IA048 – Aprendizado de Máquina
Atividade 1 – Regressão Linear
Turma A – 1º semestre de 2024
Aluno: Renato de Souza Gomes - RA: 271702

A metodologia empregada aqui é uma abordagem comum para a previsão de séries temporais usando modelos de regressão linear. A ideia é usar valores passados da série temporal (janelas de tempo) como características para prever o valor atual. A seleção do melhor número de janelas de tempo é feita com base no erro de previsão no conjunto de validação.

1. **Pré-processamento de dados:**
   - 1.1. Carrega o arquivo CSV com a base de dados
   - 1.2. Cria uma coluna Data com as colunas Year e Month e atribui o dia 1
   - 1.3. Cria uma cópia do DataFrame original, selecionando apenas as colunas 'Data' e 'Flt'
   - 1.4. Remove vírgulas dos valores da coluna Flt
   - 1.5. Converte a coluna 'Flt' para um tipo numérico, tratando erros com a opção 'coerce' que substitui valores inválidos por NaN

2. **Preparação de dados:**
   Prepara o conjunto de dados para análise de séries temporais, transformando o conjunto de dados original em um formato que é adequado para análise de séries temporais, criando novas colunas que representam valores passados da variável de interesse. Cada nova coluna é um deslocamento da coluna 'Flt', o que significa que cada valor na nova coluna é o valor da coluna 'Flt' em um ponto de tempo anterior.

3. **Divisão dos dados:**
   O conjunto de dados é dividido em conjuntos de treinamento e validação(2003 - 2019) e teste(2020 - 2023). A estratégia de validação adotada foi a holdout, sua escolha é justificada pelas seguintes razões:
   - Simplicidade: O holdout é um método de validação simples de implementar. Ele envolve dividir o conjunto de dados em um conjunto de treinamento e um conjunto de teste, sem a necessidade de ajustar parâmetros adicionais ou realizar etapas complexas de validação cruzada Aqui foi utilizada a seguinte divisão: 70% dos dados de treinamento e validação para o treinamento e 30% para validação.
   - Generalização: O holdout permite avaliar a capacidade de generalização do modelo. Ao testar o modelo em um conjunto de dados não vistos, podemos verificar se ele é capaz de fazer previsões precisas em diferentes instâncias, além de verificar se há overfitting ou underfitting.

4. **Treinamento do modelo:**
   Este passo treina um modelo de regressão linear e calcula o erro quadrático médio (RMSE) para cada valor de k no intervalo de 1 a 24.
   - 4.1. Inicialmente, uma lista vazia rsmes é criada para armazenar os valores de RMSE para cada valor de k.

4.2. Em seguida, um loop é iniciado para cada valor de k no intervalo de 1 a 24. Para cada valor de k, as colunas correspondentes de atraso (lag) são selecionadas. Por exemplo, para k=3, as colunas Lag1, Lag2 e Lag3 seriam selecionadas.

4.3. Os conjuntos de dados de treinamento e validação são filtrados para remover quaisquer linhas que tenham valores ausentes nessas colunas de atraso. As colunas de atraso formam o conjunto de recursos X e a coluna Flt (que representa o número de voos) é o alvo y.

4.4. Normaliza tanto os dados de treinamento quanto os de validação. O que pode ajudar o modelo de regressão linear a convergir mais rapidamente.

4.5. Um modelo de regressão linear é criado e treinado usando os dados de treinamento.

4.6. O modelo treinado é então usado para fazer previsões nos dados de validação

4.7. O erro quadrático médio (RMSE) é calculado entre as previsões e os valores reais de validação.

4.8. O valor de RMSE é adicionado à lista **rsmes**.

4.9. Este processo é repetido para cada valor de k, resultando em uma lista de valores de RMSE para cada valor de k. Isso permite avaliar como o número de atrasos (k) afeta o desempenho do modelo de regressão linear.

5. Encontrar o melhor K e preparar os dados:

O valor de k que resulta no menor RMSE é selecionado como o melhor k. Os dados de treinamento em conjunto com os dados de validação são então preparados para este valor de k, selecionando as colunas que contém strings no formato Lag1, Lag2, ..., até LagK.

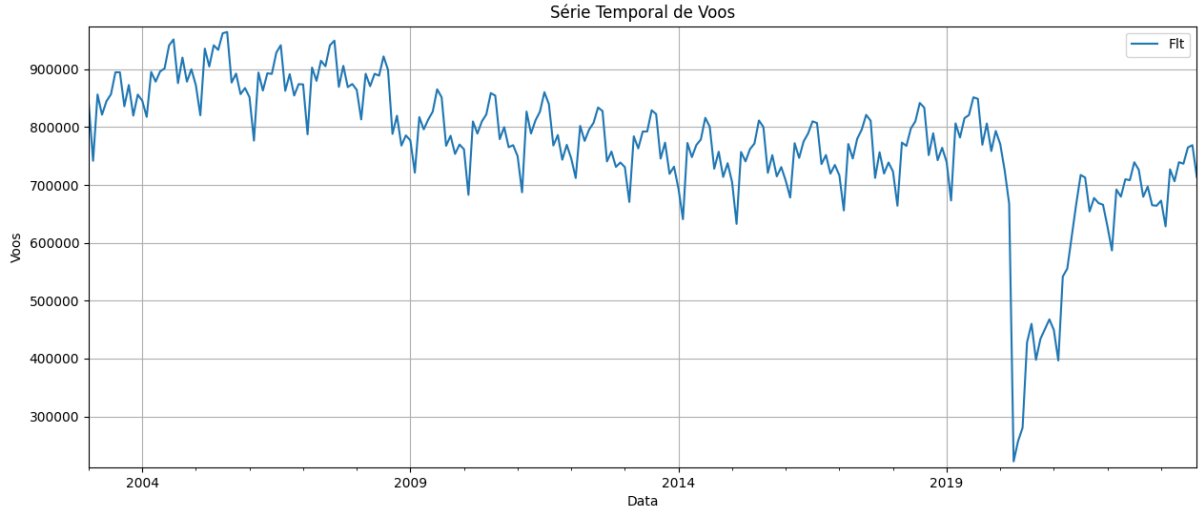
6. Normaliza os dados de entrada:

Normaliza os novos dados de treinamento, o que pode ajudar o modelo de regressão linear a convergir mais rapidamente.

7. Treinamento do modelo com o melhor K e previsão:

Treina o modelo de regressão linear nos novos dados e utiliza o modelo treinado para fazer previsões nos dados de teste.

Item a)



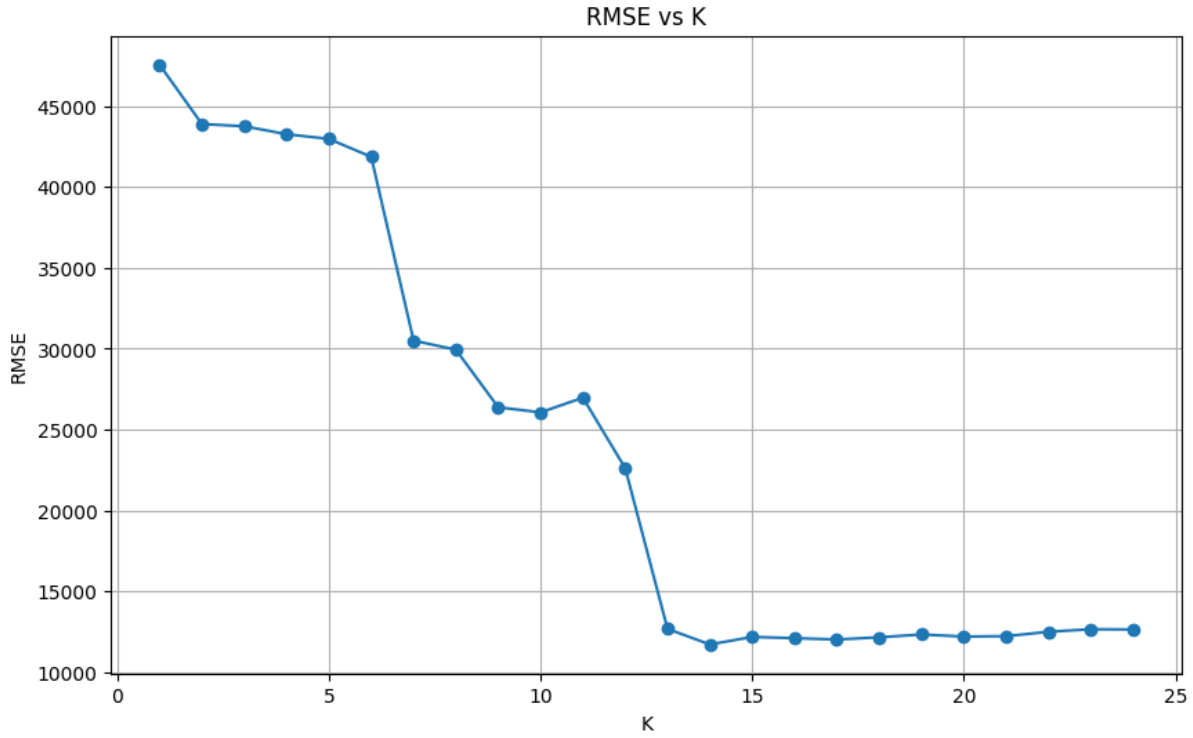
As transições de comportamento na série temporal podem ser influenciadas por diversos fatores históricos e econômicos. Aqui estão algumas possíveis razões para as três faixas distintas de comportamento:

Jan/2003 a Ago/2008: Durante esse período, a economia global estava em um período de crescimento estável após a recessão causada pela crise financeira de 2001. Vários países estavam experimentando um crescimento econômico robusto, o que pode ter levado a um aumento na demanda por transporte aéreo.

Set/2008 a Dez/2019: Essa faixa de tempo inclui a crise financeira global de 2008 e suas consequências. A crise financeira resultou em uma desaceleração econômica significativa em muitos países, levando a uma redução na demanda por viagens aéreas.

Jan/2020 a Set/2023: Essa faixa de tempo inclui a pandemia de COVID-19, que teve um impacto muito grande na demanda por transporte aéreo. As restrições de viagem, o fechamento de fronteiras e as medidas de distanciamento social resultaram em uma queda drástica na demanda por viagens aéreas.

Item b1)

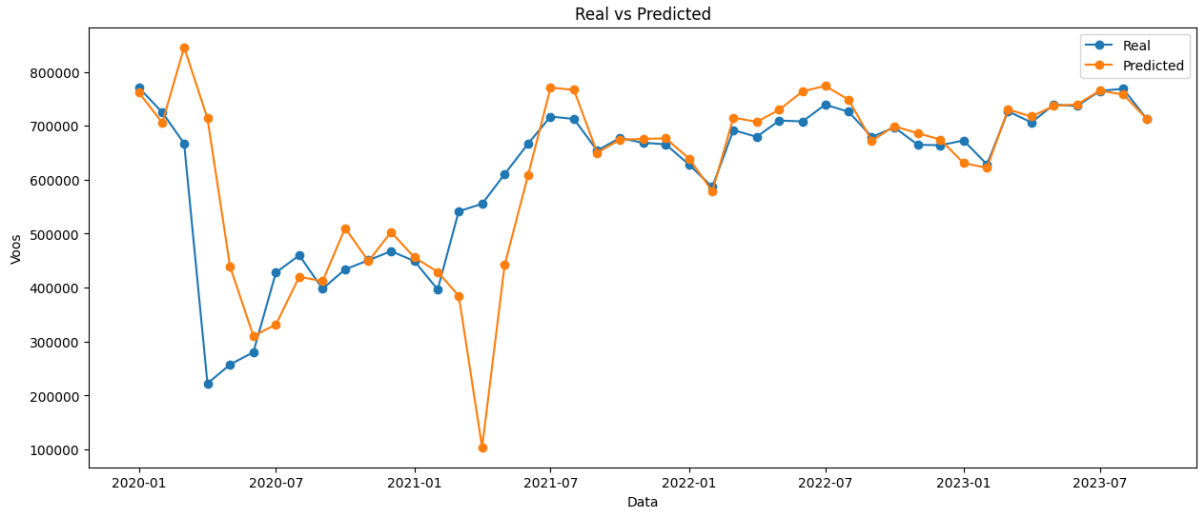


Melhor K = 14 com RMSE = 11722.431541631995

A partir dos resultados obtidos, podemos observar um comportamento interessante da métrica RMSE em relação ao número de entradas (K) do preditor. Vamos analisar algumas conjecturas sobre os motivos subjacentes a esse comportamento:

1. **K inicia com um valor alto e diminui gradualmente**: No início, com apenas uma entrada (K = 1), o modelo tem pouca informação para fazer previsões precisas, resultando em um alto valor de RMSE. Conforme aumentamos o número de entradas (K), o modelo tem mais informações históricas disponíveis para fazer previsões, o que geralmente leva a uma redução no RMSE.
2. **RMSE atinge um mínimo**: Podemos observar que o valor de RMSE atinge um mínimo em K = 14. Isso sugere que um modelo com 14 entradas é capaz de capturar bem os padrões e tendências nos dados de validação, resultando em previsões mais precisas.
3. **RMSE aumenta após atingir o mínimo**: Após o ponto mínimo, o valor de RMSE começa a aumentar gradualmente à medida que aumentamos o número de entradas (K). Isso pode indicar que adicionar mais entradas além de um certo ponto não melhora significativamente a capacidade do modelo de fazer previsões precisas. Pode haver um ponto de saturação onde o modelo começa a capturar mais ruído do que informações úteis dos dados.

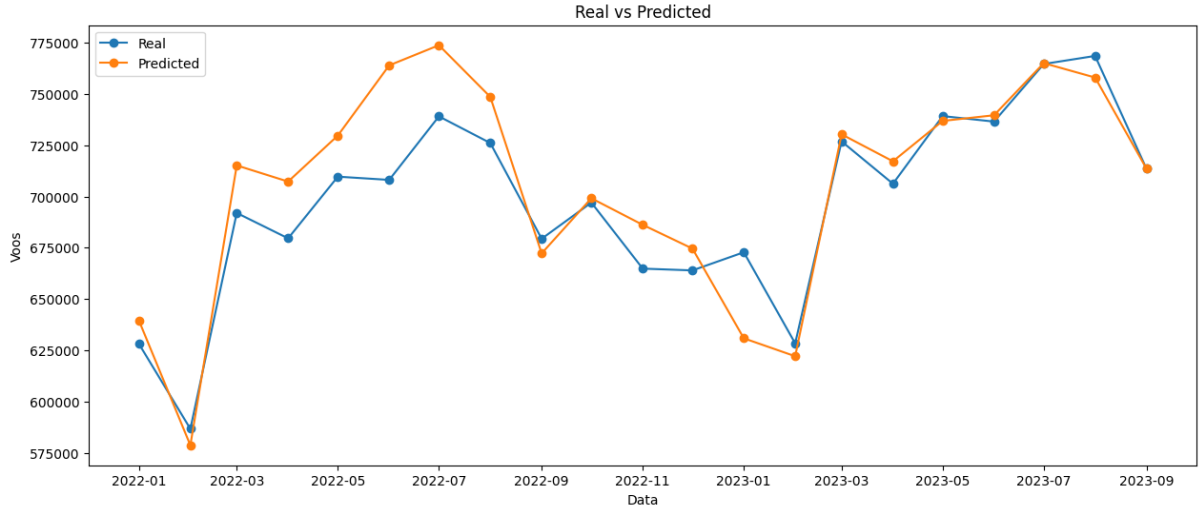
Item b2)



RMSE = 115719.37141006943
MAPE = 13.676157726430038

Os resultados obtidos indicam que o modelo de regressão linear utilizado apresenta um erro médio quadrático (RMSE) de 115719.37 e uma porcentagem média absoluta de erro (MAPE) de 13.68%. Portanto, com base nos resultados obtidos, podemos concluir que a previsão dos valores da série temporal em questão não foi muito precisa. Isso pode ser atribuído ao fato de que o modelo foi treinado antes da pandemia, enquanto os testes foram realizados nos dados do período da pandemia, no qual os números de voos caíram drasticamente. Essa mudança repentina e significativa no comportamento dos dados pode ter afetado a capacidade do modelo de fazer previsões precisas para este período de teste.

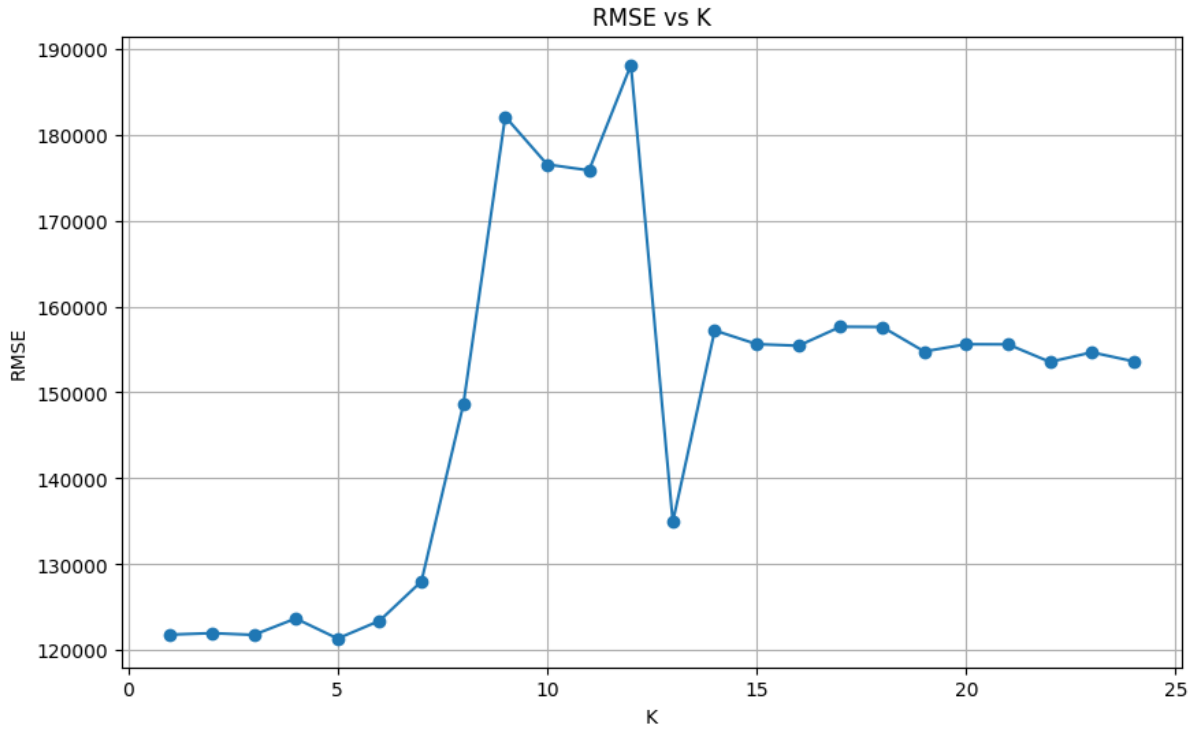
Item b3)



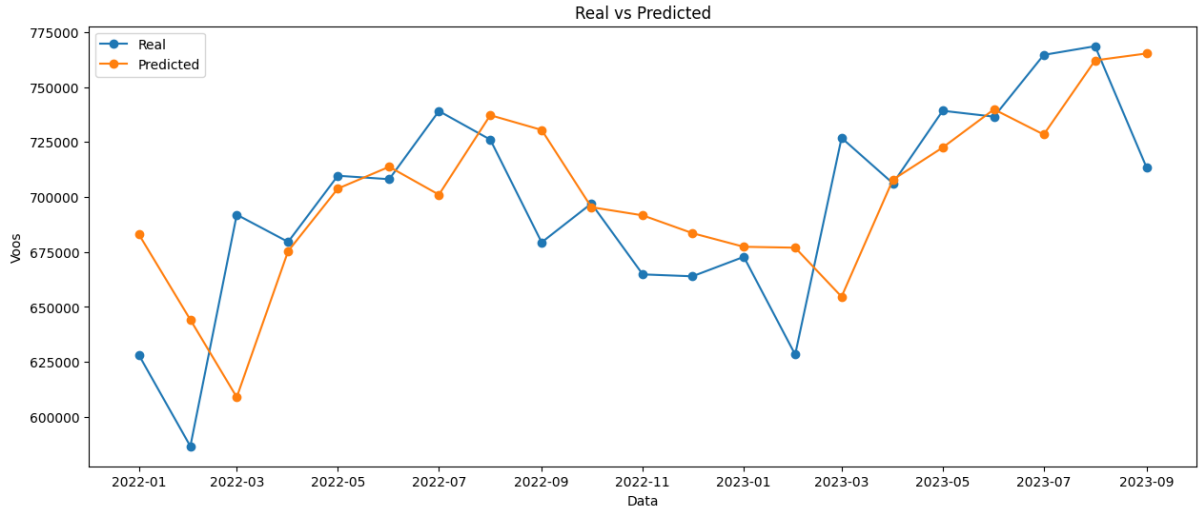
RMSE = 21184.426530544013
MAPE = 2.225182239150686

Ao aplicar o modelo aos dados de teste que incluem apenas os dois últimos anos da pandemia - um período em que a quantidade de voos se recuperou para níveis próximos aos anteriores à pandemia - observamos uma melhora significativa no desempenho do modelo. O RMSE reduziu para 21184.426530544013 e o MAPE para 2.225182239150686. Esses valores indicam que o modelo foi capaz de fazer previsões mais precisas quando aplicado a esses dados de teste, em comparação com os resultados obtidos no item b2. Portanto, podemos concluir que o modelo se adapta melhor a períodos de normalidade, onde os padrões de voo são mais consistentes e previsíveis.

Item C)



Melhor K = 5 com RMSE = 121371.60906118396



RMSE = 38127.679233339
MAPE = 4.207929673867221

O modelo foi treinado com dados anteriores à pandemia e validado com dados durante a pandemia. O resultado mostrou que o melhor valor para o parâmetro k foi 5. Isso significa que o modelo de regressão linear utilizado teve um desempenho melhor ao considerar as últimas 5 observações (Lag1 a Lag5) para prever o valor futuro.

Essa descoberta é interessante, pois indica que as últimas 5 observações têm uma influência significativa no valor futuro. Isso pode ser explicado pelo fato de que a pandemia teve um impacto significativo nas viagens aéreas, e as tendências recentes podem fornecer informações valiosas para prever o comportamento futuro. Ao contrário do item b1, em que o modelo precisa de mais informações históricas disponíveis para entender a dinâmica da série temporal para fazer previsões.

Portanto, com base nos valores de RMSE e MAPE, podemos concluir que o modelo anterior é mais preciso em prever os valores em comparação com o modelo atual.