Intro + LSTM - bea

SARIMA - fred

HoltWinters + Conclusão -tiago

**Intro**

Olá a todooos

Neste trabalho vamos analisar um dataset sobre a produção mensal de cerveja na Austrália, desde Janeiro de 1956 e agosto de 1995. Vamos estudar diferentes métodos/modelos para o forecast desta produção, incluindo SARIMA, holtwinters e lstm.

Podemos ver pelo gráfico que já possível identificar alguns picos que parecem indicar uma sazonalidade anual. Para percebemos melhor, fazemos a decomposição do gráfico em trend, parte sazonal e análise de resíduos.

Aqui podemos ver que a trend aumenta até, aproximadamente, 1975. Depois disto ela parece-nos ser mais ou menos constante (com flutuacoes minimas).

Em relacao à sazonalidade, conseguimos confirmar o que tínhamos inferido quando vimos o gráfico dos dados originais, sendo que temos uma sazonalidade anual (que se parece dar mais no final do ano, provavelmente entre outubro e dezembro). Portanto, podemos inferir que temos uma sazonalidade cíclica/constante.

Por fim, em relação aos resíduos, eles parecem estar bastante concentrados à volta da média = 0. No entanto, com o passar do tempo, eles parecem dispersar um pouco.

Podemos confirmar isto aqui neste gráfico, onde vemos que a média dos resíduos começa a mostrar flutuações maiores com o passar do tempo.

Para ser possível aplicar os nossos modelos de previsão, dividimos os nossos dados em dataset de treino e de teste, onde selecionamos os últimos 5 anos para teste.

-----------------------Next slide-------------------------

SARIMA FRED

**SARIMA**

Começamos então com o modelo SARIMA.

Para isso, aplicamos a metodologia Box-Jenkins, onde começamos por verificar a estacionariedade dos nossos dados. Para isso aplicamos o teste ADF nos dados originais. A null-hypothesis assume uma unit root. Como tivemos um p-value maior do que 0.05 não podemos rejeitar a null-hypothesis e, desta forma, não podemos assumir que a nossa série é estacionária.

Para resolver a estacionaridade, aplicamos a primeira diferença aos nossos dados, obtendo um p-value = 0 para o ADF test. Assim, já podemos assumir que a primeira diferença é estacionária, como também podemos verificar pelo gráfico.

-----------------------Next slide-------------------------

Para determinar a ordem do modelo, aplicamos ACF e PACF. Aqui mostramos a primeira diferença. Na ACF, entre 1 e o 4 parece-nos haver alguma relevância. Podemos verificar com certezas que há uma forte correlação para o lag1, mas para o 2 e o 4, como estão bastante próximos da zona de significância de estes serem 0 ficamos, inicialmente, na dúvida. Por isso, depois disso aplicamos também a função auto\_arima, para verificar estes valores.

-----------------------Next slide-------------------------

Para confirmar a ordem da sazonalidade, temos a acf e pacf apenas para a parte sazonal onde, mais uma vez, verificamos que é anual e, portanto, de 12 em 12 meses. Podemos ver na acf que de 12 em 12 lags chegamos a uma correlação de, aproximadamente 1. No entanto, novamente, como a acf e pacf sao apenas starting points, vamos confirmar estes valores com a função auto-arima.

-----------------------Next slide-------------------------

O auto-arima faz testes de diferença para escolher a ordem de diferença. Depois, testa vários modelos para valores de p e q, fazendo um processo semelhante para a parte sazonal. A seleção depende do valor que minimiza a AIC (Criteria de Informação de Akaike).

O modelo selecionado pelo auto\_arima foi (...), já que é o que tem menor AIC. No entanto, quando analisamos os coeficientes, verificamos que o segundo termo AR sazonal não era estatisticamente significante, uma vez que tinha um p-values maior o que 0.05 e uma vez que o intervalo de confiança incluía o valor 0. Assim sendo, acabamos por treinar o modelo (...), mantendo a parte não sazonal idêntica.

-----------------------Next slide-------------------------

Aqui podemos analisar visualmente as previsões feitas pelo nosso modelo, onde acabamos por obter o MSE e MAE semelhante ao modelo que obtivemos no auto\_arima, confirmando que aquele parâmetro que acabamos por retirar não era, de facto, relevante para os nossos resultados.

-----------------------Next slide-------------------------

O passo final da metodologia de Box-Jenkins consiste na análise dos resíduos, que é o que fazemos aqui. A null-hypothesis do ljung-box test assume a não correlação dos resíduos. Como obtivemos um p-value maior do que 0.05, não rejeitamos a null-hypothesis. Podemos verificar isto aqui no correlograma que não há correlação significativa entre os nossos resíduos.

A null-hypothesis do Jarque-Bera test assume a distribuição normal dos resíduos. Um p-value = 0 rejeita a null-hypothesis, o que nos faz crer, inicialmente, que os nossos resíduos não seguem uma distribuição normal. No entanto, pelo histograma, podemos ver que não têm uma distribuição assim tão distante da distribuição normal, bem como o QQ-plot, onde mesmo as tails parecem estar bastante próximas dos valores teóricos.

O Heteroskedasticity assume como null hypothesis uma variância constante nos resíduos. Como temos um p-value menor do que 0.05, rejeitamos esta hipótese, assumindo que a variância deste não é constante. Isto é possível verificar no gráfico dos resíduos, onde a variância parece estar a aumentar um pouco com o tempo. Por fim, em relação à Skew and Kurtosis, a primeira mostra que a nossa distribuição não está muito “desviada”, já a Kurtosis parece-nos crer as tails da distribuição são apena um pouco maiores do que uma distribuição normal, que assumem um valor para o Kurtosis test até 3.

Estas observações fazem-nos questionar o resultado do teste Jarque-Bera. No entanto, temos que ter em atenção de que o nosso dataset é relativamente pequeno (cerca de 416 dados) e que este teste pode não funcionar tão bem para este tipo de datasets, podendo estar a dar uma falsa avaliação da distribuição dos dados.

-----------------------Next slide-------------------------

**Holt-Winters**

O modelo de Holt Winters utiliza exponential smoothing, dividindo a trend e sazonalidade, podendo ser aditiva, multiplicativa, ou não especificada. Neste caso, aplicamos uma sazonalidade aditiva, uma vez que esta é constante. Em relação à trend não conseguimos especificar se esta é aditiva ou multiplicativa, uma vez que esta varia até um certo ponto e fica constante a partir de outro. Assim sendo, experimentamos todos, acabando por ter um melhor resultado, em termos de MSE e MAE, quando não especificamos a trend. Após a aplicação do modelo obtemos este MSE e MAE (no pp), e podemos ver, aqui, visualmente, a previsão feita pelo modelo, quando comparado com os valores de teste.

-----------------------Next slide-------------------------

Posto isto, fazemos mais uma vez a análise dos resíduos. No Ljung-Box test, obtivemos um p-values menor que 0.05, podendo rejeitar a null hypothesis, assumindo uma correlação entre os resíduos, como podemos verificar no correlograma. No teste JB obtivemos p-value menor que 0.05, rejeitando, outra vez a null-hypothesis. Assim, o teste rejeita uma distribuição normal dos resíduos. Mais uma vez, sentimos que nao podemos confiar nos resultados deste teste pois, no gráfico da distribuição dos resíduos estes não parecem ter uma distribuição muito longe da normal. Podemos confirmar o mesmo também pelo QQ-plot, onde as tails parecem estar bastante próximas do valor teórico. Comparando com os resíduos obtidos pelo SARIMA, estes parecem ter valores maiores. Em relação aos testes de Heteroskedasticity , Skew e Kurtosis, os resultados parecem ser bastante semelhantes aos que analisamos no SARIMA.

-----------------------Next slide-------------------------

**LSTM**

Em seguida realizamos testes com o modelo LSTM, que é um tipo de RNN (Recurrent neural network) como podemos observar pela imagem temos 3 portões um forget gate que tem como objetivo retirar os dados que não são relevantes, um input gate que adiciona nova informação e o output gate é onde saí as informações atualizadas. Esses portões, com as ativações, possibilitam LSTMs para regular o fluxo de informações, reter informações importantes em sequências longas e mitigar o impacto de entradas irrelevantes. Esse

torna os LSTMs adequados para capturar dependências e padrões em dados sequenciais que podem ocorrer ao longo de intervalos de tempo estendidos.

-----------------------Next slide-------------------------

Testamos a LSTM com e sem sazonalidade começando agora com sazonalidade podemos observar que os resultados obtidos não foram os mais precisos, no entanto, o modelo conseguiu seguir o movimento nas alterações de sentido do original. Obtivemos um MSE de 210.13 e um MAE de 12.39.

-----------------------Next slide-------------------------

Por vezes, a presença de sasonalidade pode interferir com a performance do modelo. Removendo a sasonalidade pode ajudar a que o modelo de foque noutras trends e padrões nos nossos dados.

Assim, como podemos observar, o resultado da previsão foi muito melhor comparativamente com o anterior, no entanto, para as métricas de MSE e MAE os valores não mostraram um diferença significante. Obtivemos um MSE de 207.80 e um MAE de 12.10.

-----------------------Next slide-------------------------

Em relação a alguns problemas em utilizar a LSTM, tivemos alguns problemas principalmente relacionados com o tamanho do dataset que estamos a utilizar, pois temos um dataset relativamente pequeno para este tipo de modelo e isto leva a uma maior variância. Consequentemente tínhamos também poucos parâmetros/features para utilizar neste modelo, que também dificulta a aprendizagem. E por último, o esforço computacional e tempo necessário para correr este modelo é considerado alto, quando comparado com os modelos anteriormente utilizados.

-----------------------Next slide-------------------------

**Conclusão**

Apresentar e comparar os valores de MSE e MAE dos vários modelos:

* SARIMA e HoltWinters com valores parecidos (HW um bocado melhor apesar dos resíduos serem maiores e apresentarem alguma correlação).
* LSTM: ambos apresentam valores mais baixos de MSE e MAE quando comparados com os métodos acima. Novamente, os resultados LSTM “normal” são surpreendentemente baixos tendo em conta a visualização das previsões.
* LSTM sem sazonalidade parece ser o melhor modelo.

-----------------------Next slide-------------------------

(talvez dizer que as metricas MAE e MSE podem, nem sempre, traduzir a verdadeira performance das forecasts e, por isso, apresentamo-las aqui visualmente)

Gráfico com todos os forecast:

* Novamente SARIMA e HW com previsões parecidas.
* LSTM sem sazonalidade parece novamente ser o melhor modelo, especialmente adequado nos pontos mais altos e mais baixos, sendo que o SARIMA e o HW parecem prever uma produção ligeiramente maior nestes casos, especialmente para os anos mais afastados.

Mais alguma palha??? Conclusão geral do Trabalho???

Foi interessante comparar modelos de análise de séries temporais mais tradicionais com modelos mais complexos como LSTM, que nós como alunos de data science já tínhamos utilizado em problemas diferentes como NLP. Interessante testar estes modelos (principalmente LSTM) para dados mais complexos.