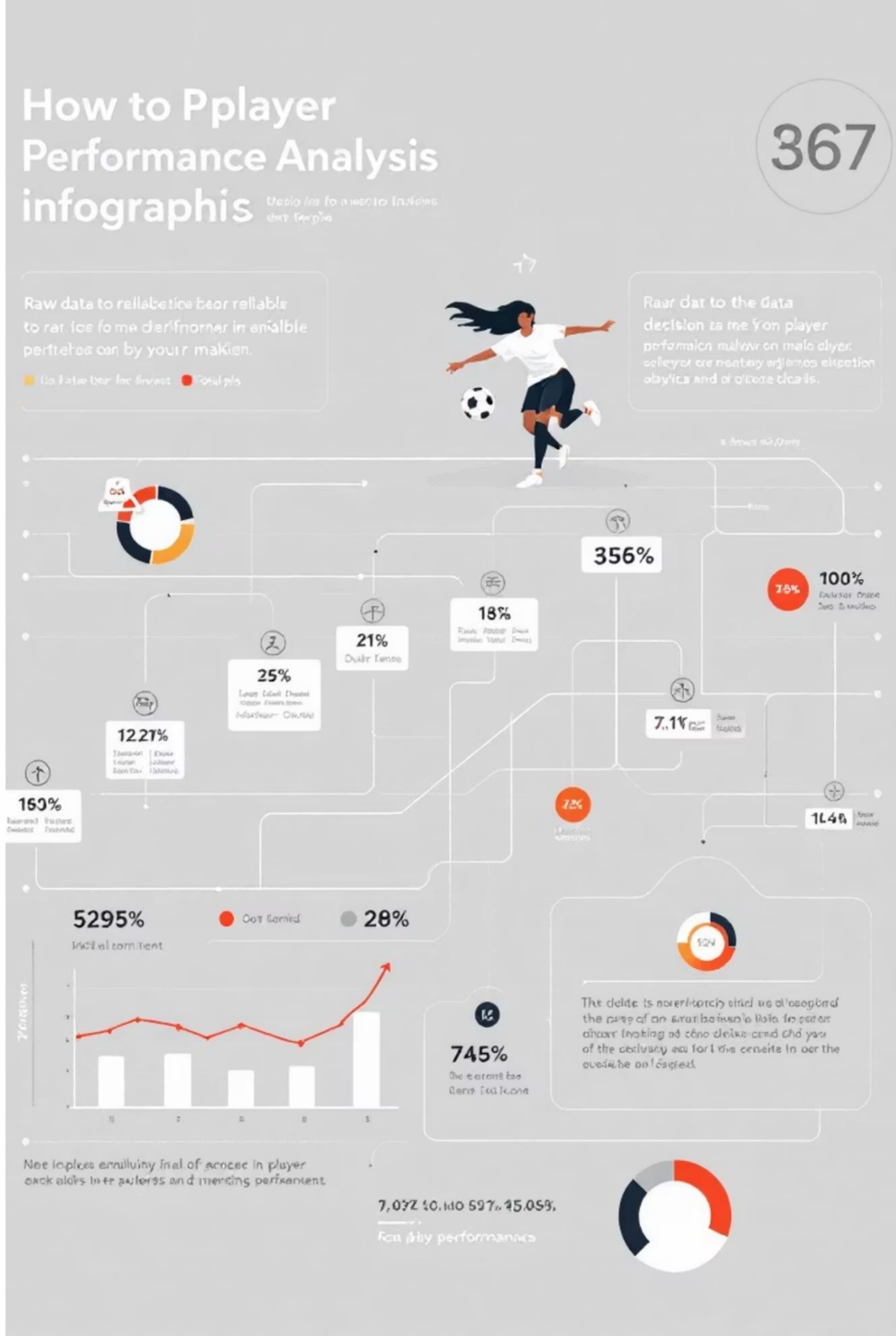
Decifrando o Desempenho de Jogadores

Dos Dados Brutos à Decisão Confiável

Maikon Evangelista, Matheus Eduardo, Rafael Arati e Vinicius Paiva

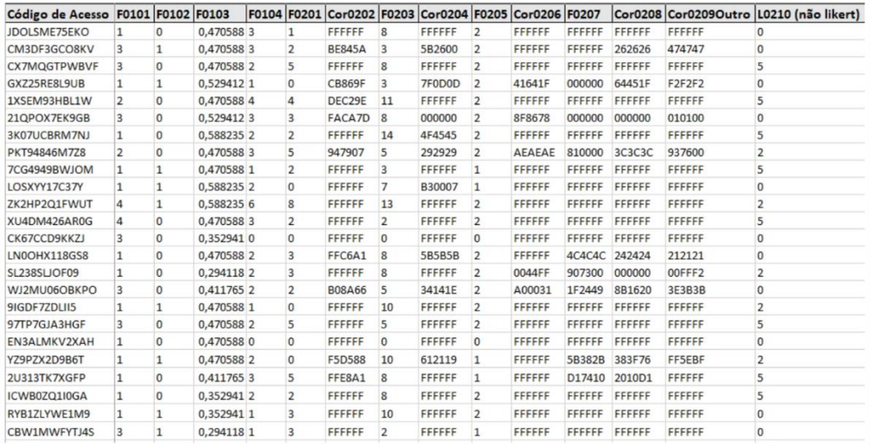


O Desafio: Três Alvos Misteriosos

Nossa jornada começou com uma missão clara e uma estratégia inicial: prever três Targets misteriosos usando clusterização para encontrar perfis de jogadores.

No entanto, o primeiro obstáculo foi imediato e fundamental: os dados falavam um idioma que não conhecíamos. As mais de 100 colunas tinham nomes enigmáticos como **F0101**, **Cor0202** e **V5Descr_0413**.

Antes de qualquer modelo de machine learning, nosso primeiro trabalho era o de um tradutor



Código de Acesso	F0101	F0102	F0103	F0104	F0201	Cor0202	F0203	Cor0204	F0205	Cor0206	F0207	Cor0208	Cor0209Outro	L0210 (não likert)
JDOLSME75EKO	1	0	0,470588	3	1	FFFFFF	8	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	0
CM3DF3GCO8KV	3	1	0,470588	3	2	BE845A	3	5B2600	2	FFFFFF	FFFFFF	262626	474747	0
CX7MQGTPWBVF	3	0	0,470588	2	5	FFFFFF	8	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
GXZ25RE8L9UB	1	1	0,529412	1	0	CB869F	3	7F0D0D	2	41641F	000000	64451F	F2F2F2	0
1XSEM93HBL1W	2	0	0,470588	4	4	DEC29E	11	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
21QPOX7EK9GB	3	0	0,529412	3	3	FACA7D	8	000000	2	8F8678	000000	000000	010100	0
3K07UCBRM7NJ	1	0	0,588235	2	2	FFFFFF	14	4F4545	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
PKT94846M7Z8	2	0	0,470588	3	5	947907	5	292929	2	AEAEAE	810000	3C3C3C	937600	2
7CG4949BWJOM	1	1	0,470588	1	2	FFFFFF	3	FFFFFF	1	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
LOSXYY17C37Y	1	1	0,588235	2	0	FFFFFF	7	B30007	1	FFFFFF	FFFFFF	FFFFFF	FFFFFF	0
ZK2HP2Q1FWUT	4	1	0,588235	6	8	FFFFFF	13	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	2
XU4DM426AR0G	4	0	0,470588	3	2	FFFFFF	2	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
CK67CCD9KKZJ	3	0	0,352941	0	0	FFFFFF	0	FFFFFF	0	FFFFFF	FFFFFF	FFFFFF	FFFFFF	0
LN0OHX118GS8	1	0	0,470588	2	3	FFC6A1	8	5B5B5B	2	FFFFFF	4C4C4C	242424	212121	0
SL238SLJOF09	1	0	0,294118	2	3	FFFFFF	8	FFFFFF	2	0044FF	907300	000000	00FFF2	2
WJ2MU06OBKPO	3	0	0,411765	2	2	B08A66	5	34141E	2	A00031	1F2449	8B1620	3E3B3B	0
9IGDF7ZDLII5	1	1	0,470588	1	0	FFFFFF	10	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	2
97TP7GJA3HGF	3	0	0,470588	2	5	FFFFFF	5	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
EN3ALMKV2XAH	1	0	0,470588	0	0	FFFFFF	0	FFFFFF	0	FFFFFF	FFFFFF	FFFFFF	FFFFFF	0
YZ9PZX2D9B6T	1	1	0,470588	2	0	F5D588	10	612119	1	FFFFFF	5B382B	383F76	FF5EBF	2
2U313TK7XGFP	1	0	0,411765	3	5	FFE8A1	8	FFFFFF	1	FFFFFF	D17410	2010D1	FFFFFF	5
ICWB0ZQ1I0GA	1	0	0,352941	2	2	FFFFFF	8	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	5
RYB1ZLYWE1M9	1	1	0,352941	1	3	FFFFFF	10	FFFFFF	2	FFFFFF	FFFFFF	FFFFFF	FFFFFF	0
CBW1MWFYTJ4S	3	1	0,294118	1	3	FFFFFF	2	FFFFFF	1	FFFFFF	FFFFFF	FFFFFF	FFFFFF	0

Primeira Vitória: Traduzindo o Idioma dos Dados

Diante de mais de 100 colunas enigmáticas e sem um dicionário de dados, agimos como **detetives de dados**. Em vez de procurar um mapa, nós o criamos.

Analisamos os padrões nos próprios nomes das colunas e, com base nessa lógica, escrevemos um código que classificou sistematicamente cada uma das 114 features em categorias funcionais, como mostra a tabela ao lado:

- **Texto:** Colunas que continham descrições ou códigos hexadecimais.
- **Likert:** Colunas que representavam escalas de opinião.
- **Tempo:** Colunas que mediam o tempo gasto em tarefas.
- **NãoLikert:** Demais variáveis numéricas e demográficas.

A lição: A estrutura dos dados estava escondida à vista de todos, nos próprios nomes das colunas. A clareza não foi um presente, foi uma conquista.

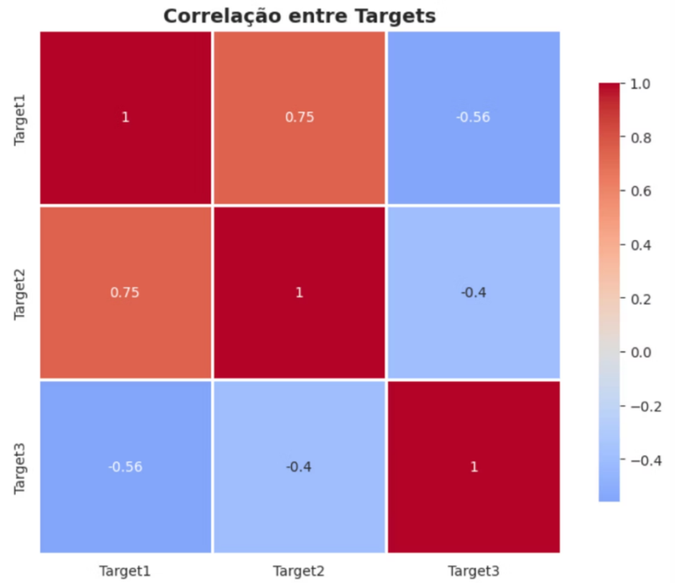
RESUMO DA CLASSIFICAÇÃO DE COLUNAS		
Categoria	Quantidade	Exemplos
Identificação	1	Código de Acesso
DateTime	1	Data/Hora Último
Targets	3	Target1, Target2
Texto	13	TempoTotalExpl
Likert	17	F1101
Tempo	32	Target3
NãoLikert	50	F0101
TOTAL	114	

Mudança de Rota: Quando os Dados Contam Sua História

A análise exploratória nos trouxe a revelação que mudou todo o rumo do projeto. Ao analisarmos a correlação entre os próprios Targets, descobrimos que eles **não se comportavam da mesma maneira**.

Como o gráfico mostra, Target1 e Target2 têm uma forte correlação positiva (**+0.75**). No entanto, ambos são **negativamente correlacionados** com o Target3.

Tentar prever os três com um único modelo "generalista" seria como um cabo de guerra, onde otimizar para um lado prejudicaria o outro.



Plano Original

❌ Um modelo generalista via clusterização

Premissa: Todos os targets compartilham padrões

Nova Estratégia

✓ Três modelos especialistas independentes

Realidade: **Cada target exige sua própria abordagem**

Este foi nosso primeiro grande pivô — e a decisão mais importante do projeto. Abandonamos a elegância de uma solução única em favor da eficácia de soluções especializadas.

A Fundação: Pré-processamento e Engenharia de Features

Antes de qualquer modelo, construímos um alicerce sólido. Nosso Notebook 01 de pré-processamento realiza toda a preparação dos dados de forma automatizada e consistente para os três modelos.

Principais Etapas:

- **Limpeza:** Remoção de colunas irrelevantes e tratamento de valores especiais (N/A, -1).
- **Engenharia de Features:** Criação de mais de 20 novas features agregadas, como Likert_Score_Medio, Tempo_Total e Performance_Score_Total, para enriquecer os dados, **como o exemplo à direita ilustra**.
- **Tratamento de Ausentes:** Imputação de valores faltantes com a mediana, garantindo que os dados estivessem completos.
- **Separação 80/20:** Os dados foram divididos em conjuntos de treino e teste, com os índices salvos para garantir que **todos os modelos fossem avaliados exatamente nas mesmas condições**.

Evidência: +20 Features Criadas no Notebook 01

```
✓ Criadas 5 features agregadas de Likert (baseadas em 17 colunas)
✓ Criadas 6 features agregadas de Tempo (baseadas em 32 colunas)
✓ Criadas 2 features de Performance (baseadas em 12 colunas)
✓ Criadas 3 features de Respostas P (baseadas em 16 colunas)
✓ Criadas 2 features de Quantidade (baseadas em 6 colunas)
✓ Criada feature 'Razao_Sono'
✓ Criada feature 'Razao_Q0413_Q0414'
✓ Criadas 3 features temporais
✓ Criada feature 'Consistencia_F07' (baseada em 9 colunas)
✓ Criada feature 'Consistencia_F11' (baseada em 8 colunas)
```

A Batalha dos Modelos: Enfrentando a Muralha do 0.700

Com os dados limpos, enriquecidos e preparados, iniciamos a 'Batalha dos Modelos'. Com a estratégia de especialistas definida, partimos para a 'Batalha dos Modelos'. Testamos mais de 15 algoritmos, incluindo os mais poderosos como XGBoost, em uma busca exaustiva pela máxima performance, visando uma meta de R^2 de 0.700.

--- Matriz de Desempenho Final dos Modelos ---

	Target	Modelo	R² Treino	R² Teste	Cumpriu Critério (≥ 0.700)
0	Target 1	Linear Regression	0.716342	-0.010000	Não
1	Target 1	XGBoost	1.000000	0.263711	Não
2	Target 2	Linear Regression	0.716342	-0.010000	Não
3	Target 2	XGBoost	1.000000	0.263711	Não
4	Target 3	Linear Regression	0.716342	-0.010000	Não
5	Target 3	XGBoost	1.000000	0.263711	Não

O resultado, como a tabela mostra, foi uma lição crucial. Atingimos a perfeição nos dados de treino, com um R^2 de **1.0**, mas a performance nos dados de teste era baixíssima, em torno de **0.26**.

Este é o fantasma que assombra todo cientista de dados: **overfitting**. O modelo não estava aprendendo, estava apenas 'decorando' as respostas.

> **Aprendemos da forma mais difícil que o melhor modelo não é o que tem a maior performance no treino, mas sim aquele que é confiável e estável no mundo real.**

A Revelação: Simplicidade Como Estratégia

Depois de semanas lutando contra a complexidade, fizemos uma pergunta radical: **e se a resposta fosse mais simples?**

Modelos Complexos

- Alta capacidade de aprendizado
- Capturam interações não-lineares
- Milhares de parâmetros ajustáveis
- Tendência a memorizar ruído
- Difícil de interpretar
- Instabilidade entre folds

Modelos Lineares Regularizados

- Simplicidade interpretável
- Regularização integrada (L1/L2)
- Controle natural de overfitting
- **Estabilidade consistente**
- **Generalização confiável**
- **Performance previsível**

A revelação foi clara: a solução não estava em mais complexidade, mas em mais controle. Adotamos os modelos lineares com forte **regularização** (Lasso, Ridge, ElasticNet) como nossa nova estratégia. Eles não apenas resolveram o problema do overfitting, como se provaram uma base estável e confiável para o refinamento que viria a seguir.

Refinamento: O Poder da Engenharia de Features Direcionada

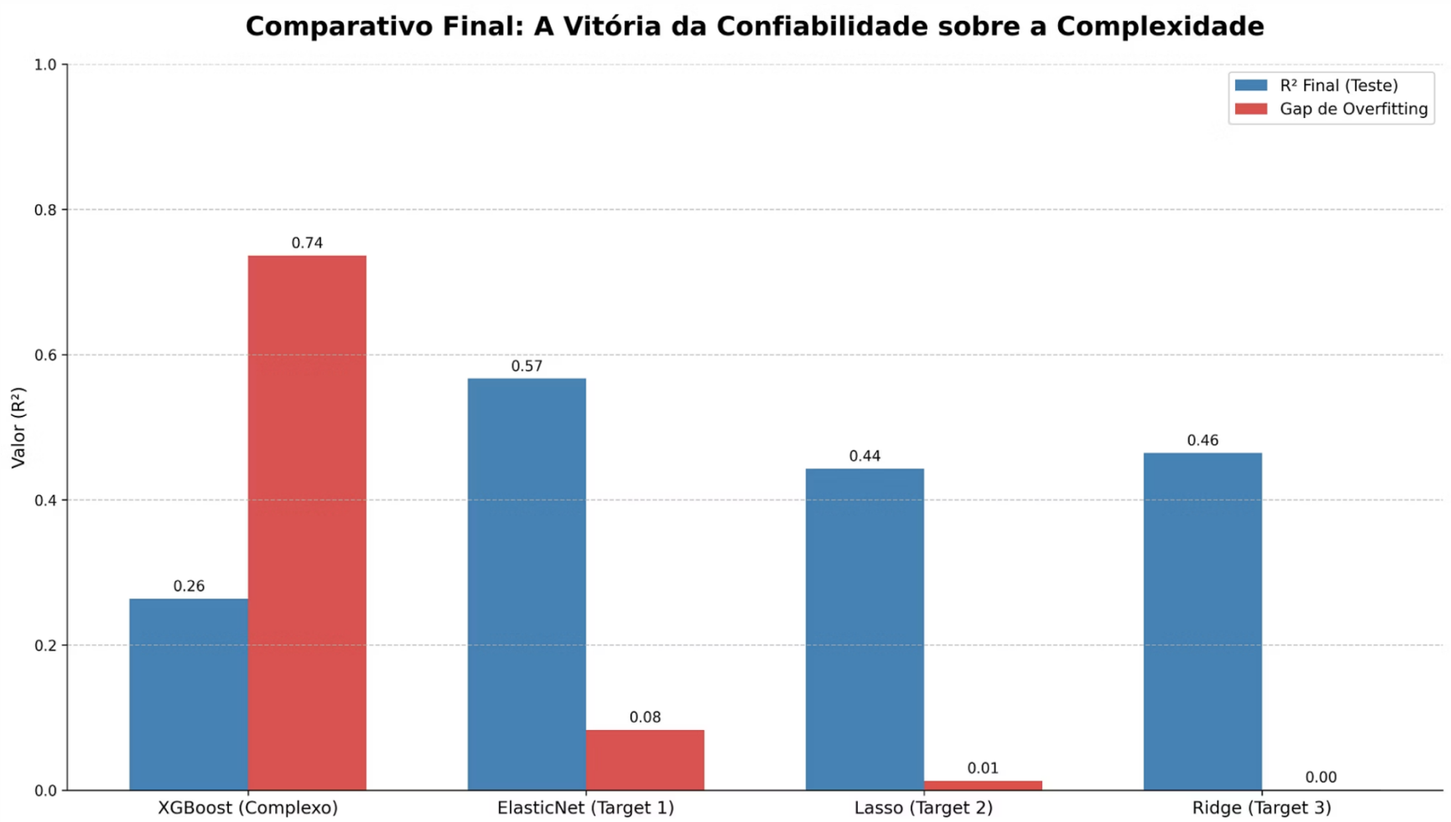
Com modelos estáveis em mãos, partimos para o refinamento. Em vez de ajustar o *modelo*, decidimos enriquecer os *dados*.

Criamos novas '**features de interação**' baseadas nas variáveis mais preditivas que já havíamos encontrado. A hipótese era que a combinação inteligente de features poderia revelar padrões que os modelos lineares não capturariam sozinhos.

```
================================================================================
==================== 🔬 CRIANDO NOVAS FEATURES DE INTERAÇÃO ====================
================================================================================

📋 As 10 features mais importantes selecionadas anteriormente foram:
    1. F1107
    2. Tempo_Total
    3. Respostas_P_Media
    4. P04
    5. F1105
    6. P08
    7. F0711
    8. Likert_Score_Min
    9. Tempo_Medio
   10. T15

✨ Criando novas features...

   ✓ Feature 'Eficiencia_Performance' criada (Performance / Tempo)
   ✓ Feature 'Atitude_Consistente' criada (Score Likert * Consistência)
   ✓ Feature 'Idade_Anos_Sq' criada (Idade ao Quadrado)
```

Comparativo Final: A Vitória da Confiabilidade sobre a Complexidade

Este gráfico resume nossa jornada. Ele não mostra apenas métricas, mas a batalha entre performance aparente e confiabilidade real.



Os números contam uma história clara: enquanto os modelos complexos sofriam com um 'gap' de overfitting **gigantesco (0.74)**, nossos modelos finais, baseados na simplicidade e regularização, entregaram uma performance sólida com um gap praticamente **nulo**.

A lição final: O melhor modelo não é aquele com o maior R², mas sim aquele em que se pode confiar.

O Modelo Mais Confiável é o Verdadeiro Vencedor

Nossa jornada através deste projeto de machine learning nos ensinou uma verdade fundamental: **em produção, confiabilidade supera complexidade**.

Começamos perseguindo performance máxima com modelos sofisticados. Terminamos encontrando sucesso sustentável através de simplicidade inteligente e forte regularização.

3

Modelos Especialistas

Cada um otimizado para seu target específico

~0.03

Gap Médio de Overfitting

Controle excepcional de generalização

~0.49

R^2 Médio em Validação

Performance sólida e confiável

> **Lição Final:** Na ciência de dados, o objetivo não é construir o modelo mais impressionante — é construir o modelo em que você pode confiar para tomar decisões importantes. E esse modelo é frequentemente mais simples do que você imagina.

Aprendizados: O Que Esta Jornada Nos Ensinou

Cada obstáculo superado trouxe lições valiosas que transcendem este projeto específico. Estas são as pérolas de sabedoria que levaremos para futuros desafios em ciência de dados.

Deixe os Dados Guiarem

Nossa análise exploratória revelou que os targets exigiam estratégias diferentes. Escutar os dados nos salvou de seguir cegamente um plano inadequado.

Simplicidade é Força

Modelos lineares regularizados superaram algoritmos complexos em confiabilidade. A solução mais simples que funciona é frequentemente a melhor.

Overfitting é o Verdadeiro Inimigo

Performance máxima no treino não vale nada se o modelo falha na validação. Controlar overfitting é mais importante que maximizar R^2.

Regularização Como Ferramenta

Lasso, Ridge e ElasticNet não são apenas técnicas — são filosofias de modelagem que priorizam generalização sobre memorização.

Contexto Importa Mais Que Algoritmos

Traduzir os dados criptografados foi tão crucial quanto escolher o modelo certo. Entender o domínio é metade da batalha.

Especialização Supera Generalização

Três modelos focados superaram um modelo generalista. Às vezes, a melhor estratégia é reconhecer que problemas diferentes merecem soluções diferentes.