Machine Learning Project – Football Match Outcome Prediction

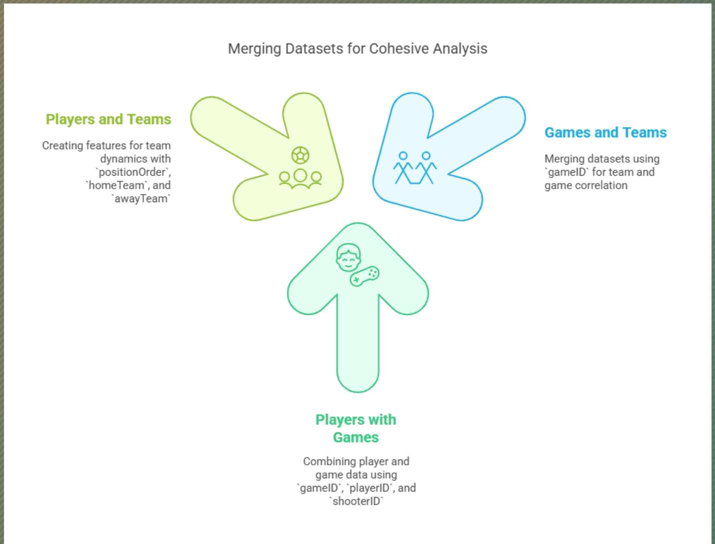
Using Machine Learning to Predict Match Results

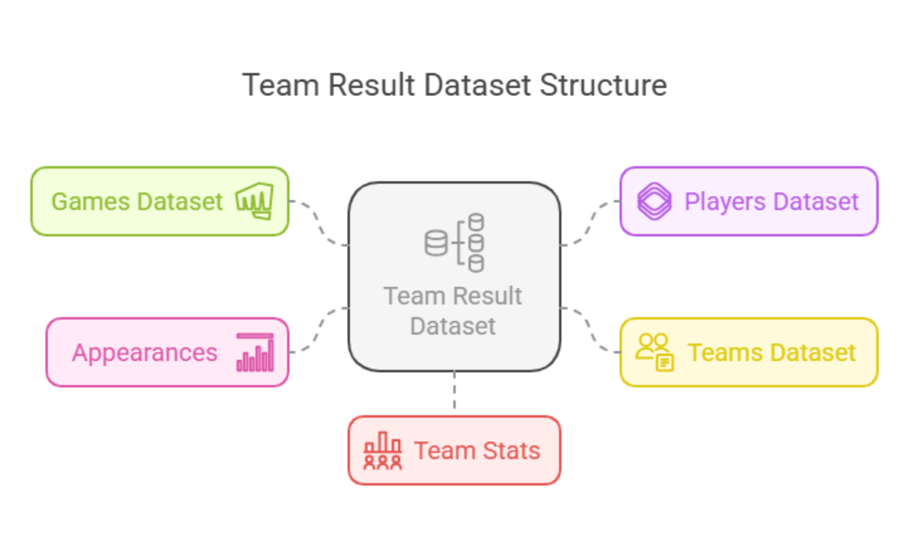
1. The importance of Football

2. The Data

- Predict games result(home team)- Win, Lose or Draw
- Dataset: games results from 2015 – 2020 (12,680 game records)
- Useful for sports analytics and game predictions
- The data Football Database <https://www.kaggle.com/datasets/technika148/football-database/data>
- 7 files (appearance, games, leagues, players, shots, teams, teamstats)
- 3 main part – Players performance; Teams statistics ;Games statistics

3. Data Preparation





- Removed duplicate columns (columns of Goals, XG etx.)
- Created Categorized columns (such as Goals 1-5 and 5+)
- Encoded target variable as numerical labels (Win=1, Draw=0, Loss=2)

3. Exploratory Data Analysis (EDA)

Performance Drive Results (How well team played)

- Goals, assists, and xG strongly linked to wins
- High pass accuracy and shots on target = better outcomes
- Top ANOVA features: goals_scored, shots_on_target

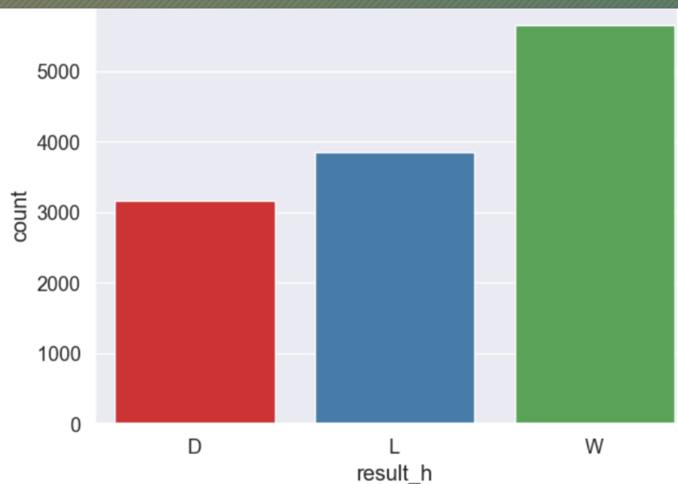
Disciplinary Factors Matter (The player behavior)

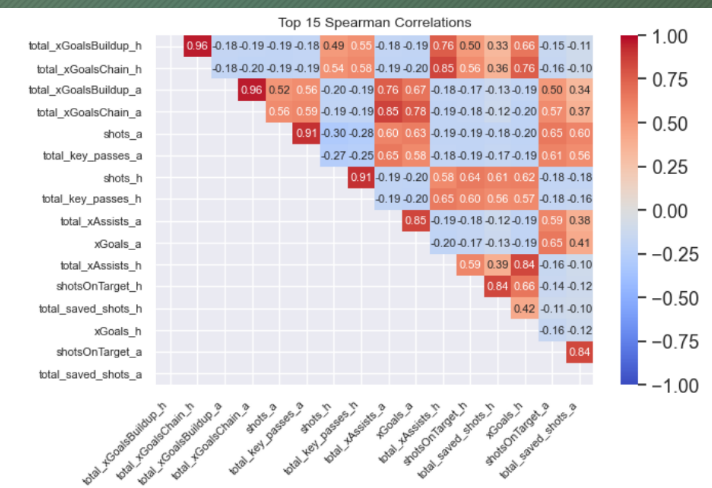
- More red cards → higher chance of losing
- Strong significance ($p < 0.001$)
- Discipline is critical to match results

Match Statistics Patterns (Data patterns)

- Most features are right-skewed
- Strong correlations: possession ↔ passes

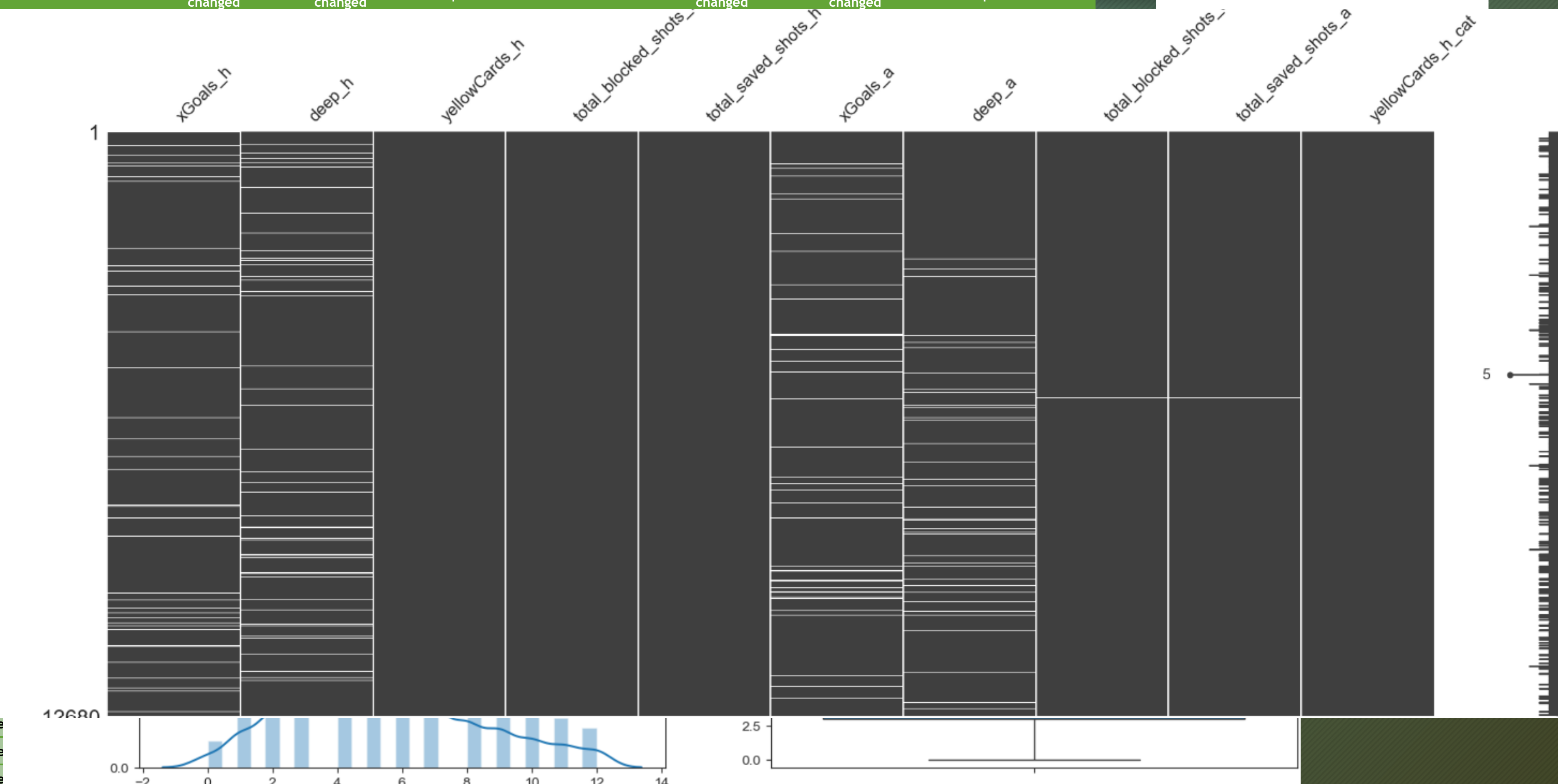
Home wins dominate





4. Outliers, Missing Data and Feature Engineering

corners_a



feature	distribution_changed	correlation_changed	drop	feature	distribution_changed	correlation_changed	drop
re							
re							
re							
season	-	-	yes				

4. Outliers, Missing Data and Feature Engineering

- Created new features derived from player and team performance metrics, including:
  - goal_difference, xGoals_chain, xAssists_total, and others.
- Rolling of average 5 years or ratio calculation
  - home_xGoals_h_rolling5 or home_win_rate_5 (wining ration in last 5 games)

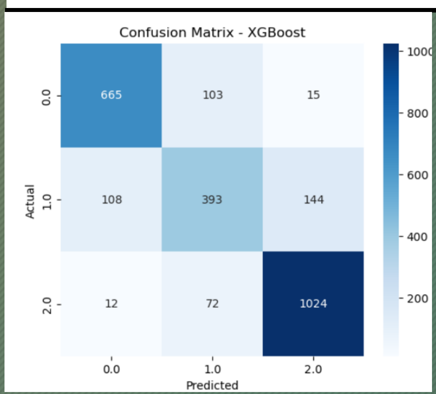
4. Model Selection & Training

- Models: Logistic Regression Decision Tree Random Forest AdaBoost Gradient Boosting XGBoost SVM Extra Trees.
- Used train-test split (80-20%).
- Fine Tuning

Parameter	Values
n_estimators	100, 200, 300
max_depth	3, 5, 7
learning_rate	0.01, 0.05, 0.1
subsample	0.7, 0.9, 1.0
colsample_bytree	0.7, 0.9, 1.0
gamma	0, 1, 5

5. Model Evaluation

Model	Accuracy	Precision	Recall	f1-score	Log-loss	AUC
Logistic Regression	0.73541	0.70354	0.69915	0.69956	0.599469	0.88263
Decision Tree	0.696372	0.669571	0.66796	0.66863	10.94385	0.75758
Random Forest	0.768533	0.741974	0.73473	0.7352	0.571391	0.90509
AdaBoost	0.73265	0.708296	0.7013	0.70392	1.028839	0.84506
Gradient Boosting	0.7847	0.763293	0.75982	0.76133	0.494114	0.9251
XGBoost	0.820978	0.801544	0.79426	0.79671	0.430875	0.9423
SVM	0.436909	0.145636	0.33333	0.20271	0.884034	0.75514
Extra Trees	0.776814	0.752512	0.74282	0.74378	0.584587	0.91064



5. Model Evaluation

- Best Parameters: {'subsample': 0.9, 'n_estimators': 300, 'max_depth': 7, 'learning_rate': 0.05, 'gamma': 1, 'colsample_bytree': 0.7}

Model	Accuracy	Precision	Recall	f1-score
XGBoost	0.821	0.802	0.794	0.797
XGBoost After Fine tuning	0.817	0.797	0.791	0.793

6. Conclusion & Model Deployment

- Key predictors: total_assists, Shots on Target, total_saved_shots , xGoals
- Model: XGBoost Accuracy: 82%
- Precision: 80%; Recall: 79%
- Useful data driven for sports analytics and game predictions, Coaches, sports scientists, media and fans.
- Future Work: Add player form, weather conditions, Try deep learning models for better predictions

Feature	Importance
total_assists_h_cat	18.332561
total_assists_a_cat	17.551327
shotsOnTarget_h	2.978227
shotsOnTarget_a	2.955934
total_saved_shots_h	2.876338
total_saved_shots_a	2.608773
xGoals_h	2.207679
xGoals_a	1.936435
total_key_passes_a	1.509871
total_key_passes_h	1.481853
total_xGoalsChain_a	1.199424
shots_h	1.136373
total_xGoalsChain_h	1.061404
shots_a	1.033608
ppda_h	1.026962
ppda_a	0.974332
corners_h	0.950514
awayTeamID	0.93714
yellowCards_h_cat	0.93661
homeTeamID	0.910703
total_blocked_shots_h	0.903287
deep_h	0.8893
deep_a	0.87133
yellowCards_a_cat	0.83383
season	0.822358

Alternative Model - prediction Before The Game (historical data)

Model	Accuracy	Precision	Recall	f1-score	Log-loss	AUC
Logistic Regression	0.460	0.408	0.353	0.268	1.061	0.530
Decision Tree	0.380	0.356	0.355	0.355	22.343	0.518
Random Forest	0.444	0.362	0.361	0.318	1.065	0.551
AdaBoost	0.465	0.302	0.367	0.300	1.076	0.549
Gradient Boosting	0.475	0.399	0.384	0.331	1.046	0.576
XGBoost	0.455	0.394	0.395	0.380	1.101	0.575
SVM	0.450	0.150	0.333	0.207	1.060	0.542
Extra Trees	0.439	0.363	0.357	0.316	1.084	0.541

- Still Beter then 0.33

Q&A