Análise Preditiva de Partidas de Futebol: Modelagem da Diferença de Gols

Relatório Completo para Apresentação

1. Introdução e Objetivo

Este projeto visa desenvolver um modelo preditivo para prever a diferença de gols em partidas de futebol (time mandante - time visitante) utilizando estatísticas de jogo como posse de bola, chutes a gol e formações táticas. O objetivo principal é explorar diferentes algoritmos de regressão para entender quais fatores têm maior influência no resultado das partidas.

2. Conjunto de Dados e Preparação

Base de dados: A análise utiliza um dataset campeonatos contendo estatísticas detalhadas de 4.161 partidas de futebol.

Variável alvo:

- Diferenca_Gols = Gols do Time 1 – Gols do Time 2
- Resultado categórico: 0 (vitória Time 2), 1 (empate), 2 (vitória Time 1)

Características do dataset:

- A diferença de gols apresenta média de 0.34
- Distribuição assimétrica positiva (favorecendo o time mandante)
- 1.886 vitórias do Time 1 (45.3%)
- 994 empates (23.9%)
- 1.281 vitórias do Time 2 (30.8%)

3. Engenharia de Features

Features derivadas criadas:

- *Diff_Chutes_Gol: Diferença em chutes a gol*
- *Diff_Posse: Diferença na posse de bola (%)*
- *Diff_Escanteios: Diferença em escanteios*
- *Diff_Faltas: Diferença em faltas*
- *Eficiencia_Ofensiva: Razão entre chutes a gol e total de chutes para cada equipe*
- *Diff_Eficiencia_Ofensiva: Diferença na eficiência ofensiva*

Tratamento de variáveis categóricas:

- *Codificação one-hot para formações táticas (Position 1, Position 2)*
- *Resultou em 74 features totais após processamento*

4. Metodologia

Pré-processamento:

Divisão treino-teste (80%-20%)

Escalonamento das variáveis numéricas (StandardScaler)

Seleção de features (SelectKBest com f_regression)

Features mais relevantes identificadas:

Diferença de chutes a gol (Diff_Chutes_Gol)

Chutes a gol do Time 1

Diferença na eficiência ofensiva

Chutes a gol do Time 2

Eficiência ofensiva dos times

Modelos implementados:

Regressão Linear

Ridge Regression

Lasso Regression

ElasticNet

Support Vector Regression (SVR)

Random Forest

Gradient Boosting

5. Resultados e Comparação de Modelos

Comparação de modelos (ordenado por R^2):

	r2	rmse	mae	evs
ElasticNet	0.237622	1.434844	1.151170	0.237638
Lasso	0.235444	1.436892	1.151208	0.235508
Ridge	0.232638	1.439527	1.156827	0.232645
Regressão Linear	0.232632	1.439533	1.156839	0.232639
Gradient Boosting	0.192352	1.476831	1.177913	0.192430
Random Forest	0.183225	1.485152	1.187725	0.183417
SVR	0.080803	1.575520	1.263256	0.080837

6. Análise dos Resultados

Principais observações:

- *Modelos lineares regularizados (ElasticNet e Lasso) apresentaram melhor performance*
- *O melhor modelo (ElasticNet) conseguiu explicar apenas ~24% da variância na diferença de gols*
- *Os erros médios absolutos ficaram em torno de 1.15 gols*
- *SVR teve desempenho significativamente inferior aos demais modelos*

Limitações identificadas:

- *R^2 relativamente baixo indica que há fatores importantes não capturados pelo modelo*
- *Alto valor de MAPE (~82%) sugere dificuldade na predição precisa da diferença de gols*

7. Conclusões e Próximos Passos

Conclusões:

- *A diferença de chutes a gol é o preditor mais importante para a diferença no placar*
- *Modelos com regularização apresentaram melhor desempenho, sugerindo a presença de multicolinearidade*
- *Prever o resultado exato de partidas de futebol é desafiador (R^2 de 0.24)*

Próximos passos:

- *Explorar features adicionais (histórico de confrontos, forma recente, fatores climáticos)*
- *Testar abordagens de ensemble e modelos mais complexos*
- *Considerar modelos específicos para previsão categórica (vitória/empate/derrota)*
- *Implementar análise temporal para capturar tendências de desempenho das equipes*

8. Implementação Técnica

O projeto foi desenvolvido em Python utilizando bibliotecas como scikit-learn 1.6.1, pandas, numpy e seaborn. A avaliação dos modelos utilizou validação cruzada e métricas diversificadas para garantir análise robusta dos resultados.