# IA048 – Aprendizado de Máquina

# Atividade 1 – Regressão Linear

Nome: José de Arimatéa Passos Lopes Júnior

RA: 271467

# Resultados

a) Exiba o gráfico da série temporal completa. Numa inspeção visual simples, é possível reconhecer ao menos três faixas distintas de comportamento aproximadamente “regular” na série: (i) Jan/2003 a Ago/2008; (ii) Set/2008 a Dez/2019; (ii) Jan/2020 a Set/2023. Discuta possíveis razões históricas / econômicas para essas transições de comportamento.

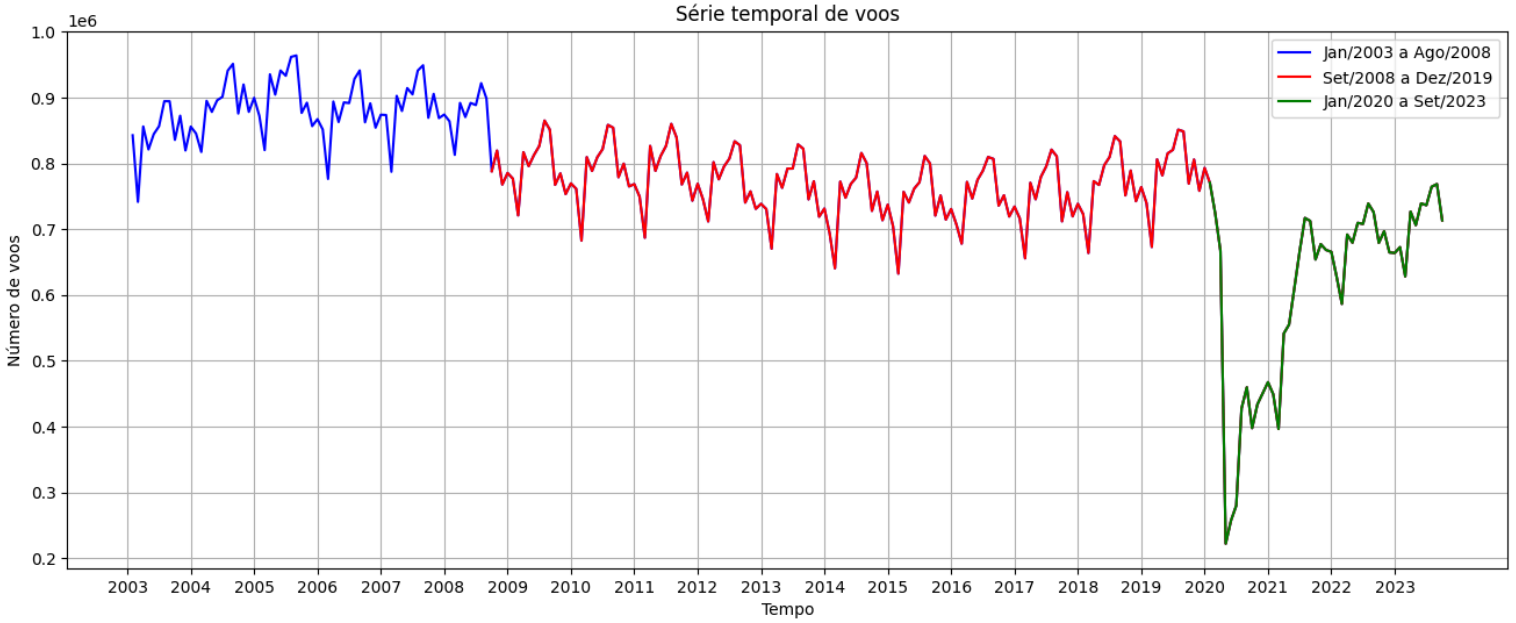


*Figura 1: Gráfico da série temporal do número total de voos, dividido por cores nos períodos (i) Jan/2003 a Ago/2008; (ii) Set/2008 a Dez/2019; (ii) Jan/2020 a Set/2023*

É possível verificar no gráfico em cor azul o período de Janeiro de 2003 até Agosto de 2008 apresentando um comportamento quase que igual em todos os anos, ficando a quantidade de voos variando entre um pouco menos de 800 mil e chegando a quase 1 milhão.

Vê-se a média anual no período:

| Período anual | Média anual de voos |
|---|---|
| 2003 | 844680.25 |
| 2004 | 891638.58 |
| 2005 | 902156.75 |
| 2006 | 876786.83 |
| 2007 | 889286.33 |
| Jan/2008 até Ago/2008 | 880120.37 |

A média anual é bem parecida durante todo o período, variando entre 850 mil e 900 mil.

A partir do período em vermelho, de Setembro de 2008 a Dezembro de 2019, vemos uma queda do número de voos, variando entre um pouco menos de 700 mil e chegando aproximadamente até 850 mil.

Vê-se as novas médias anuais:

| Período anual | Média anual de voos |
|---|---|
| Set/2008 até Dez/2008 | 790259.25 |
| 2009 | 795193.33 |
| 2010 | 791587.00 |
| 2011 | 788034.75 |
| 2012 | 772278.41 |
| 2013 | 762707.83 |
| 2014 | 746201.83 |
| 2015 | 741384.33 |
| 2016 | 752142.16 |
| 2017 | 751820.08 |
| 2018 | 771400.00 |
| 2019 | 788643.16 |

Percebe-se que a média anual caiu, não chegando a mais de 800 mil, variando entre 750 mil e 795 mil.

A razão do número de voos se deve a crise mundial que ocorreu em 2008, a crise imobiliária dos Estados Unidos. Devido a especulação imobiliária nos Estados Unidos, e a liberação desenfreada de crédito para compra de imóveis, acabou acarretando em uma crise financeira nos Estados Unidos, o que culminou na quebra de um Banco super tradicional nos Estados Unidos, o Lehman Brothers. Dessa forma, a bolsa americana acabou despencando e causando aflição mundial. Logo, a queda da bolsa afetou a economia do mundo inteiro.

Assim, com a crise financeira, era de se esperar uma queda na demanda de vários produtos e serviços, dessa forma, afetando também companhias aéreas, diminuindo o número de voos. É perceptível verificar no gráfico essa queda, e na média de voos, mostrando que está condizente com a realidade do período, em que a média de voos permaneceu constante no período, sem conseguir se recuperar totalmente.

No período em verde, correspondente ao intervalo de Janeiro de 2020 até Setembro de 2023, vê-se uma queda brusca. O número de voos caiu abaixo de 300 mil, e foi se recuperando aos poucos, chegando até um pouco acima de 700 mil.

É interessante perceber que o comportamento mensal de Janeiro a Dezembro nos anos do novo período é parecido com os dos períodos anteriores, apesar da queda brusca dos números totais de voos.

Vê-se que as médias anuais do novo período são:

| Período anual | Média anual de voos |
|---|---|
| 2020 | 463363.75 |
| 2021 | 609590.25 |
| 2022 | 681209.66 |
| Jan/2023 até Set/2023 | 717441.11 |

As médias de voos claramente caíram bastante no período.

Isso se deve a pandemia mundial do Coronavírus. Diferente da crise financeira de 2008, a crise do Coronavírus foi uma crise relacionada à saúde, com medidas de contenção da locomoção das populações dos países em geral. As pessoas se contiveram em sair de casa e principalmente viajar. Houveram campanhas de ficar em casa, e um medo grande da população em contrair o vírus ao se deslocar e ficar em ambientes com muitas pessoas como aeroportos.

Houve também uma crise financeira por fechamento de comércios e encerramento de atividades, porém com a criação das vacinas e as pessoas se imunizando, as pessoas começaram a sair mais e voltar a viajar. A vacinação começou em massa na época de 2021, onde houve um aumento bem significativo de voos. A recuperação foi mais rápida, porém não chegando ao mesmo patamar de 2019 e antes, visto também por conta da crise financeira que se instaurou. Mas em 2023 já vê-se uma média parecida com períodos anteriores a 2020.

b) Divida as séries em dois conjuntos: (i) treinamento e validação, com amostras de 2003 a 2019; (ii) teste, com amostras de 2020 a 2023. Faça a análise de desempenho do preditor linear ótimo, no sentido de quadrados mínimos irrestrito, considerando:

    b1) A progressão do valor da raiz quadrada do erro quadrático médio (RMSE, do inglês root mean squared error), junto aos dados de validação, em função do número de entradas (K) do preditor (desde K = 1 a K = 24). Apresente o gráfico obtido e busque tecer conjecturas sobre os motivos subjacentes a seu comportamento.

    Utilizou-se um modelo de Regressão Linear para realizar a predição de valores futuros. Separou-se os conjuntos de treinamento e validação

utilizando o método HoldOut, em que dedicou-se 80% dos valores iniciais para treinamento e os 20% finais para validação.

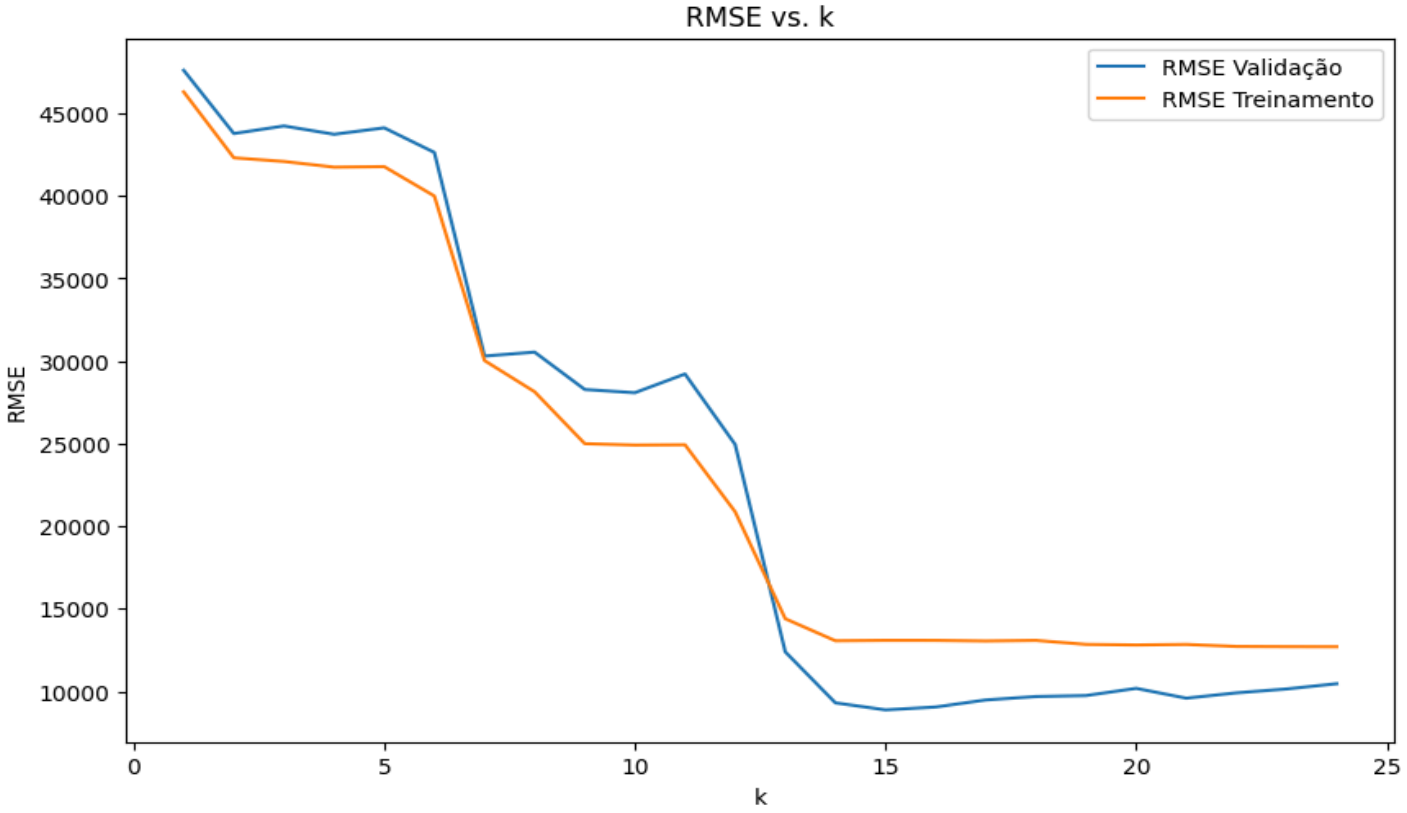


*Figura 2: Gráfico dos RMSE de Validação e Treinamento calculados de acordo com um valor específico de um k (quantidade de entradas usadas para realizar a predição do modelo)*

Obteve-se o RMSE dos dados de treinamento e validação utilizando passos de k=1 até k=24, como mostra a figura 2.

Para o treinamento, pula-se a previsão dos primeiros k meses do conjunto de treinamento, justamente porque não tem-se dados anteriores aos k primeiros meses para cada passo k. Por exemplo, para k=1, não tem-se como prever o primeiro mês do conjunto de treinamento, pois não tem valor anterior ao primeiro elemento do conjunto. Assim, descartou-se a previsão para os primeiros k meses em cada um dos treinamentos dados por cada k passos anteriores.

Dessa forma, para a os dados encontrados na figura 2, também descartamos a previsão dos primeiros k meses do conjunto de validação, começando a predizer a partir de k+1 meses, que é quando temos dados suficientes para realizar uma previsão.

Com o gráfico obtido, pode-se perceber que a partir de k=14, o RMSE do conjunto de treinamento fica praticamente constante, não apresentando mais melhora. E isso se reflete no conjunto de validação, em que a partir de k=14 o RMSE fica praticamente constante, apresentando uma leve melhora em

k=15, e depois piorando o valor de RMSE. O valor de RMSE obtido para k=15 foi de 8883,022.

Uma hipótese é de que começa a ocorrer Overfitting dos dados de treinamento a partir de k=14, ou k=15, visto que a partir daí o RMSE deve apresentar uma leve melhora quase que imperceptível visualmente, mas piora os dados de validação, ou seja, não adianta colocar uma quantidade maior de dados de entrada para treinar o modelo, pois não há mais melhora. Assim, os dados acabam ficando muito enviesados para os dados de treinamento, mas com esse efeito aparecendo de forma muito leve, já que os valores de RMSE são muito parecidos.

Na verdade, a partir de k=14 já apresentam valores muito parecidos, e dessa forma, qualquer valor a partir de k=14 poderia ser usado, levando em consideração apenas esse dado.

Assim, é lógico pensar que quanto mais dados de entrada, a predição da saída seria mais precisa, porém o modelo mostra que existe um valor ótimo de dados de entrada, que é menor que o maior valor possível. É uma interessante análise a ser feita, pois mostra que se pode obter um valor ótimo de predição, sem um gasto computacional elevado.

**Outra análise:**

Também foi feito o gráfico utilizando os k últimos dados do conjunto de treinamento para treinar os k primeiros dados do conjunto de validação. Dessa forma, foram preditos todos os valores do conjunto de validação.

Nessa nova análise obteve-se o gráfico da figura 3, e nela o melhor valor de k seria k=14, e RMSE=11162,24. Embora o resultado de k tenha sido diferente, percebe-se pelo gráfico que o comportamento é bem parecido com a análise anterior. A partir de k=14 os valores de RMSE são bem parecidos para o conjunto de validação, apresentando uma leve piora conforme se aumenta o valor de k.

Sem utilizar o conjunto de treinamento para predizer o conjunto de validação (chamaremos de **análise 1**) houve uma leve melhora em relação a análise utilizando o conjunto de treinamento para validação (chamaremos de **análise 2**) nos valores de RMSE. O conjunto de validação começa no mês 8 de 2016, e a partir dessa época começou-se um aumento no número médio de voos, mudando um pouco o comportamento do gráfico em relação aos dados do conjunto de treinamento. Logo, realmente, usando apenas os dados do conjunto de validação para predição, imagina-se que o erro médio diminua para o período.

Logo, a hipótese a se considerar para a **análise 1** ser melhor que a **análise 2**, se deve a ter dados mais característicos do período para se realizar a predição, gerando um erro médio menor.

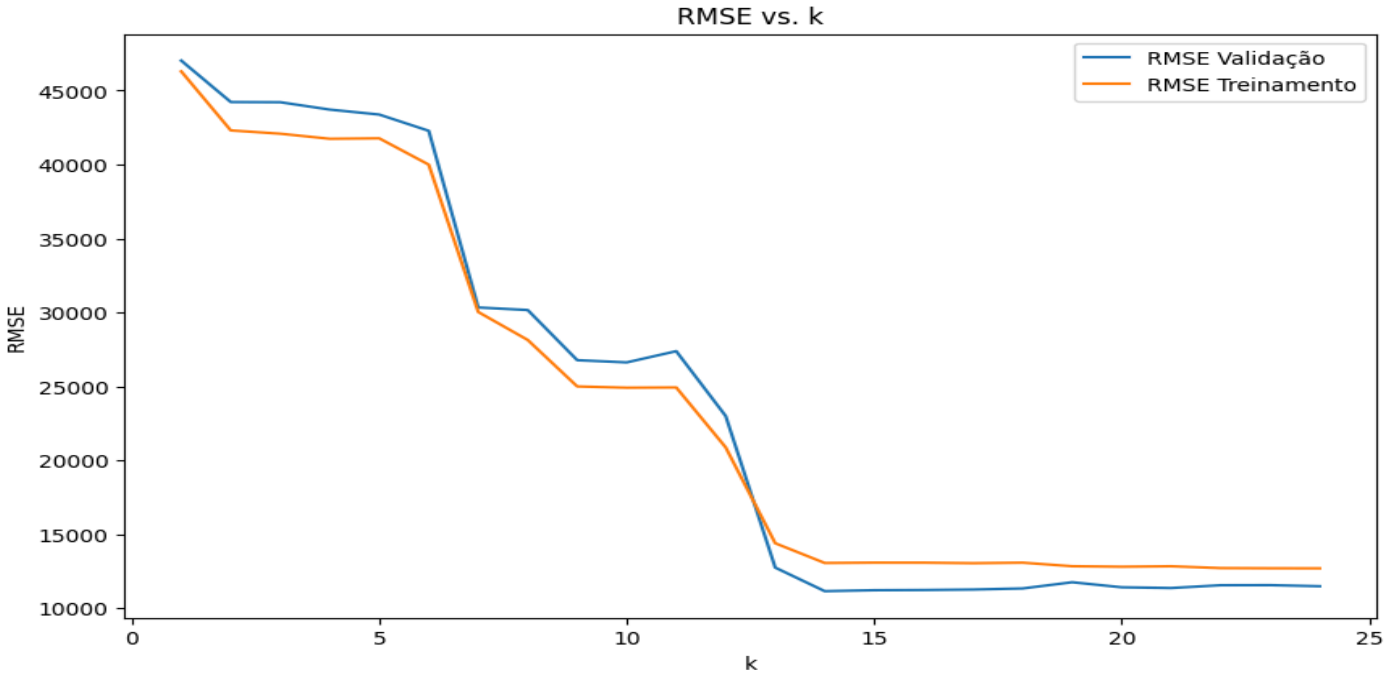


*Figura 3: Gráfico dos RMSE de Validação e Treinamento calculados de acordo com um valor específico de um k (quantidade de entradas usadas para realizar a predição do modelo), utilizando dados do conjunto de treinamento para validação*

b2) O gráfico com as amostras de teste da série temporal e as respectivas estimativas geradas pela melhor versão do preditor (i.e., usando o valor de K que levou ao mínimo erro de validação). Obtenha, também, o RMSE e o erro percentual absoluto médio (MAPE, do inglês mean absolute percentage error) para o conjunto de testes.

Utilizar-se-á apenas o gráfico gerado pelo modelo da **análise 1** para o item b2. O gráfico gerado pela **análise 2** é bem parecido.

O gráfico de predição gerado pelo modelo para o conjunto de testes está representado na figura 4.

Para **análise 1**:

O valor de RMSE foi de 110521,44

O valor de MAPE foi de 12,78%

Para **análise 2**:

O valor de RMSE foi de 112561,53

O valor de MAPE foi de 13,22%

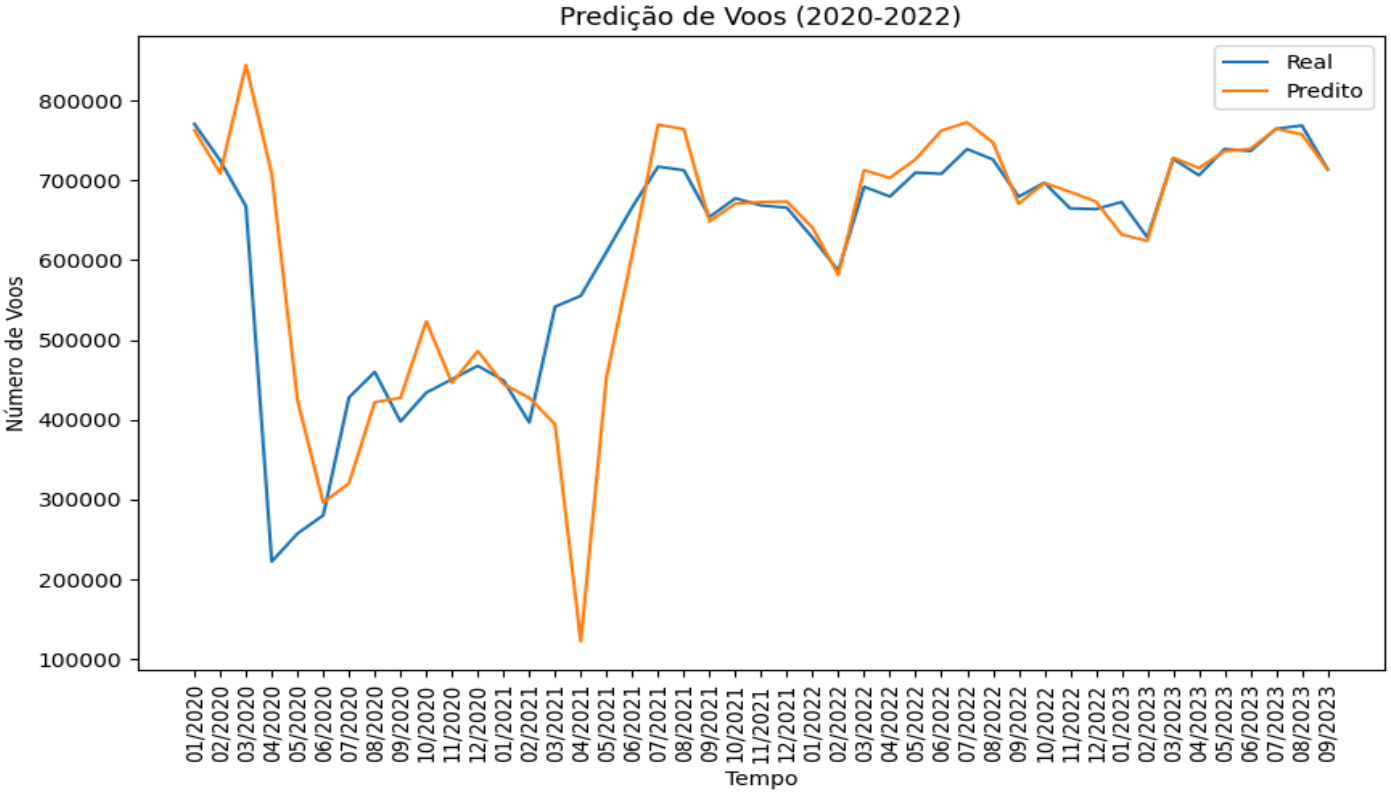


*Figura 4: Gráfico de estimativas do conjunto de testes em comparação com o valor real do número de voos no período de 2020 a 2023*

Para gerar a previsão de todos os valores do conjunto de testes, utilizou-se os últimos k valores do conjunto de validação. Acredita-se que essa seria a razão para a predição dos primeiros valores não serem tão boas, e a partir de certo ponto os valores começarem a ficar mais precisos.

b3) O gráfico com as amostras apenas dos dois últimos anos (2022 e 2023) e as estimativas geradas pelo melhor preditor, além dos respectivos valores de RMSE e MAPE.

O gráfico dos dois últimos anos está representado na figura 5.

Para **análise 1**:

O valor de RMSE foi de 44956,94

O valor de MAPE foi de 3,81%

Para **análise 2**:

O valor de RMSE foi de 45149,58

O valor de MAPE foi de 3,84%

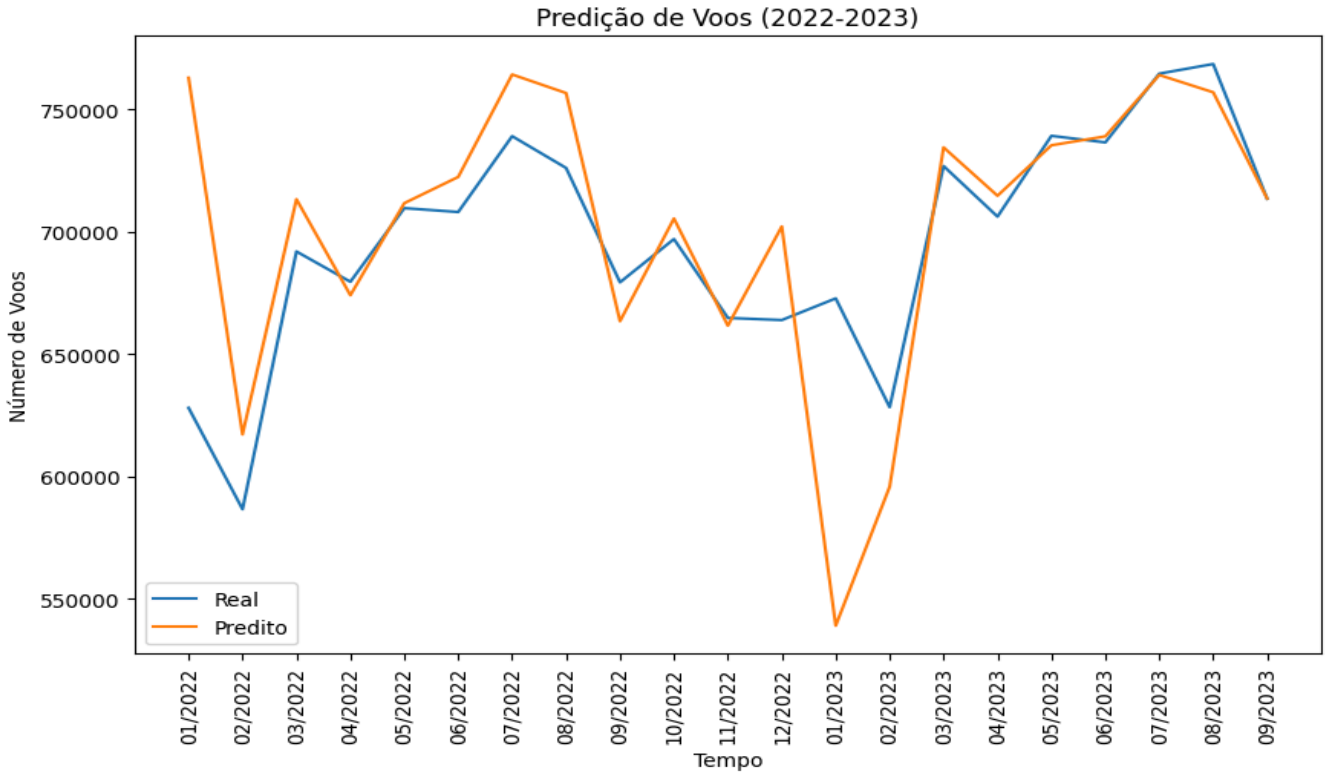


*Figura 5: Gráfico de estimativas do conjunto de testes em comparação com o valor real do número de voos no período de 2022 a 2023*

c) Repita o procedimento detalhado nos itens b1) e b2), mas adotando a seguinte divisão dos dados: (i) treinamento – amostras de 2003 a 2019; (ii) validação – amostras de 2020 e 2021; (iii) teste, com amostras de 2022 e 2023. Discuta os resultados obtidos e faça uma comparação com o cenário anterior (especialmente com o que foi obtido no item b3).

c1) Nesse caso, usamos dados do conjunto de treinamento para predizer os k primeiros meses do conjunto de validação. Viu-se que os resultados são bem parecidos.

Assim, o gráfico de RMSE dos conjuntos de treinamento e validação é dado pela figura 6.

Vê-se que desta vez os valores de RMSE apresentaram resultados bem piores em relação ao item b1. Além disso o k ótimo foi de k=5 e o RMSE para k=5 foi de 121371,61, um valor bem mais elevado que os do item b.

A hipótese para esse resultado pior para o conjunto de validação se deve ao fato que o treinamento foi realizado no período de 2003 até 2019, o período exatamente pré pandemia, enquanto o período de validação começou-se em 2020, logo quando começou-se a pandemia e o “Lock Down”. Pode-se verificar no gráfico da figura 1 que nos meses iniciais de 2020 há uma queda brusca e muito atípica da quantidade de voos, algo totalmente imprevisto e que não aconteceu no conjunto de treinamento e nem no conjunto de validação.

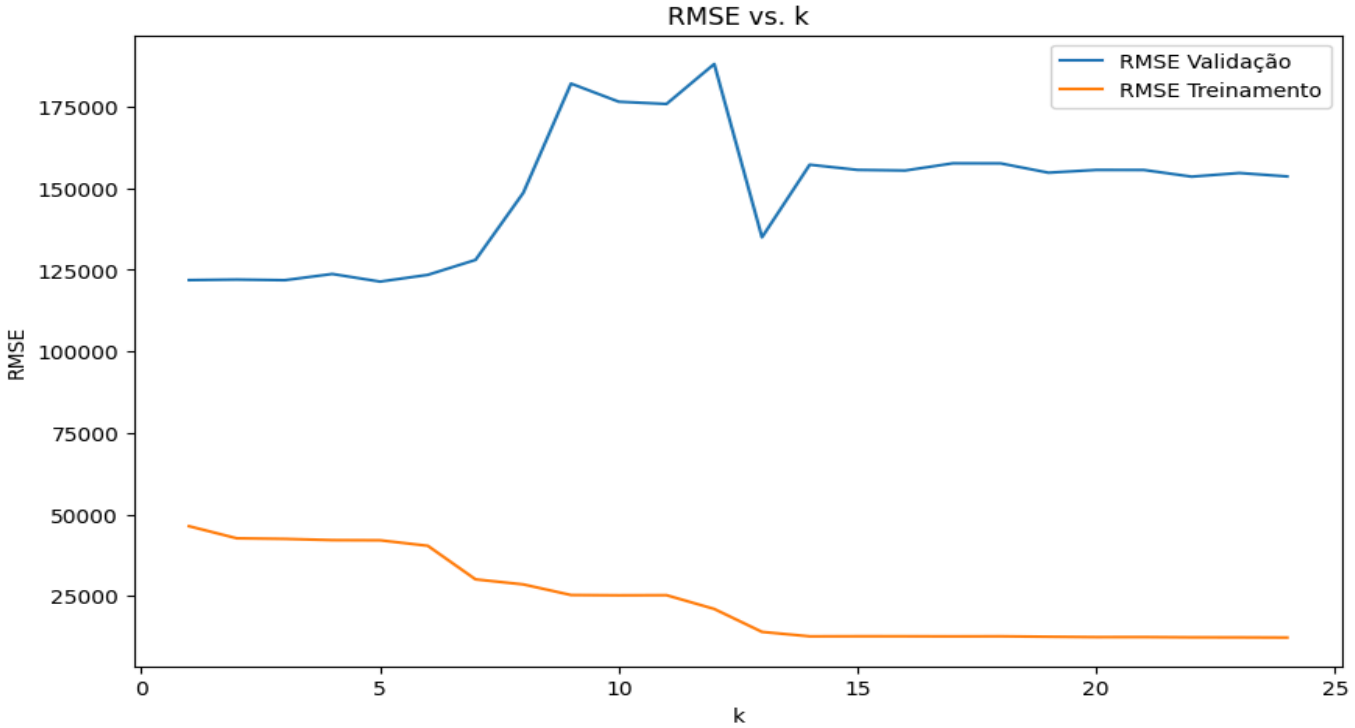


*Figura 6: Gráfico dos RMSE de Validação e Treinamento calculados de acordo com um valor específico de um k (quantidade de entradas usadas para realizar a predição do modelo), utilizando dados do conjunto de treinamento para validação, de acordo com item c.*

Dessa forma, o modelo não estava preparado para essa mudança abrupta do comportamento dos dados, o que gerou resultados bem ruins, com médias de erro acima de 120000. O melhor k foi para k=5 e o RMSE do melhor k foi de 121371,61.

Além disso, para k com valores baixos apresentou valores de RMSE melhores do que para k com valores altos. Até k=6 os valores de RMSE foram bem parecidos, começando a piorar para k=7. Isso se deve, provavelmente, porque um conjunto de dados maior para previsão acaba causando uma distorção maior, visto que os dados são completamente diferentes dos dados no período da pandemia. Logo, o modelo estava enviesado para um conjunto pré-pandemia, e prever o comportamento utilizando muitos dados enviesados acaba acarretando em erros maiores.

Durante o próprio período da pandemia o comportamento continua bem anormal. De 2020 até metade de 2021 há mudanças bem bruscas, e o comportamento começa a normalizar, seguindo um padrão, a partir da metade de 2021, como é possível ver no gráfico da figura 1.

c2) O gráfico do conjunto de teste se dá na figura 7.

O valor de RMSE foi de 39682,93

O valor de MAPE foi de 4,78%

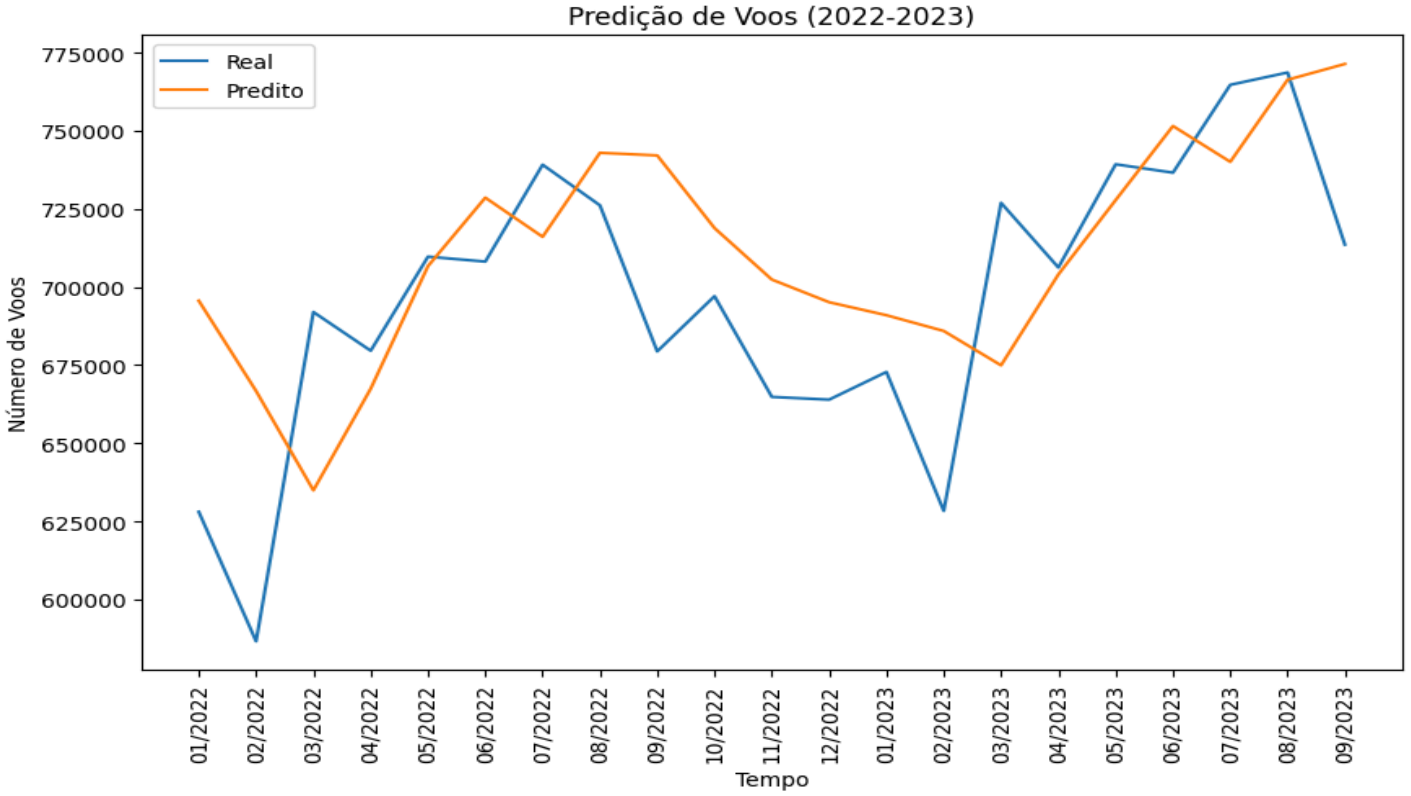


*Figura 7: Gráfico de estimativas do conjunto de testes em comparação com o valor real do número de voos no período de 2022 a 2023 do item c*

Como mencionado anteriormente, a partir da metade de 2021 o comportamento dos dados começam a apresentar um padrão, padrão este parecido com o período pré-pandemia. Em 2021 já começaram as vacinações para prevenção contra a Covid-19, e dessa forma o mundo começou a voltar às atividades normais aos poucos. Pode-se já classificar 2022 como período pós-pandemia.

Com a imunização e a volta das atividades comerciais, os dados apresentaram comportamentos parecidos ao período anterior a 2020, embora com valores ainda abaixo dos padrões de 2003 a 2019. Como o padrão voltou a ficar parecido, o modelo conseguiu resultado melhores no período de 2022 a 2023, apresentando um RMSE de 39682,93, que foi um valor até abaixo do valor obtido na análise b3, que se refere ao mesmo período, contudo, obtendo um MAPE maior, de 4,78%.

Então, no geral, a nova análise realizada no item C, apresentou um valor de RMSE pior em relação a análise feita no item b. Muito porque usamos dados pré-pandemia para treinamento, e depois utilizamos um conjunto de validação durante a pandemia, e com um modelo enviesado, os resultados do conjunto de validação acabaram ficando não tão bons. Porém, a predição dos dados do conjunto de testes, que se refere a um período pós-pandemia, apresentaram resultados bem melhores que o conjunto de validação, e esse comportamento se explica porque os dados pós-pandemia apresentam comportamento parecido com os dados de treinamento, embora os dados pós-pandemia ainda apresentarem quantidade de voos menores que os dados pré-pandemia. E dessa forma, os resultados de RMSE e MAPE do item c2 apresentam valores parecidos com os do mesmo período para o item b3.