Predição de Sucesso Eleitoral

Deputados Federais no Brasil

Validação Temporal: 2018 → 2022

Artur Garcia & Artur Saraiva
Universidade Federal do Ceará (UFC)
Disciplina: Aprendizagem de Máquina
Janeiro/2026

Contexto e Motivação

Por que prever sucesso eleitoral?

- **Democracia brasileira**: Sistema proporcional com lista aberta
- **Cenário desafiador**: Milhares de candidatos, poucos eleitos
- **Questão central**: Fatores estruturais determinam vitória?

Aplicações práticas

✅ Partidos políticos otimizarem alocação de recursos
✅ Consultorias eleitorais orientarem candidatos
✅ Cidadãos compreenderem dinâmicas de poder

Definição do Problema

Tipo de problema:

Classificação binária supervisionada

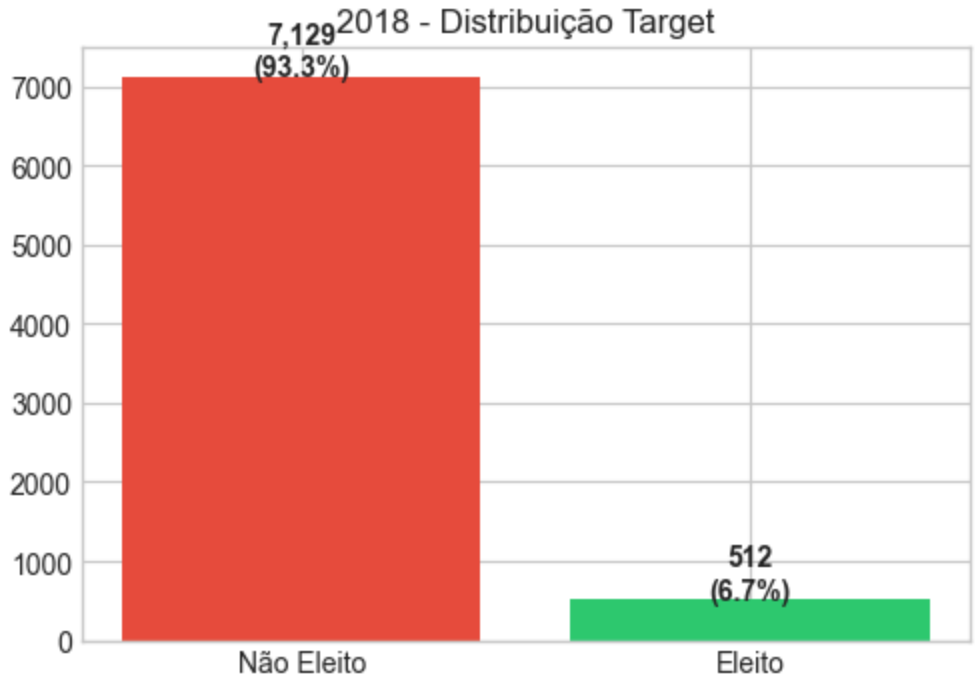
Variável alvo (Target)

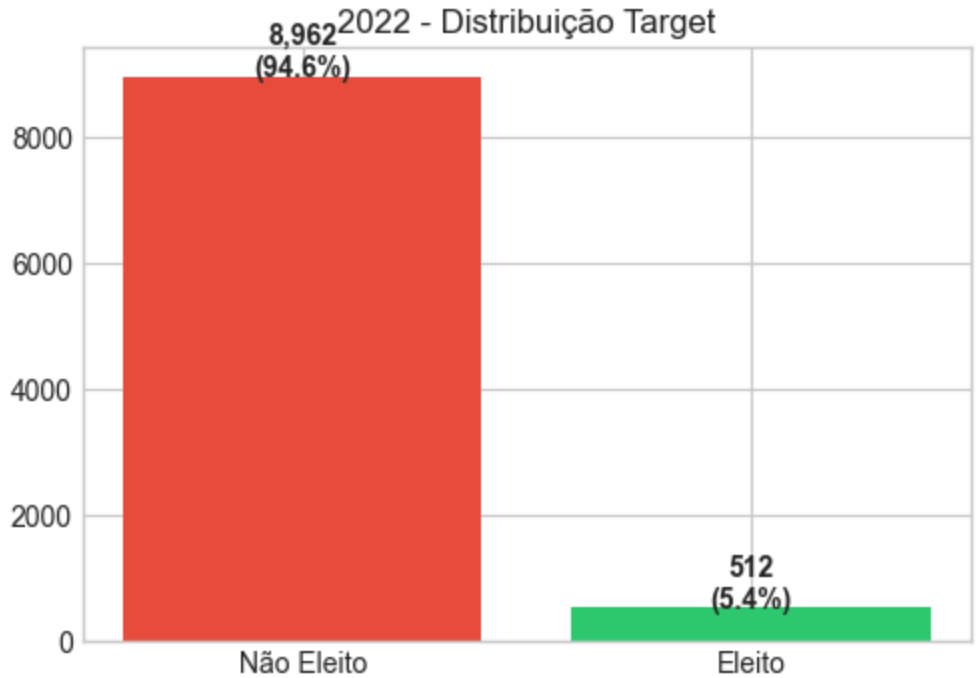
```
ELEITO = {
  1: Candidato eleito (inclui "ELEITO POR MÉDIA" e "ELEITO POR QP")
  0: Candidato não eleito
}
```

Desafio principal

Desbalanceamento extremo: ~90% não-eleitos vs. ~10% eleitos

- Razão de desbalanceamento geral: **1:~15**
- Implica em uso de métricas robustas (F1-Score, AUC-PR, ...)
- Modelo tende a favorecer classe majoritária


2018 - Distribuição Target
7,129
(93.3%)
512
(6.7%)
7000
6000
5000
4000
3000
2000
1000
0
Não Eleito
Eleito


2022 - Distribuição Target
8,962
(94.6%)
512
(5.4%)
8000
6000
4000
2000
0
Não Eleito
Eleito

Dados e Pré-processamento

Fonte dos dados

Tribunal Superior Eleitoral (TSE) - Dados públicos

Dataset	2018	2022
Candidatos	~8.000	~10.000
Features originais	50+	50+
Taxa de eleitos	6.7%	5.4%

Bases integradas

1. **Consulta de Candidatos**: Dados demográficos e políticos
2. **Complementar**: Idade, reeleição, despesa máxima
3. **Bens**: Patrimônio declarado

ETL - Pipeline de Limpeza

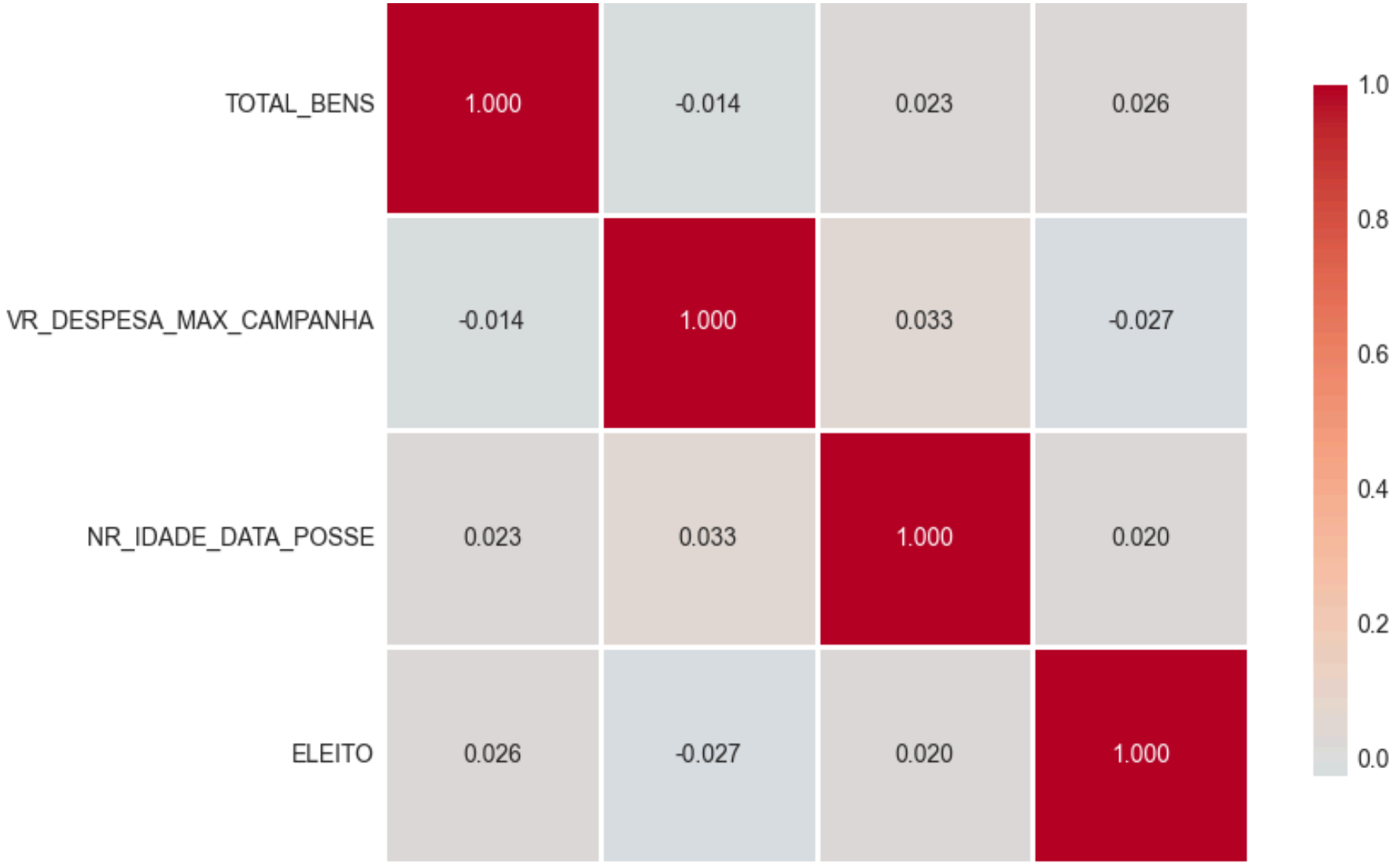
Principais transformações

1. **Definição do target**: Mapeamento de situações eleitorais
2. **Filtros de qualidade**:
    - Remover candidatos com situação irregular (CD_SITUACAO_CANDIDATURA ≠ 12)
3. **Tratamento de missings**
4. **Harmonização de partidos**: Fusões entre 2018-2022
    - `PR` → `PL`, `DEM` → `UNIÃO`, etc.

Resultado: Dataset limpo e consistente para modelagem

EDA - Correlações Entre Features

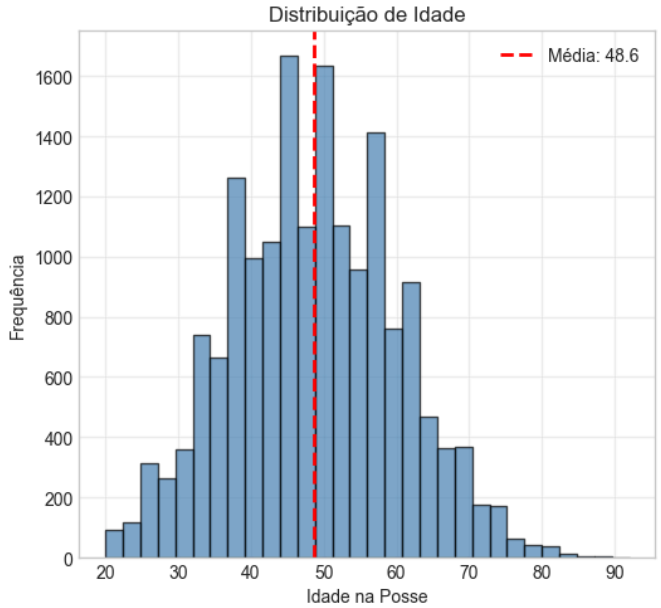
Matriz de correlação (variáveis numéricas)

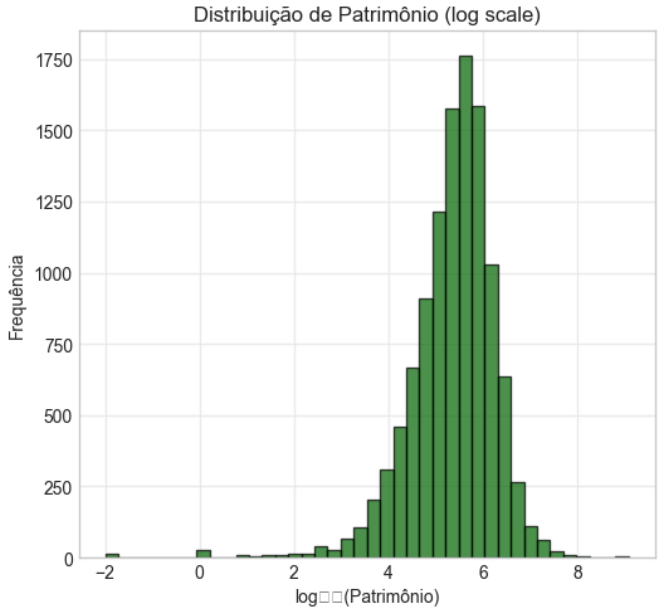


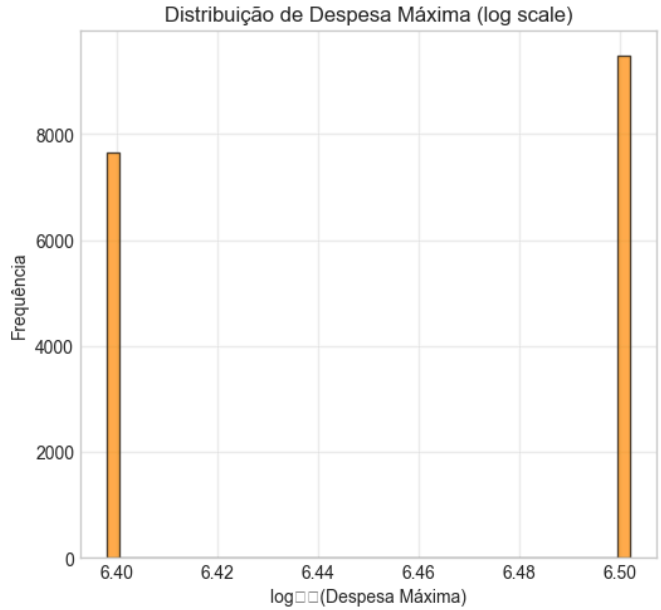
💡 Insights

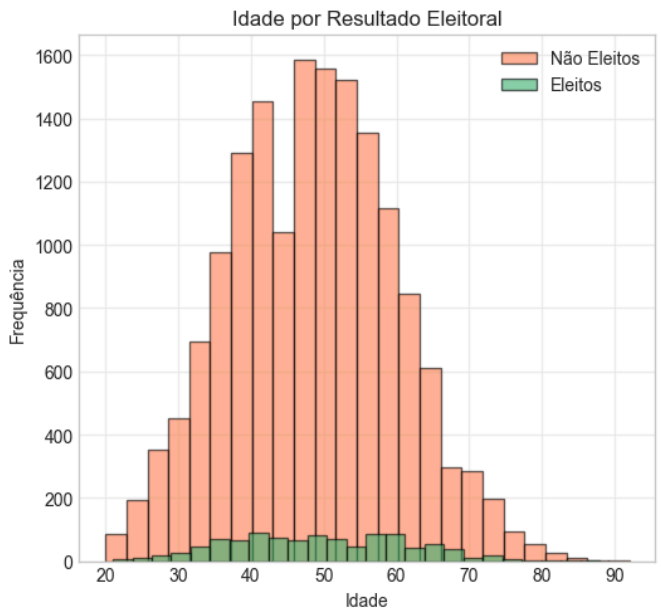
- **Correlações fracas** (<0.3): Nenhuma feature sozinha determina vitória
- **Interações complexas**: Modelos ensemble necessários

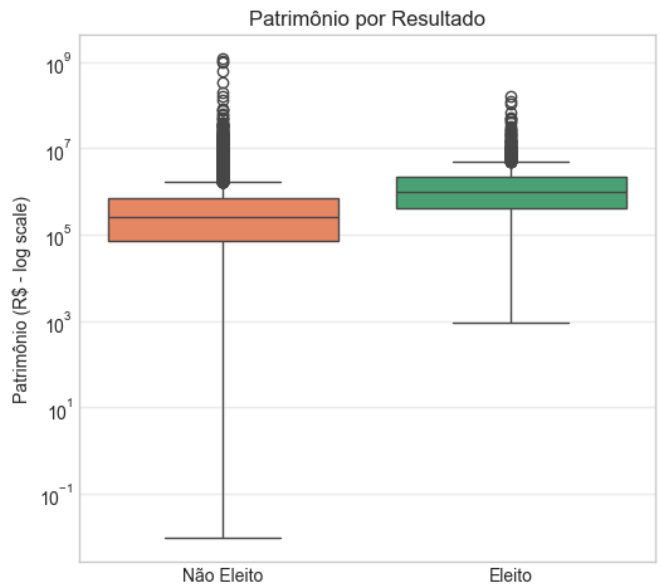
EDA - Histogramas de Distribuições

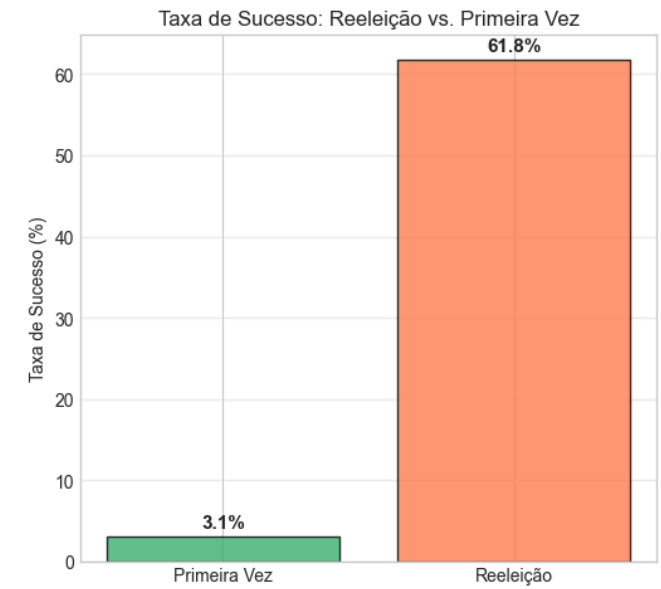

Distribuição de Idade
Média: 48.6
Frequência
Idade na Posse


Distribuição de Patrimônio (log scale)
Frequência
log□□(Patrimônio)


Distribuição de Despesa Máxima (log scale)
Frequência
log□□(Despesa Máxima)


Idade por Resultado Eleitoral
Não Eleitos
Eleitos
Frequência
Idade


Patrimônio por Resultado
Patrimônio (R$ - log scale)
Não Eleito
Eleito


Taxa de Sucesso: Reeleição vs. Primeira Vez
61.8%
3.1%
Taxa de Sucesso (%)
Primeira Vez
Reeleição

EDA - Testes de Significância

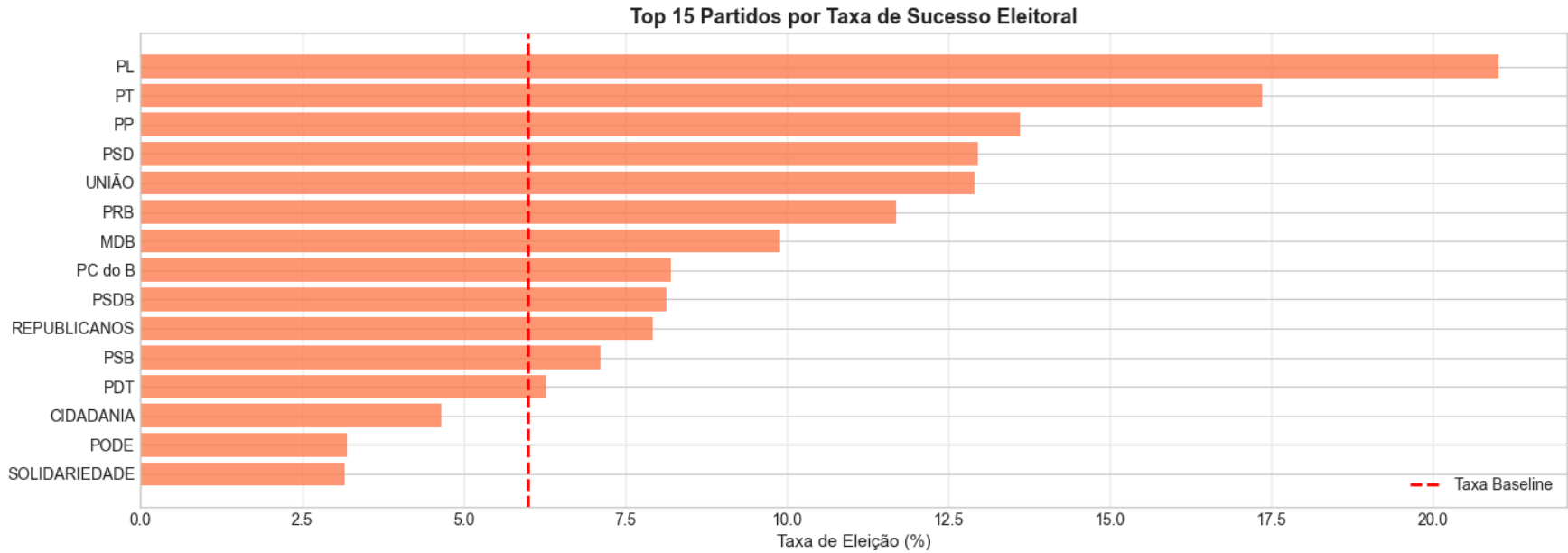
Mann-Whitney U Test (não-paramétrico)

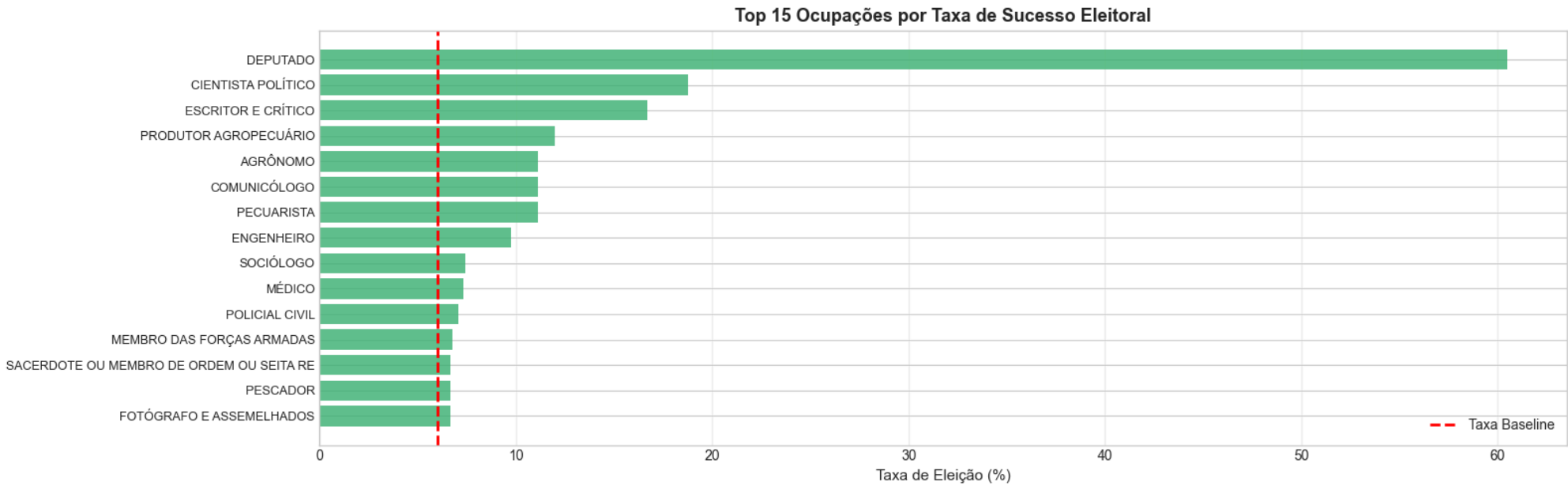
Feature	Eleitos (média)	Não-eleitos (média)	p-value	Significância
Patrimônio	R$ 2.5M	R$ 821K	p < 0.001	✅ Alta
Idade	49.5 anos	48.55 anos	p < 0.05	✅ Moderada

Interpretação

- Diferenças são **estatisticamente significativas**
- Eleitos têm, em média, **90% mais patrimônio**

EDA - Partidos e Ocupações


Top 15 Partidos por Taxa de Sucesso Eleitoral
PL
PT
PP
PSD
UNIÃO
PRB
MDB
PC do B
PSDB
REPUBLICANOS
PSB
PDT
CIDADANIA
PODE
SOLIDARIEDADE
Taxa Baseline
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Taxa de Eleição (%)


Top 15 Ocupações por Taxa de Sucesso Eleitoral
DEPUTADO
CIENTISTA POLÍTICO
ESCRITOR E CRÍTICO
PRODUTOR AGROPECUÁRIO
AGRÔNOMO
COMUNICÓLOGO
PECUARISTA
ENGENHEIRO
SOCIÓLOGO
MÉDICO
POLICIAL CIVIL
MEMBRO DAS FORÇAS ARMADAS
SACERDOTE OU MEMBRO DE ORDEM OU SEITA RE
PESCADOR
FOTÓGRAFO E ASSEMELHADOS
Taxa Baseline
0
10
20
30
40
50
60
Taxa de Eleição (%)

Baseline Models - Sanity Check

Objetivo

Estabelecer **patamar mínimo** antes de modelos complexos

Resultados (Validação 2018 → 2022)

Modelo	F1-Score	AUC-ROC	AUC-PR	Balanced Acc
Most Frequent	0.0000	0.5000	0.0540	0.5000
Stratified	0.0631	0.5021	0.0543	0.5021
Logistic (default)	0.0000	0.5323	0.0694	0.5000

Feature Engineering - Estratégia

1. Features Binárias

- `IS_REELEICAO` : Candidato busca reeleição (S/N)
- `IS_FEMININO` : Gênero feminino
- `IS_PARTIDO_GRANDE` : Partido top-6
- `TEM_BENS` : Possui bens declarados

2. Transformações Logarítmicas

- `LOG_BENS = log(TOTAL_BENS + 1)`
- `LOG_DESPESA_MAX = log(VR_DESPESA_MAX + 1)`

3. Features de Coligação

- `QTD_PARTIDOS_COLIG` : Tamanho da coligação
- `IS_COLIGADO` : Pertence a coligação (vs. partido isolado)

Feature Engineering - Target Encoding

Target Encoding com Smoothing Bayesiano

Para cada categoria (partido, ocupação, UF, etc.):

$$\text{encoded\_value} = \frac{n_{categoria} \times \bar{y}_{categoria} + \alpha \times \bar{y}_{global}}{n_{categoria} + \alpha}$$

- **α = 10** (parâmetro de smoothing validado)
- **Previne overfitting** em categorias com poucas observações
- **Preserva sinal** em categorias frequentes

Metodologia - Validação Temporal

Setup Experimental

Por que validação temporal?

✅ **Simula cenário real**: Prever eleição futura com dados passados
✅ **Mais rigoroso**: Não pode "espiar" dados de teste durante treino
✅ **Detecta concept drift**: Mudanças em padrões eleitorais

Limitação

Apenas 2 ciclos disponíveis - impossibilita análise de tendências

Estratégia de Modelagem - 3 Etapas

Etapa 1: Screening (Grid Reduzido)

- **4 modelos**: Logistic Regression, Random Forest, Gradient Boosting, XGBoost
- **Objetivo**: Identificar arquiteturas promissoras

Etapa 2: Seleção de Finalistas

- Selecionar **Top 2** por F1-Score no teste

Etapa 3: Otimização Completa

- **Grid expandido** com RandomizedSearchCV para espaços grandes
- **Métrica de seleção**: F1-Score (5-fold CV estratificado)

Tratamento do Desbalanceamento

Estratégias aplicadas

Durante Treinamento

- **Class Weights**: Penalização de erros na classe minoritária
  - Ratio: **~14x mais peso para eleitos**

Durante Avaliação

- **Métricas apropriadas**:
  - ❌ Acurácia
  - ✅ F1-Score, AUC-ROC, AUC-PR

Resultados - Screening (Etapa 1)

Grid Reduzido (2018 → 2022)

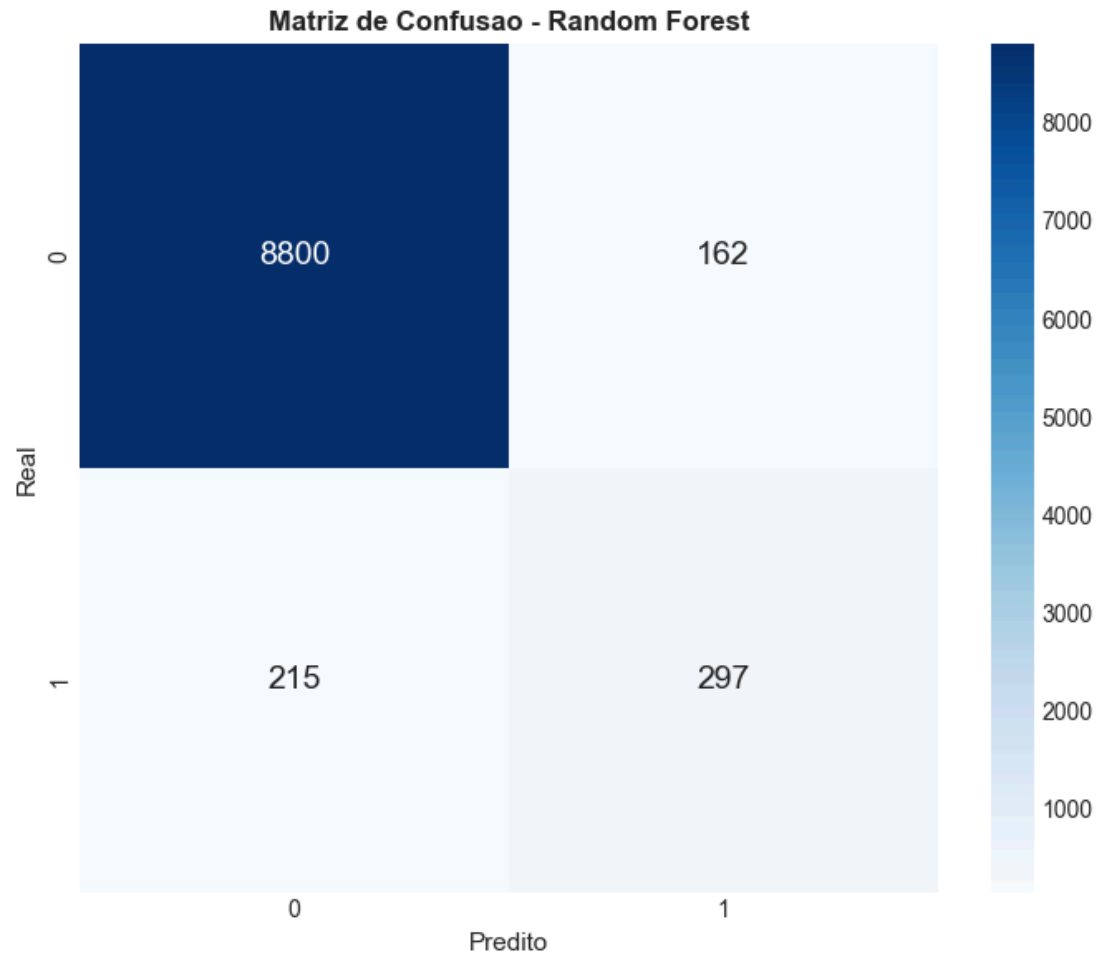
Modelo	F1-Score	AUC-ROC	AUC-PR	Precision	Recall
Random Forest	**0.5992**	**0.9312**	**0.5568**	0.6360	0.5664
XGBoost	0.5833	0.9085	0.4957	0.5513	0.6191
Gradient Boosting	0.5720	0.9267	0.5403	0.6423	0.5156
Logistic Regression	0.4491	0.9340	0.5517	0.3106	0.8105

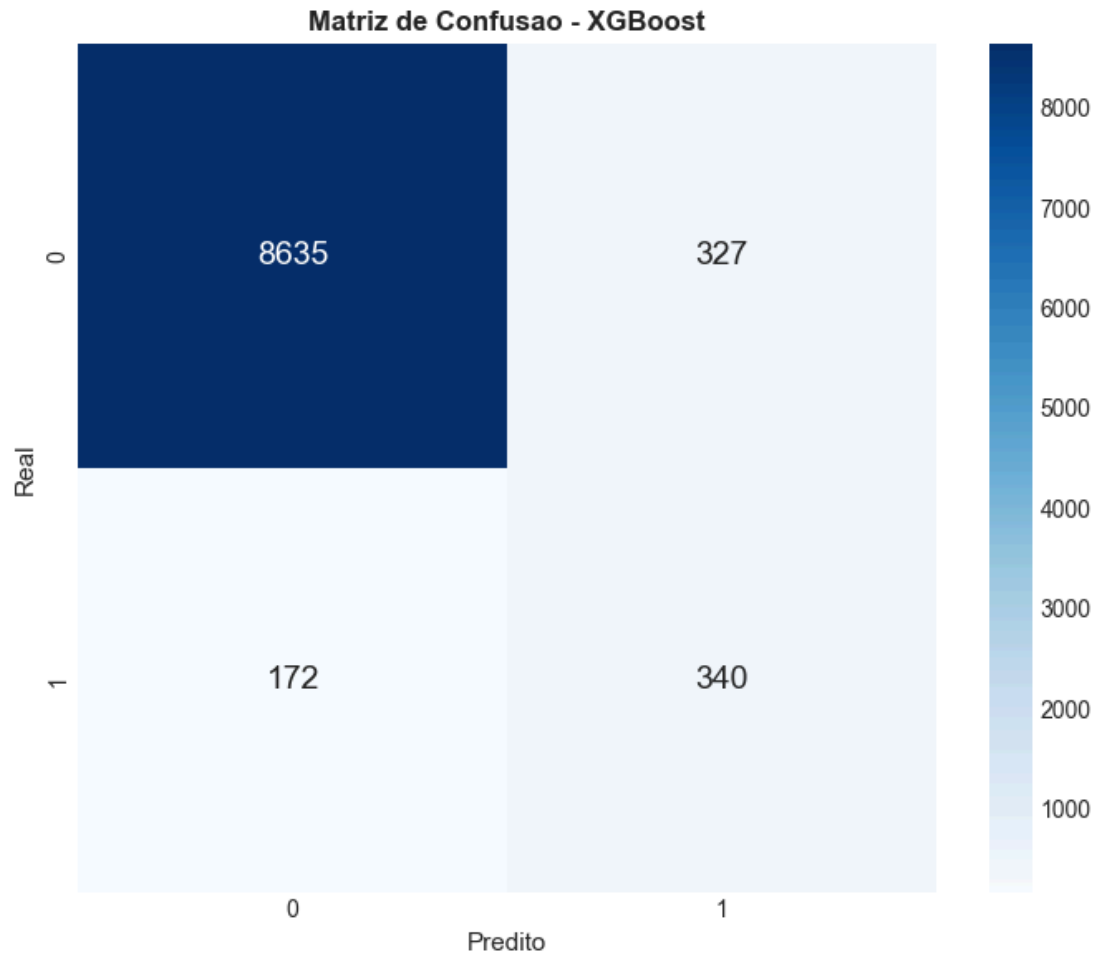
Resultados Finais - Otimização Completa

Grid Completo (Finalistas)

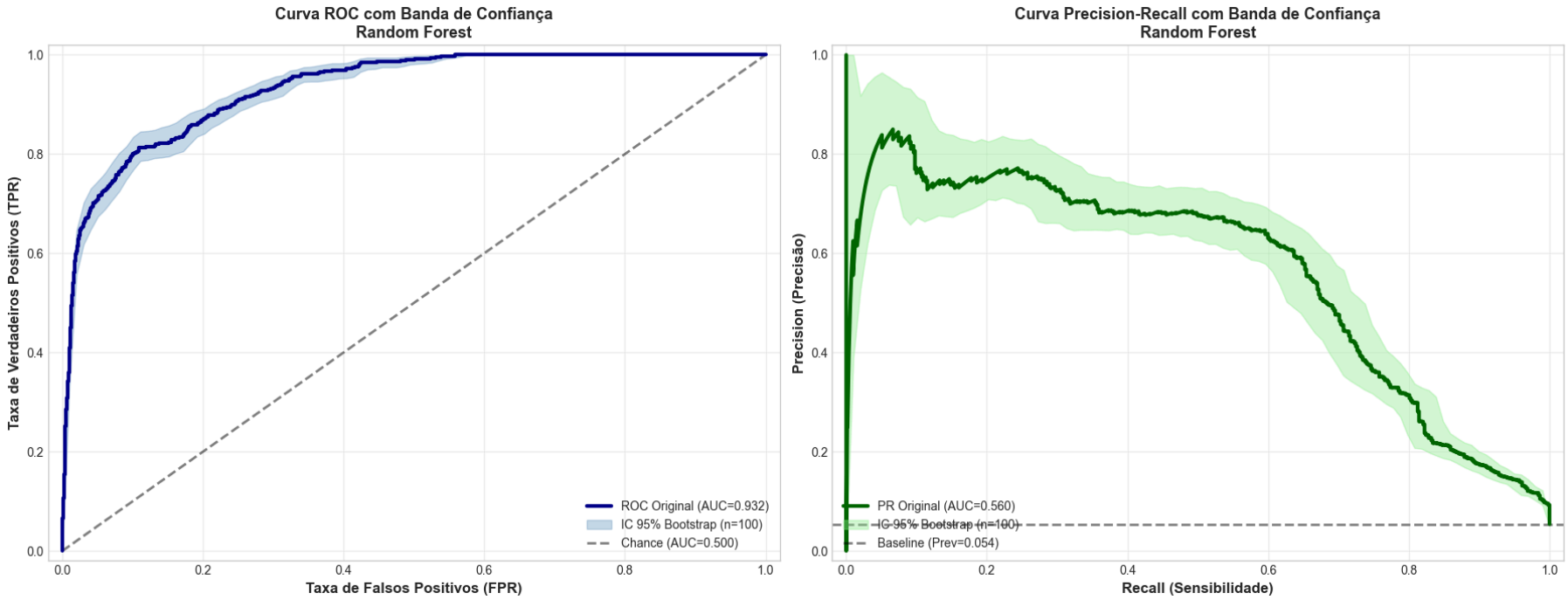
Modelo	F1-Score	AUC-ROC	AUC-PR	Precision	Recall
Random Forest	**0.6117**	**0.9324**	**0.5605**	0.6471	0.5801
XGBoost	0.5768	0.9224	0.5267	0.5097	0.6641

🏆 Modelo Vencedor: Random Forest


Matriz de Confusao - Random Forest
Real
Predito
0
1
8800
162
215
297
8000
7000
6000
5000
4000
3000
2000
1000

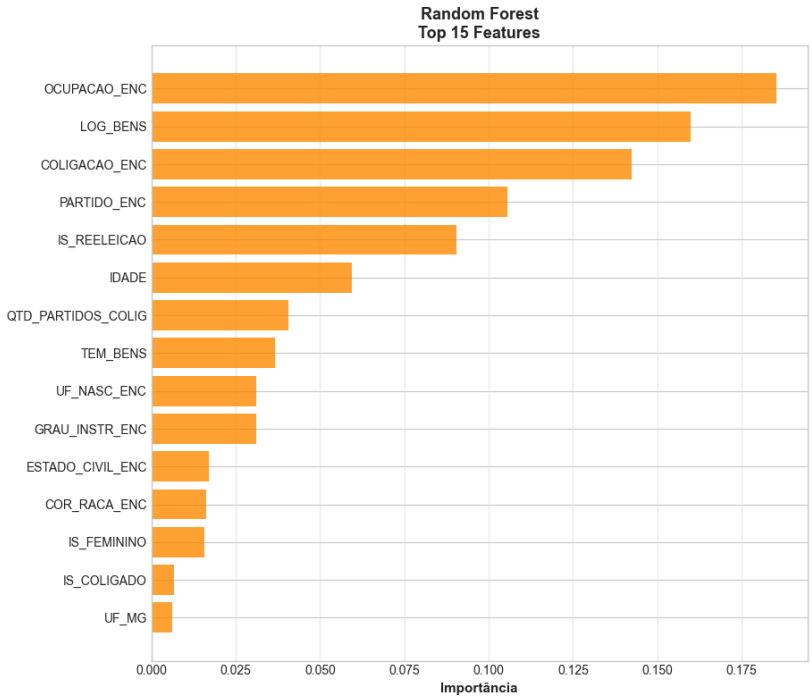

Matriz de Confusao - XGBoost
Real
Predito
0
1
8635
327
172
340
8000
7000
6000
5000
4000
3000
2000
1000

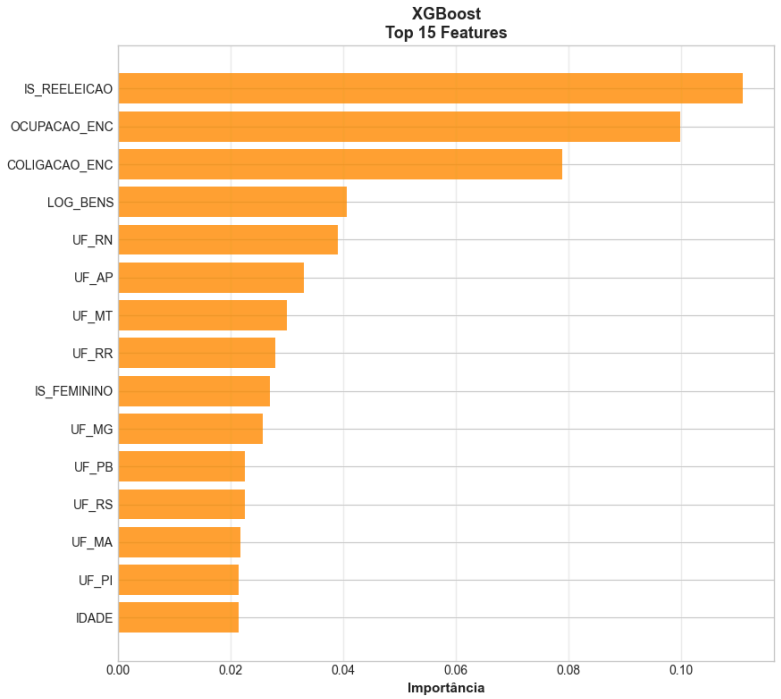
Curvas ROC e Precision-Recall



Feature Importance

Top 15 Features mais importantes

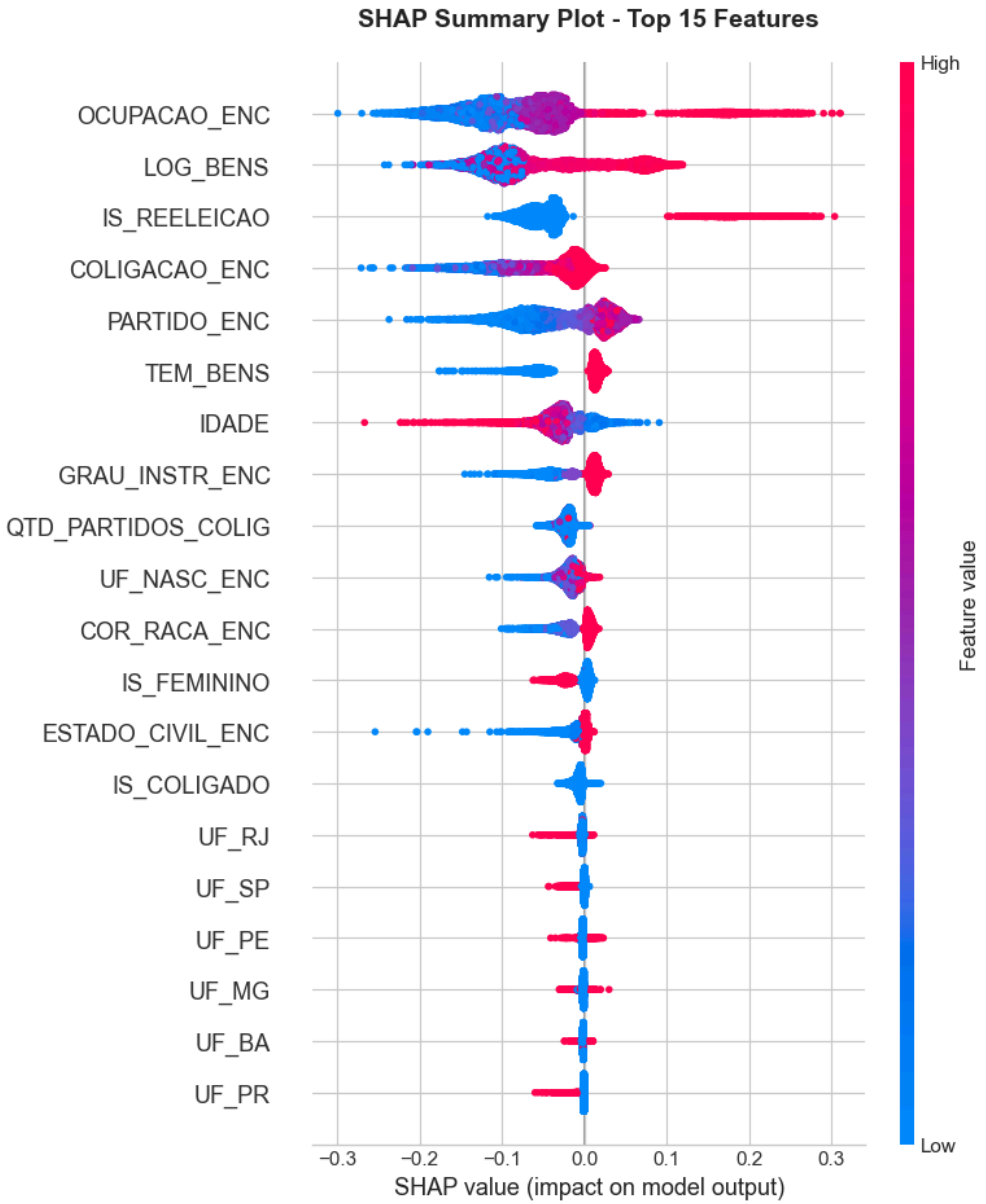




Interpretabilidade - SHAP

SHAP (SHapley Additive exPlanations)

- **Abordagem teórica de jogos** para explicar predições
- Mostra **impacto individual** de cada feature na predição
- Captura **interações** entre features


SHAP Summary Plot - Top 15 Features
OCUPACAO_ENC
LOG_BENS
IS_REELEICAO
COLIGACAO_ENC
PARTIDO_ENC
TEM_BENS
IDADE
GRAU_INSTR_ENC
QTD_PARTIDOS_COLIG
UF_NASC_ENC
COR_RACA_ENC
IS_FEMININO
ESTADO_CIVIL_ENC
IS_COLIGADO
UF_RJ
UF_SP
UF_PE
UF_MG
UF_BA
UF_PR
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low

Análise de Erros - Matriz de Confusão

Matriz Normalizada (% por linha)

	Predito: Não-Eleito	Predito: Eleito
Real: Não-Eleito	96.2% (TN)	1.8% (FP)
Real: Eleito	42% (FN)	58% (TP)

Tipos de erro

Falsos Positivos (FP): 162 candidatos

- Preditos como eleitos mas **perderam**
- **Perfil**: Alto patrimônio, partidos grandes, ocupações prestigiadas

Falsos Negativos (FN): 215 candidatos

- **Eleitos** mas modelo previu derrota
- **Perfil**: Baixo patrimônio, partidos pequenos, primeira candidatura, ocupações menos tradicionais

Falsos Positivos - Exemplos

Casos típicos (preditos eleitos mas perderam)

Candidato A

- **Perfil**: Deputado, patrimônio R$ 1.3M, UNIÃO, SP
- **Probabilidade predita**: 0.97 (alta confiança)
- **Resultado**: Não eleito

Falsos Negativos - Exemplos

Casos típicos (eleitos mas preditos como derrotados)

Candidato B

- **Perfil**: Policial Militar, patrimônio R$ 19K, PODE, RJ
- **Probabilidade predita**: 0.03 (baixa confiança)
- **Resultado**: Eleito

Conclusões

Principais Achados

✅ **Validação temporal bem-sucedida**: Generalização 2018 → 2022 (gap 4 anos)
✅ **Random Forest vencedor**: F1=0.61, AUC-ROC=0.93, AUC-PR=0.56
✅ **Features determinantes**: Partido/Coligação, incumbência, recursos financeiros

Performance contextualizada

- Supera baselines significativamente (>30% em F1)
- Desbalanceamento 1:~15 superado com sucesso
- F1~0.61 demonstra **excelente capacidade preditiva** do modelo

⚠️ Limitações

Dados Limitados

- Apenas **2 ciclos** eleitorais (2018, 2022)
- Impossível analisar **tendências de longo prazo**
- Sensível a **eventos pontuais** (pandemia, impeachment)

Variáveis Ausentes (críticas)

- Popularidade pré-campanha (pesquisas de intenção de voto)
- Engajamento em redes sociais (seguidores, likes)
- Cobertura da mídia (menções, debates)
- Escândalos e processos judiciais

Mensagem Final

Este trabalho demonstra que:

Mesmo com **features limitadas** a dados declaratórios pré-eleição, é possível desenvolver **modelos preditivos robustos** que:

1. **Superam significativamente o acaso** (AUC-ROC 0.93 vs. 0.50)
2. **Identificam fatores estruturais** de sucesso eleitoral
3. **Generalizam entre ciclos** eleitorais (2018 → 2022)

Obrigado! 🙏

```
“O modelo não é o fim — é o começo da decisão.”
```

Perguntas?

Contato

Artur Garcia & Artur Saraiva

Universidade Federal do Ceará (UFC)