Revista EduFatec: educação, tecnologia e gestão
V.2 N.5 – agosto-dezembro/2022

PREDIÇÃO DE SÉRIES TEMPORAIS PARA A TOMADA DE DECISÕES SOBRE POLÍTICAS PÚBLICAS DE SEGURANÇA NO ESTADO DE SÃO PAULO.

Clayton Suguio Hida[1]

Gilberto Sussumu Hida[2]

Resumo

As taxas de criminalidade no Brasil seguem patamares elevados, com o país ocupando a 22° posição no Índice Global do Crime Organizado (GI-TOC) de países de maior índice de criminalidade organizada. No entanto, essas taxas, apesar de elevadas, vem caindo nas últimas décadas. Esse fato poderia ser em parte explicado pelo crescente uso de dados e da inteligência artificial na tomada de decisão. Nesta direção, o presente artigo visa a construção de uma solução analítica para o auxílio na tomada de decisão no planejamento de políticas públicas de segurança do estado de São Paulo. Demonstramos a construção de uma solução analítica em forma de painel, que faz a captura, tratamento e a modelagem de séries temporais para predição das ocorrências criminais no estado de São Paulo. A ferramenta poderia fornecer informações úteis para o planejamento de políticas públicas de segurança do Estado. Em particular, poderia ser implementado com dados em tempo real, o que possibilitaria o planejamento logístico de equipes táticas.

Palavras-chave: Criminalidade. *Machine learning*. Tomada de decisão.

Abstract

Crime rates in Brazil remain at high levels, with the country occupying the 22nd position in the Global Index of Organized Crime (GI-TOC), of countries with the highest rate of organized crime. However, these rates, although high, have been falling in recent decades. This fact could be partly explained by the growing use of data and artificial intelligence in decision making. Thus, this article aims to build an analytical solution to aid in decision making in the planning of public security policies in the state of São Paulo. We demonstrate the construction of an analytical solution in the form of a dashboard, which captures, treats and models time series to predict criminal occurrences in the state of São Paulo. The tool could provide useful information for planning public security policies in the State. In particular, it could be implemented with real-time data, which would make logistical planning of tactical teams possible.

Keywords: *Criminality. Machine learning. Decision making.*

1 Introdução

[1] Professor da Fatec Praia Grande. Endereço eletrônico: clayton.hida@fatec.sp.gov.br.
[2] Mestrando em Matemática Aplicada - IME - USP. Endereço eletrônico: sussumu@ime.usp.br.

 ISSN: 2595-5942

O uso de métodos de *machine learning* para auxílio na tomada de decisão é algo que vem crescendo nos últimos anos, principalmente devido à redução dos custos de armazenamento de dados e também na melhoria da capacidade de processamento de grande quantidade de dados.

Métodos de inteligência artificial são utilizados em uma ampla gama de atividades da sociedade. Na indústria temos o uso de algoritmos de *machine learning* para controle das rotas nas entregas dos produtos; na medicina, temos o uso de aprendizado profundo em imagens médicas para a detecção de doenças; no marketing, temos o uso de redes neurais para a sugestão de produtos aos consumidores.

O uso de tais métodos no primeiro setor se faz necessário principalmente devido à escassez de recursos. Portanto, o uso racional e baseado em dados é uma necessidade que impacta a todos da sociedade. Podemos citar como exemplo a iniciativa do Ministério da Saúde com o Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde (PROADI – SUS), que desenvolve em um dos projetos ferramentas de inteligência artificial para suporte a tomada de decisão (PROADI, 2021).

Segundo Columbus (2020), o mercado de uso de ferramentas e gastos com *machine learning* subirá de $1.58 bilhões em 2017 para $20.83 bilhões em 2024.

Assim, a utilização dos métodos de *machine learning* quando aplicados adequadamente e respeitando as boas práticas podem gerar ótimos resultados.

Os custos da violência no Brasil são estimados em R$ 373 bilhões, aproximadamente 6% do PIB em 2016 (ATLAS, 2022). Assim o uso de *machine learning* pode contribuir de maneira decisiva, seja na redução dos gastos, bem como no melhor uso dos recursos para o combate à violência.

O projeto visa construir modelos de séries temporais para as quantidades de ocorrências criminais no estado de São Paulo. Os dados são da Secretária de Segurança Pública do Estado de São Paulo (SSP-SP), e estão disponíveis para consulta pública (SSP-SP, 2021), conforme a política de acesso à informação (LEI Nº 12.527). O objetivo é construir uma solução analítica que faça a captura, tratamento e apresentação dos resultados, constituindo assim uma ferramenta de suporte ao planejamento ostensivo que é praticado pela polícia militar do Estado. As projeções dos próximos trimestres permitem o planejamento da logística das equipes que compõem o policiamento das regiões.

Os dados são as quantidades de ocorrências policiais por trimestre desde o terceiro trimestre de 1995 até o quarto trimestre de 2021, das regiões da capital, interior e grande São Paulo. Devido as mudanças na apresentação dos dados, optamos por acompanhar as seguintes ocorrências: “total de delitos”, “homicídio doloso”, “homicídio culposo”, “latrocínio”, “estupro”, “roubo” e “furto”.

2 Materiais e métodos

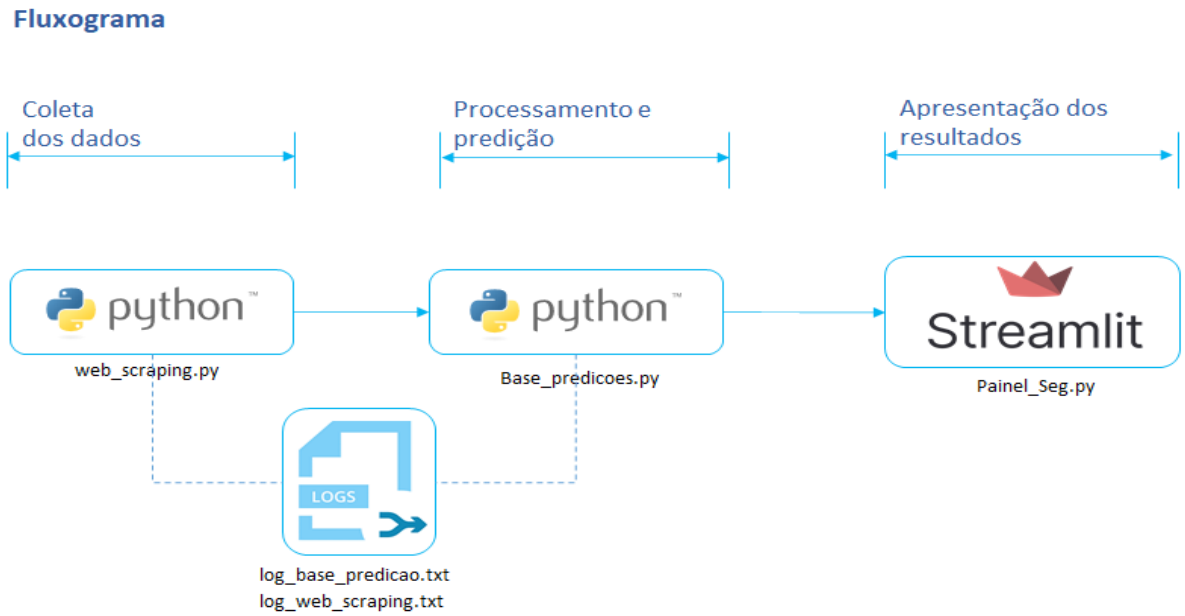
Antes de entrarmos nos detalhes do projeto, vamos apresentar a estrutura geral do projeto. Como se trata de uma solução analítica e não apenas de uma análise pontual é necessário que seja desenhado um fluxo do processo, desde a captura dos dados até a apresentação. De forma macro temos quatro blocos fundamentais:

1. Coleta de dados: Corresponde à *web scraping* que realiza a consulta ao site da SSP-SP coletando os dados de interesse.
2. Processamento e predição: Etapa que padroniza algumas variáveis (tipos de dados, *lower, upper case*) e gera as predições. Optamos por gerar as predições antes da etapa de apresentação dos dados pois o tempo de reprocessamento dos modelos seria bem maior caso optássemos pela construção da camada de modelagem junto com a visualização.
3. Apresentação dos resultados: Etapa final que corresponde a visualização tanto da parte descritiva quanto das predições e do acompanhamento da modelagem.
4. Logs: Implementamos uma camada de logs de execução das etapas 1 e 2. O uso de logs em soluções analíticas auxilia na busca de erros e na implantação de melhorias no projeto.

   A Figura 1 mostra o fluxograma do projeto:

ISSN: 2595-5942

Figura 1 - Fluxograma do projeto**.**



Fonte: Próprio autor (2022).

Vamos descrever de forma mais detalhada as etapas de coleta de dados, análise e exploração dos dados e modelagem e predição.

Coleta de dados: Corresponde ao uso da técnica de *web-scrapping*, que realiza a consulta ao site da SSP-SP coletando os dados de interesse.

Os dados foram extraídos do site da Secretaria de Segurança Pública de SP (SSP-SP, 2021). Desenvolvemos um *web-scrapping* em Python para capturar os dados de maneira automática.

Na página da SSP-SP, os dados estão organizados por trimestre, sendo que, cada trimestre possui um link para acesso. As tabelas apresentam o aspecto apresentado na Figura 2:

Figura 2 - Imagem da tabela da SSP-SP: 1º trimestre de 2022.

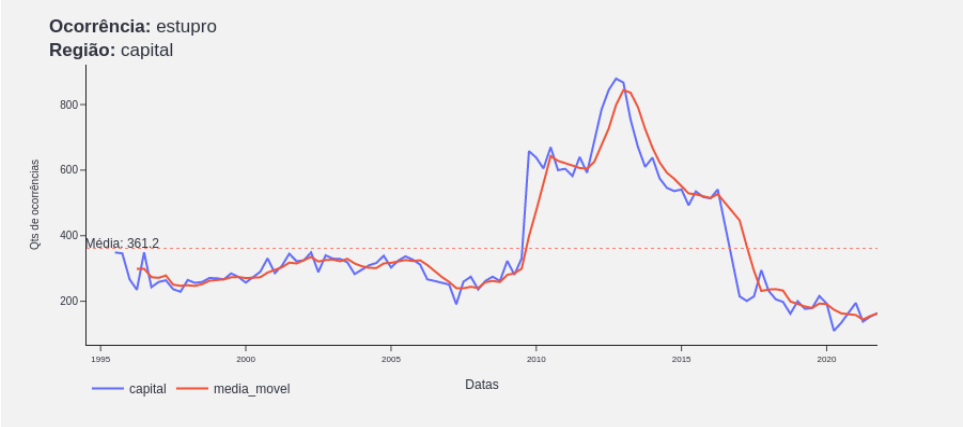
ITEM	Ocorrências policiais registradas, por natureza	Capital	Gde SP(1)	Interior
	Contra a pessoa	25.883	19.453	73.633
	Contra o patrimônio	139.780	64.811	143.969
I	Contra a Dignidade Sexual	932	839	2.842
	Entorpecentes	1.192	1.211	7.260
	Contravencionais	1.545	1.079	6.901
	Outros criminais (não inclui contravenções)	7.939	5.118	21.211
	Total de Crimes Violentos (Hom.Doloso, Roubo, Latrocínio, Estupro e EMS)	38.933	17.136	16.725
	Total de delitos	177.271	92.511	255.816
	Não Criminais	130.002	79.681	218.629

Fonte: SSP-SP (2022).

Fato importante é que devido ao longo período de coleta (1995-2020) algumas ocorrências apresentaram mudanças de nomenclatura, ou deixaram de ser monitoradas, ou foram agrupadas. Optamos por selecionar as ocorrências que apresentaram coletas de dados durante todo o período da análise.

Processamento e predição: Na primeira etapa da análise exploratória, olhamos a questão da qualidade do preenchimento e de possíveis erros na coleta dos dados. Logo em seguida, passamos para a visualização das séries. Como temos três regiões e 7 tipos de ocorrências, teremos 21 séries temporais. Assim, o uso de gráficos interativos foi necessário. Utilizamos o Plotly (PLOTLY, 2022), que consiste em uma biblioteca multiplataforma (R, Python, Java) usado para a construção de gráficos em painéis. A Figura 3 mostra as ocorrências de estupro e a curva de média móvel.

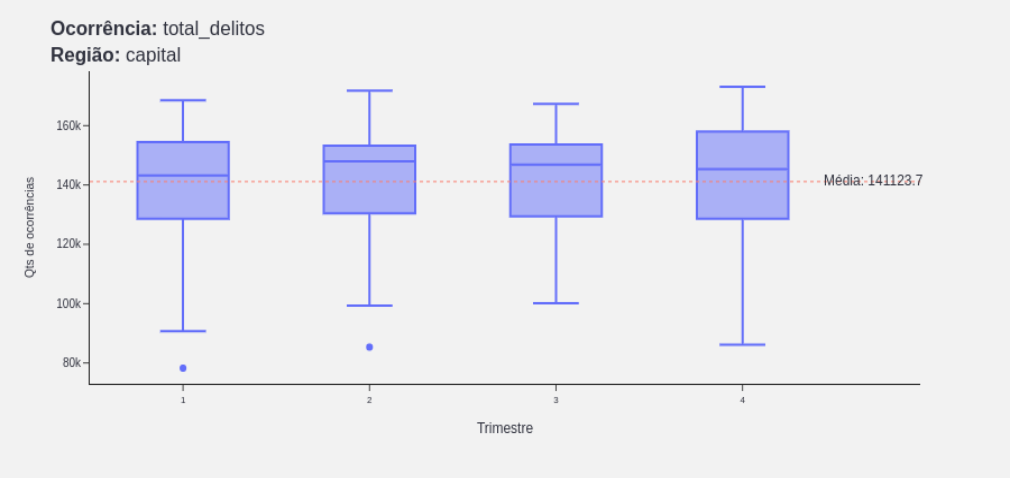
Figura 3 - Região - Capital. Ocorrência: Estupro.



Fonte: Próprio autor (2022).

Outra visualização importante para a série temporal é a observação de componentes cíclicos. No nosso caso, a nossa unidade de tempo é trimestral, sendo assim buscamos detectar possíveis comportamentos de queda ou subida ao longo de cada trimestre. Por exemplo, a Figura 4 mostra os *boxplots* trimestrais da série total de delitos:

Figura 4 - Detecção de comportamento cíclico na série.



Fonte: Próprio autor (2022).

No caso da região da capital, na ocorrência total de delitos notamos ausência de comportamento cíclico na série (observamos que os *boxplots* de cada trimestre apresentam tamanhos próximos e estão situados na mesma faixa horizontal entre 125 mil e 160 mil).

Modelagem e predição: A modelagem de séries temporais requer cuidados no sentido de olhar certas características que compõem a estrutura da série (MORETTIN, 2006). Usamos a decomposição da série temporal baseada na metodologia Box&Jenkins que decompõe a série temporal nas parcelas de sazonalidade, tendência, ciclo e componente aleatório.

A análise descritiva nos forneceu informações se as séries apresentam ou não tais componentes. Em resumo, o gráfico de evolução ao longo do tempo constitui uma ferramenta visual da presença de tendência na série, os *boxplots* por trimestre (como apresentado na Figura 4) respondem de maneira visual se a série apresenta componente cíclico (sazonalidade é mais um conceito usado para comportamento mensal, como os dados são trimestrais, não faz muito sentido falar em componente sazonal). Além disso, notamos que todas as séries são homocedásticas, ou seja, a variância é constante ao longo do tempo. Existem testes estatísticos para detecção de tendência como ADF, KPSS (FERREIRA, 2018), testes para comportamento cíclico são comparações de médias ou dependendo da distribuição testes de mediana (MORETTIN, 2006).

Adotamos o modelo usado pelo Facebook para séries temporais, o Fbprophet (FBPROPHET, 2021). Tal *framework* usa a decomposição de Box&Jenkins para

modelar a série, incorporando uma componente de eventos baseados em feriados. A vantagem do Fbprophet é a capacidade de ajuste de funções não lineares, múltiplas componentes cíclicas, indicação de pontos de quebra estrutural da série entre muitos outros parâmetros para controle do aprendizado.

Para a predição dos valores finais realizamos o treinamento para cada série temporal levando em conta a presença ou não de quebras estruturais na série. Assim, se a série apresenta quebra de estrutura, apesar do *framework* do Facebook permitir a incorporação de quebras, optamos por selecionar explicitamente a fatia da série que seria usada para o treino.

Dividimos o estudo da capacidade de generalização do modelo em duas formas: Na primeira forma, usamos a técnica de partição em 80-20, ou seja, 80% dos dados são usados para o processo de modelagem e os restantes 20% são usados para cálculos de métricas. Na segunda forma, usamos a técnica de *sliding window* (BACKTEST, 2021), que consiste em usar uma janela móvel de dados para treino e teste.

Para a primeira forma de avaliação usando split 80-20, temos uma forte semelhança com o método de validação cruzada "*holdout*" (SCHORFHEIDE, 2012) para classificação binária. A única diferença reside no fato que a partição entre treino e teste não é aleatória como no caso de classificação, pois isto destruiria a dependência temporal da série.

Para a segunda forma de avaliação usando o *sliding window*, fatiamos cada série temporal em 7 blocos, para que o volume de dados não fosse pequeno demais no treinamento (para cada série temporal temos 101 registros) agrupamos a partir do segundo passo do método dois blocos para treinamento e o posterior para predição.

Nas duas abordagens, usamos a média percentual absoluta do erro (MAPE), o desvio médio quadrático dos erros (MRSE) e a média absoluta dos erros (MAE). Ao final das duas abordagens de modelagem, obtemos os valores preditos e com os valores reais podemos calcular tais métricas.

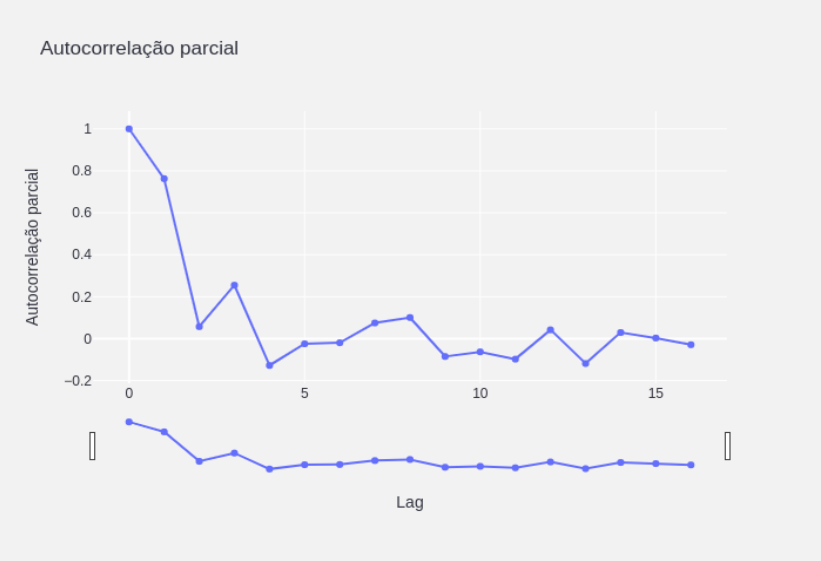
As duas formas de avaliação dos modelos buscam detectar a capacidade de generalização da abordagem (ABU-MOSTAFA, 2012), que no caso de séries temporais pode ser traduzido como a capacidade de captura do comportamento central da série juntamente com a capacidade/flexibilidade de detectar mudanças de comportamento nos dados.

Por fim, entendemos que um ajuste de um modelo de série temporal envolve a análise do resíduo (real-previsto) que deve ter o comportamento de um ruído branco (MORETTIN, 2006), ou seja, seguir o modelo normal com média e desvio padrão constante. Essa etapa pode ser vista dentro do painel para cada processo de avaliação do modelo. A Figura 5 mostra o gráfico qq-plot dos resíduos e a Figura 6 é o gráfico de autocorrelação parcial, ambos obtidos da aba Sumário dos modelos no painel.

Figura 5 - Gráfico de qq-plot para detecção de normalidade dos dados.

Fonte: Próprio autor (2022).

Figura 6 - Plot da autocorrelação em relação às defasagens.

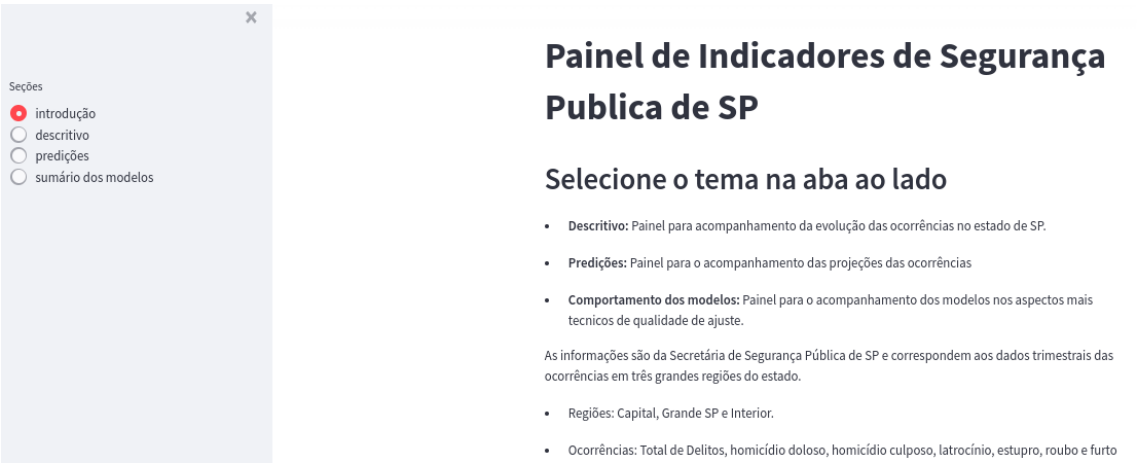


Fonte: Próprio autor (2022).

3 Resultados e discussão

A ferramenta final é um painel (*dashboard*) que permite acompanhar as ocorrências policiais das três regiões de São Paulo tanto do ponto de vista descritivo quanto da predição. Utilizamos a biblioteca Streamlit (STREAMLIT, 2022) para a construção do painel. A Figura 7 abaixo mostra a página inicial de boas-vindas:

Figura 7 - Página inicial do Painel.



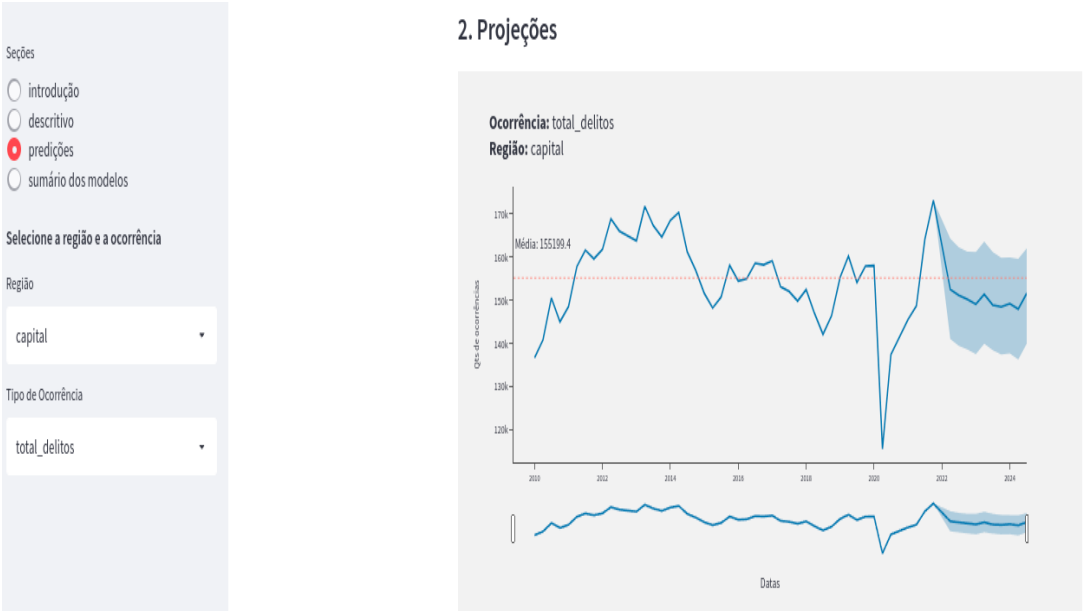
Fonte: Próprio autor (2022).

Na página inicial, podemos escolher a análise descritiva dos dados, como mostra a Figura 8 e as predições, como mostra a Figura 9:

Figura 8 - Aba da análise descritiva.

Fonte: Próprio autor (2022).

Figura 9 - Predições com limites superior e inferior.



Fonte: Próprio autor (2022).

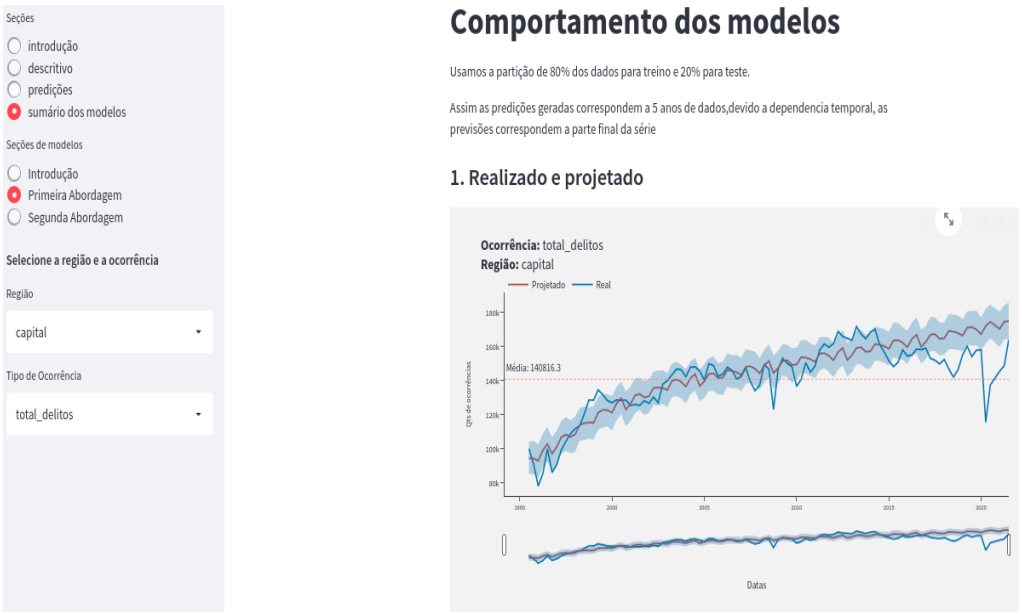
Para o acompanhamento dos modelos, temos a divisão em dois itens, como mostra a Figura 10, lembrando que a Primeira abordagem se refere ao método de *split* 80-20 e a Segunda abordagem refere-se ao método de *sliding window.*

Figura 10 - Acompanhamento da modelagem.

Fonte: Próprio autor (2022).

Selecionando por exemplo a Primeira abordagem, somos levados ao painel mostrado na Figura 11, que primeiramente mostra o valor real da série (em azul) e o valor predito (em vermelho).

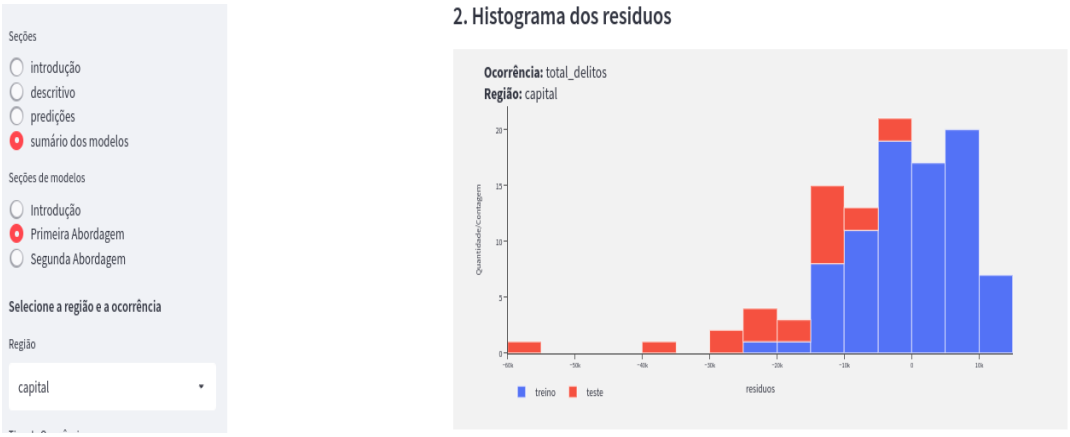
Figura 11 - Primeira abordagem para análise do modelo - Split 80-20.



Fonte: Próprio autor (2022).

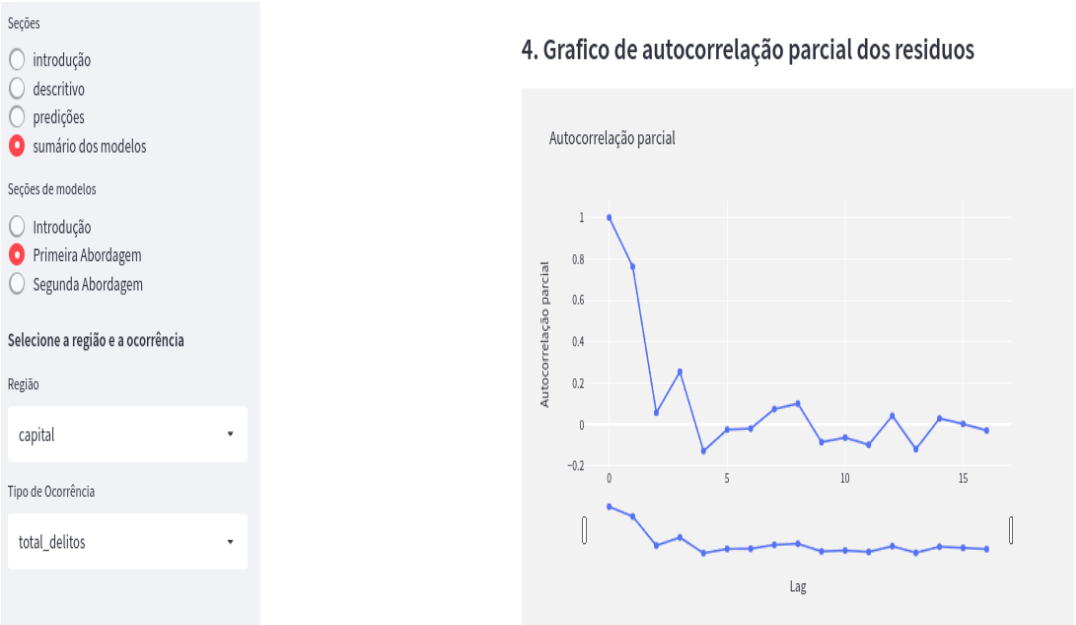
Para acompanhamento da qualidade do ajuste do modelo, podemos visualizar o histograma dos resíduos, como mostra a Figura 12, e a autocorrelação dos resíduos como mostra a Figura 13.

Figura 12 - Histograma dos resíduos e qq-plot.



Fonte: Próprio autor (2022).

Figura 13 - Funções de autocorrelação.



Fonte: Próprio autor (2022).

Para a segunda abordagem, temos painéis com as mesmas funções e *plots*. A Figura 14 mostra a comparação entre realizado e projetado para a segunda abordagem:

Figura 14 - Segunda abordagem - *Sliding windows*.

Fonte: Próprio autor (2022).

Para estabelecer os resultados analíticos das duas abordagens, o dashboard conta com a visualização de tabelas com as medidas estatísticas. Abaixo um exemplo com as principais medidas estatísticas:

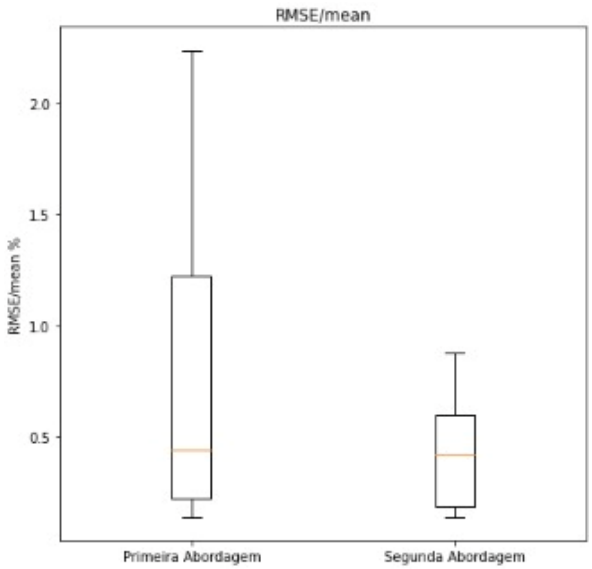
Figura 15 - Tabela obtida do Dashboard.

	regiao	ocorrencia_padro	media_teste	rmse_teste	mae_teste	mape_teste	rmse_mean_teste	mae_mean_teste
0	capital	total_delitos	140,816.2885	19,668.1800	15,724.1200	0.1100	0.1400	0.1100
1	capital	homicidio_doloso	612.2212	255.9600	171.1900	0.4300	0.4200	0.2800
2	capital	homicidio_culposo	201.6842	105.0200	61.9300	0.3100	0.5200	0.3100
3	capital	latrocinio	33.3654	22.8500	18.8800	0.8900	0.6800	0.5700
4	capital	estupro	367.5200	322.1300	216.8700	0.8100	0.8800	0.5900
5	capital	roubo	28,443.1579	4,723.6900	3,527.6700	0.1200	0.1700	0.1200
6	capital	furto	39,605.2212	8,008.5000	6,724.5000	0.1700	0.2000	0.1700

Fonte: Próprio autor (2022).

Notamos que o uso da segunda abordagem produz erros menores quando comparado com a abordagem clássica de *split* 80-20, como mostra a Figura 16. Esse resultado era esperado, uma vez que ao fatiar a série, permitimos que os modelos gerados possam captar mudanças de comportamento estrutural da série, ao rodar apenas um único modelo para a série completa o modelo tende a equilibrar possíveis mudanças de comportamento para obter uma média de custo baixa ao longo de todo o treinamento.

Figura 16. Comparativo das abordagens.

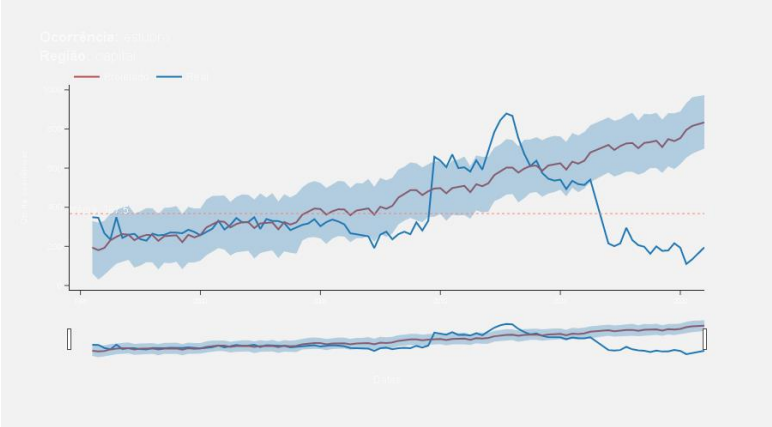


Fonte: Próprio autor (2022).

Observamos que a segunda abordagem foi eficiente na modelagem dos totais de delitos da região do Grande ABC, com média absoluta percentual (MAPE) de 8%, e na modelagem do total de furtos na região interior, com MAPE também de 8%.

No entanto, a segunda abordagem não foi capaz de modelar a mudança abrupta de algumas séries, por exemplo, a série ocorrência de estupros. Isso pode ser observado pela Figura 17, com o comparativo real e predito, bem como pelo alto valor do MAPE que foi de 0.81.

Figura 17. Real versus predito.



Fonte: Próprio autor (2022).

Como versões aprimoradas da ferramenta, poderíamos implementar outros modelos de *machine learning* para a modelagem das séries. Por exemplo, modelos de redes neuras recorrentes, como o *Long Short Term Memory* (LSTM).

Considerações finais

O uso de métodos de *machine learning* para suporte a tomada de decisão faz mais sentido ainda em um ambiente de recursos escassos que é o caso das políticas públicas de segurança.

O projeto desenvolvido poderia ser usado para o planejamento das rondas ostensivas que são realizadas pela polícia militar do Estado. Claramente há a necessidade de aumentar o nível de granularidade dos dados, no que diz respeito às regiões, permitindo o acompanhamento das ocorrências por cidade e por bairro.

Do ponto de vista de *deploy* (implementação), a solução foi desenvolvida seguindo as boas práticas de desenvolvimento (PRESSMAN, 2021) usando ambientes de desenvolvimento e camada de logs permitindo assim o uso de maneira contínua. O uso de orquestradores como o Apache Airflow ou Microsoft Azure Pipelines poderiam ser implementado adicionando um *Docker* básico para a sustentação da operação e também para que não haja interferências.

A adoção do Fbprophet como solução para a modelagem foi interessante por incorporar a presença de componentes cíclicos diferentes da trimestral, pois os dados possuíam período de 3 meses (dados trimestrais). Observamos ainda que não se trata totalmente de um modelo de caixa preta, pois poderíamos reproduzir o método do Fbprophet desenvolvendo os dados de forma tabular em que cada coluna seria a série temporal diferenciada em várias ordens mais variáveis *dummies* sobre outras características, assim uma regressão corresponderia ao Fbprophet com ajuste linear.

Referências

ABU-MOSTAFA, Y. S.; ISMAIL, M. M.; LIN, H. T. **Learning from data**. New York. Amlbook, 2012.

ATLAS DA VIOLÊNCIA. **Estudo: custo da violência equivale a percentual do PIB gasto com educação.** Disponível em <https://www.ipea.gov.br/atlasviolencia/noticia/57/estudo-custo-da-violencia-equivale-a-percentual-do-pib-gasto-com-educacao>. Acesso em: 10 set. 2022.

BACKTESTING. **Cross validation for time series**. Disponível em <https://www.kaggle.com/cworsnup/backtesting-cross-validation-for-timeseries>. Acesso em: 2 set. 2022.

COLUMBUS, L. **Roundup of Machine Learning Forecasts and Market Estimates**, 2020. Disponível em <https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/?sh=6b3019795c02>. Acesso em: 15 set. 2022.

DOCSTRINGS, **Numpy**. Disponível em <https://numpydoc.readthedocs.io/en/latest/format.html>. Acesso em: 20 Ago. 2022.

FBPROPHET. **Quick Start**. Disponível em <https://facebook.github.io/prophet/docs/quick_start.html>. Acesso em: 10 set. 2022.

FERREIRA, P. **Análise de séries temporais em R: Curso introdutório.** São Paulo, SP: GEN Atlas, 2018.

GI-TOC. **Global Organized Crime Index.** Disponível em < https://ocindex.net/ >. Acesso em: 25 out. 2022.

MORETTIN, P. A.; TOLOI, M.C. **Análise de Séries Temporais**. São Paulo. Blucher. 2006.

PLOTLY. Disponível em <https://plotly.com>. Acesso em: 28 ago. 2022.

PRESSMAN, R. S.; BRUCE R. M. **Engenharia de software.** McGraw Hill Brasil, 2021.

PROADI–SUS. **Conheça nossos projetos do Triênio 2021-2023**. Disponível em <https://hospitais.proadi-sus.org.br>. Acesso em 10 mar. 2021.

SCHORFHEIDE, F.; WOLPIN, K.I. On the Use of Holdout Samples for Model Selection. **The Amercain Economic Review**, v.102, n. 3, p.477-481, 2012.

SSP-SP. **Secretaria de Estado da Segurança Pública - Governo do Estado de São Paulo**. Disponível em <https://www.ssp.sp.gov.br/Estatistica/Trimestrais.aspx>. Acesso em: 29 mar. 2021.

STREAMLIT. Disponível em <https://streamlit.io/>. Acesso em: 20 ago. 2022.