Goal Oriented Forecasting

MAS Thesis Presentation

Sheng (Tim) Chen

March 5, 2024

Introduction

A professional soccer match comprises two teams, each consisting of 11 players, engaging in a 90-minute game. Teams are categorized into leagues based on the teams' level of professionalism and their respective countries. The league structures vary across countries, typically featuring 16 to 20 teams.

Key Questions

1. Which (team) feature plays the most important role in goal differences?
2. Can we predict future game's goal difference?

Prior Work

1. StatsBomb **Expected Goals (xG)** is a soccer metric assessing the probability of a shot resulting in a goal
   - assigning values based on distance, angle, and type of assist.

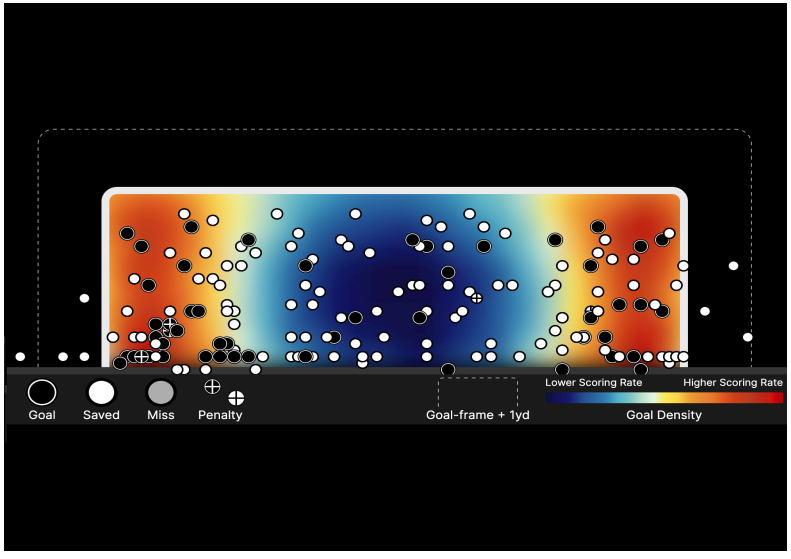


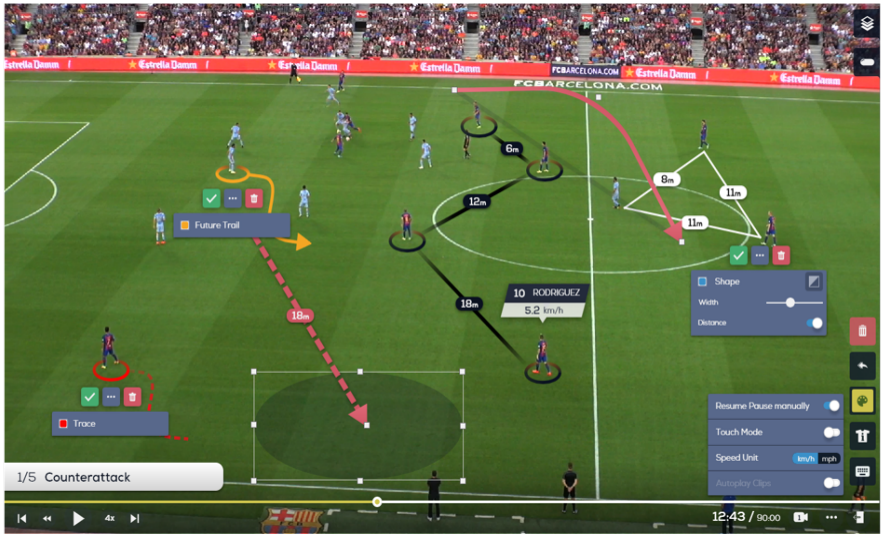
Figure: XG assigned to Penalty Shots

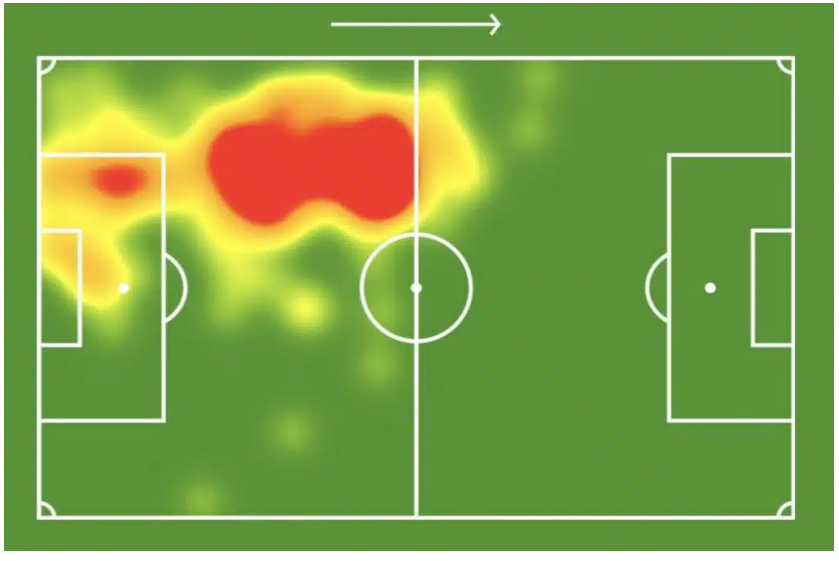
2. GNN for soccer analytics, relying on exact player positions as input, help detecting (Semi-)automating Patterns [Bauer and Anzer, 2021]

Background

Common Types of Raw Data in Sport Analytics

1. Visual data (image, video) containing game scenes.
   - Rich Information with hierarchical representations
   - Multimodal Analysis
2. Spatial-Temporal data (heatmaps and passing maps)
3. Textual data (commentary and news report)
4. Tabular data containing team performance and match outcomes.
   - Scalable, Comprehensible, Queryable





Data Retrieval

Manchester City Player Stats Glossary

Summary | Passing | Pass Types | Defensive Actions | Possession | Miscellaneous Stats

						Touches							Take-Ons				
Player	#	Nation	Pos	Age	Min	Touches	Def Pen	Def 3rd	Mid 3rd	Att 3rd	Att Pen	Live	Att	Succ	Succ%	Tkld	Tkld%
Erling Haaland	9	NOR	FW	23-043	90	17	2	2	6	9	4	16	1	0	0.0	1	100.0
Phil Foden	47	ENG	LW,RW	23-097	89	51	0	5	27	19	5	51	4	2	50.0	2	50.0
Oscar Bobb	52	NOR	RW	20-052	1	4	0	0	1	3	1	4	2	2	100.0	0	0.0
Jeremy Doku	11	BEL	RW	21-098	75	36	0	2	19	16	0	36	4	2	50.0	1	25.0
Sergio Gómez	21	ESP	LW	22-363	15	11	0	1	4	6	1	11	2	2	100.0	0	0.0
Julián Álvarez	19	ARG	AM	23-214	89	41	0	0	23	19	2	41	1	1	100.0	0	0.0
Rico Lewis	82	ENG	AM	18-285	1	4	0	0	1	3	1	4	0	0		0	
Mateo Kovačić	8	CRO	DM	29-119	83	58	2	6	39	14	0	58	0	0		0	
Kalvin Phillips	4	ENG	DM	27-274	7	14	1	1	11	2	0	14	1	1	100.0	0	0.0
Rodri	16	ESP	DM	27-072	75	97	3	19	75	4	0	97	1	1	100.0	0	0.0
Bernardo Silva	20	POR	DM	29-023	15	22	0	2	13	7	0	22	0	0		0	
Nathan Aké	6	NED	LB	28-196	90	101	2	33	58	10	2	101	0	0		0	
Manuel Akanji	25	SUI	CB	28-045	90	114	7	29	81	5	2	114	0	0		0	
Rúben Dias	3	POR	CB	26-111	90	112	8	53	58	1	0	112	0	0		0	
Kyle Walker	2	ENG	RB	33-097	90	109	4	22	67	20	2	109	1	1	100.0	0	0.0
Ederson	31	BRA	GK	30-016	90	47	35	47	0	0	0	47	0	0		0	
16 Players					**990**	**838**	**64**	**222**	**483**	**138**	**20**	**837**	**17**	**12**	**70.6**	**4**	**23.5**

Figure: Example of Data Collected from fbref.com

Data Matrix

$$\mathbf{X} = \begin{array}{c} \\ \textit{Game}1 \\ G_2 \\ G_3 \\ \vdots \\ G_{n-1} \\ G_n \end{array} \begin{array}{c} \begin{array}{ccccccc} \textit{Touch} & \textit{Block} & \textit{Dist} & & \textit{Tac} & \textit{RedC} \end{array} \\ \begin{bmatrix} 801 & 6 & 1472.8 & \dots & 23 & 0 \\ 566 & 15 & 893.5 & \dots & 19 & 0 \\ 561 & 9 & 1592.7 & \dots & 24 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 395 & 5 & 801.4 & \dots & 19 & 0 \\ 726 & 11 & 1630.2 & \dots & 10 & 0 \end{bmatrix} \end{array}, \quad \mathbf{y} = \begin{array}{c} \textit{Goal Diff} \\ \begin{bmatrix} 4 \\ -2 \\ 1 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \end{array}$$

Data Matrix $\mathbf{X} \in \mathbb{R}^{8975 \times 234}$, Outcome Variable $\mathbf{y} \in \mathbb{R}^{234}$

Explanatory modeling

Explanatory modeling used to gain a profound understanding of which features within team statistics contribute significantly to the observed patterns and behaviors on the field.

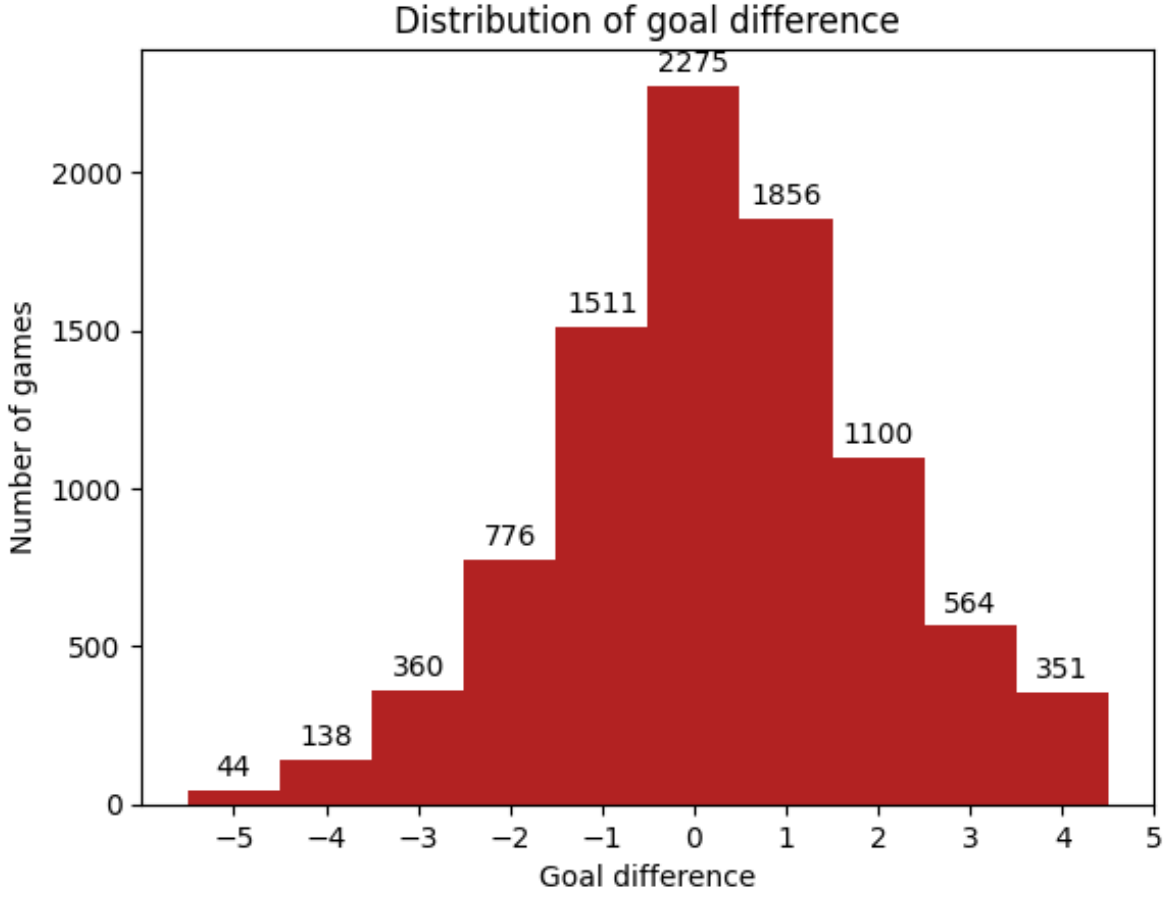
- **GLM Coefficient**
  Change in the linear predictor (e.g., log-odds) for a one-unit change in the corresponding outcome variable
- **Random Forrest**
  How much each feature contributes to the overall decrease in impurity across all the trees

Mean Decrease in Impurity

$$g(f_k) = \frac{1}{N_t} \sum_{t=1}^{N_t} \sum_{i \in \text{nodes in tree } t} \mathbb{1}(\text{f}_k \text{ is used at node}_i) \times \Delta Gini(i)$$

Imbalanced Classification Problem

Imbalanced Classification Problem usually occurs when one class is more expensive in time, cost, computation, or other resources.



- Oversample minority class
- Downsample majority class
- Weighted loss function

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE generates synthetic examples by interpolating between neighboring instances of the minority class. [Chawla et al., 2002]

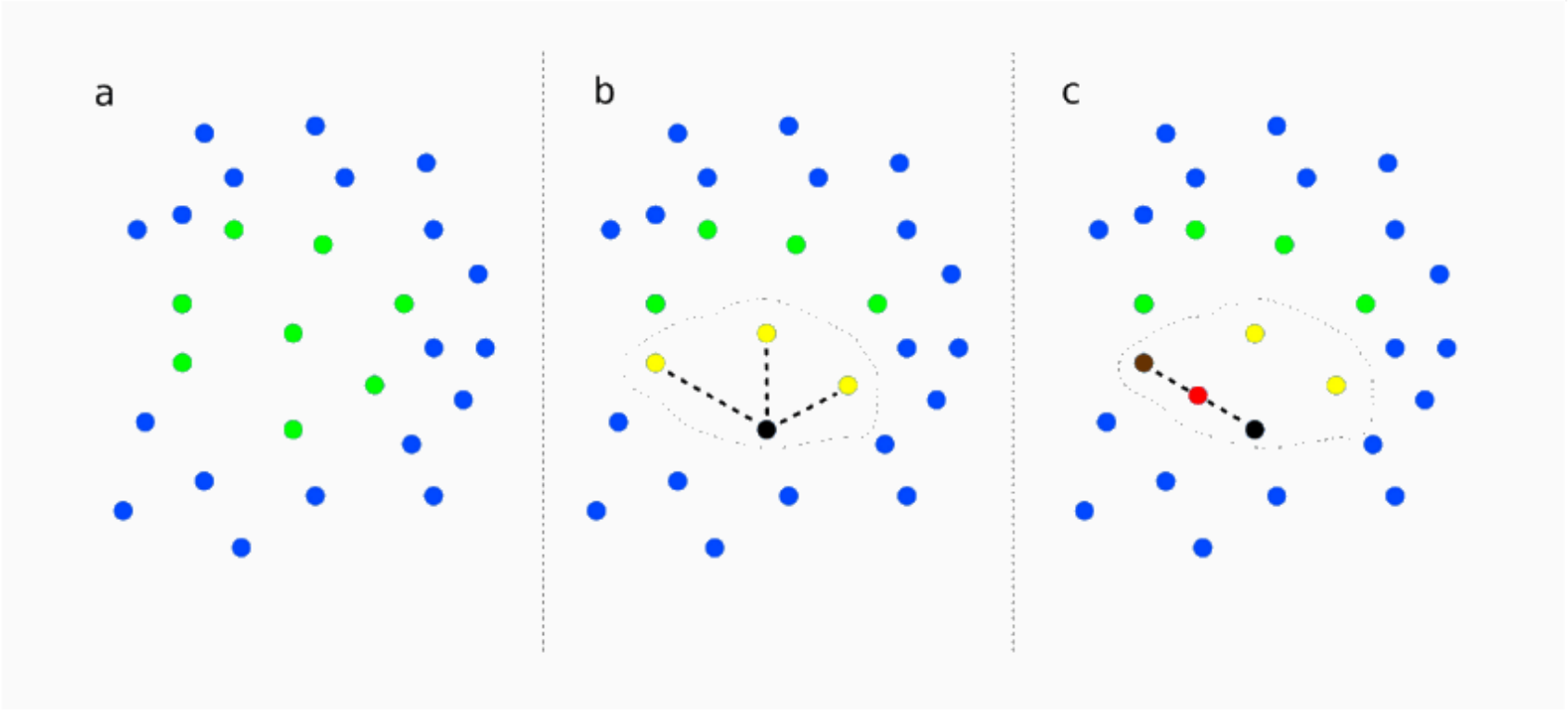


Figure: Example of generating new points using SMOTE

Tabular GAN(TGAN)

TGAN is a generative adversarial network designed specifically for generating tabular data [Xu and Veeramachaneni, 2018]

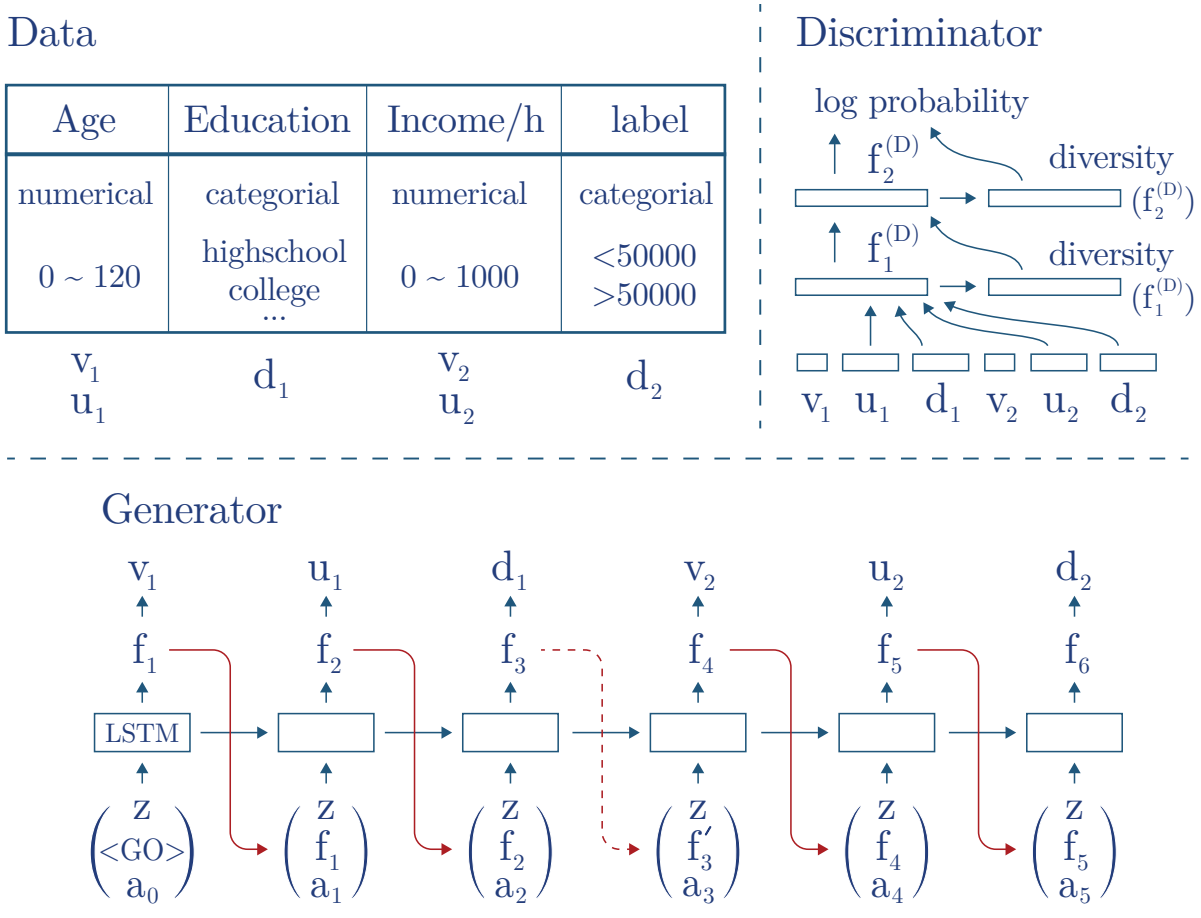


Figure: Example of using Tab GAN to generate a simple census table.

Feature Selection

- Filter
  - Information Gain
  - Correlation Coefficient
- Wrapper
  - Sequential Selection
- Embedded
  - L1 Lasso
  - Decision Tree

Mutual Information

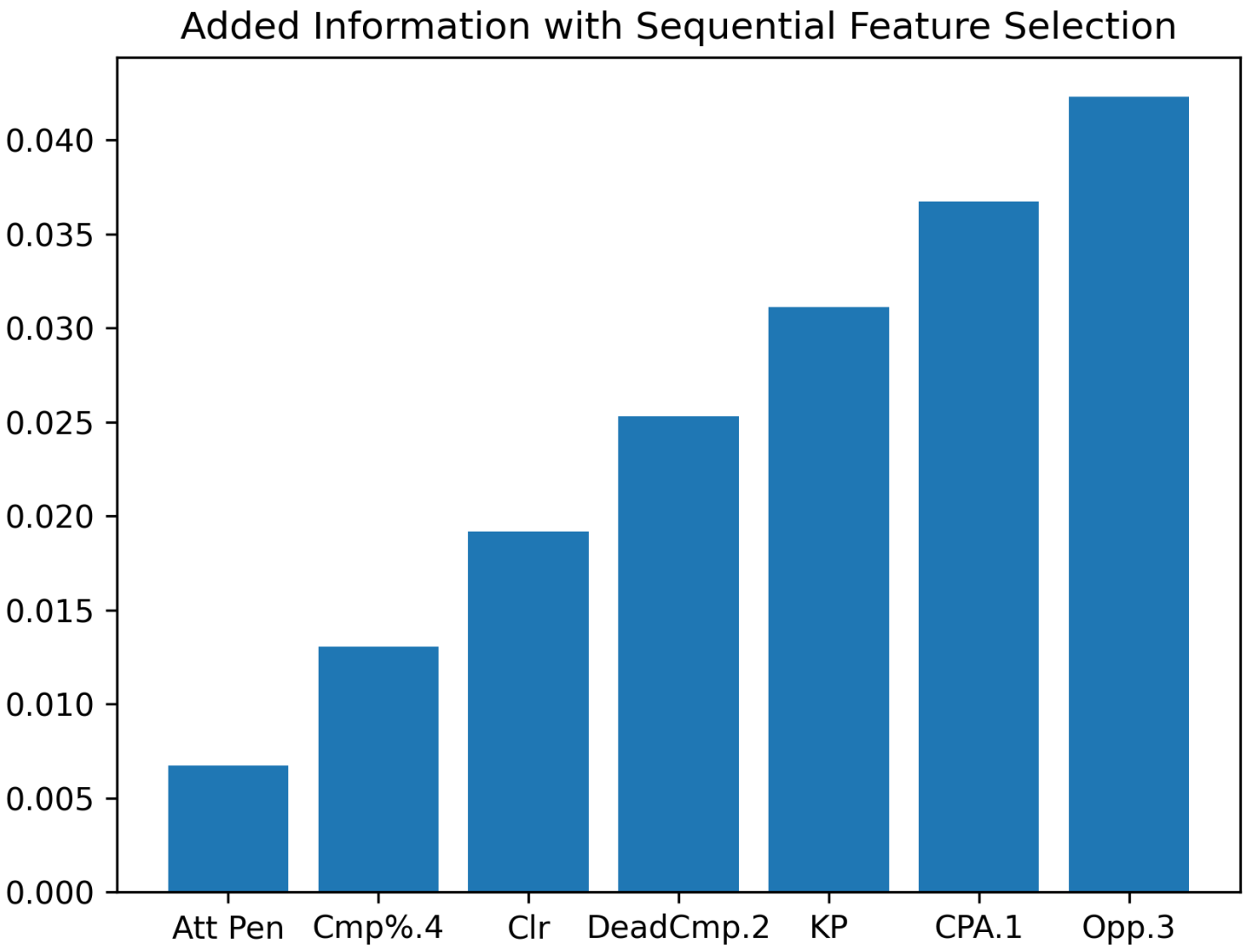
Mutual Information

Mutual Information (MI) measures the amount of information shared between two random variables. Defined as:

$$
\begin{aligned}
I(X;Y) &= D_{\mathrm{KL}}(P_{(X,Y)} \,||\, P_X \cdot P_Y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} P_{(X,Y)}(x,y) \log \left(\frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)} \right) dxdy
\end{aligned}
$$

Since MI quantifies the degree to which the knowledge of the value of one variable reduces uncertainty about the other, we can compare added Mutual Information to decide on which features to add in a sequential search.

Mutual Information



Selected Features

Feature id	Meaning
Att Pen	Touches in Attacking Penalty Area
Cmp%.4	Forward Pass Completion Rate
Clr	Clearance
DeadCmp.2	Dead Ball Completed
KP	Key Passes
CPA.1	Carries into Penalty Area
Opp.3	Attempted Crosses

Table: Glossary of Feature Selected by added MI

Feature id	Meaning
Blocks.2	Number of times standing in the ball path
TkIW.2	Tackles Won

Table: Glossary of Additional Features Selected by Lasso

Prediction

- We converted matrix using a weighted sum of previous k games data from each team, and use the previous model to predict on goal differences.
- Naive Approach with flaws
  - Opponent Strength
  - Assumes Linearity

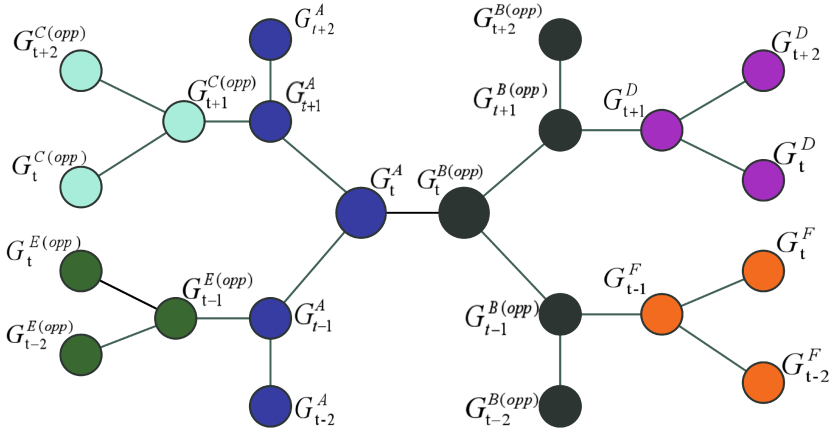
Graph Neural Network Representation(GNN)

Convert to a node classification task using Graph Convolution Networks. Node h_i has Goal Differences as label and contains feature information. Edges between the nodes represent recent games related to the teams and the current opponent of the current game.

Convolutional Layer$^{(l+1)}$ has the following update:

$$h_i^{(l+1)} = \sigma\Big(\sum_{j \in N_i} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)}\Big)$$

Weight Matrix $W \in \mathbb{R}^{100 \times 7}$, $h_i \in \mathbb{R}^7$, c_{ij} normalizing constant, and N_i denotes neighboring nodes with node i.



Result

Step1	Step2	Method	Accuaracy
/	MI	Logistic Regression	0.2380
SMOTE	MI	Logistic Regression	0.2093
SMOTE	MI	KNN	0.1567
SMOTE	MI	XGB	0.2450
MI	SMOTE	XGB	0.2344
MI	TGAN	XGB	0.2621
TGAN	MI	XGB	0.2767
TGAN	MI	GNN	0.2890
		Betting Agency	0.2734

Table: Model Result with Combination of Methods

Future Work

Tracking Data Analysis combines high-resolution videos and computer vision techniques, enhancing precision in player movement and object tracking.

Figure: Challenge Posed by Tracking: Player movements during Corner

References


G. Anzer, P. Bauer, U. Brefeld, and D. Faßmeyer.
Detection of tactical patterns using semi-supervised graph neural networks.
03 2022.


N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer.
Smote: Synthetic minority over-sampling technique.
Journal of Artificial Intelligence Research, 16:321–357, June 2002.


L. F. Kozachenko and N. N. Leonenko.
Sample estimate of the entropy of a random vector".
Probl. Peredachi Inf, 23(2):95–101, 1987.


A. Kraskov, H. Stögbauer, and P. Grassberger.
Estimating mutual information.
Phys. Rev. E, 69:066138, Jun 2004.


P. Xenopoulos and C. Silva.
Graph neural networks to predict sports outcomes.
In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1757–1763, 2021.