Final Project
IS SOCCER ALL ABOUT STATS?

by Al Yazid Bensaid, Utkarsh Prasad, Sahil Shaik

Introduction

Ongoing debate: Is soccer really unpredictable?

Motivation:

- Develop a model with practical benefits for sports bettors
- Recognizing diverse goals
- Guiding users to a classifier aligned with their risk tolerance and betting strategy.

Research Goals:

- Primary Goal: Build a predictive model that will actually predict the outcome of a match for a new dataset
- Secondary Goal: Yield reliable interpretative insights about the nature of the relationship between all the variables

Presentation of the dataset

Source: Dataset from Kaggle, capturing all 64 matches of the 2022 FIFA World Cup.

Data Collection:

- Each row corresponds to a unique match
- Detailed tracking data recorded by multiple cameras, synchronized with match footage.
- Enables diverse exploratory analysis and visualization techniques.
    - Dataset allows us to conduct analysis and create visualizations

Response Variable:

- Binary indicator representing match outcome (win or not).

Explanatory Variables:

- Numerical Variables:Possession, Passes, Goals Scored, Total Attempts, On-Target Attempts.
- Categorical Variable : Would a team be considered defensive?

Dataset Cleaning:

Implicit Missing Values: Identify and demonstrate the existence or absence of implicit missing values.

Handling Missing Values:

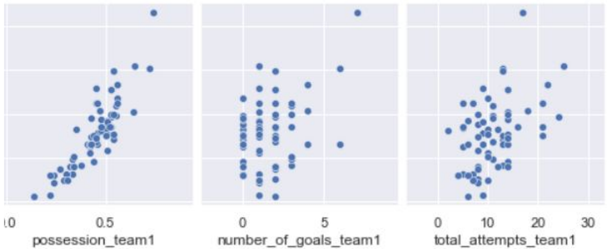
- Specify the strings representing these missing values.
- dropping rows to address missing values.

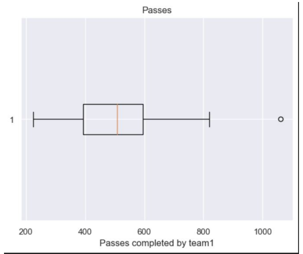
Sample Size Cleaning:

- Assess if cat. explanatory variables contain levels with few observations.

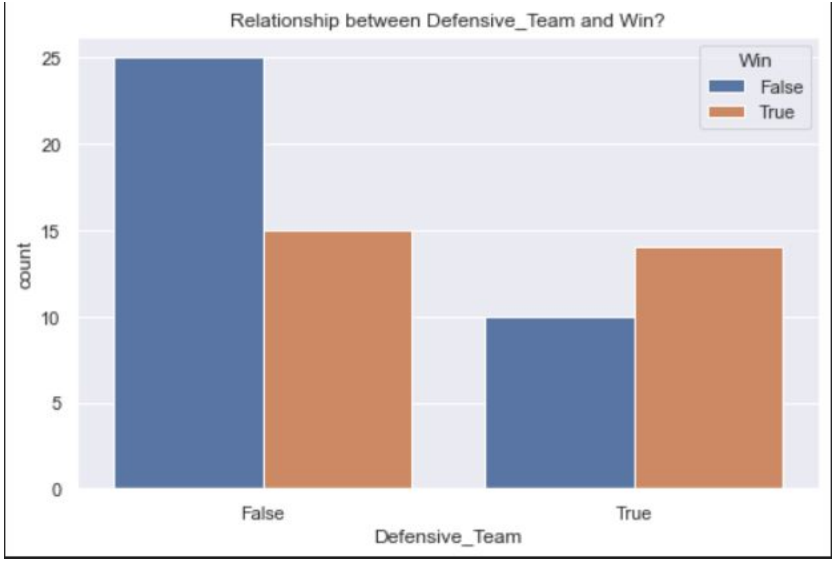
Outlier Inspection:

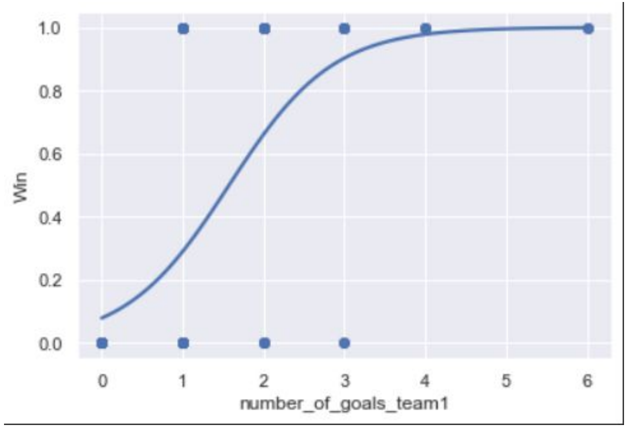
- Create scatterplots for numerical explanatory variable pairs.
- Identify and evaluate outliers

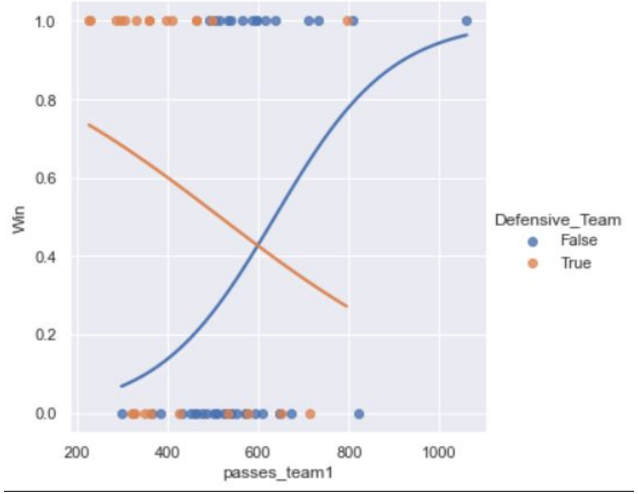




Preliminary Analysis







Model Technique Used

0/1 Variable:

- Created the indicator response variable using the ‘Win’ column

Scaling:

- Used the StandardScaler() library
- Scaled the numerical explanatory variables so that there is a focus on the secondary research goal of interpretability.

Algorithm:

- Utilized the Backwards Elimination Algorithm with k=5 cross validation

AUC Comparison

Full Model:

- Mean Test AUC Score: 0.8501587301587301

Model 1 (passes_team1 removed):

- Mean Test AUC Score: 0.8561904761904762

Model 2 (on_target_attempts_team1 removed):

- Mean Test AUC Score: 0.8834920634920636

Best Model!

Model 3 (total_attempts_team1 removed):

- Mean Test AUC Score: 0.8512698412698413

AUC Comparison Cont.

Model 4 (number_of_goals_team1 removed):

- Mean Test AUC Score: 0.643015873015873

Model 5 (possession_team1 removed):

- Mean Test AUC Score: 0.8501587301587301

Model 6 (Defensive_Team removed):

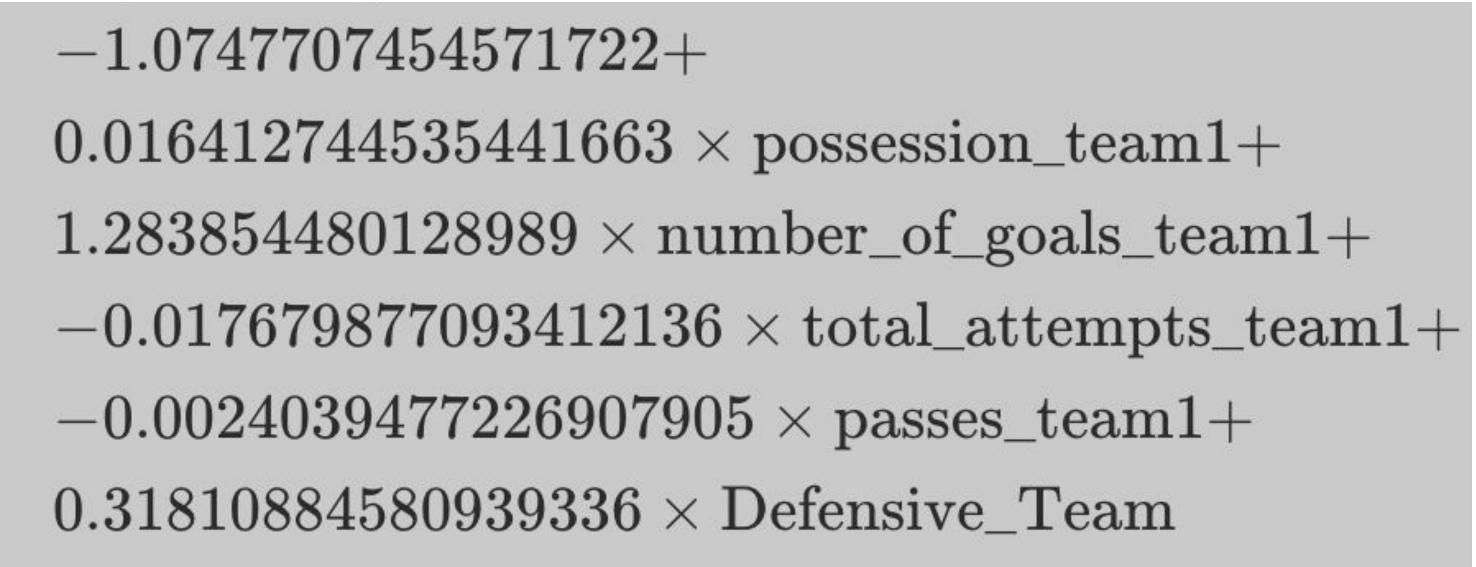
- Mean Test AUC Score: 0.8407936507936509

Comparing all AUC values, Model 2 has the one closest to 1, being 0.88

Best Logistic Regression Model Equation

$P\hat{}(Win = 1) = 1 / (1+e^{X})$

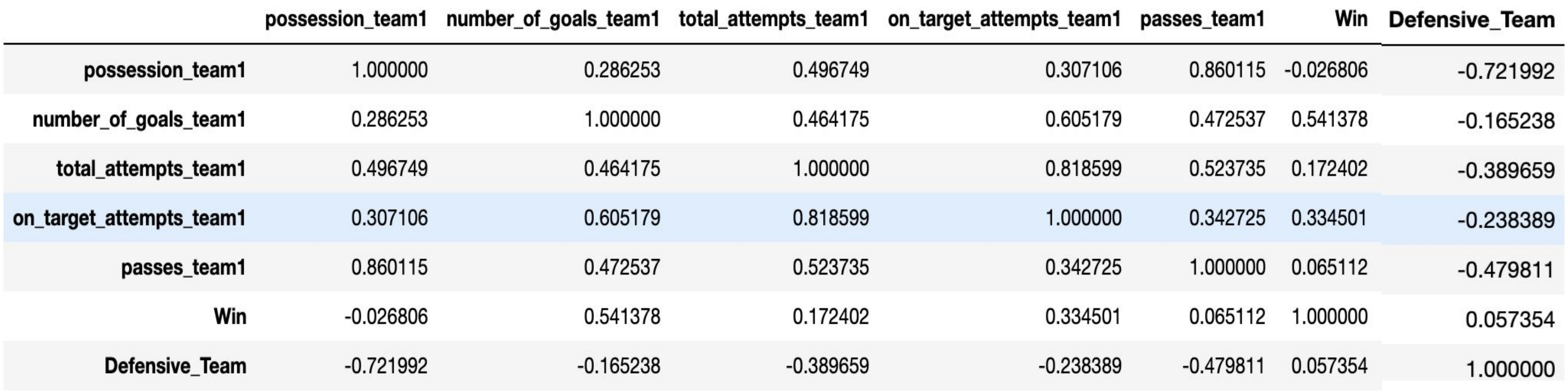
$X =$



$$\begin{aligned}
&-1.0747707454571722+\\
&0.016412744535441663 \times \text{possession\_team1}+\\
&1.283854480128989 \times \text{number\_of\_goals\_team1}+\\
&-0.017679877093412136 \times \text{total\_attempts\_team1}+\\
&-0.0024039477226907905 \times \text{passes\_team1}+\\
&0.31810884580939336 \times \text{Defensive\_Team}
\end{aligned}$$

Insights from Best Model Discussion

df.corr() table:



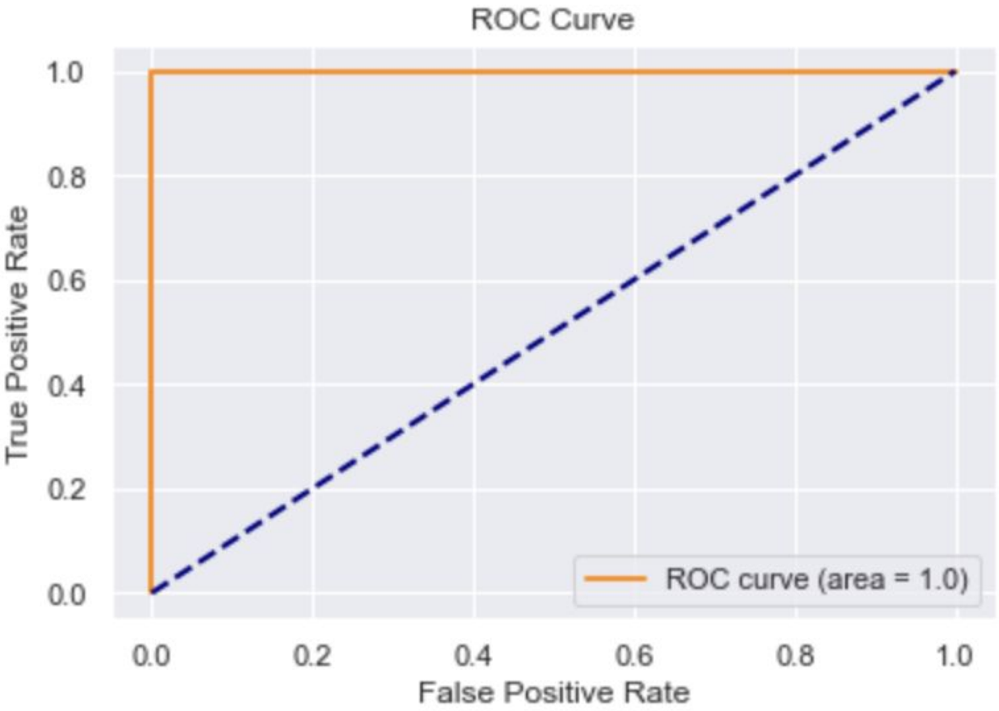
	possession_team1	number_of_goals_team1	total_attempts_team1	on_target_attempts_team1	passes_team1	Win	Defensive_Team
possession_team1	1.000000	0.286253	0.496749	0.307106	0.860115	-0.026806	-0.721992
number_of_goals_team1	0.286253	1.000000	0.464175	0.605179	0.472537	0.541378	-0.165238
total_attempts_team1	0.496749	0.464175	1.000000	0.818599	0.523735	0.172402	-0.389659
on_target_attempts_team1	0.307106	0.605179	0.818599	1.000000	0.342725	0.334501	-0.238389
passes_team1	0.860115	0.472537	0.523735	0.342725	1.000000	0.065112	-0.479811
Win	-0.026806	0.541378	0.172402	0.334501	0.065112	1.000000	0.057354
Defensive_Team	-0.721992	-0.165238	-0.389659	-0.238389	-0.479811	0.057354	1.000000

- passes_team1 + possession_team1
- on_target_attempts_team1 + total_attempts_team1

Insights from Best Model Discussion Cont.

ROC Curve:

- AUC = 1
- ROC Curve peaks at left corner, creating a right angle
- Either perfect prediction rate or overfitting variables

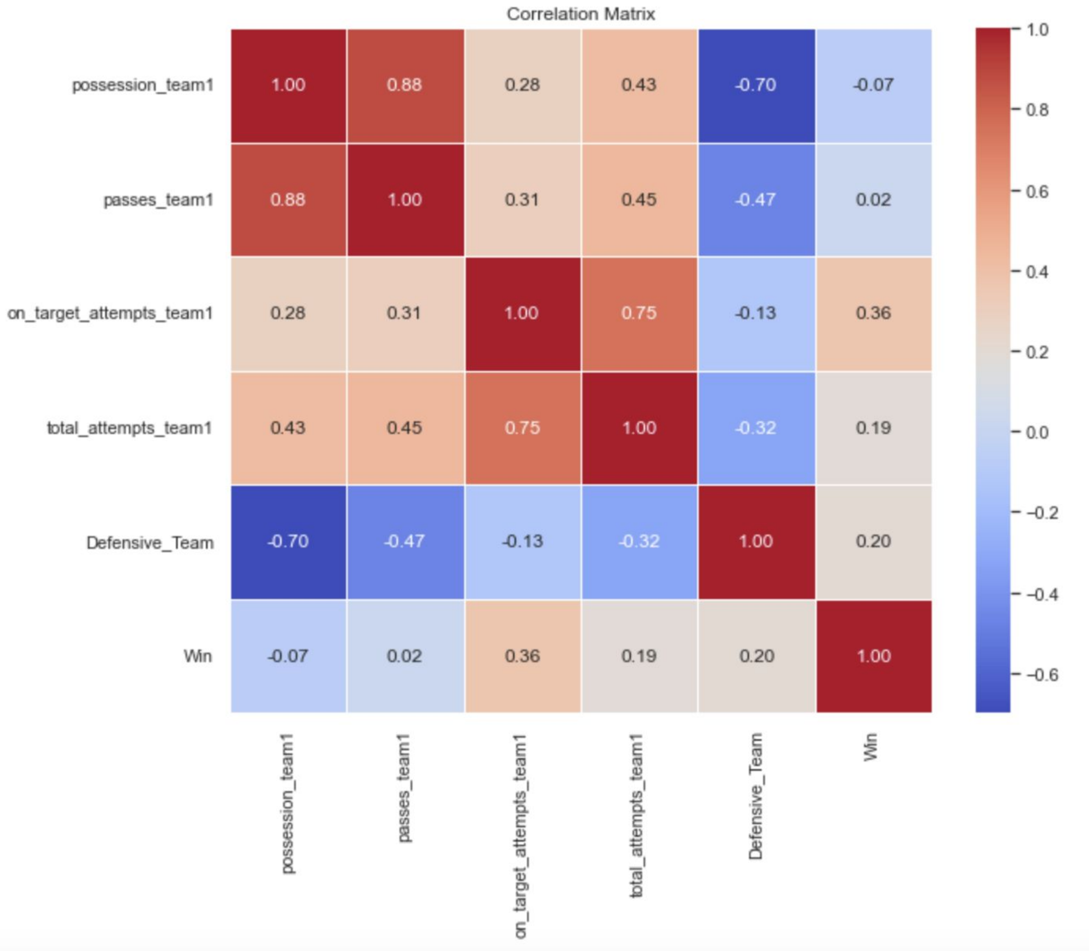


Additional Visualization: HeatMap

Defensive_Team + possesion_team1

On_target_attempts_team1 + total_attempts_team1

Passes_team1 + possession_team1



Correlation Matrix

	possession_team1	passes_team1	on_target_attempts_team1	total_attempts_team1	Defensive_Team	Win
possession_team1	1.00	0.88	0.28	0.43	-0.70	-0.07
passes_team1	0.88	1.00	0.31	0.45	-0.47	0.02
on_target_attempts_team1	0.28	0.31	1.00	0.75	-0.13	0.36
total_attempts_team1	0.43	0.45	0.75	1.00	-0.32	0.19
Defensive_Team	-0.70	-0.47	-0.13	-0.32	1.00	0.20
Win	-0.07	0.02	0.36	0.19	0.20	1.00

Conclusion + Shortcomings

- Best model w/ average test AUC of 0.88
    - Would recommend this model to sports betters
- Backward elimination doesn’t guarantee best model
    - Simpler option like LASSO Regularization
        - Help prevent overfitting as it would 0 out certain variable slopes
    - Does not capture interaction effects between variables
- Some variables were similar
    - On_target_attempts_team1 + total_attempts_team1
    - Lead to overfitting of the full model
- Future work
    - Utilize more data cleaning techniques and more accurate models in order to increase the average test AUC
    - Proceed to create a sports betting app