Aplicação de Regressão Linear a Série Temporal*

Tiago C A Amorim (RA: 100675)[a], Taylon Xxxxxxx (RA: xxxxxx)[b]

[a] *Departamento de Engenharia de Petróleo da Faculdade de Engenharia Mecânica, UNICAMP, Campinas, SP, Brasil*
[b] *Departamento XXXX, UNICAMP, Campinas, SP, Brasil*

Abstract

xxxxxxx

Keywords: Regressão Linear, Séries Temporais, Validação Cruzada

1. Introdução

2. Tarefa Proposta

Trabalhar com a base de dados U.S. Airline Traffic Data, a qual contém informações referentes ao tráfego aéreo mensal norte-americano no período de 2003 a 2023, disponibilizadas pelo *U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics*. Em particular, vamos explorar a série temporal do número total de voos (domésticos e internacionais).

Explorar um modelo linear para a previsão considerando que o horizonte de predição é L = 1 (passos à frente da série temporal).

(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.
   (iii) Jan/2020 a Set/2023.

   Discuta possíveis razões históricas/econômicas para as transições de comportamento.

(b) Divida a série 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:

   (1) 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.
   (2) O gráfico com as amostras de teste da série temporal e as respetivas 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 teste.
   (3) O gráfico com as amostras apenas dos dois últimos anos (2022 e 2023) e as estimativas geradas pelo melhor preditor, além dos respetivos valores de RMSE e MAPE.

(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**: 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).

3. Aplicação

Toda a avaliação foi feita em um único *notebook* Jupyter, em Python. Foi feito o uso da biblioteca *Scikit-learn* [1] para fazer as diferentes manipulações nos dados.

3.1. Avaliação do Conjunto de Dados

O conjunto de dados utilizado nesta avaliação é o de tráfego aéreo dos EUA disponibilizado no Kaggle [2]. Segundo o autor este conjunto de dados fornece o tráfego aéreo mensal dos EUA de 2003 a 2023, incluindo o número de passageiros, o número de voos, as milhas de passageiros pagantes, as milhas de assentos disponíveis e o fator de ocupação.

Nesta avaliação o foco é no número total de voos. Observa-se na figura 1 que o período de janeiro/2003 a agosto/2008 (azul) é de certa normalidade na economia americana. O início do período em verde (setembro/2008) é marcado

*Relatório número 01 como parte dos requisitos da disciplina IA048: Aprendizado de Máquina.

pela crise financeira do *subprime*. Esta crise foi consequência do estouro da bolha imobiliária nos EUA devido a empréstimos sem lastro e preços de imóveis inflacionados. O período em vermelho corresponde à crise sanitária causada pela pandemia de Covid-19 e os anos que se seguiram. A queda mais significativa, que inicia em janeiro/2020, corresponde ao período de maior restrição de voos imposta durante a crise sanitária.

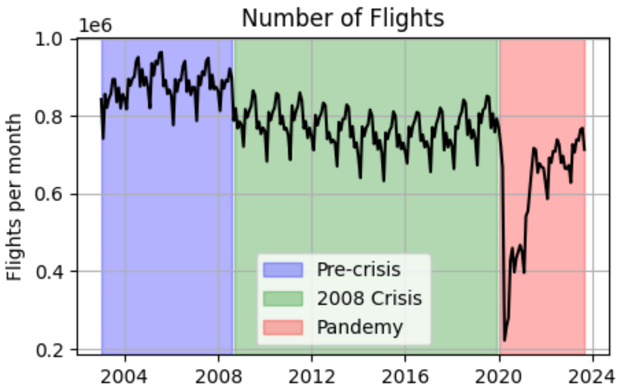

Figura 1: Série com total de voos nos EUA.

3.1.1. Pré-processamento

Alguns exemplos na figura 2 mostram que a série original apresenta um caráter cíclico (com frequência anual) e *serrilhado* (não-suave). Foram avaliadas duas alternativas para deixar a série temporal mais suave e facilitar o processo de regressão. Como os meses do ano não tem o mesmo número de dias, uma primeira tentativa foi de calcular o número médio de voos por dia para cada mês da série. A segunda tentativa foi de calcular o número de voos diários considerando apenas os dias úteis de cada mês.

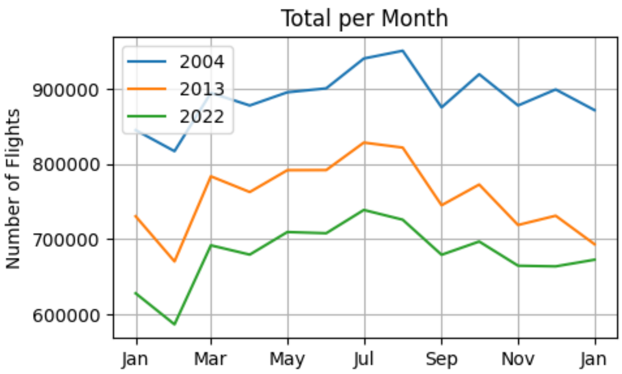

Figura 2: Exemplos de total de voos nos EUA ao longo do ano.

A figura 3 mostra que a série temporal de média diária de voos por mês se mostra mais suave que a série original. O efeito é mais pronunciado no mês de fevereiro. A série temporal de total mensal de voos dividido pelo número de dias úteis, na figura 4 não teve uma reposta adequada.

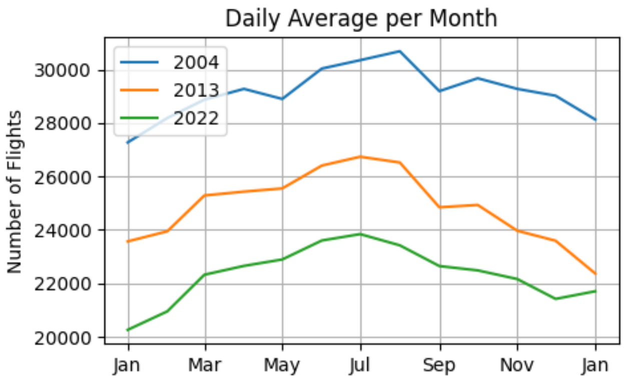

Figura 3: Exemplos de média diária do total de voos nos EUA.

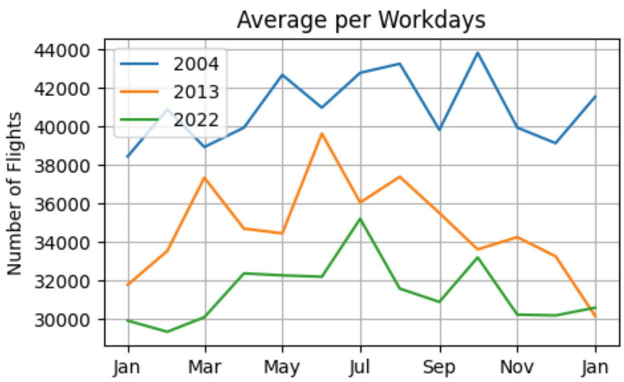

Figura 4: Exemplos do total de voos nos EUA pelo número de dias úteis no mês.

Optou-se por utilizar a média mensal de voos por dia nas análises posteriores (figura 5). Todos os valores de RMSE e MAPE foram calculados em função desta nova série.

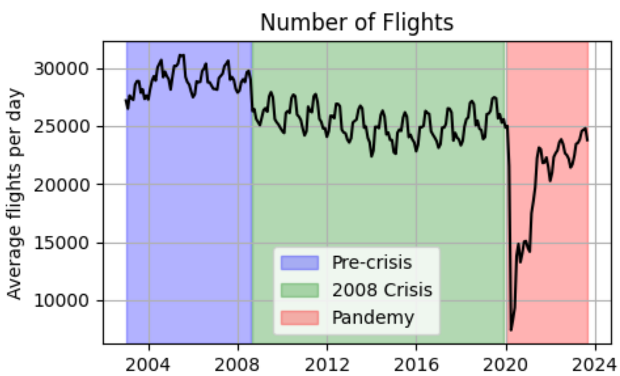

Figura 5: Média diária do total de voos nos EUA.

3.2. Primeiro Modelo de Regressão

Antes de separar os dados em treino+validação e teste, foram construídas séries temporais com os valores passados associado a cada entrada na série de voos[1]. Definindo a série de média diária de voos como $\mathbf{y}$, as séries dos valores passados foram definidas como:

$$\mathbf{x_k}(t) = \mathbf{y}(t-k) \quad \text{com } k = 1, \ldots, 24 \tag{3.1}$$

Onde t são os índices dos dados da série temporal.

De modo a ter uma melhor comparação entre os diferentes modelos que serão construídos, foram retiradas todas as linhas com valores não definidos (as 24 primeiras entradas da tabela de dados). Desta forma todos os treinos e validações serão feitos com os mesmos conjuntos de dados. Com este filtro o total de dados passa de 249 para 225.

Aplicando os limites propostos no item b1, o conjunto de dados de treino+validação fica com 180 elementos e o de teste com 45. Os intervalos de dados de treino+validação e teste apresentam um comportamento cíclico, com suaves variações ao longo dos anos (figura 6). A exceção é em 2008, quando eclodiu a crises financeira nos EUA. O intervalo de teste corresponde ao período crítido de restrição de vôos da pandemia de Covid-19 e os anos que se seguiram. É esperado que o modelo apresente dificuldades em prever a série temporal no período mais crítico da pandemia.

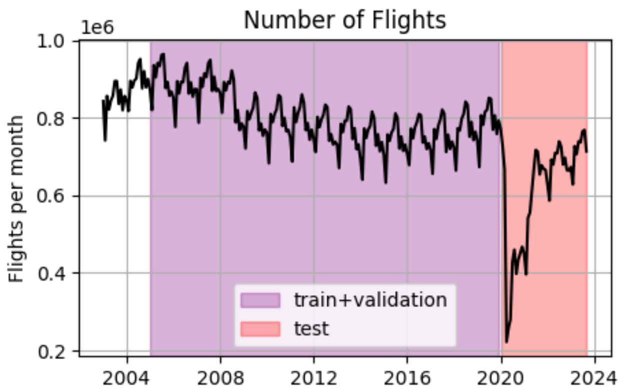

Figura 6: Períodos de treino+validação e teste.

3.2.1. *Normalização dos Dados*

Como todas variáveis de entrada para os modelos lineares são de mesma natureza, uma modificação na escala dos dados não terá qualquer impacto nos resultados dos ajustes dos modelos lineares. Foi realizado um rápido teste assumindo um modelo linear sem regularização. Como conjunto de teste foram utilizados os primeiros 140 dados do conjunto de treino+validação, e os demais como conjunto de validação. Foram utilizados $\mathbf{x_1}$ e $\mathbf{x_2}$ como variáveis do problema.

A tabela 1 mostra que, mesmo gerando modelos distintos, os resultados de RMSE são os mesmos quando as variáveis de entrada estão no intervalo $[0; 1]$[2] e quando estão com os valores originais.

RMSE	Valores Originais	Com Mudança de Escala
Treino	788.91	788.91
Validação	854.33	854.33

Tabela 1: Efeito da normalização dos dados de entrada.

Durante as iterações da rotina de busca do melhor modelo linear o algoritmo teve dificuldades de convergência com valores alto de $\mathbf{K}$ (>15). Ao modificar a escala dos dados de entrada o algoritmo não apresentou problemas. Deste modo, as análises seguintes foram feitas com os dados de entrada escalados para o intervalo $[0; 1]$.

3.2.2. *Validação Cruzada*

Como existe dependência entre os dados observados, o processo de validação cruzada *clássico*, que define os conjuntos de treino e validação sem levar em conta a sua ordem, não pode ser utilizado. Os dados de validação ($I(v)$) são definidos em índices posteriores aos dos dados de treinamento ($I(t)$) [3]:

$$\min_{i \epsilon I(t), j \epsilon I(v)} |i-j| > h > 0 \tag{3.2}$$

Foi utilizada a rotina **TimeSeriesSplit**, implementada no pacote *Scikit-learn*, e um total de 4 pastas (*folds*). A figura 7 mostra a divisão dos dados entre treino e validação para cada pasta.

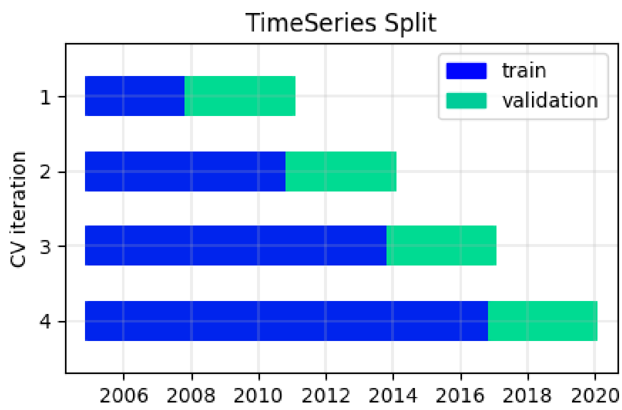

Figura 7: Divisão entre treino e validação para cada pasta.

3.2.3. *Modelos Simples*

Antes de seguir com o teste de diversos modelos lineares, foram propostos três modelos simples (*naïve*):

1. Igual ao passo de tempo anterior:
   $\hat{\mathbf{y}}(t) = \mathbf{y}(t-1)$

[1] Ao longo destes tópicos a série que se busca ajustar é a da média diária de voos.

[2] Foi utilizado o objeto **MinMaxScaler** do *Scikit-learn*

2. Igual ao passo de tempo 12 meses atrás:
   $\hat{\mathbf{y}}(t) = \mathbf{y}(t-12)$
3. Igual ao passo de tempo 12 meses atrás mais variação entre 24 e 12 meses atrás:
   $\hat{\mathbf{y}}(t) = \mathbf{y}(t-12) + [\mathbf{y}(t-12) - \mathbf{y}(t-24)]$

Os modelos *naïve* foram propostos com base na natureza cíclica da série temporal em estudo. Os valores de RMSE e MAPE destes modelos servirão como referência para os demais modelos a serem construídos.

Foi aplicada validação cruzada a cada um dos modelos e computada a média dos RMSE. Os resultados de RMSE são relativamente próximos: 861.25, 765.83 e 950.38, para o primeiro, segundo e terceiro modelos *naïve* propostos, respectivamente. Em uma inspeção gráfica (figura 8) a primeira proposta de modelo parece ser melhor que as outras duas. Uma avaliação do gráfico do erro cometido por estes modelos (figura 9) demonstra que na verdade os modelos tem níveis de erro parecidos, mas com diferentes frequências.

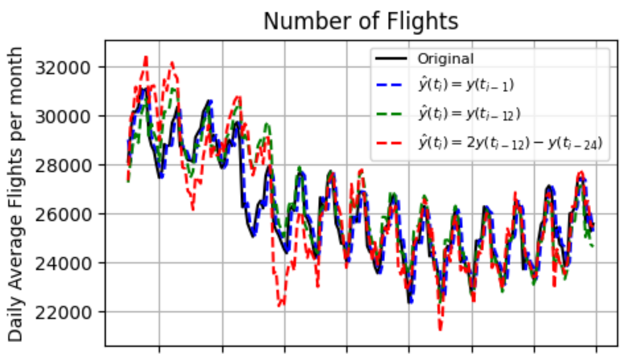

Figura 8: Modelos simples (*naïve*).

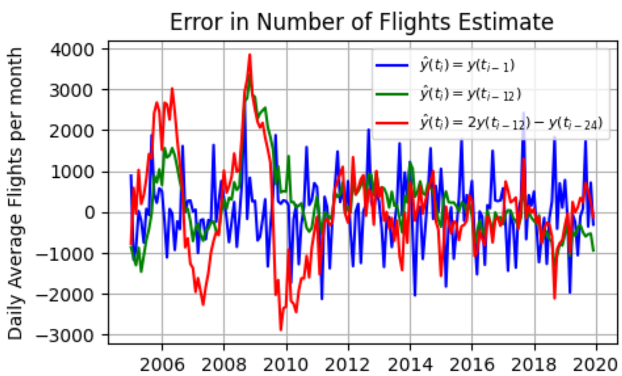

Figura 9: Erros dos modelos simples.

Por ter o menor valor de RMSE médio e apresentar um comportamento de erro melhor *comportado* que as outras alternativas, foi escolhido o segundo modelo para ser utilizado como base para comparação com os demais modelos.

3.2.4. Busca pelo Melhor Modelo

Em um primeiro momento foi definido utilizar um modelo linear regularizado do tipo *Elastic-Net*. A busca pelo melhor modelo para cada valor de **K** foi feita com uma busca em grid por parâmetros ótimos de **alpha** e **l1_ratio**. O parâmetro de busca do melhor modelo foi a média do RMSE.

Todos os 24 modelos encontrados utilizaram o menor valor de *alpha* que fez parte da busca (0.01). Em face destes resultados a avaliação foi refeita utilizando um modelo linear simples (sem regularização). Como o novo modelo não tem parâmetros, não foi preciso realizar uma busca em grid, e foi testado apenas o valor de **K**.

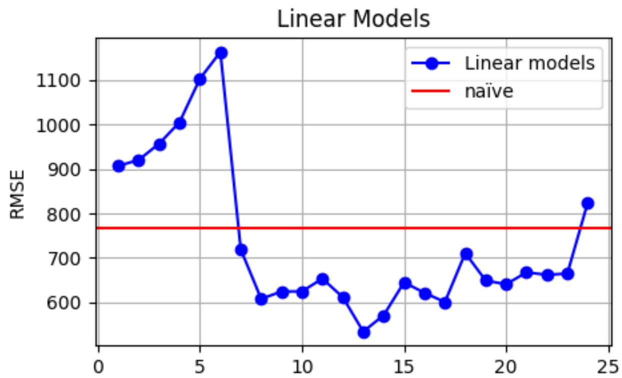

Figura 10: RMSE de validação médio dos modelos lineares em função de **K**.

O melhor resultado foi com **K** igual a 13 (validação cruzada com RMSE = 534.25 e MAPE = 0.02). Com o melhor valor de **K** definido, foi utilizado todo o conjunto de dados de treino+validação para ajustar os parâmetros do modelo escolhido.

Foi ajustado um segundo modelo com **K** igual a 13, mas sem normalizar as variáveis. Fica mais fácil analisar os valores dos coeficientes do modelo desta forma (figura 11). O modelo linear pode ser interpretado como uma média ponderada dos valores nos meses anteriores, já que as variáveis e o valor de saída tem todos a mesma natureza. A soma dos coeficientes com o valor relativo da constante[3] é próxima da unidade: $0.9687 + 0.0305 = 0.9992$.

O gráfico com os valores dos coeficientes utilizados mostra que as variáveis de maior peso foram o passo de tempo anterior ($\mathbf{x_1}$), um ano antes ($\mathbf{x_{12}}$) e o anterior ($\mathbf{x_{13}}$). Este resultado guarda certa correspondência com dois dos modelos naïve propostos.

3.2.5. Erros com os Dados de Teste

O modelo construído teve um resultado ruim nos dados de teste, com RMSE = 3187.39 e MAPE = 0.1126. Na figura 12 observa-se que o conjunto de teste (área em

[3]Foi definido valor relativo da constante como a divisão do termo constante do modelo linear pela média dos valores de saída.

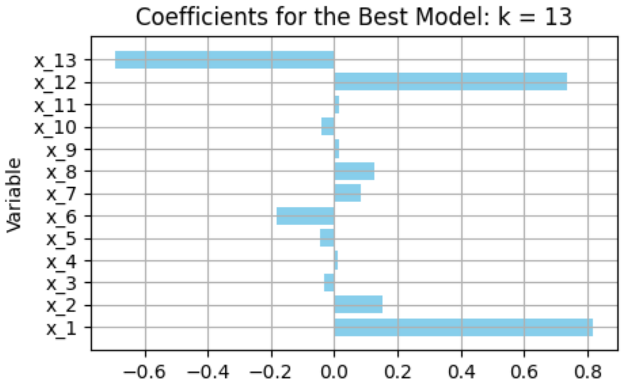

Figura 11: Coeficientes do melhor modelo linear ($\mathbf{K} = 13$).

vermelho na figura) coincide com o período da pandemia e os meses posteriores. Por se tratar de um evento singular e que não apareceu nos dados de treino e validação, já era esperado que o modelo tivesse dificuldade em prever o comportamento da série temporal. O alto valor de RMSE associado ao conjunto de teste demonstra a dificuldade do modelo.

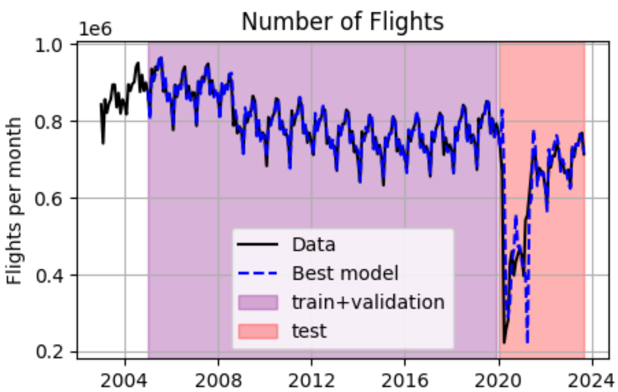

Figura 12: Ajuste do melhor modelo linear ($\mathbf{K} = 13$).

Quando o período de teste é limitado aos anos de 2022 e 2023 os resultados são bem melhores (RMSE = 575.85 e MAPE = 0.0190). O período de 2022 a 2023 corresponde ao final da pandemia de Covid-19, quando muitas das barreiras sanitárias já estavam sendo levantadas. O comportamento da série temporal neste intervalo já se assemelha mais ao comportamento observado nos dados de treino e validação (figura 13). Um valor de RMSE mais próximo do observado no processo de validação cruzada mostra que o modelo teve um resultado aceitável para este período (figura 14).

3.3. Segundo Modelo de Regressão

4. Resultados

O código foi implementado em Python em um *notebook* Jupyter. Pode ser encontrado em https://github.com/TiagoCAAmorim/machine_learning.

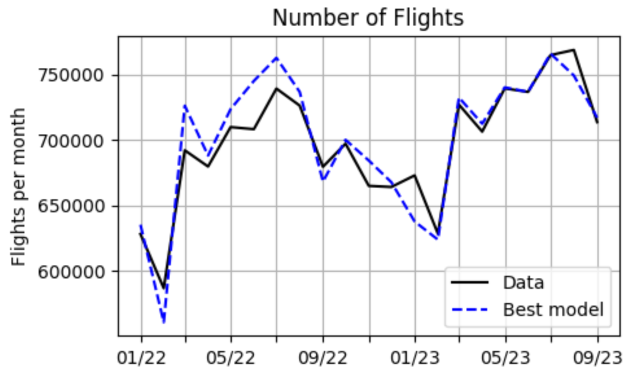

Figura 13: Ajuste do melhor modelo linear ($\mathbf{K} = 13$) ao anos de 2022 e 2023.

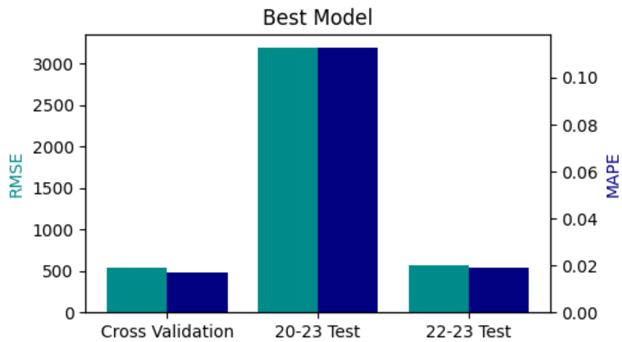

Figura 14: Resumo dos erros do melhor modelo linear ($\mathbf{K} = 13$).

5. Conclusão

Apêndice A. Lista de Variáveis

Referências

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[2] YYXian, U.s. airline traffic data (2003-2023), acessado: 22/03/2024 (Jan 2024).
URL `https://www.kaggle.com/datasets/yyxian/u-s-airline-traffic-data/data`

[3] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Statistics Surveys 4 (none) (2010) 40 – 79. `doi:10.1214/09-SS054`.
URL `https://doi.org/10.1214/09-SS054`