SCi-XP: uma Ferramenta para Previsão Séries Temporais utilizando Técnicas de Inteligência Artificial.

SCi-XP: a Tool for Time Series Forecasting using Artificial Intelligence Techniques.

Gabriela Mota de Lacerda Padilha Schettini [1,2] https://orcid.org/0000-0003-3103-5794

Paulo Salgado Gomes de Mattos Neto [3] https://orcid.org/0000-0002-2396-7973

[1] Pós-graduação em Inteligência Artificial Aplicada, Escola Politécnica de Pernambuco, Pernambuco, Brasil,
[2] Mestrado em Inteligência Computacional, Universidade Federal de Pernambuco, Pernambuco, Brasil,
[3] Centro de Informática, Universidade de Pernambuco, Recife, Brasil.

E-mail do autor principal: Gabriela M. L. P. Schettini gabriela.lacerda@poli.br

Resumo

Diversas ferramentas de extração de informações de dados baseadas em Inteligência Artificial (IA) têm sido desenvolvidas. Contudo, a necessidade de programar ou de prototipar um código para um teste rápido faz com que pesquisadores com pouco, ou nenhum conhecimento em programação evitem utilizar técnicas de Inteligência Artificial. Este trabalho propõe uma ferramenta simples em formato de website com o objetivo de introduzir conceitos e modelos de Inteligência Artificial na tarefa de previsão de séries temporais a pessoas com pouco conhecimento em programação. A ferramenta visa tornar a IA uma área mais atraente para pesquisadores de outros ramos do conhecimento.

Palavras-Chave: Previsão de Séries Temporais; Ferramenta; Aprendizado de Máquina.

Abstract

In this fast development world it is common to hear in academic environments about new tools for extracting information out of data. Unfortunately these tools that exists are made for people with previous knowledge of computer programming or coding, and in some way it excludes a little science from the practical field. Given the amount of technical tools for information mining this paper proposes a tool as an attempt of trying to join this both worlds that sometimes seems to be so distant to a single website for prototyping time series forecasting models, to people with little to no knowledge of programming, in instants.

Key-words: *Prototype; Time Series Forecasting; Tool; Machine Learning.*

DOI: 10.25286/repa.v5i2.1339

1 INTRODUÇÃO

A ciência vem progredindo de forma rápida e se aproximando cada vez mais do campo prático, e portanto da indústria, pelos resultados rápidos e cada vez mais precisos em comparação com as formas convencionais de serem feitas. Sabendo disto este trabalho tem por objetivo tentar aproximar pessoas de fora da área da computação, logo sem experiência com programação, com uma ferramenta simples e prática para prototipação de algoritmos baseados em Inteligência Artificial na tarefa de previsão de séries temporais.

A ferramenta é 100% autoral e depende da decisão do usuário para tomar as melhores decisões diante das opções dadas. Esta ferramenta também pode ser utilizada por estudantes curiosos a fim de visualizar, comparar e buscar entender alguns comportamentos das técnicas disponibilizadas através de imagens e métricas típicas da área.

1.1 Motivação

Este trabalho tem como causa a conclusão da especialização em inteligência artificial, e portanto nele será visto aplicações práticas dos conteúdos vivenciados em sala de aula. Tendo como uma das motivações a união do campo teórico com o prático e a difusão dos princípios da Inteligência Artificial e suas mais recentes técnicas na tarefa de previsão de séries temporais.

1.2 Problema

A inteligência artificial modificou bastante a forma com que lidamos com a resolução dos problemas. Antes, para resolver um problema de forma da melhor forma, bastava contratar especialistas. Agora está se tornando comum entregar exemplos de comportamento do que se quer analisar para uma inteligência artificial e obter insights de extrema relevância.

Ainda hoje, saber sobre como as técnicas de inteligência artificial funcionam não é o suficiente para utilizá-las. Parece simples, mas códigos acabam se tornando um empecilho para sequer tentar uma resolução que utilize IA.

Este trabalho baseou na disposição de aproximar pessoas das mais diversas áreas a experimentar uma forma simples e prática da inteligência artificial através de uma ferramenta. A pergunta inicial pensada foi: o quão simples podemos deixar uma ferramenta para que ela se torne acessível para pessoas que não tem muito conhecimento na área e que também não conseguem programar? E a ideia desenvolvida neste trabalho, então, concretizou essa primeira questão implementando uma ferramenta mais simples, no sentido de que: o usuário final não precisa - necessariamente - utilizar a programação e criar as lógicas da Inteligência Artificial para testar a previsão de uma série utilizando IA.

2 FERRAMENTAS RELACIONADAS

Para fins comparativos são expostas nesta seção ferramentas mais conhecidas, bem como suas características.

2.1 Python, Jupyter Notebook & Bibliotecas

Abordagem comum utilizando uma linguagem de programação para desenvolver as soluções com um apelo visual mais agradável do que a linha de comando. Possui diversas bibliotecas que ajudam na visualização dos resultados. Porém ainda não possui versão com interação exclusiva via interface.

2.2 Matlab

Ferramenta com sua linguagem de programação própria, também chamada de Matlab. Muito utilizada por estudantes de engenharia e cientistas [7]. Possui alguns pacotes específicos para o desenvolvimento de resoluções problemas de aprendizado de máquina. Não possui versão com interação exclusiva via interface para soluções em séries temporais.

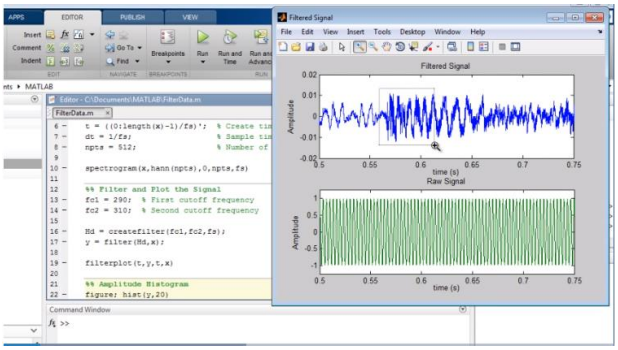

Figura 1: Exemplo da ferramenta em funcionamento.
Fonte: Mathworks: What Is MATLAB? [13]

2.3 R Studio

Ferramenta de cunho mais visual em cima da linguagem de programação R que tem como foco abordagens estatísticas. Não possui versão com interação exclusiva via interface.

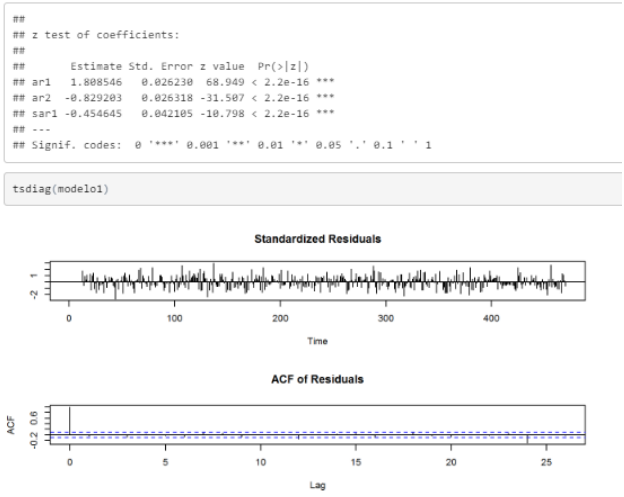

Figura 2: Trecho do código como exemplo da ferramenta em funcionamento.
Fonte: Oliveira, J. A. B. (2016). [6]

3 CONTEXTO

3.1 Definição de previsão de séries temporais

Uma série temporal pode ser entendida como uma sequência de valores □□, observados e registrados, geralmente, de forma equidistante no tempo t (P. J. Brockwell and R. A. Davis), dada por:

$$X_t = \{ x_t \in \Re \mid t = T_0 + n \cdot \Delta t,\ n = 1, 2, \ldots, N \} \quad (1)$$

onde □ é o índice temporal, □$_0$ é o instante inicial das observações, □□ é o intervalo de tempo de coleta entre as observações, e □ é a quantidade de observações obtidas.

A previsão de valores futuros (□$_{□+1}$) de uma série é feita a partir de uma ou mais observações passadas. O objetivo por trás da previsão é identificar padrões temporais nos dados, de forma a construir um modelo capaz de estimar, com precisão, os valores futuros da série.

3.1.1 Análise de Regressão

A análise de regressão visa entender como se relacionam uma ou mais variáveis através do uso de dados históricos. Os dados históricos são utilizados pela crença de que existem relações fortes e possivelmente repetitivas que podem existir com o tempo [2].

3.2 Unidimensionalidade das séries temporais

Um série temporal pode não se limitar a apenas uma observação em cada intervalo de tempo. Por exemplo, um médico querendo monitorar um paciente pode coletar a cada 10 min sua temperatura e pressão arterial. Por tanto 2 observações diferentes para cada instante de tempo observado. Essa série seria considerada uma série temporal bidimensional. Para esta ferramenta iremos apenas utilizar de séries temporais que possuem apenas uma informação para cada intervalo de tempo, portanto uma série temporal unidimensional.

4 FERRAMENTA PROPOSTA

Visando iniciar a experimentação de um novo conceito - aprendizagem de máquina - por um usuário iniciante, a ferramenta proposta é composta de tomadas de decisão imprescindíveis por um cientista de dados. São passos simples e básicos, mas que influenciam diretamente no resultado final da predição. Desta vez a proposta é introduzir uma interação total através de uma interface.

Inicialmente e após o usuário fazer *upload* da base de dados que deseja prever, o mesmo faz escolhas comuns à técnicas utilizadas no campo das previsões de séries temporais da estatística.

Após a etapa de decisão dos parâmetros relativos à série temporal o usuário faz a escolha dos possíveis modelos que ele deseja testar para realizar a previsão. Os modelos pré-programados foram: perceptron de múltiplas camadas (ou MLP), máquinas de vetor de suporte, a técnica conhecida como Random Forest e por fim a técnica conhecida como XGBoost, sendo todas as técnicas adaptadas para previsão.

Por fim o usuário consegue visualizar a série de teste original e previsão em formato de gráfico interativo e o resultado de 3 métricas: MSE, MAE e MAPE.

DOI: 10.25286/repa.v5i2.1339

4.1 Uso da Ferramenta

Alinhado com a proposta de trazer simplicidade à ferramenta, esta possui apenas 3 telas: carregamento da série temporal, escolha dos parâmetros e resultados.

4.1.1 Tela de Carregamento da Série Temporal

Nesta tela será possível carregar a série temporal, onde o formato de entrada deve ser *.csv* e deverá conter no máximo 2 colunas. A primeira coluna será considerada a sequência ordem temporal e a segunda coluna será considerada os dados que se deseja prever. Caso o documento contenha apenas uma coluna, subentende-se que é de dados a serem previstos.

Figura 3: Tela inicial da ferramenta proposta.

4.1.2 Tela de Escolha dos Parâmetros

Após fazer o carregamento dos dados na ferramenta são sugeridos 4 opções de controle sob a previsão a ser realizada: o percentual de treinamento, o tamanho da janela, qual a ocorrência futura que se deseja prever e qual o modelo de previsão se deseja utilizar.

- **Percentual de treinamento:** indica a quantidade relativa de dados que se deseja utilizar para que o modelo tenha capacidade de aprender por si.
- **Tamanho da janela:** Como a série temporal parte do pressuposto da relação dos dados passados para prever o futuro, normalmente é escolhida uma quantidade de dados passados que se deseja fazer relação com o futuro.
- **Qual a ocorrência futura que se deseja prever:** este atributo está medido na unidade em que o dado é inserido, portanto se o dado inserido for semanal, se deseja-se prever a semana seguinte ocorrência futura que se deseja prever é a próxima ocorrência futura, portanto é a primeira (1).
- **Qual o modelo de previsão se deseja utilizar:** O usuário terá a opção de escolher entre 4 possíveis modelos predefinidos: MLP, SVR, Random Forest ou XGBoost. Os modelos serão brevemente explicados na seção 4.2.

Figura 4: Tela de escolhas dos parâmetros da execução da ferramenta proposta.

4.1.3 Tela de Resultados

Na tela de resultados será possível conferir os parâmetros escolhidos na tela anterior e também o resultado de 3 métricas: erro quadrático médio (MSE), erro médio absoluto (MAE) e a média percentual do erro absoluto (MAPE). As métricas serão brevemente explicadas na seção 4.3. Por fim, mas não menos importante, um gráfico interativo e comparativo da série original e da previsão são mostradas nesta última tela.

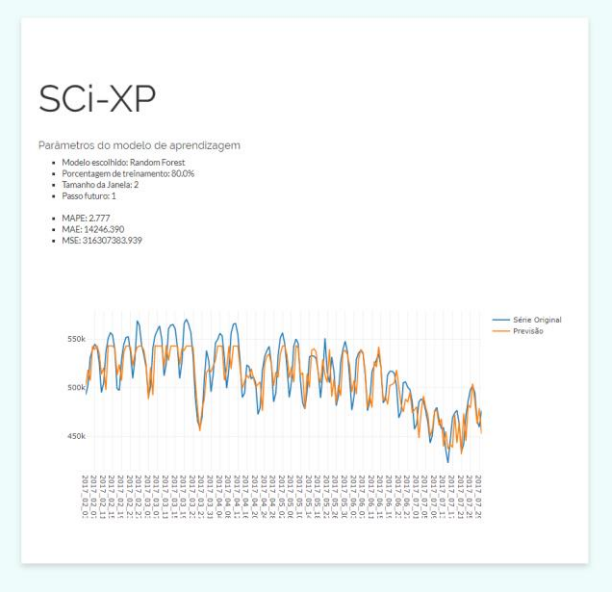

Figura 5: Tela de resultados da ferramenta proposta.

4.2 Modelos Disponíveis na Ferramenta

De forma a aproximar o usuário das técnicas mais utilizadas e outras mais recentes da academia foram disponibilizados na ferramenta 4 modelos de aprendizagem de máquina para a realização da previsão de séries temporais. Sendo a MLP a técnica mais antiga - porém ainda muito utilizada -, SVR que é uma técnica mais robusta do que a MLP e duas outras técnicas muito utilizadas baseadas em árvore de decisão: Random Forest e XGBoost.

4.2.1 Multilayer Perceptron (MLP)

O perceptron de múltiplas camadas é um sistema capaz de resolver problemas além dos linearmente separáveis [3] e que tem como unidade básica o Perceptron [4]. Ao contrário do Perceptron que possui apenas 1 camada e resolve apenas problemas linearmente separáveis, a MLP é uma rede neural artificial que apresenta ao menos 1 camada intermediária[5].

Com passar dos anos o algoritmo dessa estratégia passou por diversas evoluções a partir de proposições de mudança com a pura finalidade de melhorar o seu desempenho [3] e atualmente ainda é muito explorada por ser simples. Ao fazer pequenas adaptações dentro do Perceptron - como na função de ativação, forma de aprendizado ou organização da disposição dos perceptrons e suas camadas-, este modelo conseguiu atingir resultados inacreditáveis com as famosas *Deep Learnings*, e nas mais diversas aplicações como processamento de imagem e vídeo.

O algoritmo utilizado para a versão deste trabalho teve como base a biblioteca para a linguagem de programação Python Scikit Learn v0.21.3, função MLPRegressor com todos os parâmetros originais dessa versão, incluindo o algoritmo de aprendizado.

4.2.2 Regressão por Vetores de Suporte (SVR)

Embora os algoritmos sejam bastante diferentes, a forma de divisão do espaço de soluções da MLP e do SVR linear são bem parecidos. A diferença pode ser descrita com a motivação da criação da Máquina de Vetores de Suporte (SVM) - versão de classificação do SVR - para problemas de classificação binária, onde o algoritmo consegue posicionar um hiperplano de separação de forma a maximizar a distância entre as classes distintas e minimizar o erro de classificação nas proximidades da fronteira de decisão [5].

Após a adaptação deste algoritmo para problemas de resultados reais e contínuos, o SVR se destaca cada dia mais na resolução ótima dos problemas de previsão de séries temporais (citar trabalhos). Essa estratégia ainda pode contar com uma função de base radial para extrapolar as formas de divisão das fronteiras do espaço de decisão.

O algoritmo utilizado para a versão deste trabalho teve como base a biblioteca para a linguagem de programação Python Scikit Learn v0.21.3, função SVR com todos os parâmetros originais dessa versão, incluindo o algoritmo de aprendizado.

4.2.3 Random Forest

Esse algoritmo tem como base um outro mais conhecido chamado de árvore de decisão. Uma árvore de decisão é obtem seu aprendizado através de um processo indutivo de forma a

DOI: 10.25286/repa.v5i2.1339

resultar em um conjunto de regras de decisão [10].

O algoritmo Random Forest se inspira em sistemas de múltiplos modelos ou *ensembles* que utiliza da estratégia da sabedoria das multidões para tomar melhores decisões. Além disso ele cria árvores de decisão a partir de diferentes e aleatórios subconjuntos dos atributos disponíveis, para assim criar um conjunto diverso de árvores com perspectivas diferentes do mesmo problema. Utilizando por fim uma votação para chegar em um consenso, um resultado [11].

O algoritmo utilizado para a versão deste trabalho teve como base a biblioteca para a linguagem de programação Python Scikit Learn v0.21.3, função RandomForestRegressor com todos os parâmetros originais dessa versão.

4.2.4 XGBoost

Um outro algoritmo que vem tendo grande destaque em competições de aprendizado de máquina é o XGBoost [12].

As bases deste modelo também são árvore de decisão e sistema de múltiplos modelos. Porém, possui o grande diferencial de ser apoiado por diversos cálculos matemáticos de otimização visando melhores resultados e para suportar bases de dados grandes.

O objetivo do XGBoost estar neste trabalho é trazer um modelo que está sendo bastante utilizado para ter seu desempenho comparado com outras técnicas, além de democratizar o uso dela.

O algoritmo utilizado para a versão deste trabalho teve como base a biblioteca para a linguagem de programação Python XGBoost, sem versão identificada, e sem nenhuma alteração de parâmetro da versão disponibilizada até a data do último acesso [8].

4.3 Métricas de Erro

As métricas descritas abaixo têm como melhores resultados aqueles que mais se aproximam de zero.

4.3.1 Erro Quadrático Médio (MSE)

Esta métrica usa da potência para mitigar problemas com os resultados negativos da subtração. Quando utilizada em séries não normalizadas entre 0 e 1 pode resultar em valores muito altos sem que indique um erro tão exorbitante.

O cálculo dessa métrica é realizado através do vetor de entrada □, um modelo regressor □(□) e do resultado do alvo da previsão □ da seguinte forma:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (g(x_i) - y_i)^2 \quad (2)$$

4.3.2 Erro Médio Absoluto (MAE)

O cálculo dessa métrica é realizado através do vetor de entrada □, um modelo regressor □(□) e do resultado do alvo da previsão □ da seguinte forma:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} (|g(x_i) - y_i|) \quad (3)$$

O erro médio absoluto tem como forma de mitigar o valor negativo do erro a operação de valor absoluto. Embora essa métrica não utilize a potência, em casos que os valores da série são muito altos o resultado dessa métrica ainda pode ter um valor alto sem que o erro seja tão significativo.

4.3.3 Média Percentual do Erro Absoluto (MAPE)

Esta métrica consegue mitigar os possíveis problemas do erro ser negativo e também traz uma interpretação mais intuitiva, principalmente em termos de erro relativo [9].

O cálculo dessa métrica é realizado através do vetor de entrada □, um modelo regressor □(□) e do resultado do alvo da previsão □ da seguinte forma:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \left(\left|\frac{g(x_i) - y_i}{y_i}\right|\right) \quad (4)$$

É importante ressaltar que esse resultado é mais interessante quando os valores de entrada não estão normalizados entre 0 e 1. Resultando apenas em um valor relativo e obtendo um mesmo valor de MAPE para duas séries idênticas preditiva e real, sendo uma normalizada e outra não.

5 Validação

Pensado para aproximar públicos diferentes, esse trabalho concluir seu objetivo em ser utilizado sem necessidade de aprender uma nova linguagem de programação, utilizando puramente a interface como intermediário da previsão da série temporal. Ao dar a liberdade ao usuário de utilizar a série que deseja, a

ferramenta fica interessante e útil para experimentações com previsões. Além de aproximar a academia do público em geral ao disponibilizar, na ferramenta, algoritmos de bastante relevância ao alcance de poucos cliques.

6 CONCLUSÃO

É inegável que no mercado existam diversas ferramentas que podem auxiliar a tomada decisões utilizando alguma tecnologia. Umas mais complexas, outras que exigem bastante conhecimento prévio e assim por diante.

Este trabalho foi pensado a partir da premissa de que qualquer pessoa, de qualquer área do amplo conhecimento, que contenha uma série temporal em mãos, possa realizar uma previsão a partir da Sci-XP. A contribuição deste trabalho é tornar a inteligência computacional acessível para pessoas que não possuem conhecimentos profundos de programação ou de estatística. Além de criar uma ponte entre os conteúdos teóricos e práticos, aproximar a inteligência computacional da sociedade e a outros campos do saber através da simplificação dos seus passos.

7 Trabalhos Futuros

Como uma ferramenta que quer proporcionar simplicidade na utilização para pessoas com poucas habilidade técnicas em relação a uma pessoa da área de computação e aproximar a comunidade acadêmica de estudantes das mais diversas áreas e também de pessoas comuns sugerimos outras ideias além de apenas colocar novos modelos.

- Como uma ferramenta de experimentação dar a opção do usuário escolher alguns parâmetros internos de cada modelo;
- Ter a possibilidade de carregar modelos de previsão na ferramenta;
- Ter algumas bases de dados pré-carregada para teste.

8 Agradecimentos

Agradeço a todos os órgãos que participaram e contribuíram na minha formação e na dos demais colegas, portanto não posso deixar de mencionar a UPE, Fitec, FACEPE, CMA-Parqtel, SETIC, além dos professores e monitores que participaram da nossa formação, e de toda a administração que estava presente nos auxiliando com o que precisávamos durante a nossa formação. Agradeço aos meus familiares e amigos pela força neste ano tão complicado que foi para mim. Ao meu orientador, Paulo Salgado, por toda a paciência. E acima de tudo e primordialmente importante, Deus, que ouviu minhas súplicas e me acompanhou durante esse processo me dando muita força nessa dura e longa caminhada.

9 Referências

[1] BROCKWELL, P. J.; DAVIS, R. A. **Introduction to Time Series and Forecasting**. Livro, 2ª edição, Springer, 2002.

[2] NUGUS, S. **Financial Planning Using Excel: Forecasting Planning and Budgeting Techniques.** Livro, 2ª edição, Elsevier, 2009.

[3] BRAGA, A. P; CARVALHO, A. C. P. L. F; LUDEMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. Livro, LTC, 2000.

[4] ROSENBLATT, F. **The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain**. Psychological Review, 65, 386. 1958.

[5] HAYKIN, S. O. **Neural Networks and Learning Machines.** Livro, 3ª edição, Pearson, 2008.

[6] **Aula prática em R. Oliveira, J. A. B.** <https://rpubs.com/GutoBarros/ast>. Acesso em: 16 nov. 2019.

[7] **Ferramenta MATLAB**. Disponível em: <https://www.mathworks.com/products/matlab.html>. Acesso em: 15 nov. 2019.

DOI: 10.25286/repa.v5i2.1339

[8] **Biblioteca XGBoost em Python.** Disponível em: <https://xgboost.readthedocs.io/en/latest/python/python_intro.html>. Acesso em 12 nov. 2019

[9] MYTTENAERE, A; GOLDEN B.; LE GRAND, B.; ROSSI, F. **Mean Absolute Percentage Error for regression models.** Neurocomputing, Elsevier, 2016, Advances in artificial neural networks,machine learning and computational intelligence - Selected papers from the 23rd European Symposium on Artificial Neural Networks (ESANN 2015), 192, pp.38 - 48.

[10] QUINLAN, J. R. **Induction of Decision Trees**. Kluwer Academic Publishers. Machine Learning 1: 81-106, 1986, Boston.

[11] BREIMAN, L. **Random Forests.** Kluwer Academic Publishers. Machine Learning, 45, 5–32, 2001.

[12] CHEN T. E GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System.** KDD'16, August 13-17, 2016, San Francisco, CA, USA.

[13] **MATHWORKS: What Is MATLAB?** <https://www.mathworks.com/videos/matlab-overview-61923.html> Acesso em: 16 nov. 2019.