PREDIÇÃO DE TEMPERATURA MÉDIA DO CLIMA DIÁRIO EM DELHI

SSC0277 - Competições de Ciências de Dados

Nome: Victor Kendi Arakaki N°USP: 11219092

Sumário

1. Descrição do problema

O dataset disponibilizado possui dados climáticos de Delhi, Índia no período entre 1 de janeiro de 2013 até 24 de abril de 2017. Estes dados possuem o histórico diário das seguintes métricas: temperatura média, umidade, velocidade do vento, pressão média, além da data correspondente. O objetivo do problema é criar um preditor de temperatura média a partir das outras métricas em uma análise de série temporal.

2. Análise dos dados

2.1 Explicação das variáveis

O dataset foi coletado pelo Weather Undergroud API e possui 5 variáveis, os quais foram citados anteriormente. A explicação de cada variável consta a seguir:

- date: data no formato YYYY-MM-DD
    - Numérico
    - 5314 valores únicos
- **meantemp**: Temperatura média calculada a partir de múltiplos intervalos de 3 horas num dia.
    - Numérico
- humidity: Valor de umidade para o dia (as unidades são gramas de vapor de água por metro cúbico de volume de ar).
    - Numérico
- wind_speed: Velocidade do vento medida em km/h.
    - Numérico
- meanpressure: Leitura da pressão climática (medida em atm).
    - Numérico

A variável target do problema é a **meantemp**, enquanto os outros atributos são utilizados na predição.

2.2 Análise dos dados

O dataset não possui nenhum dado faltante em nenhum atributo. Assim, não precisou-se de utilizar técnicas para suprir a falta de dados como preenchimento ou exclusão das linhas, por exemplo. Além disso, percebe-se que todos os atributos são do tipo numérico, mais especificamente do tipo float, o que permite serem utilizados na análise da série temporal. Vale notar que o dataset disponibilizado no Kaggle já estava separado em conjunto de treinamento e de teste, o que facilitou ainda mais no desenvolvimento do problema.

Nos gráficos temporais abaixo, verifica-se que os dados tendem a apresentar uma sazonalidade em todos os casos, menos no campo meanpressure que aparentemente possui outliers. Escolheu-se preservar estes valores estranhos deste campo por praticidade, visto que possui pouquíssimos valores deste tipo. Nota-se que para gerar os gráficos, foi concatenado os conjuntos de treinamento e de teste para ter visualização total do conjunto de dados. A variável **target** possui a melhor percepção de sazonalidade, o que pode aparentar uma melhor facilidade em sua predição.

Gráfico de humidity:

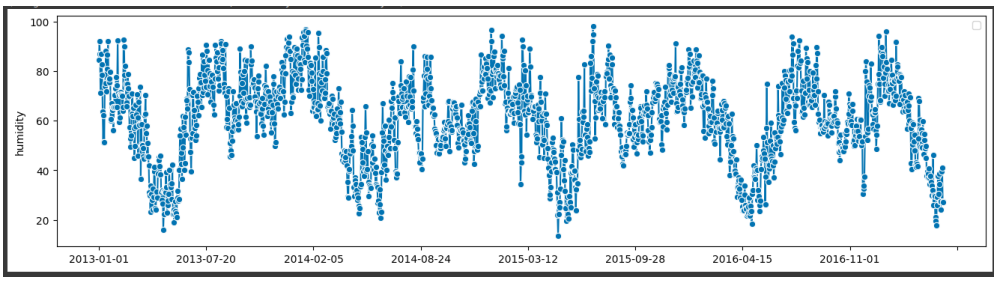


Gráfico de wind_speed:

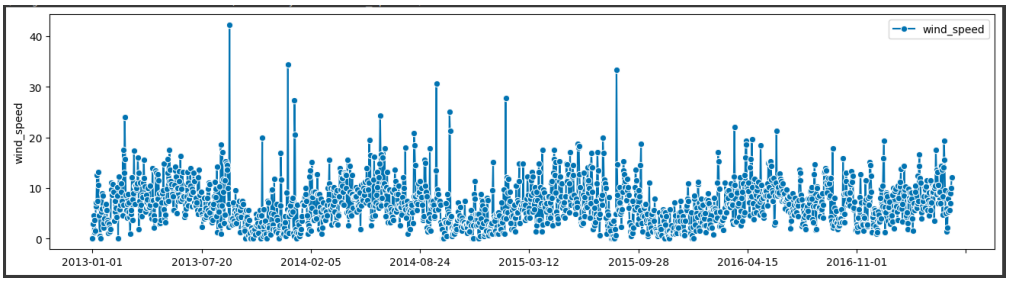


Gráfico de meanpressure:

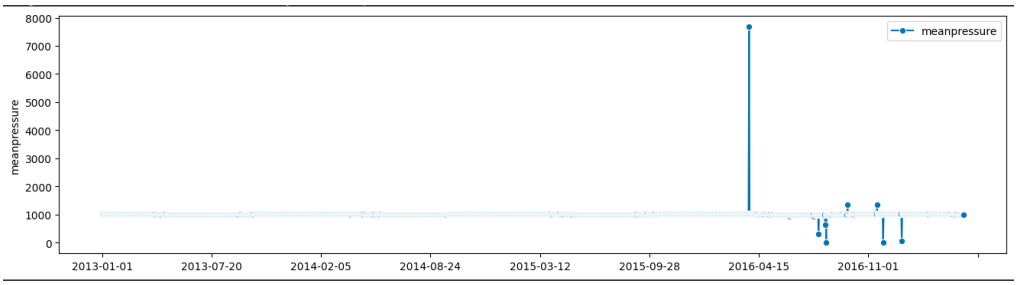
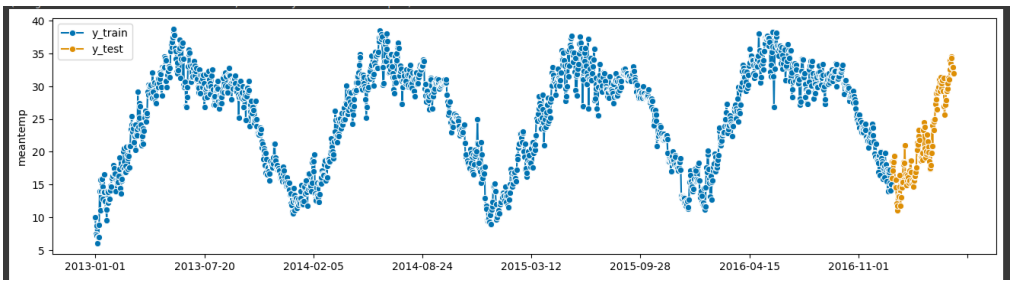
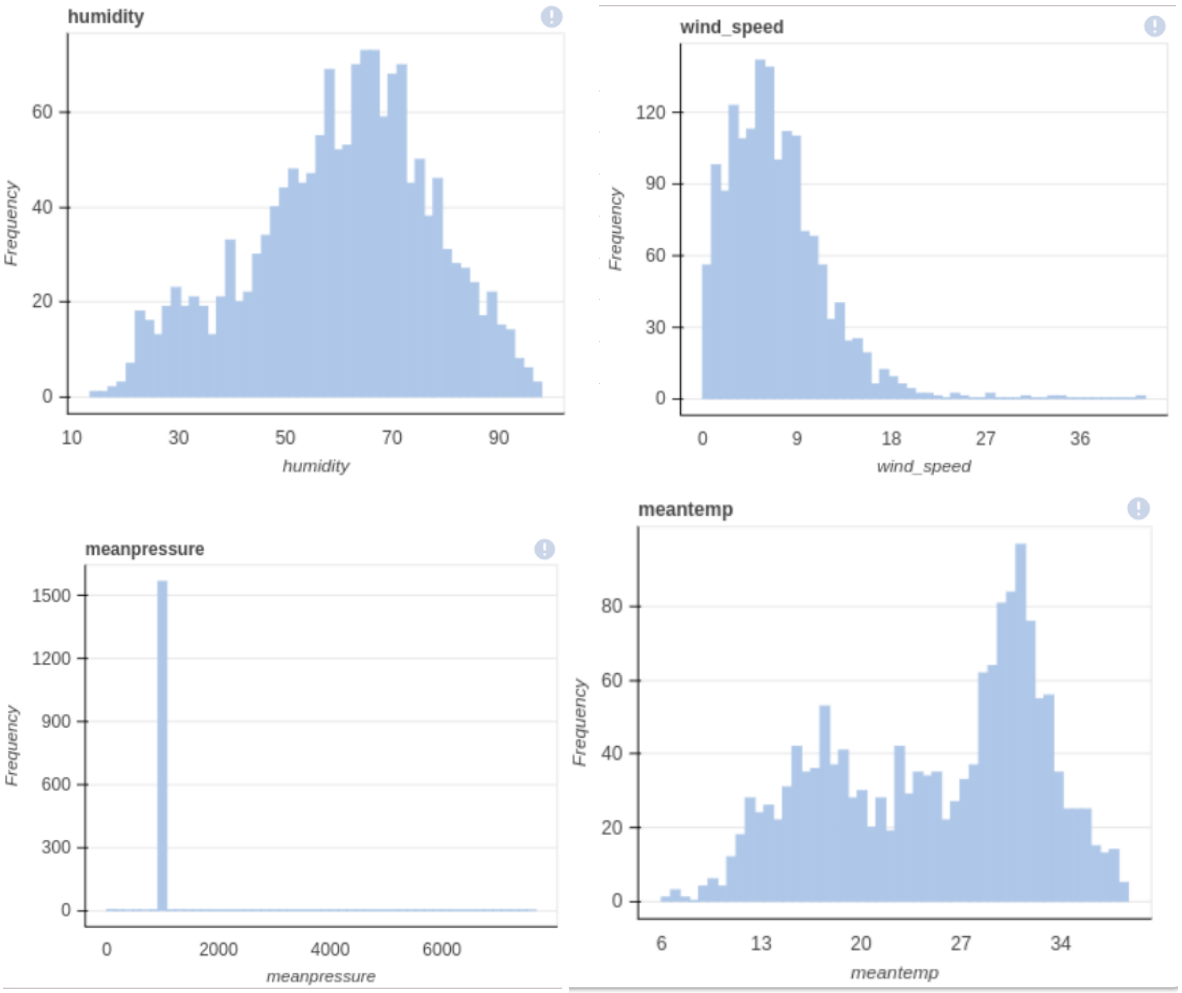


Gráfico de meantemp:



Nos gráficos seguintes, pode-se notar que os atributos humidity e wind_speed parecem seguir uma distribuição normal, com wind_speed com valor alto de skewness e humidity com um valor um pouco baixo. Já a variável meanpressure, nota-se que a sua distribuição parece ser sem variação, ou seja, possui valores bem próximos ou iguais. Apesar da dificuldade encontrada em visualizar por causa dos valores outliers, é possível perceber o padrão citado tanto no gráfico temporal quanto no gráfico de distribuição de valores. Já a variável **target**, apesar de não parecer tanto seguir uma distribuição normal, ela parece se assemelhar ao gráfico de humidity, talvez possuindo certa correlação.



A seguir está a matriz de correlação de Pearson. Nota-se que a maior correlação é entre humidity e meantemp, como já era esperado pela distribuição similar dos valores e pela sazonalidade notada em ambos os casos nos gráficos temporais. Parece que há uma certa correlação também das duas variáveis com wind_speed, o que já não se pode dizer sobre o atributo meanpressure que parece que não há correlação nenhuma com nenhum atributo.

2.3 Benchmarks

A métrica utilizada para analisar os algoritmos de predição de séries temporais foi:

- Erro Percentual Absoluto Médio (MAP):

O erro percentual médio absoluto (MAP - Mean Absolute Percentage Error) é uma medida de precisão comumente usada para avaliar a acurácia de previsões ou estimativas em relação aos valores reais. Ele é calculado como a média dos valores absolutos dos erros percentuais individuais. O resultado do MAPE é expresso como uma porcentagem, indicando o erro médio absoluto em relação aos valores reais. Quanto menor o valor do MAPE, mais precisa é a previsão ou estimativa.

3. Descrição e resultados das técnicas utilizadas

3.1 Técnicas de Predição em Séries Temporais

Todas as técnicas utilizadas possuíram como parâmetro o valor de sazonalidade sp=365, ou seja, o ciclo de 1 ano. Além disso, foi passado como parâmetro a variável target em conjunto das outras variáveis para uma melhor predição e o valor de ForecastingHorizon, ou seja, os valores absolutos das datas os quais vão ser preditos os valores de meantemp.

A seguir uma descrição sobre as técnicas utilizadas e resultados obtidos:

3.1.1 Naive Forecaster

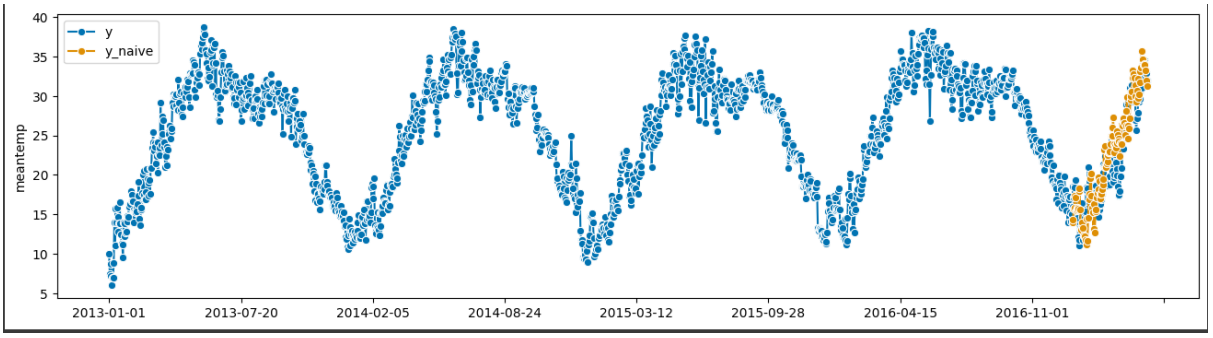
Naive Forecaster é uma técnica de previsão simples que assume que o valor futuro de uma série temporal será igual ao valor atual. Ela não leva em consideração nenhuma informação adicional, como tendências ou sazonalidade, e é adequada apenas para séries temporais estáveis e sem padrões claros. Apesar de sua simplicidade, pode fornecer resultados razoáveis em cenários de curto prazo ou quando não há dados suficientes para modelos mais complexos.

OBS: Apesar do modelo inicialmente não levar em conta a sazonalidade, optou-se por passar como parâmetro para obter um melhor resultado.

3.1.1.1 Estratégia Last

Esta estratégia utiliza-se como base o último valor conhecido pelo modelo, ou seja, do conjunto de treinamento.

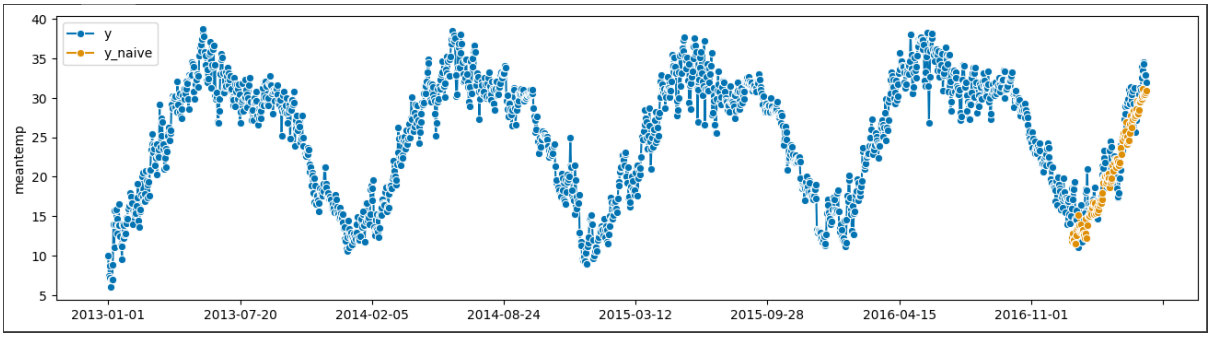
Métrica	**Naive Forecaster Last**
MAP	0.13628



3.1.1.2 Estratégia Mean

Esta estratégia utiliza-se como base o valor médio.

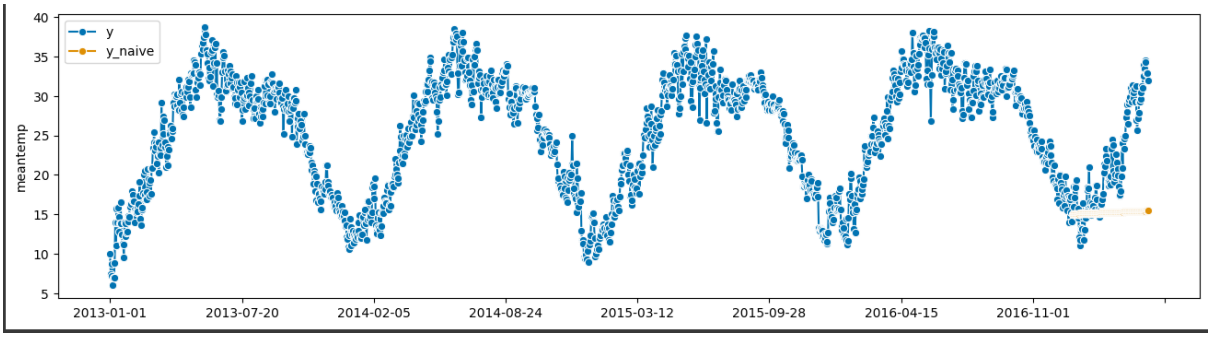
Métrica	Naive Forecaster Mean
MAP	0.11029



3.1.1.3 Estratégia Drift

Esta estratégia utiliza-se como base o último valor conhecido pelo modelo, ou seja, do conjunto de treinamento.

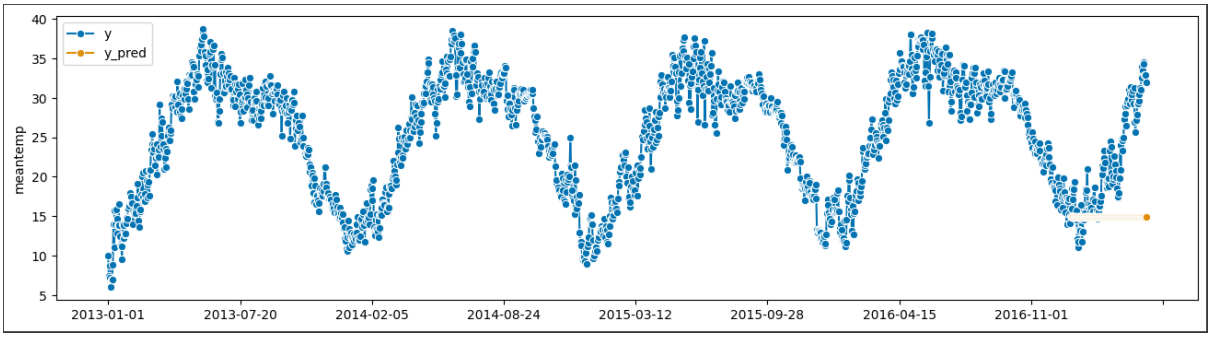
Métrica	Naive Forecaster Drift
MAP	0.26938



3.1.2 Auto ETS

Auto ETS (Error-Trend-Seasonality) é uma técnica automatizada de previsão que utiliza um algoritmo avançado para identificar automaticamente os melhores modelos de previsão para séries temporais. Ela incorpora três componentes principais - erro, tendência e sazonalidade - e ajusta automaticamente os parâmetros desses componentes para obter a melhor previsão possível. O Auto ETS é útil quando as séries temporais possuem padrões complexos e é capaz de lidar com diferentes tipos de sazonalidade e tendências de forma eficiente. Ele é amplamente utilizado em análise de demanda, previsão de vendas e outras áreas que dependem de projeções precisas de séries temporais.

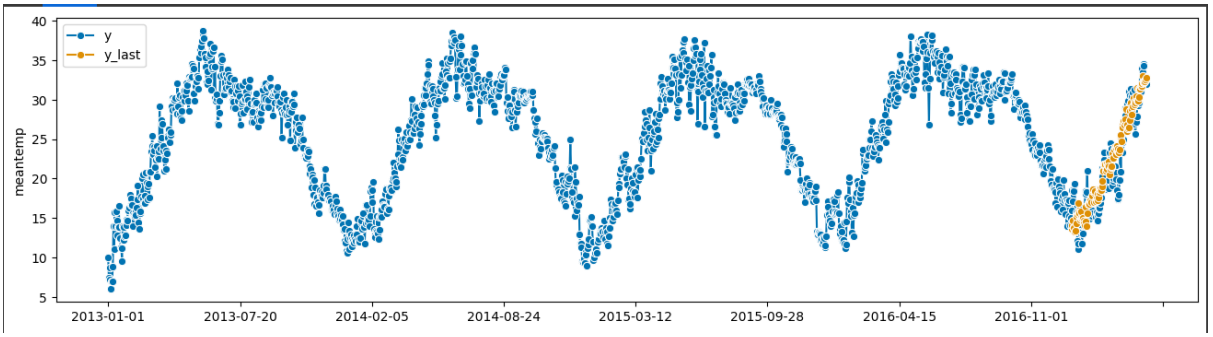
Métrica	Naive Forecaster Drift
MAP	0.28133



3.1.3 Exponential Smoothing

Exponential Smoothing (Suavização Exponencial) é uma técnica de previsão que atribui pesos decrescentes exponencialmente aos valores passados de uma série temporal. Ela calcula a média ponderada dos valores anteriores, atribuindo mais peso aos valores mais recentes. Essa abordagem permite capturar tendências e padrões de curto prazo na série temporal. Existem diferentes variantes do Exponential Smoothing, como a suavização simples (Simple Exponential Smoothing), a suavização com tendência (Holt's Exponential Smoothing) e a suavização sazonal (Seasonal Exponential Smoothing). Cada variante é adequada para diferentes tipos de séries temporais e padrões de dados

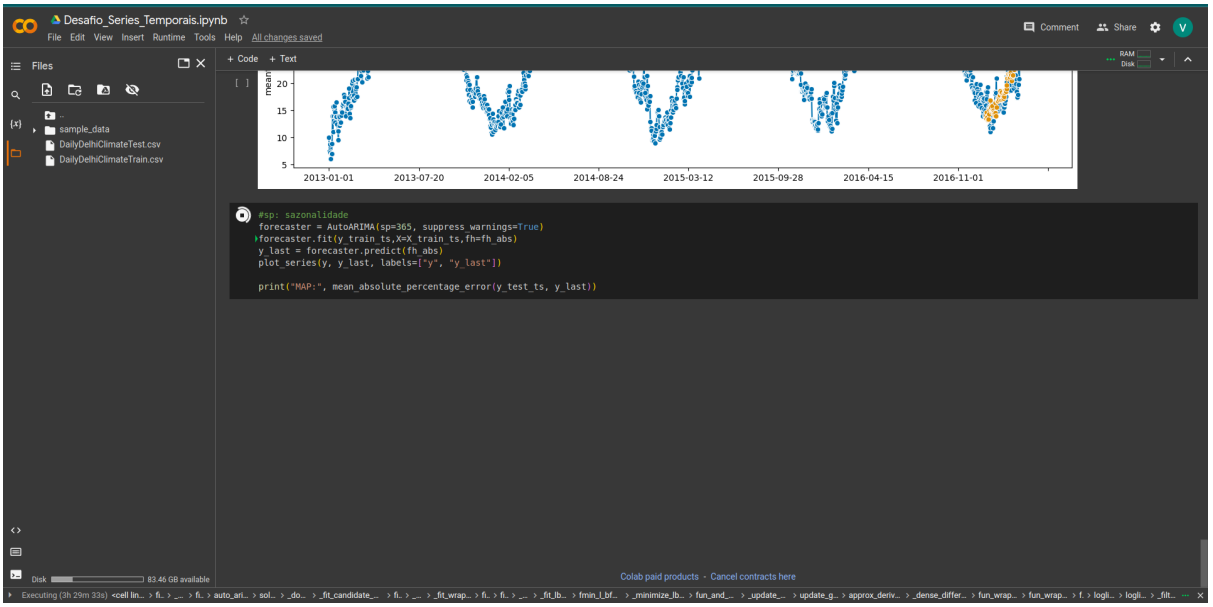
Métrica	Naive Forecaster Drift
MAP	0.11498



3.1.4 Auto ARIMA

Auto ARIMA (Auto Regressive Integrated Moving Average) é uma técnica automatizada de previsão que busca automaticamente o modelo ARIMA mais adequado para uma série temporal. O ARIMA é um modelo estatístico que leva em consideração a autocorrelação dos dados, a tendência e a sazonalidade para fazer previsões. O Auto ARIMA realiza uma busca exaustiva pelos melhores parâmetros do modelo ARIMA, como a ordem do componente auto regressivo (AR), a ordem do componente de média móvel (MA) e a ordem de diferenciação (I) necessária para tornar a série temporal estacionária. Essa técnica é amplamente utilizada para previsão em séries temporais com diferentes padrões e é especialmente útil quando não se possui conhecimento prévio sobre a série.

Após a execução ininterrupta de aproximadamente 3h30min, o modelo ainda assim não havia sido executado. Dessa maneira, optou-se pela desistência de sua execução. Havia sido testado o modelo com valores bem menores do parâmetro “sp”, mas gerou respostas bem insatisfatórias.

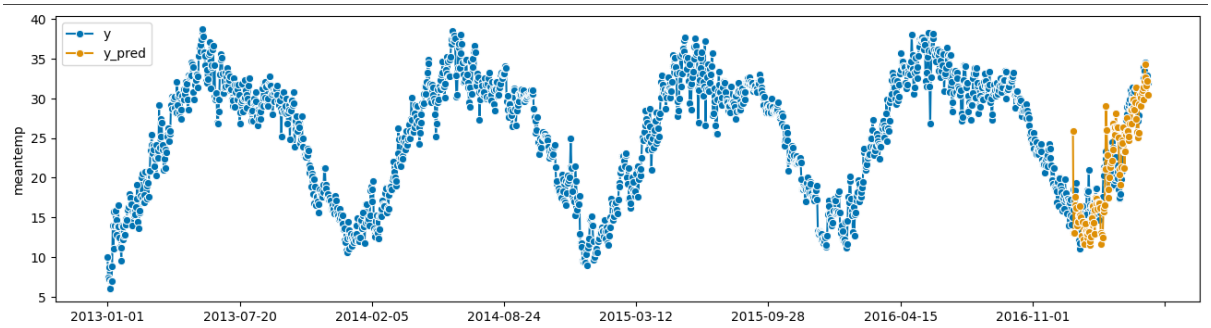


3.1.5 XGBoost

O algoritimo escolhido para tentar superar as técnicas do baseline foi o XGBoost, que apesar de não ser exatamente para série temporal, no último trabalho demonstrou ser um poderoso método para problemas de regressão.

Utiliza árvores de decisão como modelos fracos e faz um ajuste iterativo dos pesos para melhorar a precisão das previsões. Durante cada iteração, o XGBoost ajusta uma nova árvore de decisão aos resíduos (diferenças entre os valores reais e as previsões atuais) do modelo anterior. O XGBoost também incorpora algumas técnicas avançadas para melhorar o desempenho e evitar overfitting, como a regularização L1 e L2 nos pesos das árvores, a limitação da profundidade das árvores e a amostragem estocástica das instâncias de treinamento. Uma das principais vantagens do XGBoost é a sua capacidade de lidar com conjuntos de dados grandes e complexos. Ele é eficiente em termos computacionais e pode lidar com milhões de instâncias e características. Além disso, o XGBoost possui mecanismos para lidar com valores ausentes e possui recursos incorporados para selecionar automaticamente as melhores características.

Métrica	Naive Forecaster Drift
MAP	0.12874



3.2 Descrição dos resultados obtidos

Observa-se nos resultados obtidos que a técnica mais simples, ou seja, o Naive Forecaster demonstrou o melhor resultado, a partir da estratégia Mean com um valor muito bom de 0.11029 na métrica MAP. Nota-se que a estratégia Last também se demonstrou satisfatória, com resultado de 0.13628 em MAP. A mesma coisa não ocorreu com a estratégia Drift, pois esta estratégia não permite a utilização do parâmetro sp, o mesmo que permitiu resultados excelentes para os modelos citados. Pela execução do modelo e pelo resultado, a técnica Auto ETS provavelmente foi afetada por algum erro de execução e não conseguiu performar bem. Já o modelo com Exponential Smoothing atingiu um resultado excelente com valor de MAP igual a 0.11498, sendo pior apenas para a estratégia Mean do Naive Forecaster. Por último, a técnica escolhida para superar o baseline, ou seja, o XGBoost, demonstrou resultados muito bom com valor de 0.12874, o que pode ser levado em consideração que não é uma técnica especializada para séries temporais e seria possível fazer uma avaliação melhor sentando outros valores de parâmetros que a técnica possui.

4. Conclusão

Dessa forma, pode-se concluir que a melhor técnica para este problema foi o Naive Forecaster com estratégia Mean, pois possuiu o melhor resultado e por ser de procedência mais simples, necessitando de menos desempenho. Apesar da técnica do Auto ARIMA não ter executado, esta só seria considerada superior apenas se o resultado fosse muito superior ao conquistado, beirando ao erro minimo. Além disso, vale ressaltar novamente que a técnica XGBoost possui muitos parâmetros que podem ser testados para tentar garantir resultados melhores que o obtido.

5. Referências

Código desenvolvido:
https://colab.research.google.com/drive/1p4DQyG5iJYktwjnEHAloZCWuTJtINnIb?usp=sharing
Dataset:
https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data

6. Apêndice

Resumo das técnicas de predição de séries temporais:

6.1 Forecasting at Scale

O artigo fornece insights valiosos e abordagens práticas para melhorar a capacidade de previsão. Além disso, aborda os desafios associados à produção de previsões confiáveis e de alta qualidade, especialmente quando há uma variedade de séries temporais. O modelo de regressão modular proposto é flexível o suficiente para uma ampla gama de séries temporais de negócios, mas pode ser configurado por não especialistas que podem ter conhecimento de domínio sobre o processo de geração de dados, mas com pouco conhecimento sobre modelos e métodos de séries temporais.

O documento também descreve um procedimento para automatizar a avaliação do desempenho das previsões, comparando vários métodos e identificando as previsões em que a intervenção manual pode ser necessária. Recomenda-se o uso de previsões de linha de base ao avaliar qualquer procedimento de previsão. Há a sugestão também do uso de modelos simplistas (último valor e média amostral), bem como os procedimentos de previsão automatizados.

Na maioria das configurações realistas, um grande número de previsões será criado, necessitando de meios eficientes e automatizados para avaliá-las e compará-las, bem como para detectar quando é provável que tenham um desempenho ruim. Há a citação de um sistema de avaliação de previsões que usa previsões históricas simuladas para estimar o desempenho fora da amostra e identificar previsões problemáticas para que um analista humano entenda o que deu errado e faça os ajustes necessários no modelo.

6.2 SepTr: Separable Transformer for Audio Spectrogram Processing

O arquivo PDF discute o Separable Transformer (SepTr), uma nova arquitetura projetada para o processamento eficiente de espectrogramas de áudio. O SepTr utiliza módulos de atenção separáveis para capturar informações contextuais locais e globais da sequência de entrada, permitindo o tratamento eficaz de dependências de longo prazo na fala e no ruído. Os experimentos realizados em conjuntos de dados de referência demonstram que o SepTr supera os modelos inspirados em transformadores de visão, bem como outros métodos de referência. A SepTr obtém resultados significativamente melhores em todos os benchmarks, demonstrando seu alto nível de eficácia. Além disso, o SepTr reduz o número de parâmetros aprendidos em comparação com os transformadores de visão, o que o torna uma opção mais eficiente. Ou seja, o SepTr apresenta uma nova arquitetura baseada em blocos de transformadores separáveis, projetada especificamente para o processamento eficiente de espectrogramas.

6.3 Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case

O artigo apresenta uma nova abordagem para a previsão de séries temporais usando modelos de aprendizado de máquina baseados em Transformer. O paper se concentra no estudo de caso da previsão da prevalência da gripe e demonstra a eficácia do método proposto. Os autores comparam sua abordagem com outros métodos populares, como ARIMA, LSTM e Seq2Seq com modelos de atenção. Os resultados mostram que o modelo baseado no Transformer supera os outros métodos em termos de RMSE, com uma redução relativa de 27% e 8,4% em comparação com os modelos LSTM e Seq2Seq com atenção, respectivamente. Os coeficientes de correlação são muito semelhantes entre as abordagens de aprendizagem profunda, com o modelo baseado em Transformer sendo ligeiramente superior ao LSTM e ao Seq2Seq com modelos de atenção. O artigo também discute as vantagens do uso de modelos baseados no Transformer para a previsão de séries temporais, como sua capacidade de capturar dependências de longo prazo e sua escalabilidade para grandes conjuntos de dados. Desse modo, o método proposto apresenta resultados promissores e tem potencial para ser aplicado a outros problemas de previsão de séries temporais.