A2 - Séries Temporais

Ari Oliveira André Costa Breno Marques Azevedo
Daniel Jacob Tonn Gustavo Ramalho Vinicius Hedler

27 de novembro de 2024

1. Objetivo

O objetivo deste trabalho é modelar a variável `Consumption` do conjunto de dados `us_change`, aplicando conceitos de estatística, *machine learning* e séries temporais. Inclui a análise de métricas, transformação de variáveis, ajuste de modelos e avaliação de diferentes abordagens, como modelos de suavização exponencial, regressão linear múltipla, SARIMA, redes neurais convolucionais (1D) e recorrentes (LSTM), utilizando Python. Os notebooks utilizados para a construção desse relatório estão disponíveis no Github.

2. Métricas e métodos de avaliação

Utilizaremos o Erro Absoluto Médio (MAE) como principal métrica de avaliação, devido à sua facilidade de interpretação e simplicidade. Adicionalmente, utilizaremos a Raiz do Erro Quadrático Médio (RMSE) para identificar possíveis discrepâncias significativas entre os valores previstos e os valores reais. Como os valores reais y_t aproximam-se de zero, o uso do Erro Percentual Absoluto Médio (MAPE) não é adequado. Por esse motivo, adotaremos o Erro Escalado Absoluto Médio (MASE) para a análise de erros percentuais, garantindo uma avaliação mais robusta. Por fim, também avaliaremos o R^2 para compreender quão bem nosso modelo se ajusta aos dados.

3. Análise exploratória

3.1. Interpretação das covariáveis

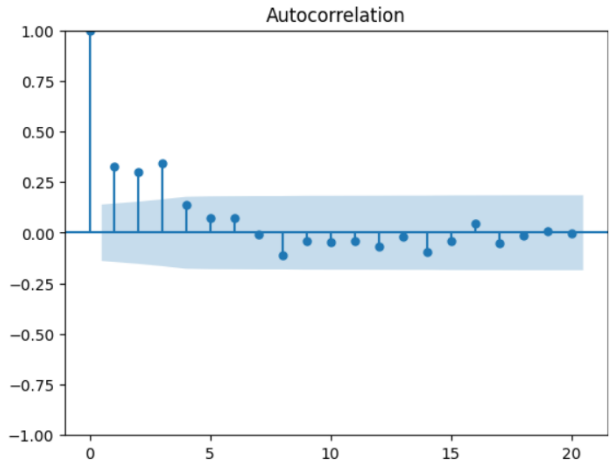
Realizando uma análise exploratória em torno das variáveis, observamos que `Savings` possui amplitude mais alta, com valores que variam de aproximadamente -57 e 41. Isso sugere que alguma transformação logarítmica para adequar as escalas possa ser interessante em alguns modelos. Além disso, a normalização *Z-Score* nas variáveis `Comsumption`, `Income`, `Production` e `Unemployment` pode ser valiosa em alguns contextos, principalmente em modelos sensíveis a escalas, como redes neurais. Isso será melhor explorado no ajuste de modelos.

3.2. Decomposição sazonal

Utilizamos a função `seasonal_decompose` para separar os componentes principais da série temporal - tendência, sazonalidade e resíduo. A imagem pode ser encontrada aqui.

A tendência não é tão clara, mas é possível notar que possui alguns ciclos. Por exemplo, um claro aumento de 1980 até 1985, depois um declínio até o início da década seguinte. As covariáveis presentes no dataset podem revelar novas informações que capturem melhor essa tendência.

A sazonalidade parece ter um pico maior em um dos trimestres, mas ainda não está tão claro se está ou não presente. O gráfico de ACF revela novas informações:



O gráfico de autocorrelação na figura 1 revela uma alta correlação em efeitos de curto prazo (lags 1, 2 e 3) - o que pode ser uma informação útil no desenvolvimento de modelos SARIMA, e também reforça a ideia de não sazonalidade. Também pode indicar não-estacionaridade, algo que será conferido por meio do *teste de Dickey-Fuller*.

Figura 1: Gráfico de Autocorrelação

Além disso, os resíduos da série possuem alguns picos ao longo dos anos, mas parecem se comportar normalmente em torno de 0.

Por fim, aplicamos o *Teste de Dickey-Fuller* para compreendermos se nossa série é ou não estacionária. Como nossa estatística de teste foi menor do que todos os valores críticos, rejeitamos a hipótese nula de não estacionariedade.

3.3. Correlação das covariáveis

Analisamos a multicolinearidade das covariáveis, pois havia a possibilidade de interdependência entre elas. Foi observado que `Savings` possui alto grau de correlação com `Income` (0.72) e `Unemployment` possui correlação negativa com `Production` (-0.77). Ambos os resultados parecem ter sentido, pois são relações econômicas muito diretas. A matriz de correlação completa pode ser encontrada aqui.

4. Ajuste dos modelos

4.1. Transformação de variáveis

A análise inicial das variáveis do dataframe indicou que todas apresentavam distribuições que se assemelhavam a uma normal. Por esse motivo, não foi necessário realizar ajustes para corrigir a distribuição. Em seguida, foram testadas algumas técnicas para lidar com valores extremos, incluindo a Winsorização na regressão linear. Além disso, avaliamos a aplicação da padronização por meio do z-score nos modelos. No entanto, essa técnica não trouxe diferenças significativas, uma vez que as distribuições já eram normais e centradas em zero. Essa técnica foi mais útil na LSTM e na rede neural convolucional.

4.2. Modelos utilizados

4.2.1. Baselines

Foram utilizados modelos ingênuos como benchmarks iniciais. Separamos os dados em treino (80%) e teste (20%). Escolhemos o `Mean Method`, `Naive`, `Seasonal Naive` e `Drift Method` para esta etapa. Esses modelos simples não terão grandes mudanças nos dados, como transformações ou técnicas adicionais. Isso serve para entendermos os resultados da maneira mais simples possível.

Modelo	MAE	RMSE
Mean Method	0.29	0.36
Naive	1.01	1.05
Seasonal Naive	1.18	1.25
Drift Method	1.14	1.19

Tabela 1: Métricas dos baselines

Pelos resultados, a série temporal parece não ter uma tendência linear forte. O método da média, que funciona bem para séries sem tendência ou sazonalidade forte, mostrou-se melhor que os demais. O `Naive Method` também foi melhor que o `Seasonal Naive Method`, o que reforça essa premissa. Os gráficos com todas as predições podem ser encontrados aqui.

4.2.2. Suavização Exponencial

Utilizamos também modelos de suavização exponencial, que utilizam pesos exponencialmente decrescentes para observações passadas, priorizando observações mais recentes. Pode ser útil para o caso de nossos dados mudarem ao longo do tempo, mas com alguma regularidade.

Inicialmente, aplicamos um modelo de suavização exponencial simples, e depois seguimos para métodos mais avançados, como Holt-Winters (que podem capturar tendência e sazonalidade).

Modelo	MAE	RMSE	MASE
Simples	0.95	1.0	1.55
Holt	1.1	1.15	1.8
Holt-Winters	1.08	1.14	1.77

Tabela 2: Métricas das suavizações exponenciais

Os métodos de Holt assumem a tendência e o nível da série temporal. Como discutido anteriormente na sessão de análises, a série não possui uma tendência clara e esses métodos não são os melhores para a situação.

Portanto, os resultados foram piores até mesmo que a suavização exponencial simples. Em comparação com um modelo base, os métodos também se mostraram piores, como indica O MASE > 1.

4.2.3. Regressão Linear Múltipla

Modelos de regressão foram ajustados usando variáveis preditoras significativas. Fizemos uma transformação na variável "Quarter" para funcionar nessa modelagem como uma covariável. Apenas adaptamos os valores para serem números sequenciais, que vão de 1 a 198, representando cada ano/trimestre. Os seguintes cenários foram considerados:

- Modelo com todas as covariáveis inclusas;
- Modelo com as covariáveis que apresentavam menor erro (`Income, Savings, Quarter_order`);
- Modelo utilizando Winsorização no limite superior de `Savings`;
- Log-Sign Transformation em `Savings`.

Curiosamente, as variáveis que apresentaram menor erro conjuntamente são `Savings` e `Income` - que possuem uma multicolinearidade considerável. Entretanto, isso sugere que ambas possuem informações distintas que o modelo utiliza para as previsões e podem estar correlacionadas com Consumption de diferentes modos.

Mas o melhor desses modelos foi utilizando Winsorização com limite superior de 3%, ou seja, valores que passam desses 3% são limitados ao valor limite. Testamos com outros limites, mas esse foi o melhor ajustado. Isso sugere que os valores extremos inferiores são importantes para as predições, mas os superiores não. As métricas serão exibidas no final do relatório. Abaixo, a visualização dessa melhor predição, e as outras podem ser encontradas neste link.

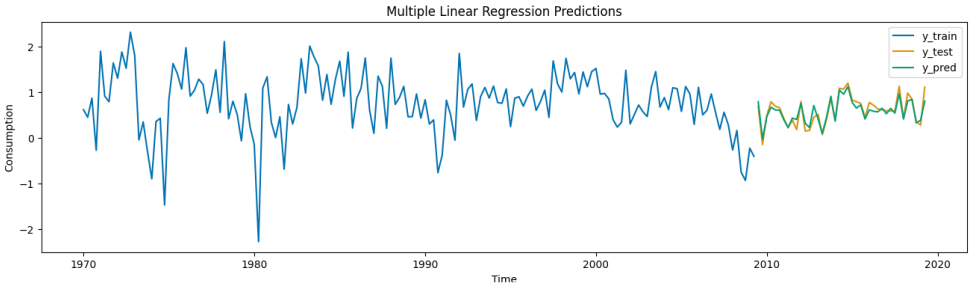


Figura 2: Melhor modelo de regressão linear múltipla

Os resultados indicam que, embora a regressão linear múltipla simples e com seleção de variáveis tenham apresentado um desempenho semelhante ($R^2 = 0.76$), a winsorização ajudou a reduzir a influência de valores extremos, melhorando o ajuste do modelo ($R^2 = 0.87$).

4.2.4. Modelos SARIMA

Em prosseguimento, modelos SARIMA também foram ajustados aos dados. Como supracitado, o gráfico ACF sugeriu a aplicação de modelos como uma abordagem interessante. Nesse sentido, foram aplicadas as variantes SARIMA e SARIMAX.

Primeiramente, o modelo SARIMA foi ajustado sem diferenciação e com ambas as partes autorregressivas e média móvel sendo 3. Os resultados ficaram aquém do esperado, obtendo `MAE` 0.82, `RMSE` 0.89 e `MASE` 1.35. Em seguida, foi ajustado o modelo SARIMAX, com os mesmos parâmetros que o anterior, mas agora temos a possibilidade de usar as outras covariáveis. O modelo obteve `MAE` 0.11, `RMSE` 0.14 e `MASE` 0.18, e o gráfico com predições pode ser encontrado neste link.

4.2.5. Convolução 1D e RNN (LSTM)

A arquitetura específica das LSTMs permite que esse tipo de modelo regule o fluxo de informações ao longo da rede, priorizando dados importantes. Com exceção de `Quarter`,

os dados foram normalizados usando uma transformação `MinMaxScaler`. A rede ajustada é composta por duas camadas: uma camada LSTM com 50 unidades de saída e ativação *ReLU*, seguida por uma camada densa que serve como saída do modelo, produzindo a previsão final. O modelo teve resultados um pouco promissores, com `MAE` 0.33 , `RMSE` 0.44 e `MASE` 0.56.

A arquitetura de CNN utiliza três camadas `Conv1D`, cada uma contendo 4 filtros, kernel de tamanho 2, ativação *leaky ReLU*, regularização L2 (`0.001`) e `padding='same'`. Foi utilizado um `Dropout` de `0.2` e `GlobalAveragePooling1D` ao final para agregar os filtros. Aqui usamos as mesmas técnicas de trnasformação dos dados anteriormente utilizadas, e em adição foi utilizado OneHotEncoding na variável `Quarter`, coletando como uma classe o trimestre do dado. Foi testada a utilização ou não de MinMaxScaler em `Consumption`, ao final não sendo encontradas grandes diferenças no resultado com ou sem ele. Ao final, os resultados de CNN ficaram piores que os baselines, com as previsões pontuais muito próximas da média dos dados.

5. Conclusão

A Tabela 3 apresenta uma comparação das métricas do melhor desempenho de cada modelo avaliado.

Modelo	**MAE**	**RMSE**	**MASE**	**R^2**
Baseline (mean method)	0.29	0.36	-	-
Suavização Exponencial	0.95	1.0	1.55	-9.26
Regressão Linear Múltipla	**0.09**	**0.11**	**0.14**	**0.87**
SARIMAX	0.11	0.14	0.18	0.79
Convolução 1D	0.64	0.80	0.87	-1.09
RNN (LSTM)	0.33	0.44	0.56	-0.92

Tabela 3: Comparação dos Modelos com Métricas de Desempenho

Os dois modelos com melhor desempenho são o de Regressão Linear Múltipla e Sarimax. A aplicação da Winsorização no limite superior da variável Savings melhorou significativamente o desempenho do modelo de regressão linear. Isso sugere que os valores extremos superiores de Savings estavam influenciando negativamente o modelo, talvez introduzindo ruído ou distorcendo a relação real entre as variáveis. Já o SARIMAX aproveitou tanto as dependências temporais intrínsecas da série (componentes autorregressivos e de média móvel) quanto as informações das covariáveis exógenas (Income, Savings, etc.), resultando em previsões mais precisas.