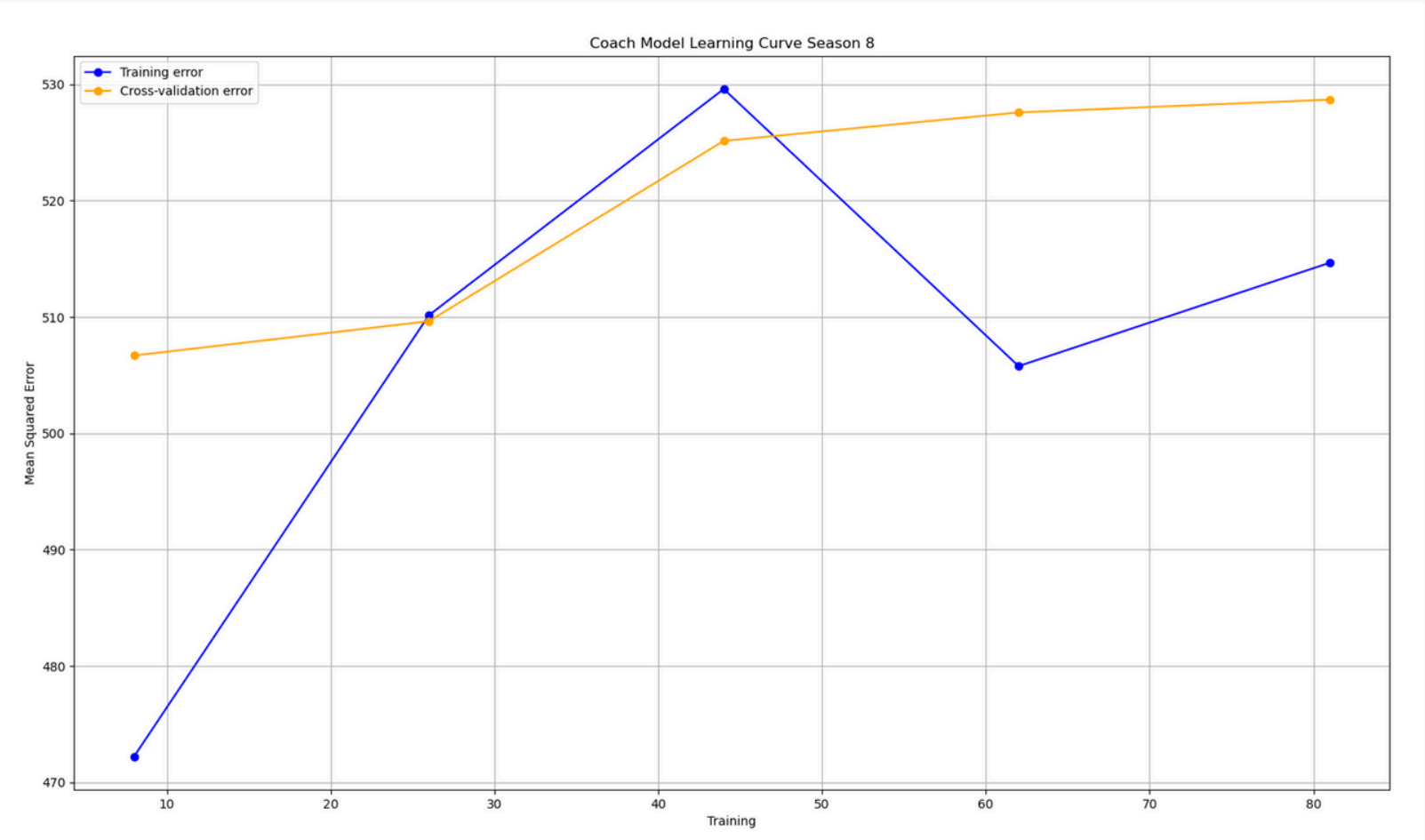


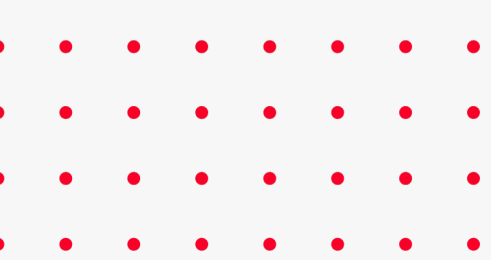
Annexes

Performance Training Data Season 8 – Prediction Team Rating



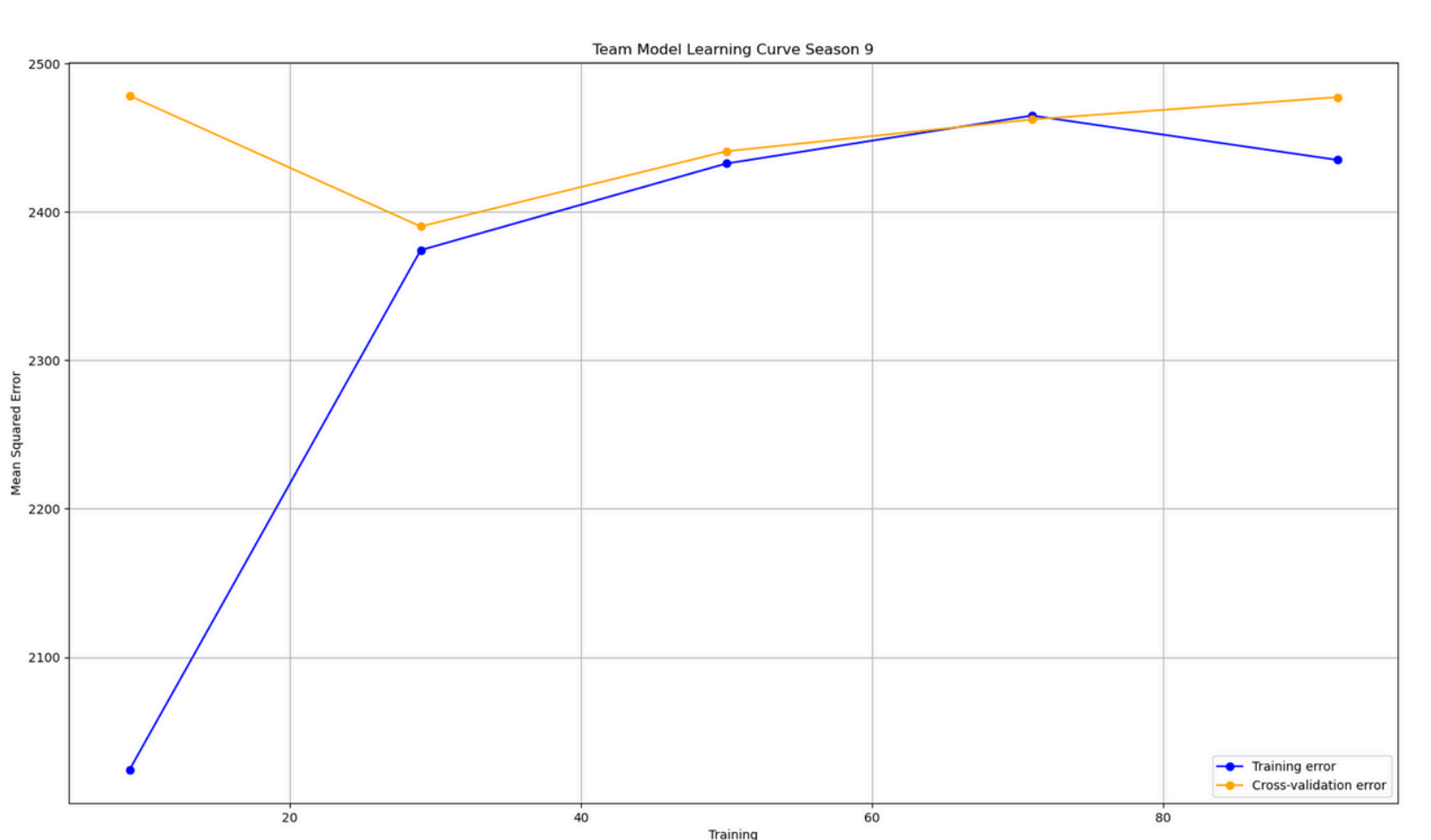
Performance Training Data Season 8 – Prediction Coach Rating



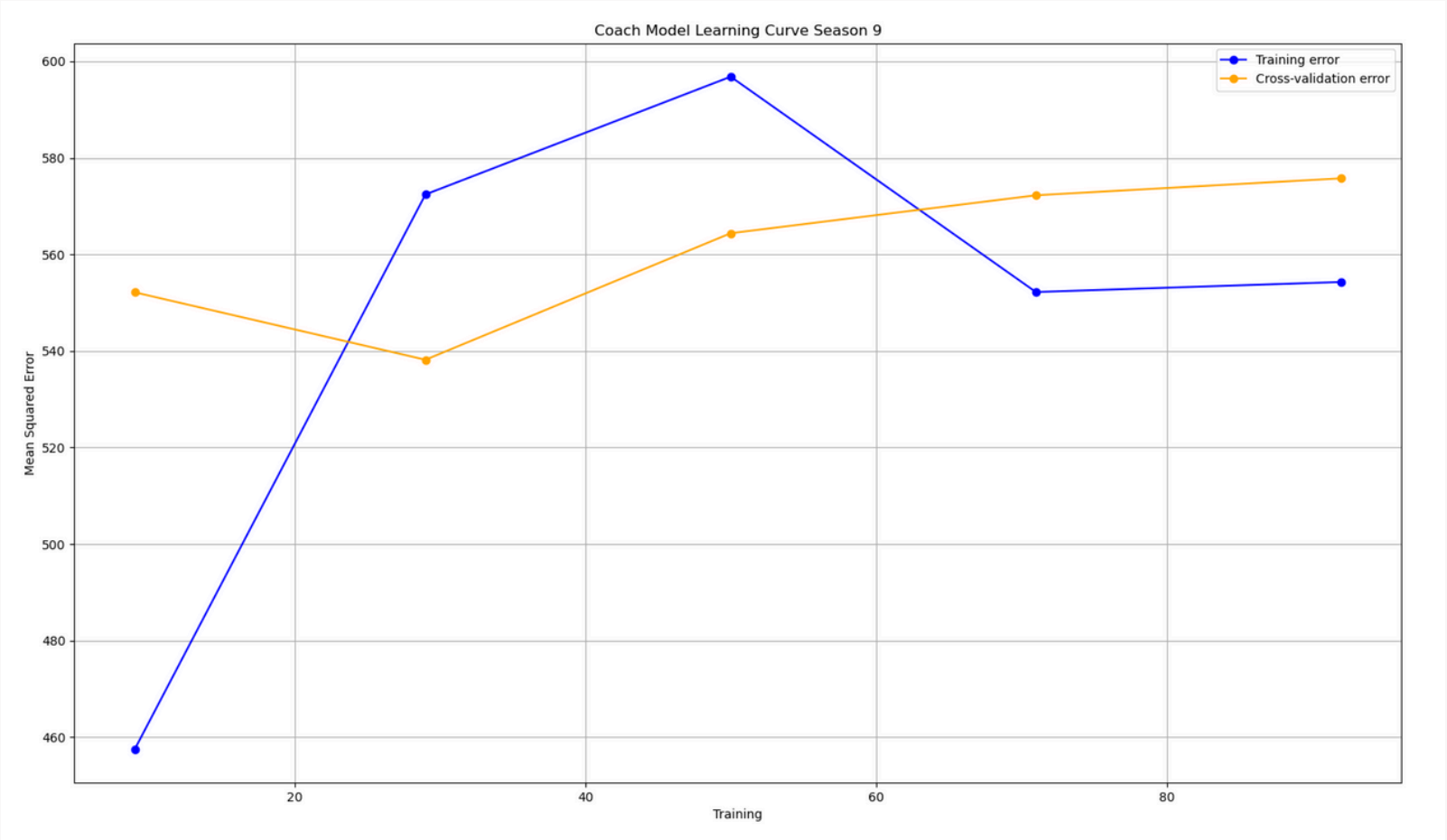


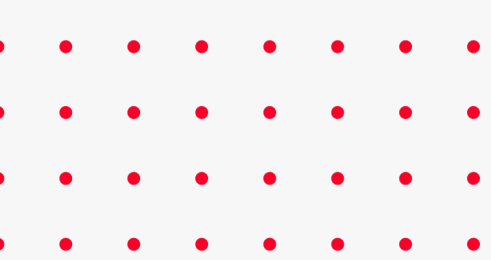
Annexes

Performance Training Data Season 9 – Prediction Team Rating



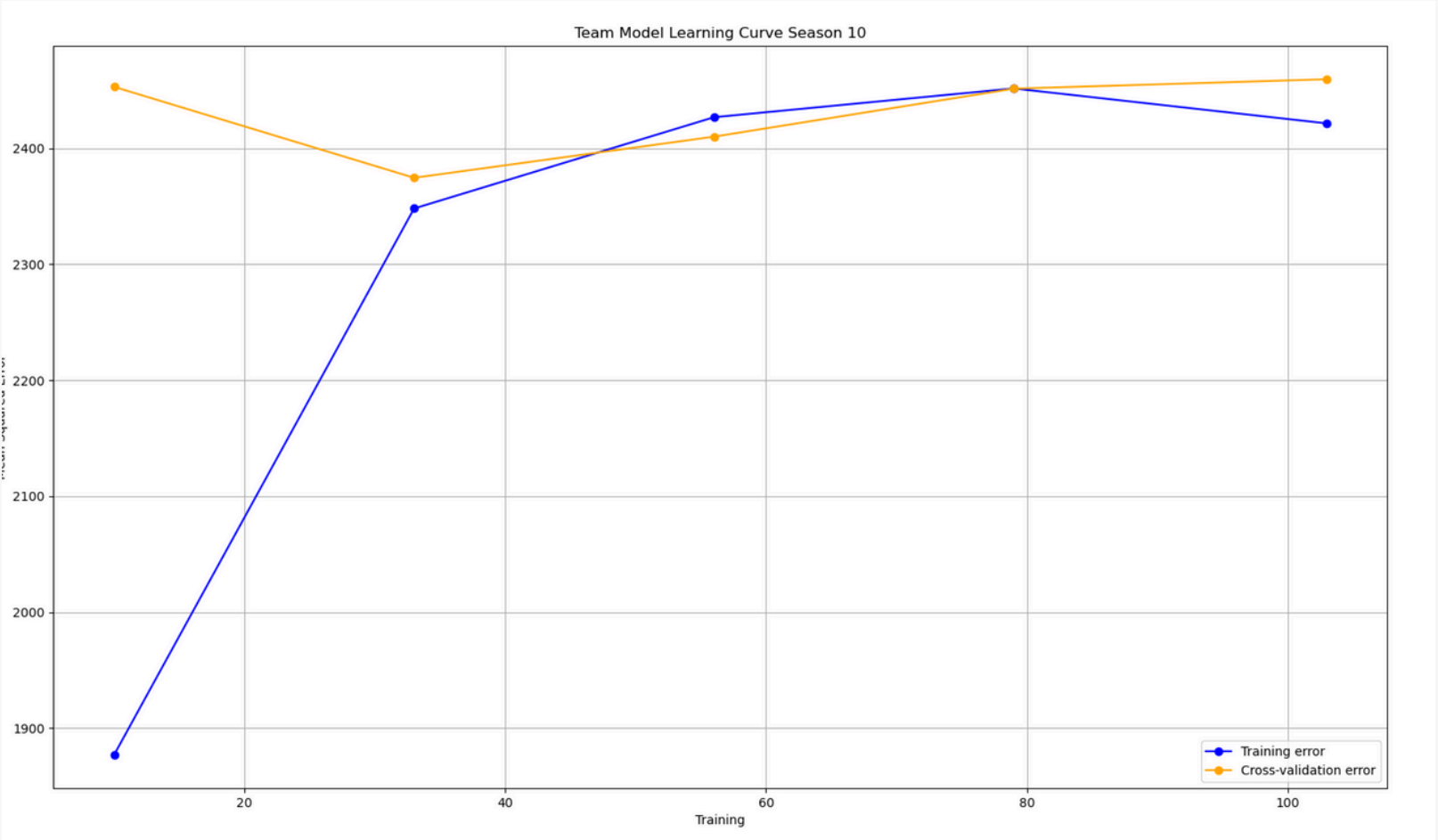
Performance Training Data Season 9 – Prediction Coach Rating





Annexes

Performance Training Data Season 10 – Prediction Team Rating



Performance Training Data Season 10 – Prediction Coach Rating

