NYPD SHOOTING LETHALITY PREDICTION

Artur Pigari Prata

Contexto e problema

Motivação

- Violência armada é um dos desafios persistentes das grandes metrópoles.
- Mesmo com queda geral da criminalidade em NYC, tiroteios continuam frequentes e perigosos.
- A letalidade (morte vs não morte) é um desfecho crítico para segurança pública.

Perguntas de pesquisa

- Quais fatores estão associados à letalidade dos tiroteios em NYC?
- É possível estimar a probabilidade de um tiro resultar em morte a partir das informações disponíveis no momento do crime?

Contexto e problema

Hipóteses

- **H1:** Letalidade varia com horário, borough e precinct.
- **H2:** Perfil demográfico da vítima e do agressor influencia nos desfechos.
- **H3:** Modelos baseados em árvores (RF/XGBoost) superam modelos lineares.

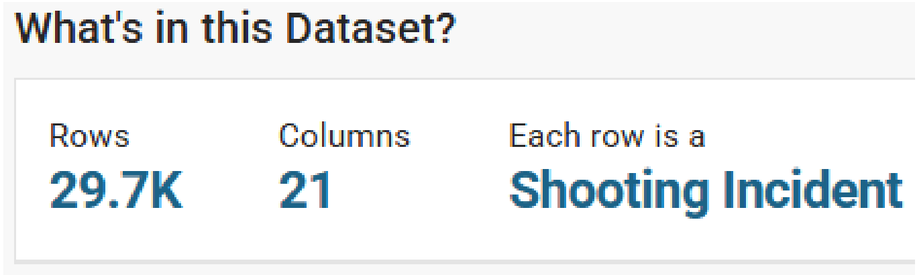
Demonstração dos dados

Fonte

- NYPD Shooting Incident Data (Historic)

Variáveis principais

- **Temporais:** ano, hora, período do dia.
- **Geográficas:** borough, precinct, coordenadas X/Y.
- **Demográficas:** idade, sexo e raça de vítimas e agressores.
- **Target:** STATISTICAL_MURDER_FLAG (0/1).



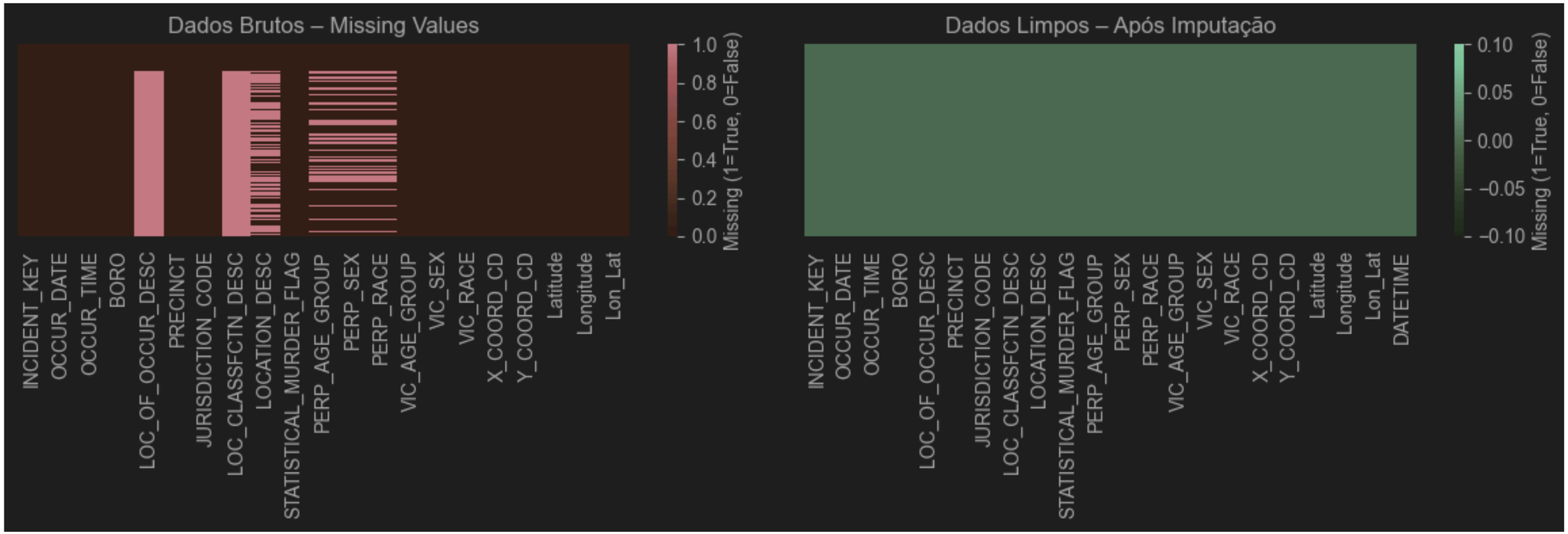
Demonstração dos dados

Pré-processamento

- Remoção de duplicatas
- Padronização e limpeza de categorias
- Conversão de datas
- Criação de novas features (ex: período do dia)

BORO	LOC_OF_OCCUR_D	PRECINCT	JURISDICTION_COD	LOC_CLASSFCTN_D	LOCATION_DESC	STATISTICAL_MU
BROOKLYN	OUTSIDE	69	0	STREET	(null)	FALSE
BROOKLYN	OUTSIDE	69	0	STREET	(null)	FALSE
BRONX	OUTSIDE	52	0	STREET	(null)	FALSE
BRONX	OUTSIDE	47	0	STREET	(null)	FALSE
BRONX	OUTSIDE	47	0	STREET	(null)	FALSE
BROOKLYN	OUTSIDE	60	2	HOUSING	MULTI DWELL - PUE	FALSE
BRONX	INSIDE	41	0	DWELLING	MULTI DWELL - APT	TRUE
BRONX	OUTSIDE	47	0	STREET	(null)	FALSE
MANHATTAN	OUTSIDE	23	0	STREET	(null)	FALSE
BRONX	OUTSIDE	43	0	STREET	(null)	FALSE
MANHATTAN	INSIDE	18	0	DWELLING	MULTI DWELL - APT	TRUE

Demonstração dos dados



Metodologia

Visão geral do pipeline

1. Diagnóstico dos dados
2. Análise exploratória (EDA)
3. Testes estatísticos (Qui-quadrado + V de Cramér)
4. Preparação para modelagem (OneHot, scaling, undersampling)
5. Modelos supervisionados
6. Avaliação e comparação

Testes estatísticos

- Qui-quadrado para associação
- V de Cramér para força do efeito
- Hipóteses H1 e H2 avaliadas aqui

Metodologia

Modelos implementados

- Regressão Logística (baseline + tuning)
- Random Forest (baseline + tuning)
- XGBoost (baseline + tuning + threshold CV)

Validação

- Treino/teste
- GridSearchCV
- Métricas: F1 classe letal, Recall, ROC-AUC, PR-AUC

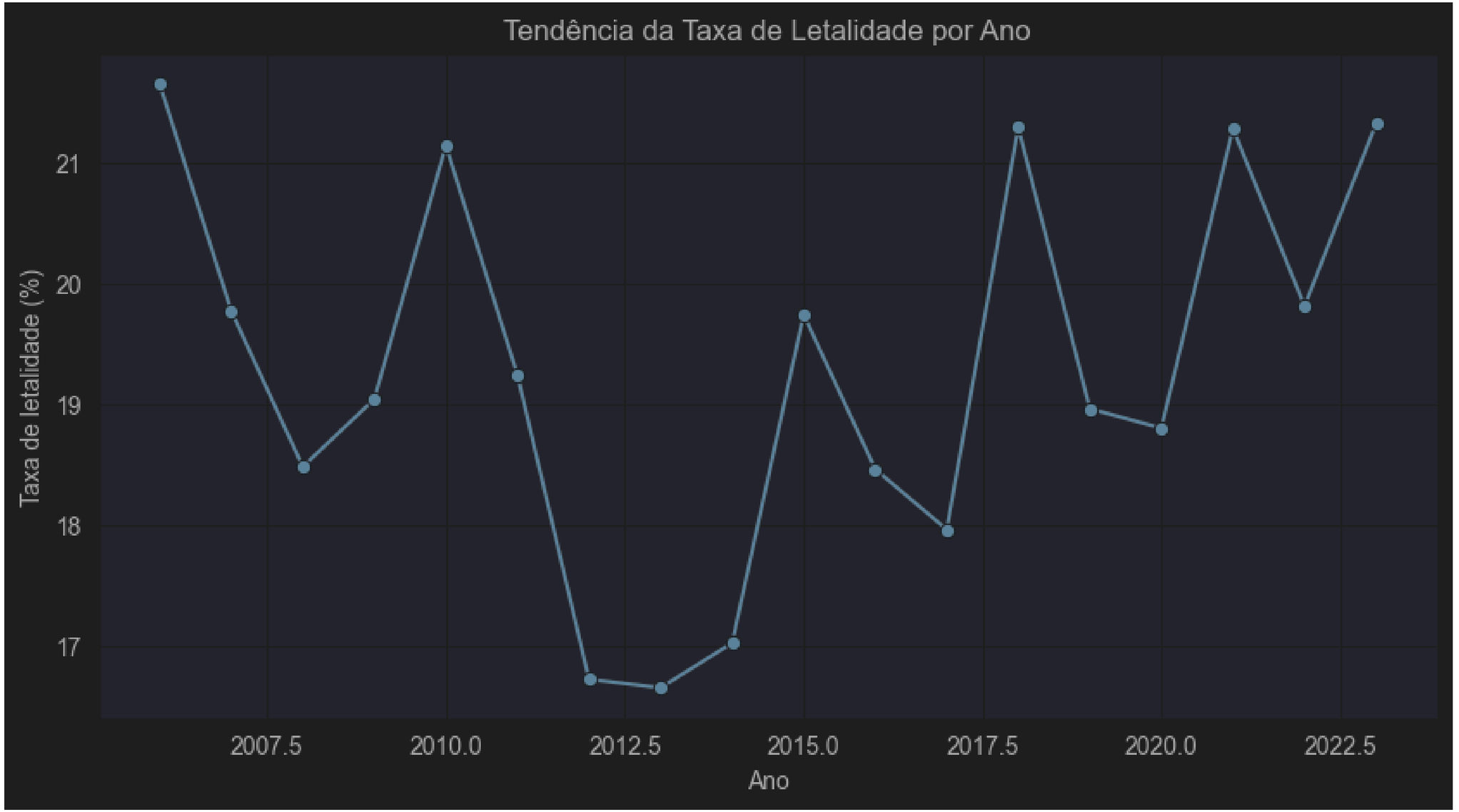
Resultados principais

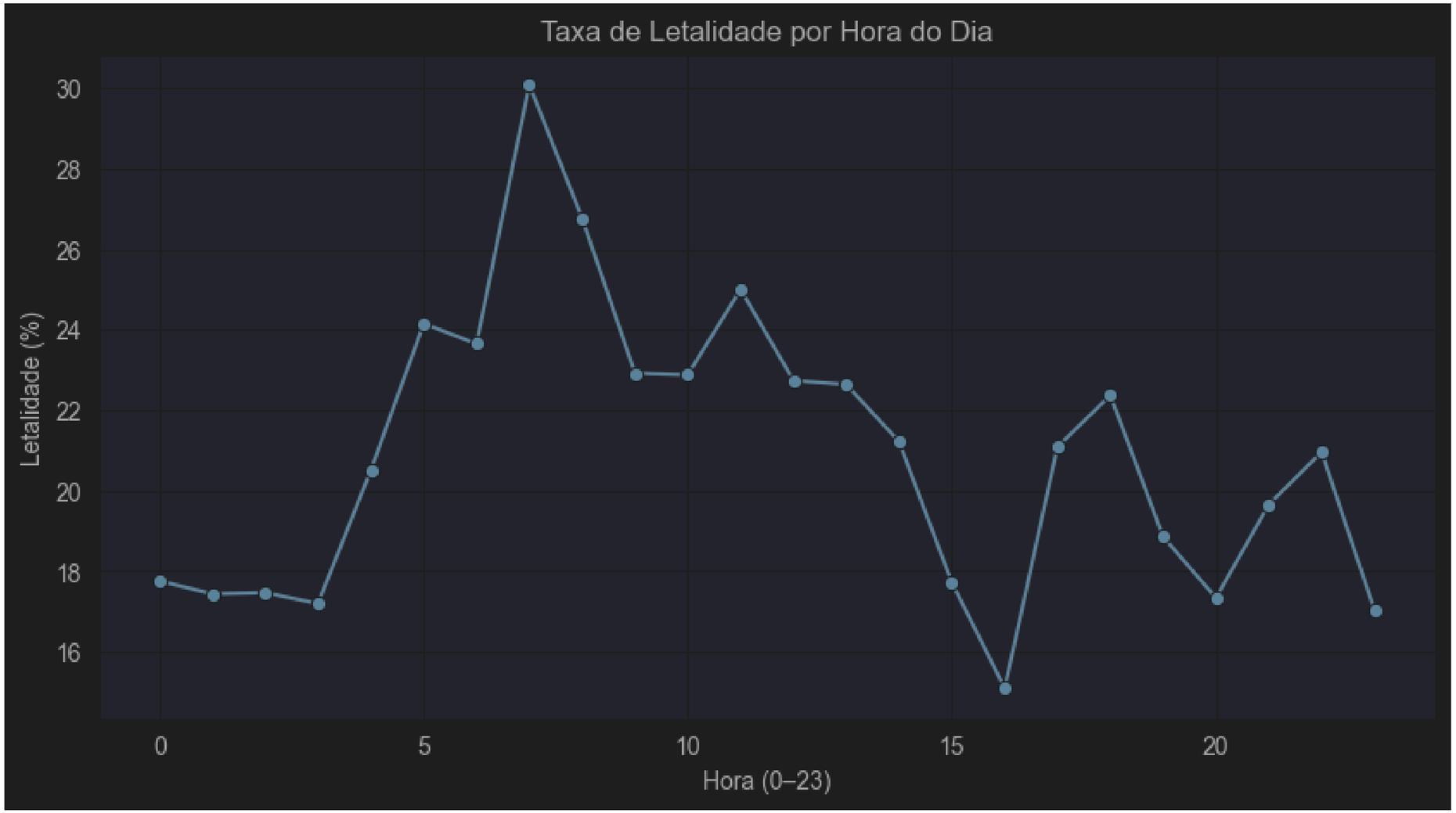
Achados exploratórios

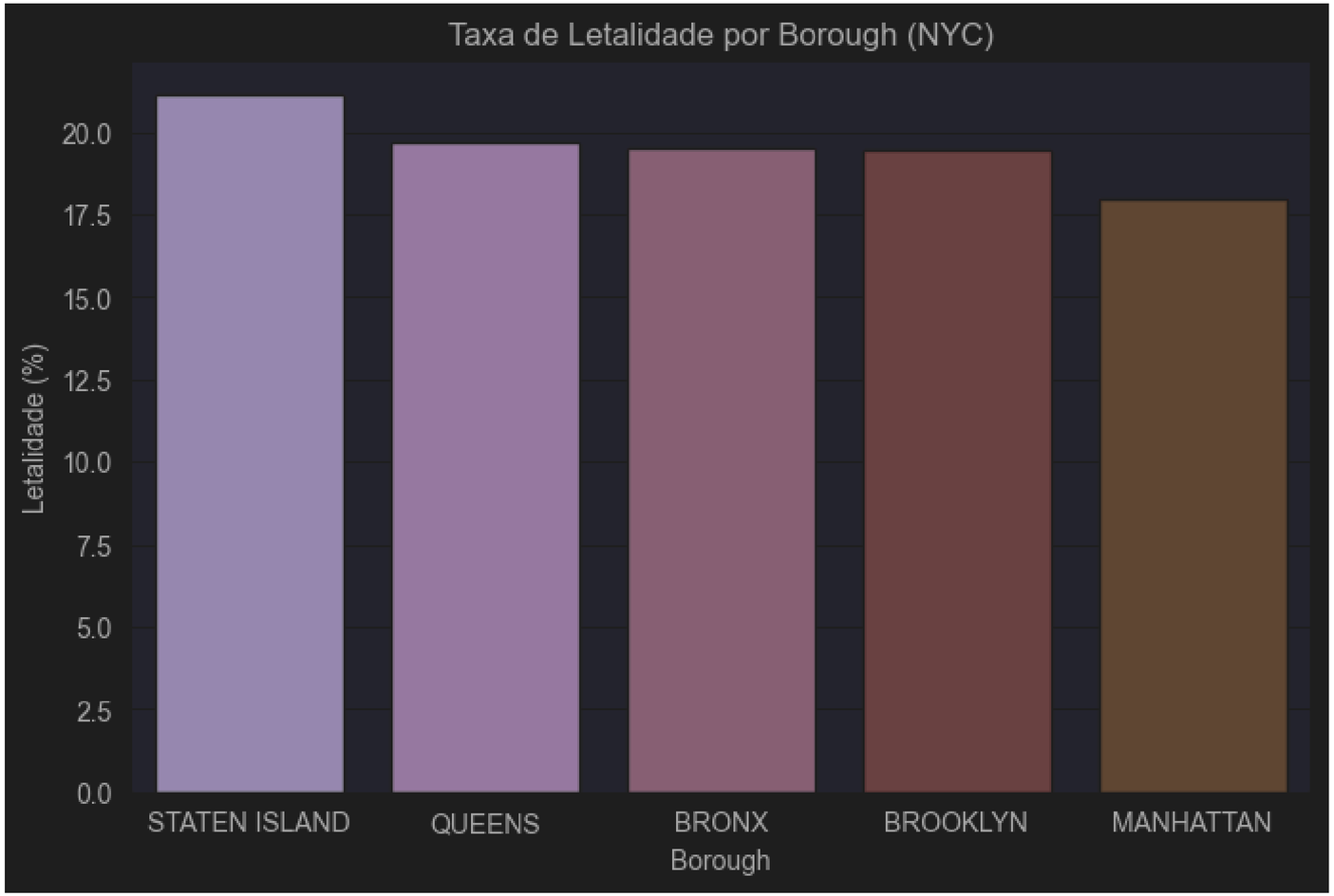
- Letalidade estável (17%–22%) ao longo dos anos.
- Picos de letalidade pela manhã.
- Boroughs parecidos → não significativos.
- Precincts muito diferentes → significativos.
- Vítimas mais velhas têm maior risco.
- Sexo e raça do agressor também significam associação.

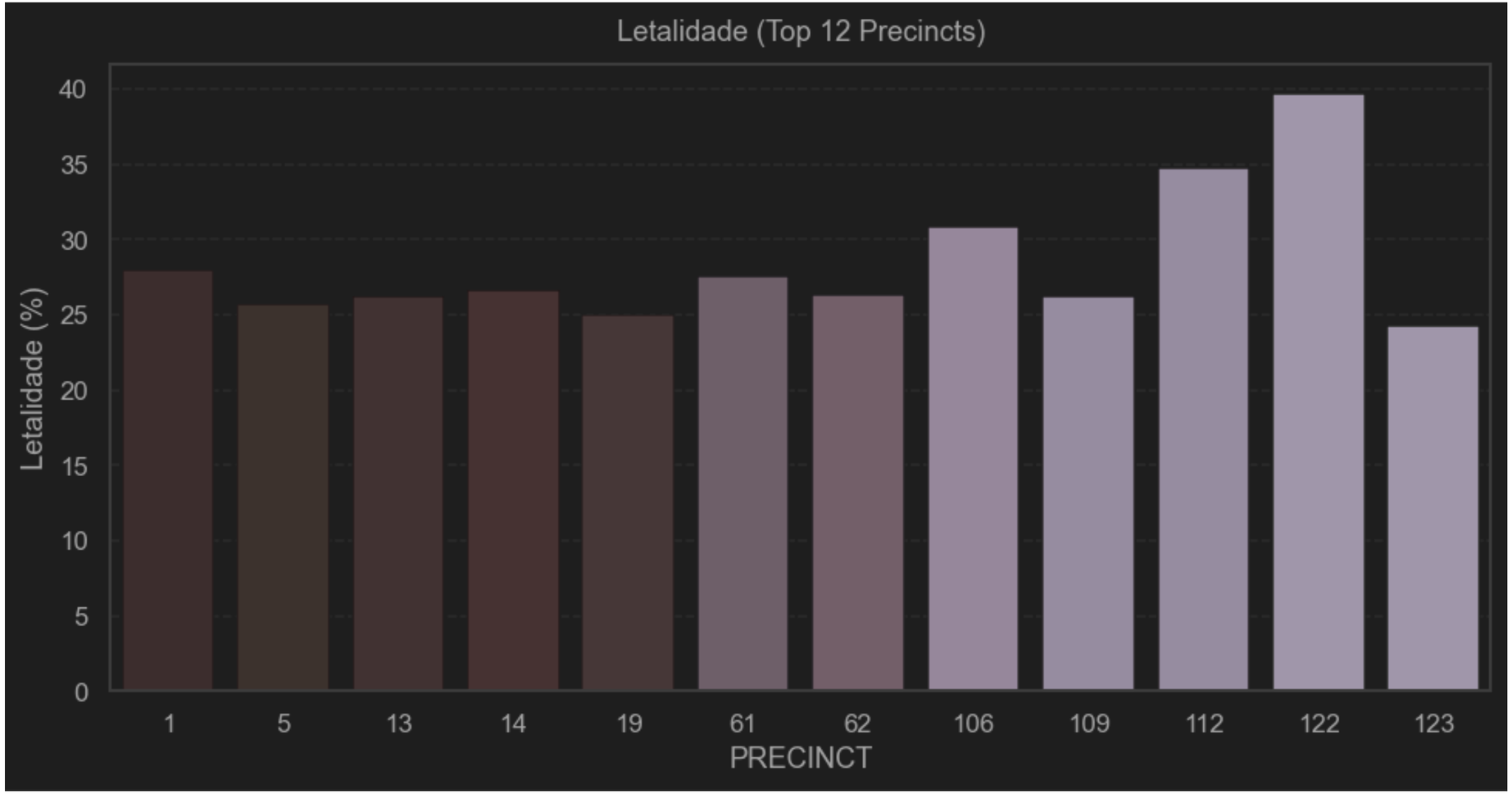
Testes estatísticos — resumo

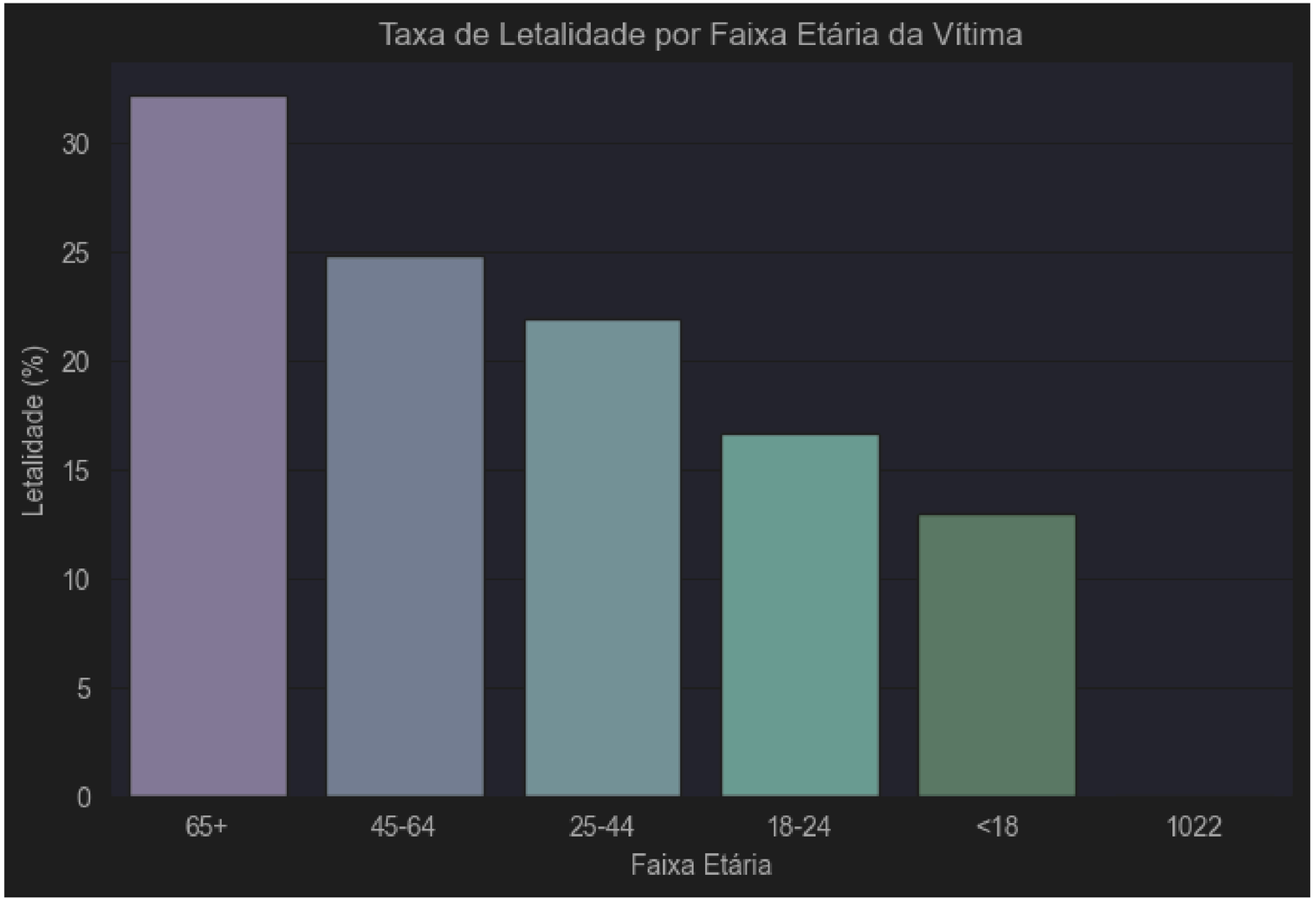
- Significativos: período do dia, precinct, idade da vítima, raça da vítima, sexo do agressor, raça do agressor.
- Não significativos: borough e sexo da vítima.

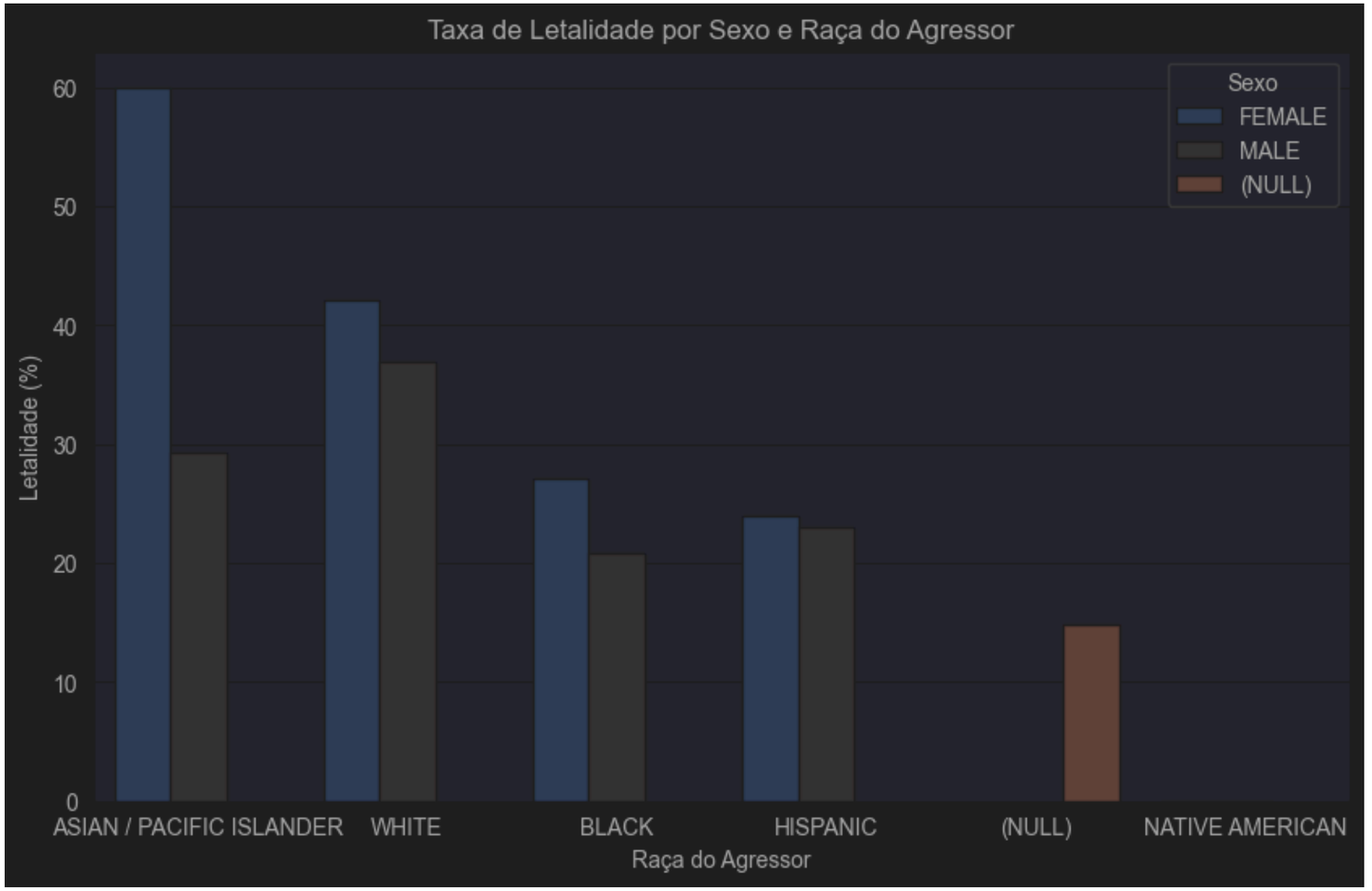

Tendência da Taxa de Letalidade por Ano
Taxa de letalidade (%)
21
20
19
18
17
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5
Ano


Taxa de Letalidade por Hora do Dia
Letalidade (%)
30
28
26
24
22
20
18
16
0
5
10
15
20
Hora (0–23)


Taxa de Letalidade por Borough (NYC)
Letalidade (%)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
STATEN ISLAND
QUEENS
BRONX
BROOKLYN
MANHATTAN
Borough


Letalidade (Top 12 Precincts)
Letalidade (%)
40
35
30
25
20
15
10
5
0
1
5
13
14
19
61
62
106
109
112
122
123
PRECINCT


Taxa de Letalidade por Faixa Etária da Vítima
Letalidade (%)
0
5
10
15
20
25
30
65+
45-64
25-44
18-24
<18
1022
Faixa Etária


Taxa de Letalidade por Sexo e Raça do Agressor
Sexo
FEMALE
MALE
(NULL)
Letalidade (%)
0
10
20
30
40
50
60
ASIAN / PACIFIC ISLANDER
WHITE
BLACK
HISPANIC
(NULL)
NATIVE AMERICAN
Raça do Agressor

Resultados principais

Modelagem

Modelo	Precision (classe 1)	Recall (classe 1)	F1-score (classe 1)	ROC-AUC	PR-AUC
LogReg (baseline)	0.45	0.09	0.15	0.68	NaN
LogReg + UnderSampling + Tuning	0.28	0.62	0.39	0.67	NaN
Random Forest (baseline)	0.15	0.04	0.07	0.64	NaN
Random Forest + Tuning	0.28	0.63	0.38	0.66	0.30
XGBoost (baseline)	0.48	0.06	0.10	0.68	NaN
XGBoost + Tuning + Threshold CV	0.28	0.59	0.38	0.67	0.33

Interpretação

- Letalidade não é aleatória.
- Há padrões aprendíveis (apesar das limitações do dataset).
- Modelos baseados em árvores confirmam H3 parcialmente.

Recomendações

Territorialização de políticas

- Precincts de maior risco → prioridade de patrulhamento e ações sociais.

Grupos vulneráveis

- Vítimas mais idosas
- Comunidades com maior concentração de fatores associados

Uso futuro de modelos

- Ferramenta complementar a equipes de inteligência.
- Identificação de áreas críticas.
- Suporte à análise, não ao policiamento automatizado.

Lições aprendidas

O que funcionou?

- Pipeline consistente
- Testes estatísticos claros
- Tuning e undersampling melhoraram muito o desempenho
- XGBoost foi o melhor modelo

O que não funcionou?

- Forte desbalanceamento da variável alvo
- Modelos lineares não capturam as relações complexas
- Variáveis disponíveis limitam previsões

Lições aprendidas

Próximos passos

- Incluir dados socioeconômicos
- Testar SMOTE e outros métodos
- Avaliar modelos probabilísticos e deep learning
- Estudos comparativos com outras cidades

Referências

CITY OF NEW YORK. NYPD Shooting Incident Data (Historic). Disponível em: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8. Acesso em: 15 out. 2025.

CHEN, Tianqi; HE, Tong; BENESTY, Michael; KHOTILOVICH, Vadim; TANG, Yuan; CHO, Hyunsu; CHEN, Kailong; MITCHELL, Rory; CANO, Ignacio; ZHOU, Tianyi; LI, Mu; XIE, Junyuan; LIN, Min; GENG, Yifeng; LI, Yutian; YUAN, Jiaming; CORTES, David; et al. xgboost: Extreme Gradient Boosting — Python package. versão 3.2.0.0, 2025. Disponível em: https://xgboost.readthedocs.io/en/stable/python/python_intro.html. Acesso em: 26 nov. 2025.

SCIKIT-LEARN developers. RandomForestClassifier — scikit-learn 1.7.2 documentation. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acesso em: 25 nov. 2025.

SCIKIT-LEARN developers. LogisticRegression — scikit-learn documentation. Disponível em: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression. Acesso em: 25 nov. 2025.