Universidade Estadual de Campinas

Faculdade de Engenharia Elétrica e de Computação

IA048 – Aprendizado de Máquina 27 de março de 2024

Docentes: Levy Boccato & Romis Attux

Discente:

– Gabriel Toffanetto França da Rocha – 289320

Atividade 1 – Regressão Linear

Sumário

1 Apresentação da série temporal

De acordo com a base de dados *U.S Airline Traffic Data*, carregada por meio do arquivo `air_traffic.csv`, é possível obter diversas informações sobre o tráfego aéreo norte-americano no período de 2003 até 2023, com uma atualização mensal, e dados como número de voos e de passageiros.

Extraindo desses dados a série temporal do número total de voos em cada mês, é possível observá-la por meio do gráfico da Figura 1. Analisando visualmente os dados, observa-se que a distribuição dos voos não é constante, e que tem um aspecto periódico, apresentando um formato parecido que se repete ano a ano. Esse comportamento é esperado, uma vez que nas férias de verão, as famílias tendem a viajar mais a turismo, e no inverno tendem a voar menos devido ao mal tempo.

Porém, mesmo com uma característica regular, é possível observar faixas que se distinguem da demais, causando irregularidades no padrão da série temporal. Isso pode ser notado principalmente em dois pontos, destacados na Figura 2, em Agosto de 2008, onde o padrão da série decaí, e em Janeiro de 2020, onde o número de voos despenca drasticamente, com um transitório lento de recuperação, divergindo totalmente o que se observava no comportamento do número de voos.

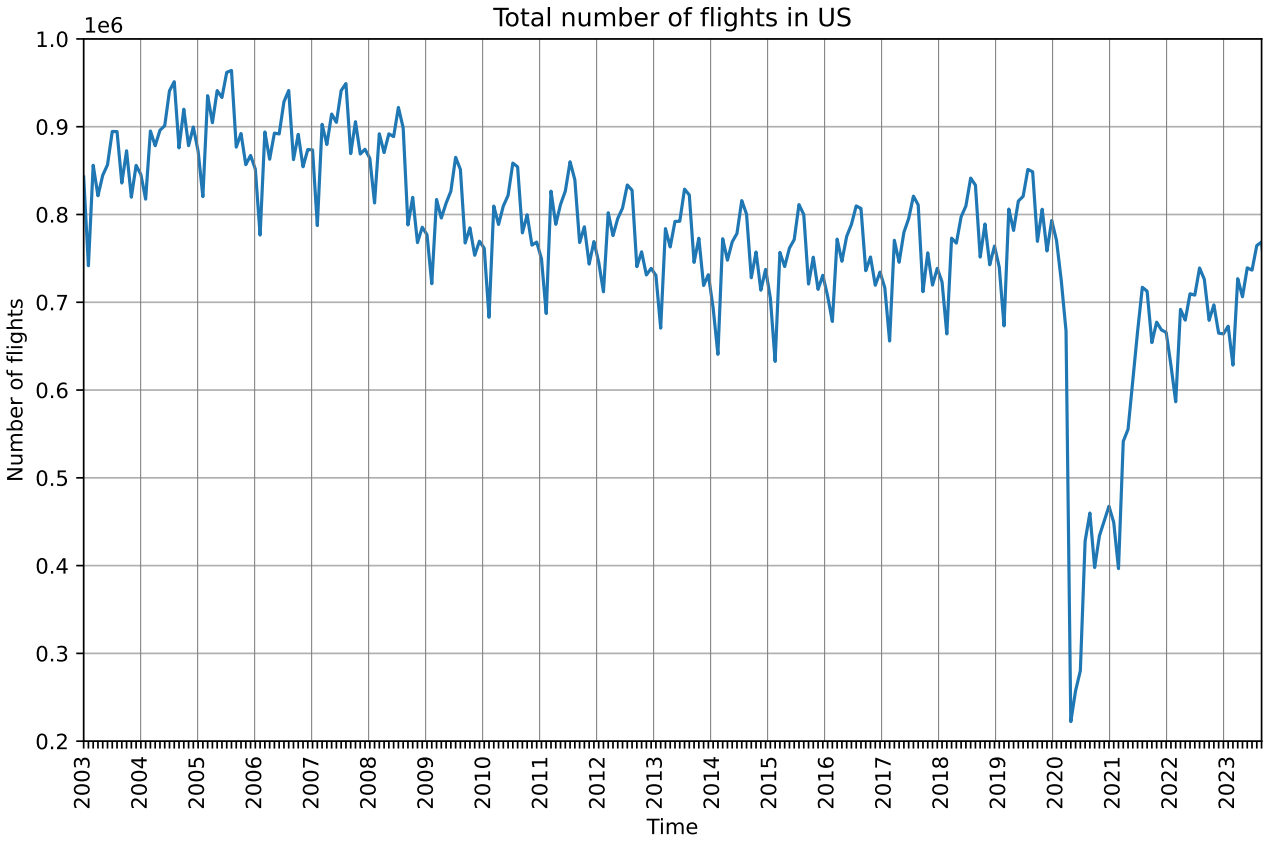

Figura 1: Série temporal do número total de voos nos EUA entre 2003 e 2023.

A primeira mudança de regularidade da série se da em Agosto de 2008, época que o mundo enfrentou grande crise econômica, e os EUA estavam em estado de recessão.

A queda brusca do número de voos em Janeiro de 2020 até aproximadamente Julho de 2021 é facilmente explicada, ocorrendo devido a pandemia do vírus Covid-19 enfrentada pelo Mundo na época, que levou ao fechamento de inúmeros aeroportos até que fosse possível buscar formas de prevenir a doença, e principalmente a imunização da população com a vacina. Dessa forma, se observa uma primeira queda muito forte, ocorrida com o início da pandemia e o fechamento do aeroportos. Porém, entre o meio de 2020 e o início de 2021, os aeroportos passaram a ser reabertos, principalmente para voos domésticos, resultando em uma elevação da série. Com a vacina em 2021, ocorreu um salto no número de voos, com a abertura de fronteiras e o retorno a normalização do padrão da série a partir do meio de 2021, com o número de voos se elevando em 2022 e 2023.

Com isso, observa-se que mesmo podendo detectar padrões na série temporal do número de voos nos EUA, que se repetem com os anos, existem fatores externos, imprevisíveis, como crises econômicas e pandemias globais que influenciam diretamente na variável que pode ser de interesse sua previsão futura. Dessa forma, durante o comportamento regular da série, se espera que a previsão seja mais fácil, porém, ao ocorrer eventos imprevisíveis, a estimação tende a ser mais errônea.

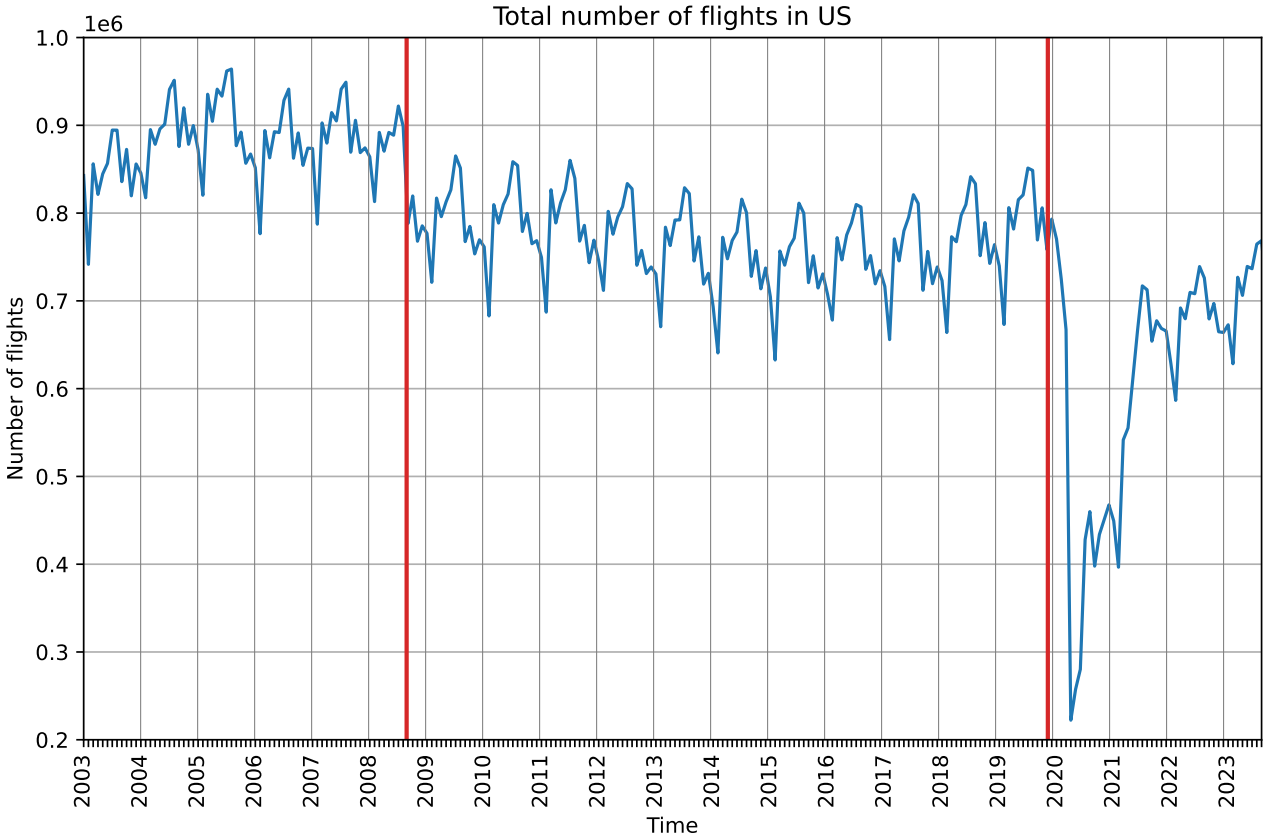

Figura 2: Série temporal do número total de voos nos EUA entre 2003 e 2023.

2 Modelo de predição da série temporal

Considerando a predição L passos a frente, utilizando K amostras passadas, o modelo do preditor linear utilizado como estrutura do problema é dada por (1), sendo o vetor de entradas $\mathbf{x}$ dado por (2).

$$y(n) = \mathbf{w}^{\mathrm{T}}\mathbf{x}(n) \tag{1}$$

$$x(n) = 1 + x(n-L) + x(n-L-2) + \cdots + x(n-L-K+1) \tag{2}$$

Fazendo $L = 1$, o problema pode ser reescrito matricialmente como (3).

$$y(n) = x(n) = \begin{bmatrix} w_0 & w_1 & \dots & w_{n-K} \end{bmatrix} \times \begin{bmatrix} 1 \\ x(n-1) \\ \vdots \\ x(n-K) \end{bmatrix} \tag{3}$$

Uma vez que $x(n)$ e os K $x(n-i)$, $i = 1, 2, \dots, K$ são conhecidos a partir dos dados utilizados para o treinamento do modelo, é possível utilizar o critério dos quadrados mínimos para a obtenção de $\mathbf{w}$ por meio da pseudo-inversa de $\mathbf{x}$, como feito em (4).

$$\mathbf{w} = \texttt{pinv}[\mathbf{x}(n)] \cdot y(n) \tag{4}$$

Para implementação, dado todos os dados disponíveis para treinamento, expande-se $y(n)$ na forma $\mathbf{y}(n)$, onde $\mathbf{x}(n)$ se torna uma matriz contendo o vetor de entradas para cada uma das entradas do vetor $\mathbf{y}$, de acordo com (5), sendo Q o comprimento do vetor de saída, e resolvida de forma similar a (4).

$$\begin{bmatrix} x(n) \\ x(n+1) \\ \vdots \\ x(n+Q) \end{bmatrix}^T = \begin{bmatrix} w_0 & w_1 & \dots & w_{n-K} \end{bmatrix} \times \begin{bmatrix} 1 & 1 & \dots & 1 \\ x(n-1) & x(n) & \dots & x(n+Q-1) \\ \vdots & \vdots & \ddots & \vdots \\ x(n-K) & x(n-K+1) & \dots & x(n+Q-K) \end{bmatrix} \tag{5}$$

A solução do vetor $\mathbf{w}$ para (5) é realizada via código por meio da função `trainingModel`, descrita abaixo.

```
def trainingModel(data, k):
    y = data[k:len(data)]
    x = np.ones([len(y), k+1])
    for i in range(1,k+1):
        x[:, i] = data[i-1:len(y)+i-1]
    w = np.linalg.pinv(x).dot(y)
    return w
```

2.1 *Dataset* 1

Para a primeira base de dados de treinamento, validação e teste, foi realizada a divisão dos dados considerando: Dados de treinamento e validação de Janeiro de 2003 até Dezembro de 2019; Dados de teste de Janeiro de 2020 até Setembro de 2023. Foi considerado o método *holdout* para a realização do treinamento do modelo, onde os dados de treinamento e validação foram divididos sendo 80% dos dados para treinamento e 20% para validação, sendo os de validação os dados finais do conjunto.

Para os dados de validação, assim como os de teste, foram aproveitados dados passados para inicialização do preditor com dados de grupos passados, conseguindo um melhor aproveitamento na análise da saída $y(n)$ e da saída estimada $\hat{y}(n)$ em cada grupo.

A Figura 3 mostra o gráfico da raiz do erro quadrático médio(RMSE) para os dados de validação e os dados de treinamento, de acordo com a progressão do número de entradas do preditor, de 1 até 24. Pode-se observar que o erro de treinamento e validação caminham de forma bem similar, apresentando platôs e mínimos locais, como pode ser visto em $K \in [7, 8]$. A partir de 14 entradas para o preditor, observa-se um grande platô para o RMSE de validação, mostrando que o ajuste do hiper-parâmetro K está chegando perto do ideal, quando o modelo não consegue mais reduzir o erro na etapa de validação, e pode entrar em regime de *overfitting*. Como já é conhecido o perfil e o comportamento da série temporal, sabe-se que os dados de validação e de treinamento possuem o mesmo perfil de comportamento, e dessa forma, o RMSE de Validação e de Treinamento decaem juntos.

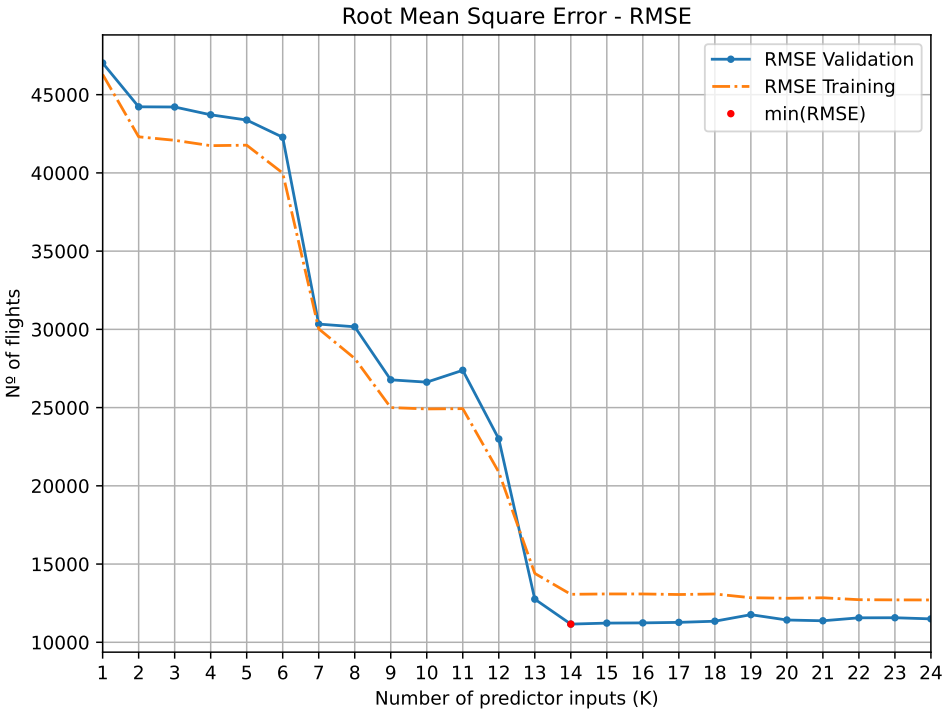

Figura 3: Progressão da raiz quadrada do erro quadrático médio de treinamento e validação em função do número de entradas do preditor (K).

Ao captar o valor de K que minimiza o RMSE, obtém-se o argumento de 14 entradas para

o preditor ótimo, cujo gráfico de comparação entre a saída real $y(n)$ e a saída estimada $\hat{y}(n)$ consta na Figura 4, resultando em um RMSE de 11162,235, e um erro percentual absoluto médio(MAPE) de 1,102%.

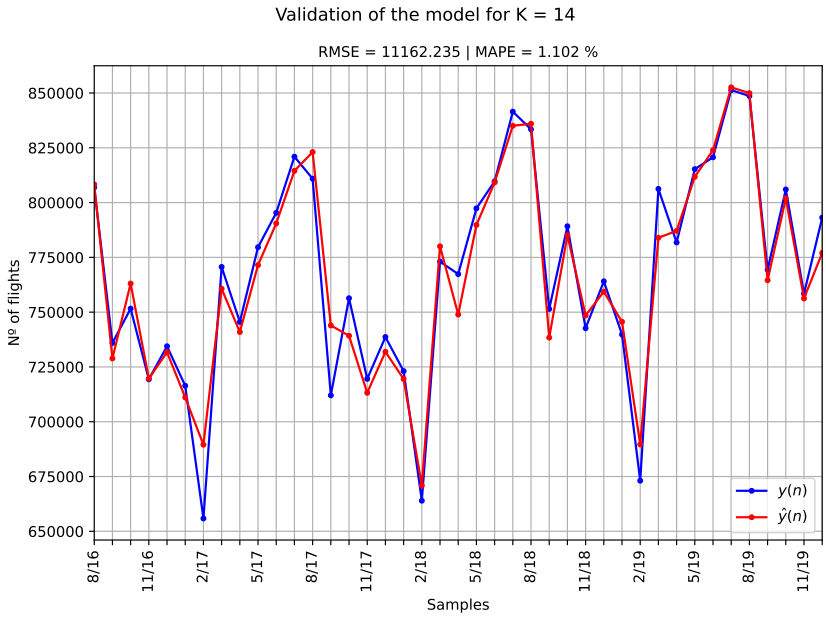

Figura 4: Validação da predição da série temporal para o preditor com menor RMSE.

Ao aplicar todos os dados de teste sobre esse modelo, obtém-se o gráfico da Figura 5, observa-se facilmente que o preditor possuiu uma grande dificuldade de acompanhar o número de voos, chegando a um RMSE de 112561,530, e um MAPE de 13,222%. Observa-se em Fevereiro de 2020, que a série temporal começou a decrescer fortemente, e o preditor promoveu uma saída crescente, passando a decrescer penas em Maio. O preditor também apresenta um grande pico negativo em Abril de 2021, onde a série original apresenta um grande crescimento. Com isso, observa-se que o preditor não teve capacidade com os dados treinados, de representar com fidelidade os dados ocorridos entre Fevereiro de 2020 até aproximadamente Setembro de 2021. Porém, cabe explicitar que os dados utilizados para teste contemplam episódios inéditos na história da série temporal, como a pandemia ocorrida em 2020, que trás uma grande irregularidade à serie em tal época, e por consequência, provoca um maior erro na saída do preditor, que não teve a oportunidade de aprender com esse tipo de irregularidade durante seu treinamento.

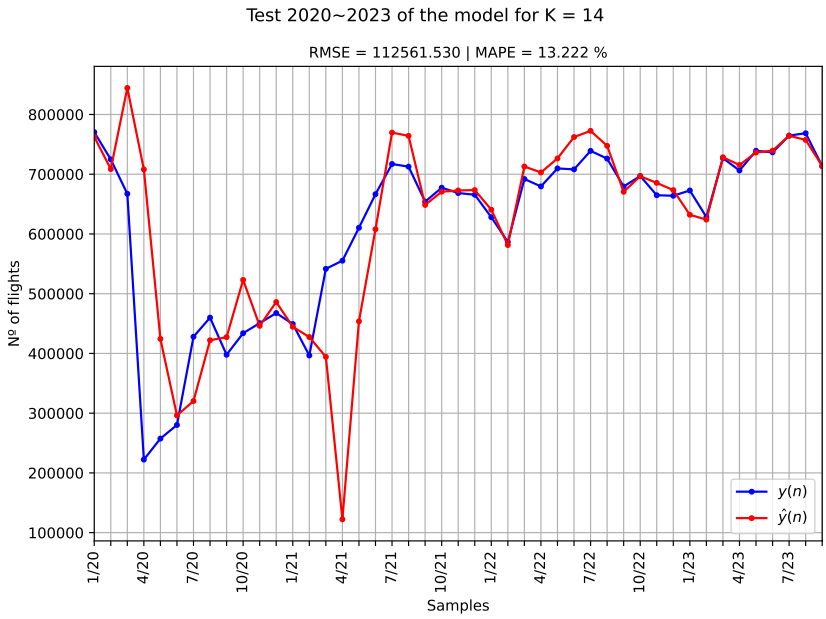

Figura 5: Teste da predição da série temporal para o preditor com menor RMSE utilizando todos os dados de teste.

Aplicando ao preditor os dados de teste que contém apenas os números de voos de 2022 e 2023, observa-se o comportamento mostrado na Figura 6. O erro já é consideravelmente menor, resultando em um RMSE de 20054,297 e um MAPE de 2,057%, uma vez que a série entre 2022 e 2023 tende a ter uma regularidade mais similar ao período pré 2020, ou seja, conforme os dados com que o preditor foi treinado.

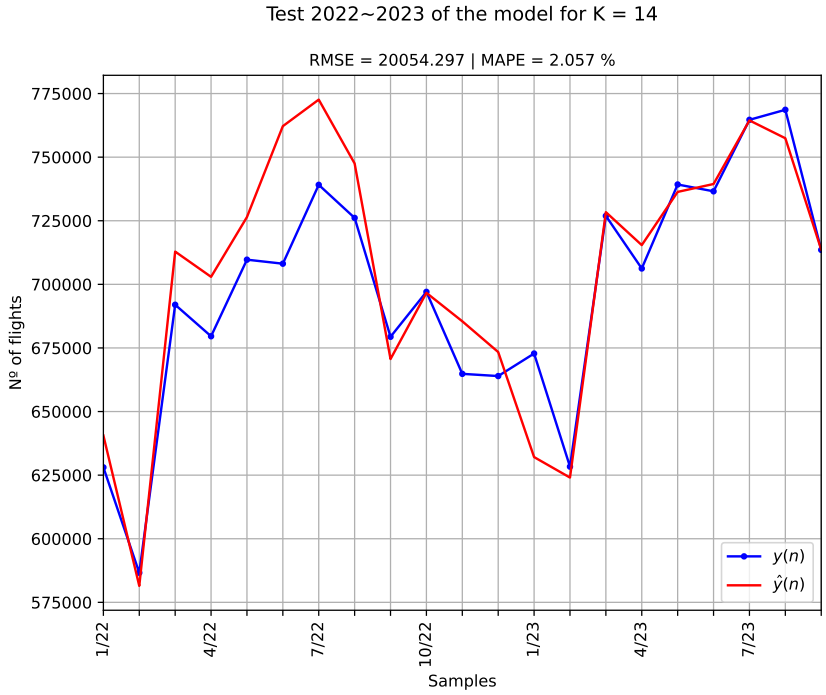

Figura 6: Teste da predição da série temporal para o preditor com menor RMSE utilizando os dados de teste entre 2022 e 2023.

2.2 *Dataset 2*

O *dataset* 2 possuí uma divisão diferente de dados de treinamento, validação e teste, sendo: Dados de treinamento de 2003 até 2019; Dados de validação de 2020 até 2021; Dados de teste de 2022 até 2023. *A priori* já se observa que esses dados tem uma característica diferente dos escolhidos no conjunto de dados anterior, onde o treinamento é feito em uma faixa de comportamento da série temporal, e a validação é realizada em uma faixa de comportamento irregular.

A Figura 7 mostra a progressão do erro de treinamento e validação conforme o aumento do número de entradas do preditor. A princípio, já se observa que o erro de validação não acompanha o de treinamento, seja no caso inicial, ou durante a progressão. Enquanto o erro de treinamento caí durante o aumento do número de atrasos, até saturar 14 entradas, o erro de validação aumenta, apresentando picos máximos e mínimos, até saturar em um valor muito mais alto que o erro de validação.

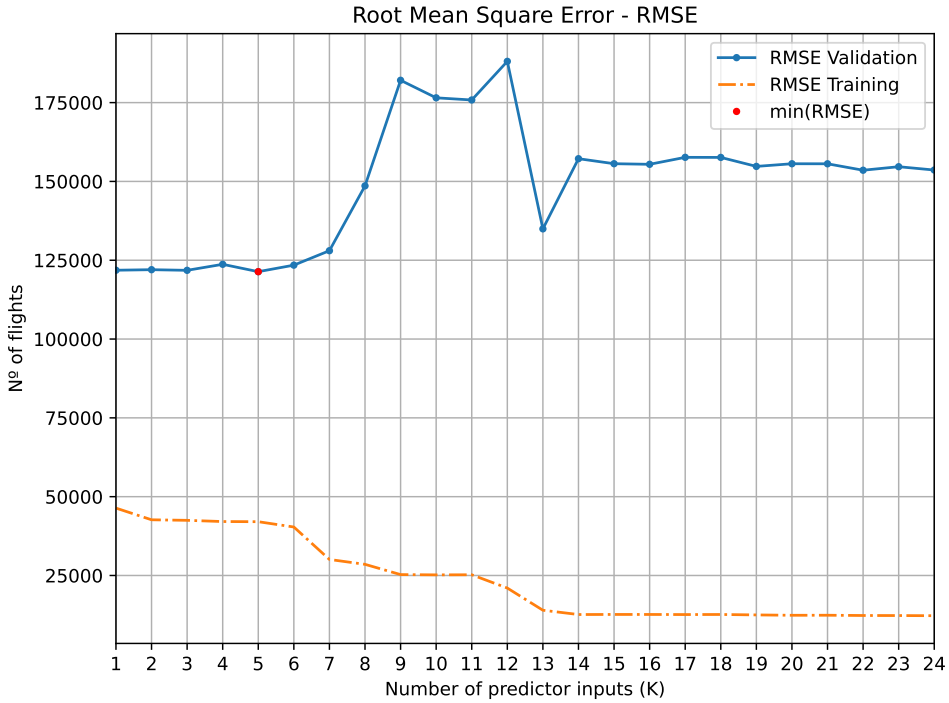

Figura 7: Progressão da raiz quadrada do erro quadrático médio de treinamento e validação em função do número de entradas do preditor (K).

O argumento $K = 5$, foi encontrado como o que minimiza o RMSE, e seu gráfico de validação é mostrado na Figura 8. A raiz do erro quadrático médio é de 121371,609 voos, e o erro percentual absoluto médio é de 21,439%. Esses valores de erro chegam a ser dez vezes maiores do que os resultados obtidos para o *dataset 1*, devido a inserção do período de irregularidade no grupo de validação do modelo de predição.

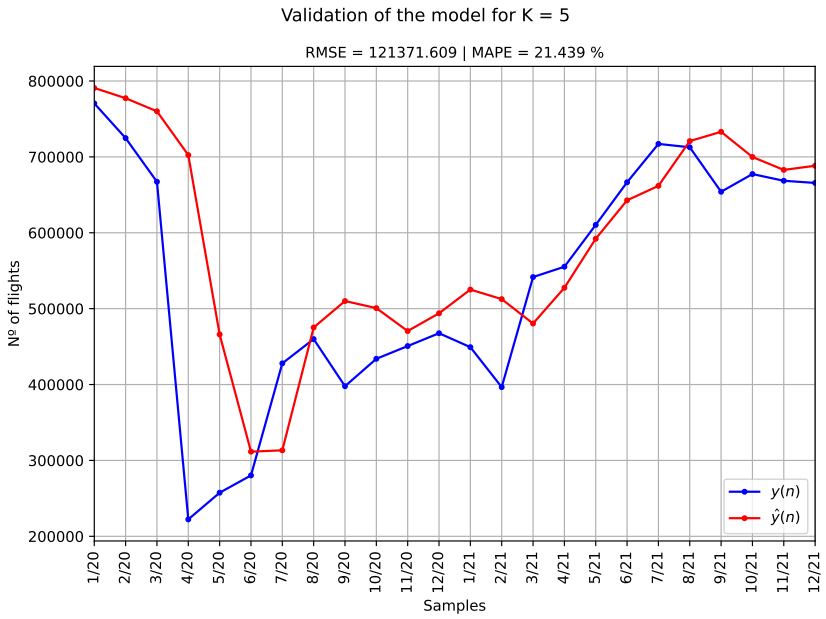

Figura 8: Validação da predição da série temporal para o preditor com menor RMSE.

Ao aplicar os dados de teste ao modelo treinado do preditor, obtém-se o gráfico da Figura 9. O RMSE se aproxima do erro de treinamento para o mesmo número de entradas, enquanto o MAPA é de 4,782%, uma vez que esse período da série tende a retornar a regularidade utilizada durante o treinamento. Como foi escolhido um modelo de predição com apenas 5 entradas, observa-se que a curva em vermelho, que se refere a saída estimada, não tem a capacidade de se contorcer o suficiente para a companhar a série temporal, apresentando uma curva mais constante e com isso impondo um erro maior.

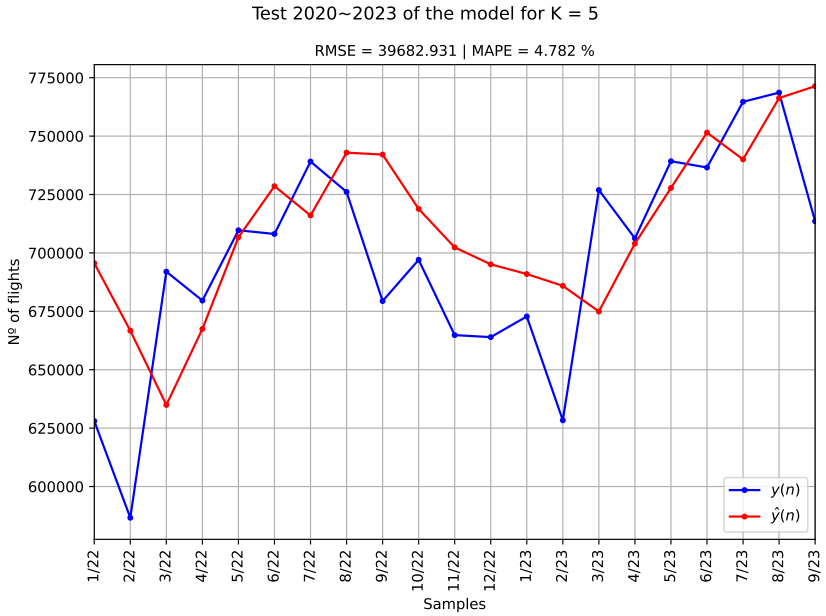

Figura 9: Teste da predição da série temporal para o preditor com menor RMSE utilizando todos os dados de teste.

3 Discussão

A predição de uma série temporal com base em seus dados passados é uma tarefa que tem grande dependência da regularidade dessa série, uma vez que ao ocorrer eventos que tragam um comportamento inédito à série, o preditor pode não conseguir realizar uma boa previsão, devido a não ter sido treinado com dados que abordem aquela situação em específico. A série que está sendo trabalhada engloba aproximadamente 20 anos de dados, onde nesse período podem ser observadas duas principais irregularidades, sendo uma mais suave em 2008, e uma muito abrupta em 2020.

Ao treinar a série com os dados regulares de 2003 à 2019, o preditor adquire grande capacidade de estimação de dados com comportamento similares à esse intervalo de tempo, onde ao validar o preditor nesse mesmo intervalo, significa obter o melhor modelo para o conjunto de dados com tal característica. Isso pode ser visto claramente para o primeiro conjunto de dados, onde na Figura 3, o erro para os dados de treinamento acompanham os dados para os dados de validação, mostrando que os dados de validação são uma boa generalização dos dados utilizados para obter os parâmetros do preditor ($\mathbf{w}$).

Já na situação de treinamento com os dados regulares, de 2003 à 2019, porém, realizando a validação com dados de 2020 à 2021, sabe-se pelo gráfico da série temporal (Figura 1) que os dados utilizados para validação não possuem a mesma característica dos dados de treinamento. O mesmo se comprova pela Figura 7, onde os erros de treinamento e validação são muito disjuntos, uma vez que o comportamento dos dados utilizados para validar o treinamento, não são representados no conjunto de treinamento.

Analisando os testes realizados para o modelo treinado, para o primeiro conjunto de dados, observa-se uma dificuldade do preditor para os dados irregulares de 2020, como visto na Figura 5, enquanto para os dados restringidos de 2022 à 2023, onde a regularidade da série retorna, o desempenho é superior (Figura 6). Como foi escolhido um preditor de 14 entradas, a saída estimada se contorce tentando acompanhar a série temporal. Já para o segundo *dataset*, o desempenho para o conjunto de testes é superior comparado ao erro de validação, porém como foi escolhido um preditor com apenas 5 entradas, é claro a dificuldade da saída estimada de se contorcer para acompanhar a série, com isso, a aproximação apresenta um erro considerável mesmo com dados menos inesperados da série (comparados aos dados regulares de treinamento).

Desta forma, é evidente como a escolha e a divisão dos dados impacta diretamente no modelo do preditor e no seu desempenho, sendo que a capacidade de generalização do preditor depende tanto dos dados de treinamento, como o quanto os dados de validação estão contidos no conjunto de treinamento, onde, ao treinar com dados de característica a e validar o modelo em um conjunto de característica b faz com que a estrutura possa ser escolhida erroneamente, o que se reflete diretamente nos dados de teste. Também se mostra muito claro como a ocorrência de comportamentos inesperados da série influência no desempenho da predição, onde uma vez que um certo evento não foi utilizado para treinamento e validação, o modelo não saberá como tratá-lo em teste ou diretamente na aplicação final.

Anexos

Códigos fonte

Todos os códigos fonte e arquivos de dados utilizados para a elaboração deste documento podem ser encontrados no repositório do GitHub no link: github.com/toffanetto/ia048.