Machine Learning Project

PREDICTING TEAM PERFORMANCE AT THE FIFA WOMEN'S WORLD CUP

BY CAMÉLIA HELAL

Table of Contents

Introduction

- Objective : Build a model that can estimate how far a team will progress in the tournament
- Bases on statistics : FIFA ranking, goals scored or conceded, possession, and other match features
- Machine Learning : EDA → Preprocessing → Models → Evaluation → Decision Support

I - Dataset Selection

DATASET

Women's World Cup statistics with 32 teams (synthetic .mat file)

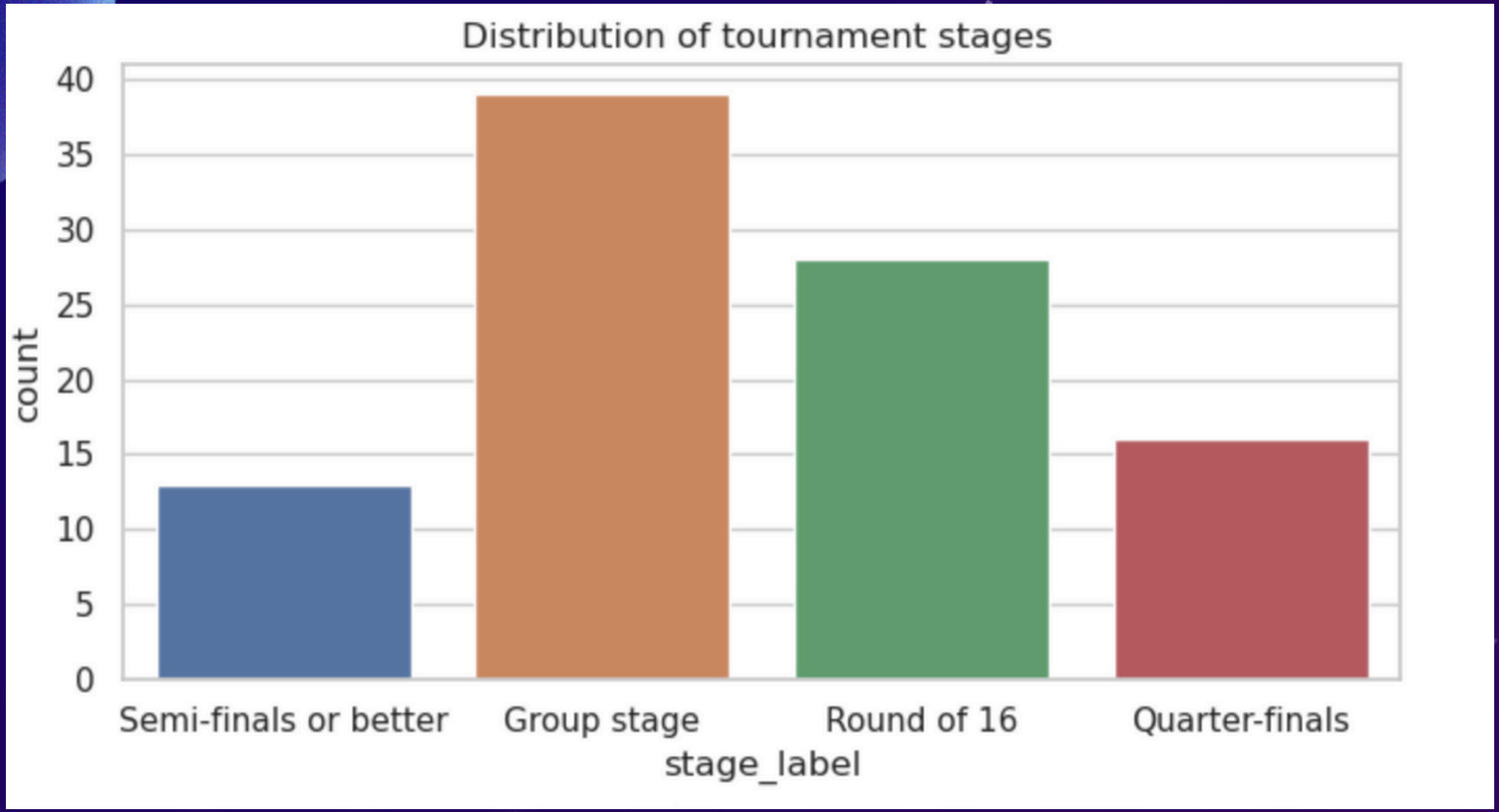
TARGET

Stage reached (Group, Round of 16, Quarterfinal, Semifinal, Final, Winner)

FEATURES

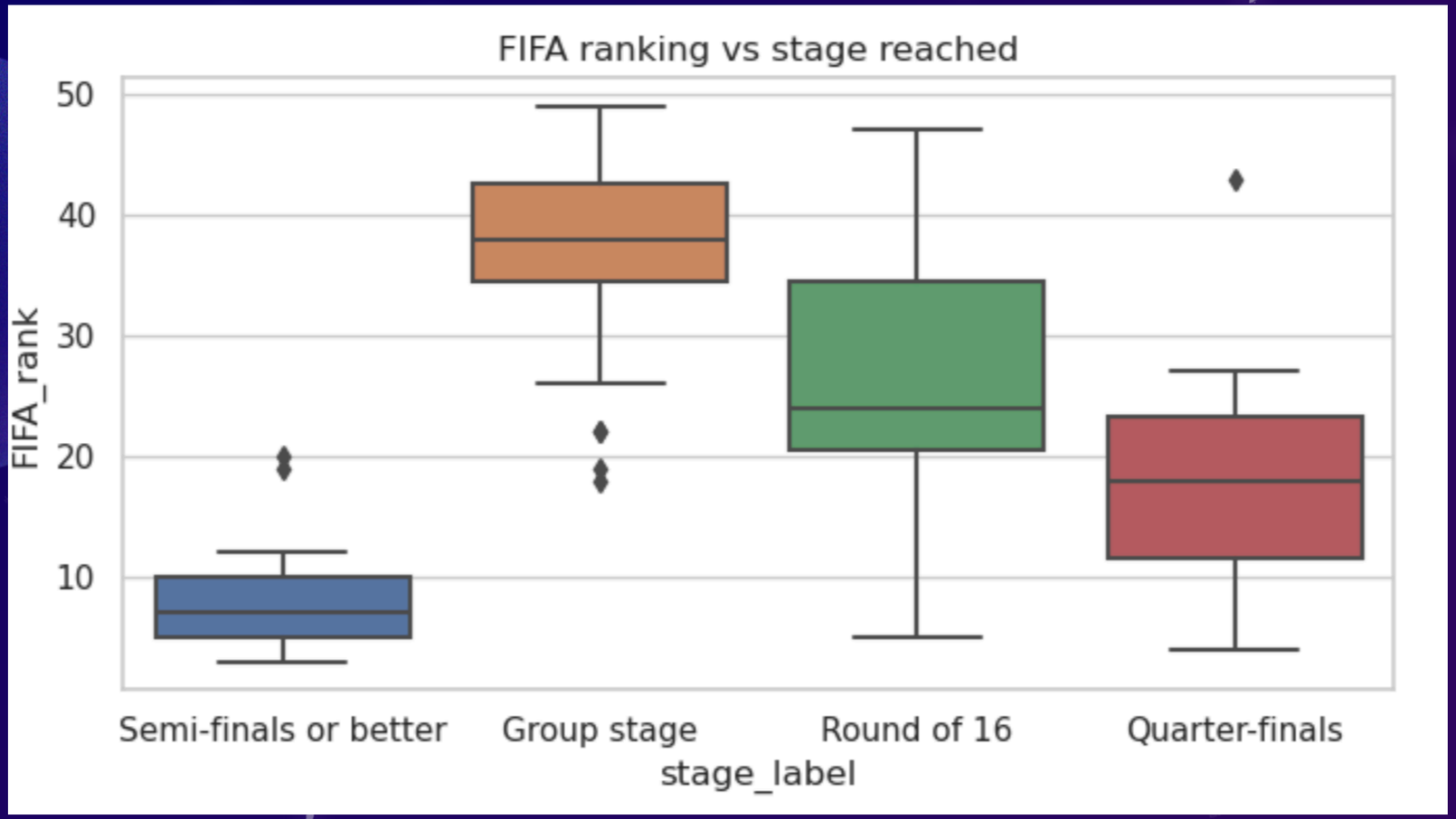
	FIFA_rank	avg_goals_scored	avg_goals_conceded	total_shots_per_game	possession_pct	pass_accuracy_pct	host_nation	confed_code	stage_reached
0	5.0	2.21	0.00	8.8	52.3	72.8	0.0	2.0	3
1	39.0	1.76	1.36	10.9	42.3	80.4	0.0	1.0	0
2	33.0	2.08	0.71	9.5	46.6	77.3	1.0	1.0	1
3	22.0	1.65	0.95	12.6	54.5	68.8	1.0	4.0	1
4	22.0	1.80	0.83	6.0	61.2	82.6	0.0	3.0	1

II - Exploratory Data Analysis & Preprocessing



Imbalance visible : more teams eliminated in group stage

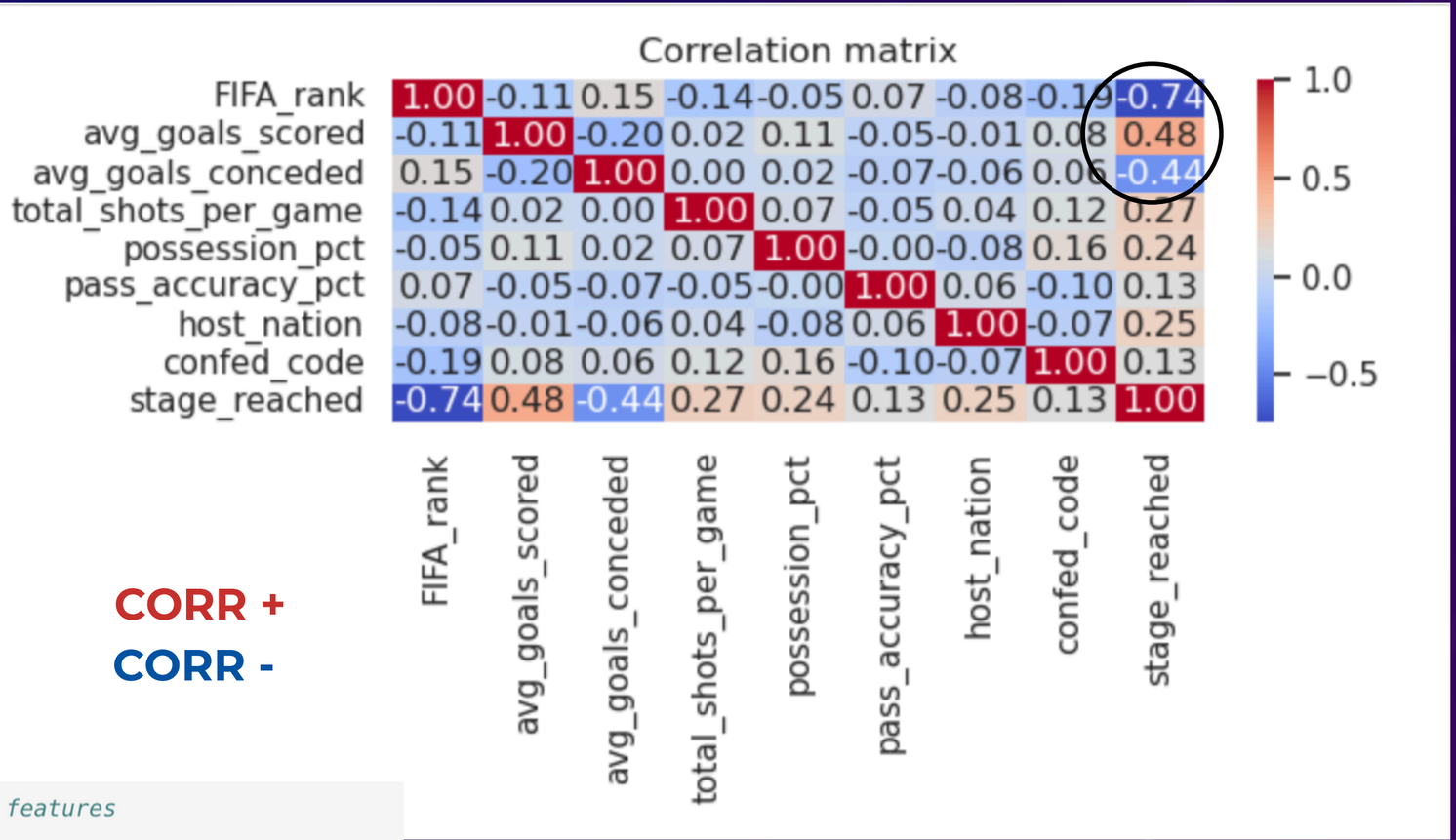
```python
# Distribution of the target variable
stage_mapping = {
    0: "Group stage",
    1: "Round of 16",
    2: "Quarter-finals",
    3: "Semi-finals or better"}

df["stage_label"] = df["stage_reached"].map(stage_mapping)
stage_counts = df["stage_label"].value_counts().sort_index()
stage_counts

# Plot the distribution of the tournament stages
sns.countplot(x="stage_label", data=df)
plt.title("Distribution of tournament stages ")
plt.tight_layout()
plt.show()
```



Lower FIFA ranking numbers
=> reach higher stages

Negative correlation

```
# Relationship between FIFA ranking and stage reached
sns.boxplot(x="stage_label", y="FIFA_rank", data=df)
plt.title("FIFA ranking vs stage reached")
plt.tight_layout()
plt.show()
```



CORR +

CORR -

```python
# Correlation matrix between features
num_cols = feature_names
corr = df[num_cols.tolist() + ["stage_reached"]].corr()

sns.heatmap(corr, annot=True, fmt=".2f", cmap="coolwarm")
plt.title("Correlation matrix")
plt.tight_layout()
plt.show()
```

To succeed, a team's FIFA ranking matters most, followed by scoring goals and defending well

Preprocessing & train / test split

TRAINING SET 60%

→ to train the model

VALIDATION SET 20%

→ to adjust the hyperparameters

TEST SET 20%

→ to evaluate the final performance

Standardization of X values :

→ Same scale for all variables
→ Faster converging
→ No variable dominates

```
Training set size : (57, 8)
Validation set size : (19, 8)
Test set size : (20, 8)
```

WHY WE DO THIS ?

III - Model Training & Comparison

Parameters

- model : Classifier with fit / predict methods
- X_tr : Training features
- y_tr : Training labels
- X_val : Validation features
- y_val : Validation labels model_name

	Logic Regression (multinomial)	**Random Forest** (ensemble of decision trees)	**k-Nearest Neighbors** (k=7)
Accuracy	0.789	0.579	0.632
F1-score	0.764	0.419	0.601

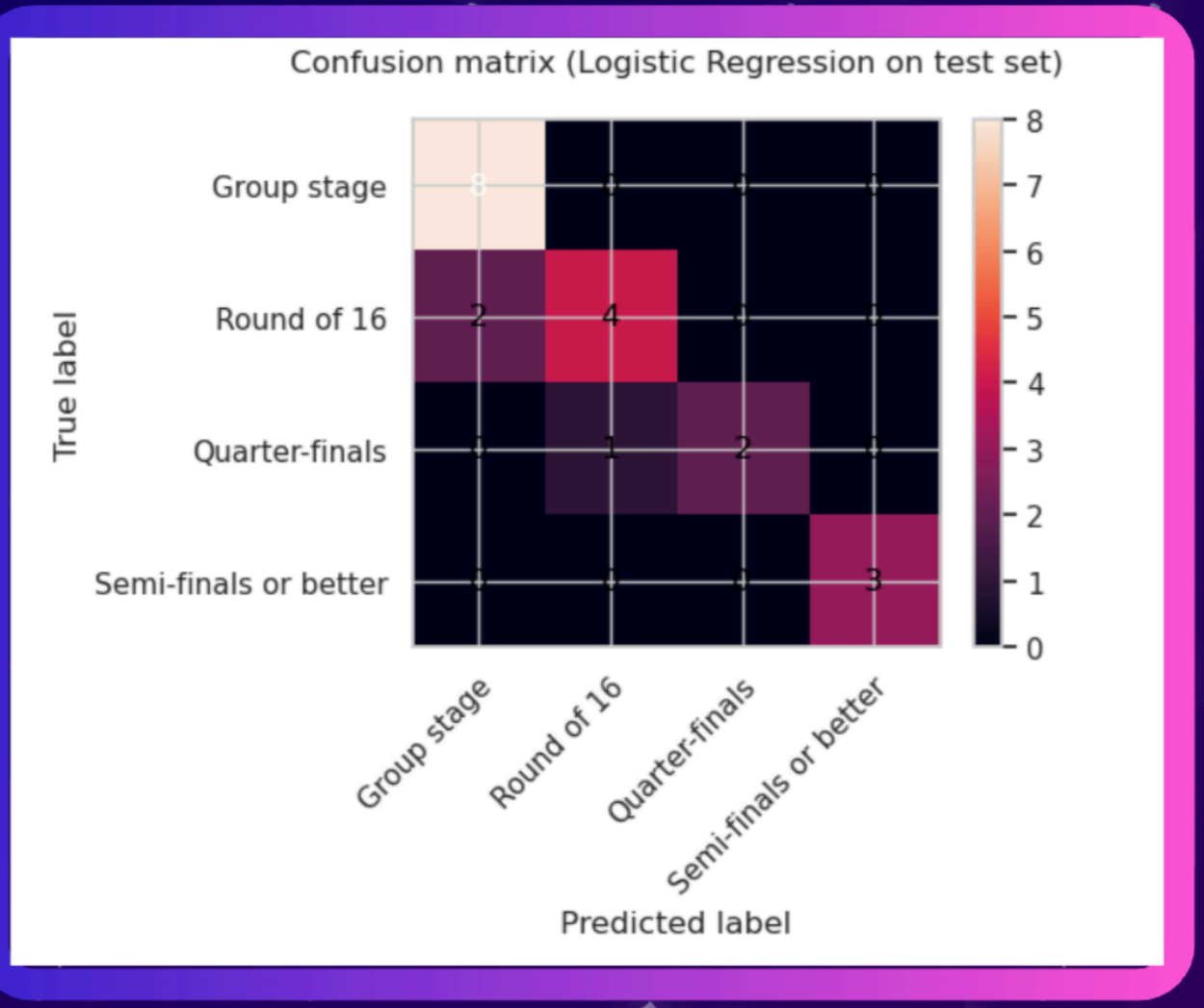
IV - Model Evaluation & Visualization

Best model according to validation F1-score : Logistic Regression

Test set performance :

	precision	recall	f1-score	support
0	0.80	1.00	0.89	8
1	0.80	0.67	0.73	6
2	1.00	0.67	0.80	3
3	1.00	1.00	1.00	3
accuracy			0.85	20
macro avg	0.90	0.83	0.85	20
weighted avg	0.86	0.85	0.84	20

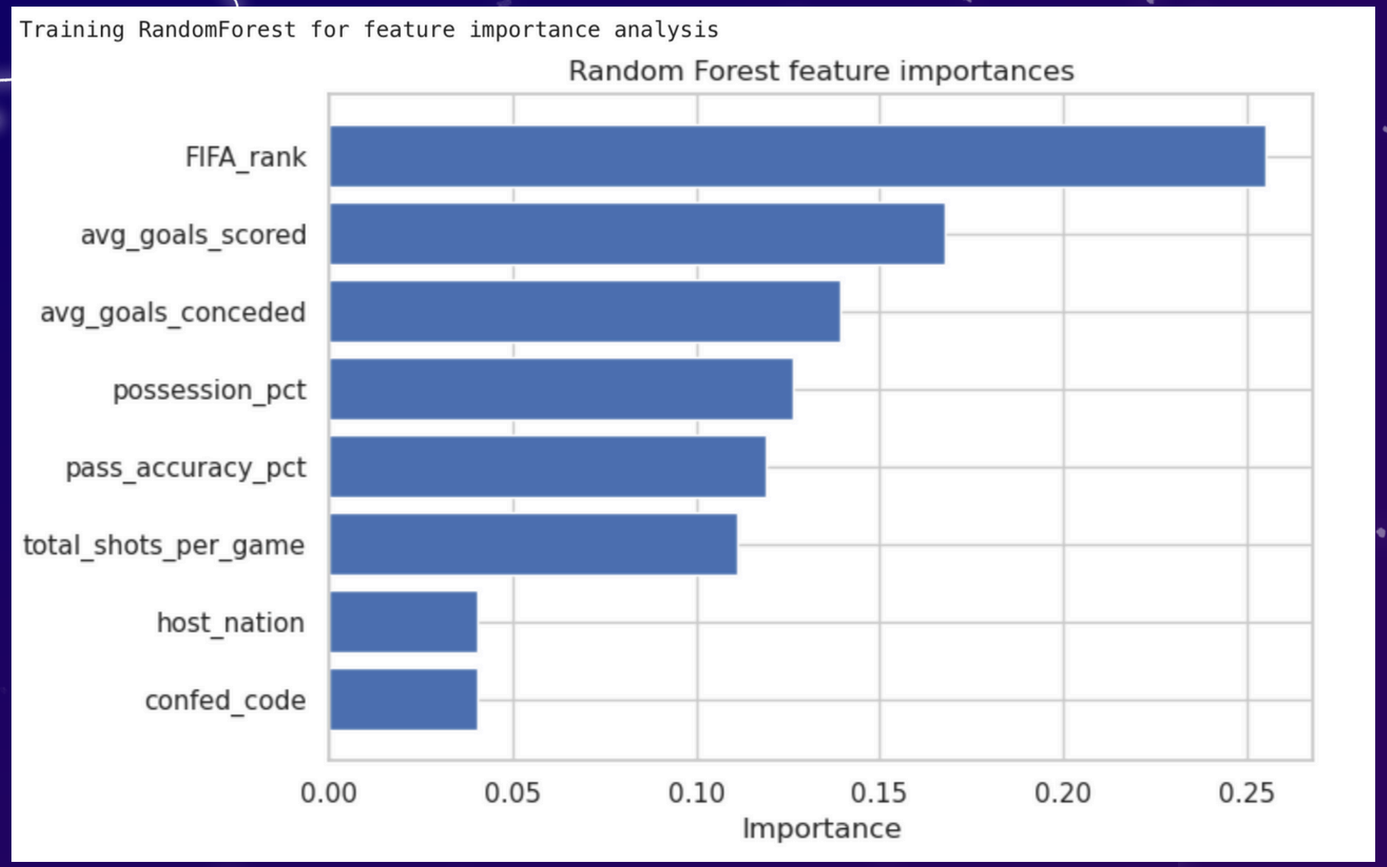
We select the model with the best validation F1-score and evaluate it on the held-out test set, which simulates new, World Cup tournaments.



We can see that the predictions are mostly correct, but adjacent stages are often confused, which makes sense because their statistical differences are small.

```python
# Confusion matrix on the test set
class_names = list(stage_mapping.values())
plot_confusion_matrix(
    y_test, y_pred,
    class_names=class_names,
    title=f"Confusion matrix ({best_name} on test set)\n")
```

V - Model Interpretation

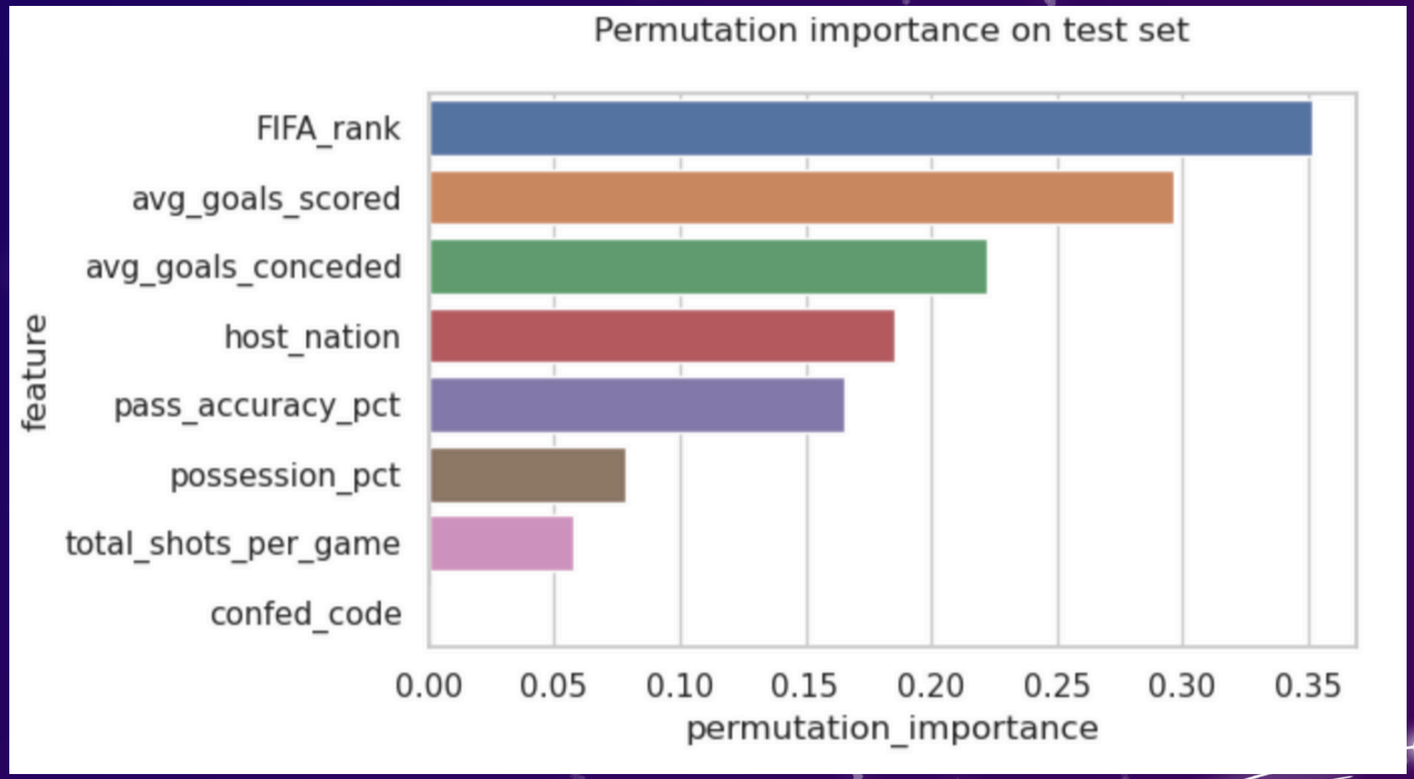


- FIFA ranking → most important feature
- Goals scored → second most important
- Goals conceded → third most important
- Other stats → very little impact
- Host nation advantage → negligible

FEATURE IMPORTANCES

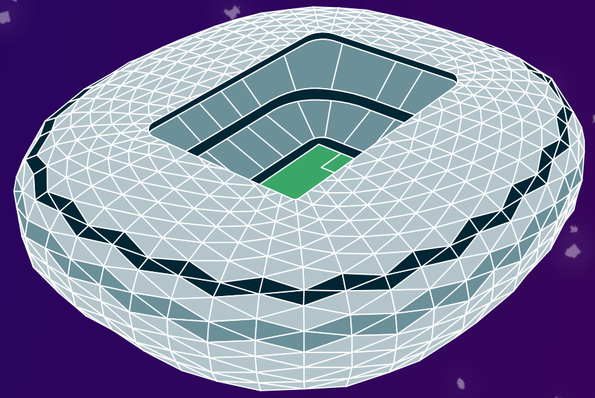
Show how each feature pushes a prediction towards a higher or lower tournament stage

	feature	permutation_importance
0	FIFA_rank	0.351750
1	avg_goals_scored	0.296554
2	avg_goals_conceded	0.222007
6	host_nation	0.185534
5	pass_accuracy_pct	0.165874
4	possession_pct	0.078727
3	total_shots_per_game	0.057789
7	confed_code	0.000000



PERMUTATION IMPORTANCES

VI - Decision Support



How can a federation use these results ?

- FIFA ranking and offensive power matter a lot
- Defensive solidity is also important
- Ball possession and passing accuracy
- Host nation advantage

Conclusion

Before the World Cup draw, the staff could :

- Estimate the current features of their team (goals, possession, etc.)

- Use the model to compute the probability of each World Cup stage

- Decide on priorities for player selection, focusing on the features that the model ranks as most important

THANK YOU

Any questions ?

BY CAMÉLIA HELAL