B2W Challenge

Nilton G. Duarte[1]

[1]nilton.gduarte@gmail.com

1. Etapas e Estratégias

A arquitetura de software utilizada foi a de contêineres em Docker, sendo um contêiner para o banco de dados MySQL e um contêiner para o código em Python3, versionado em repositório privado no GitHub. Essa arquitetura foi escolhida pela facilidade de configuração tendo em vista que já havia uma configuração prévia do Docker no sistema. O serviço Docker estava sendo executado em um servidor Ubuntu próprio, mas por problemas de falha de hardware teve de ser migrado para uma VM em meu computador pessoal. A VM também é Ubuntu, versão 18.04.

Sendo assim, não foi necessário nenhuma configuração extra, apenas ajustes nos arquivos de configuração `docker-compose.yaml` e no arquivo de inicialização do MySQL `/db/init.sql`. A configuração `docker-compose.yaml` cria os serviços (contêineres) `challenge_app` e `mysqldb`, ambos na mesma sub-rede e com acesso para internet e acesso da (minha) rede local para os contêineres.

1.1. Modelo de Predição

Com os dados no banco, o primeiro passo foi fazer um gráfico (Figura 1) das informações das vendas. A vendas foram agrupadas pela data.

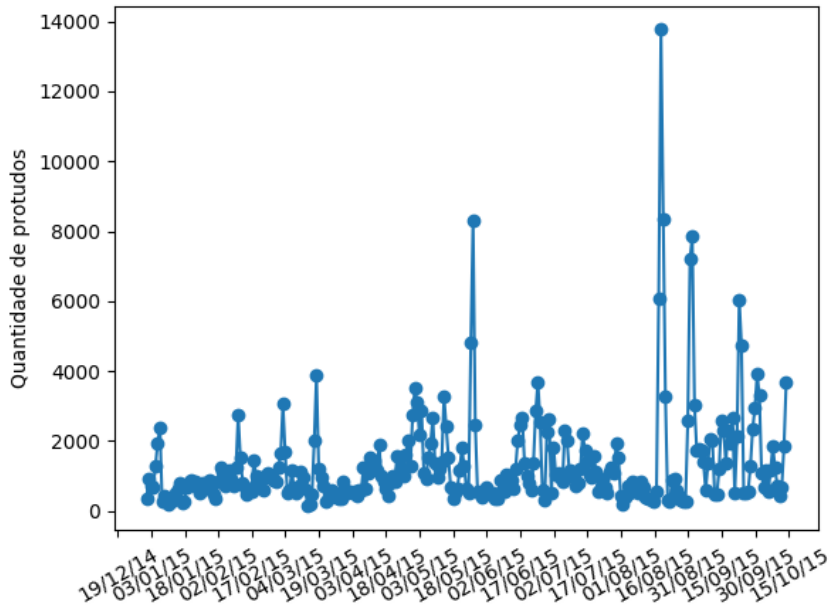

Figura 1. Quantidade de produtos vendidos.

O gráfico parece mostrar uma relação da quantidade de produtos vendidos e o dia do mês e/ou semana. Para fazer uma análise dessa relação, o segundo passo foi calcular a transformada de Fourier [de Figueiredo 1977] para fazer a análise do espectro dos dados. O resultado está na Figura 2. O período de 7 dias é a frequência dominante na série, sendo seus harmônicos 14, 21 e 28 também de muita influência. Essa análise me indica que o dia da semana é um fator importante para a quantidade de produtos vendidos.

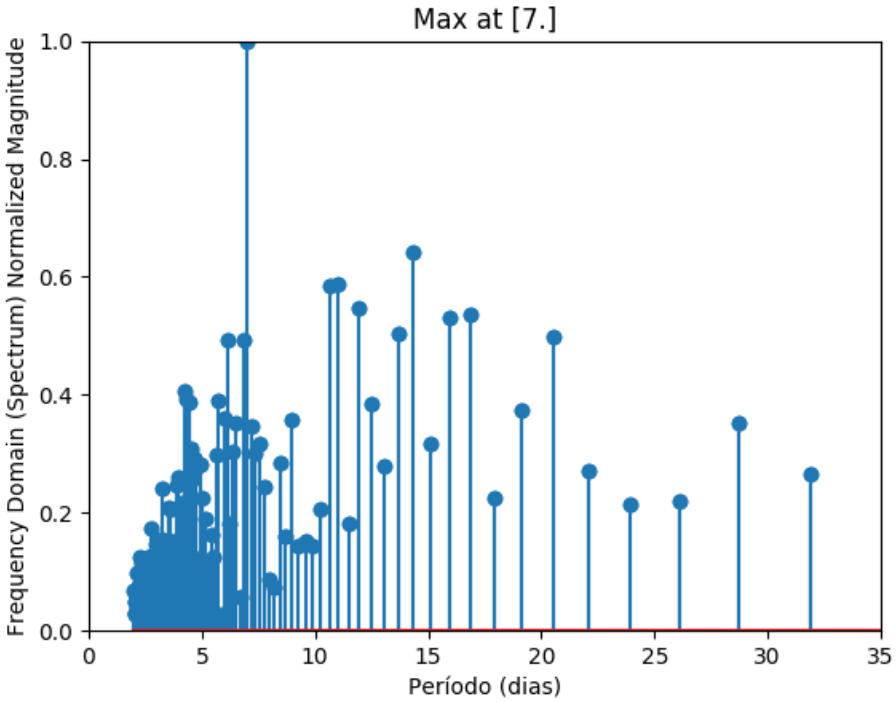

Figura 2. Transformada de Fourier da base de dados das vendas.

Para as figuras anteriores, que não faz separação por produto, não há dia sem que haja algum registro venda. Quando é feita a separação por produto, existem dias em que não há venda de alguns produtos. Para verificar esse fator de periodicidade nos produtos foi necessário preencher os dias nos quais não tínhamos registro de vendas. De modo geral, essa influência permanece. Nas Figuras 3 e 5, verificamos que na primeira o máximo está próximo de 63 que também é múltiplo de 7. Na segunda o máximo está em 7, assim como no caso geral. Outras imagens podem ser encontradas na pasta `/app`.

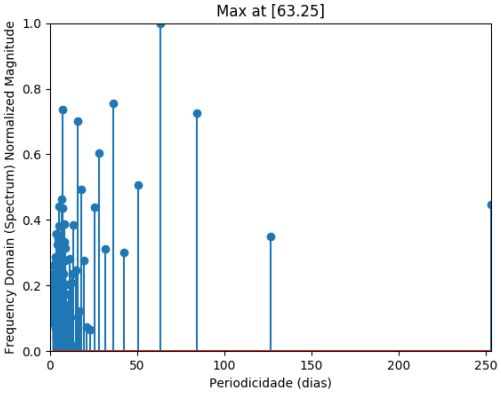

Figura 3. Transformada de Fourier para P1

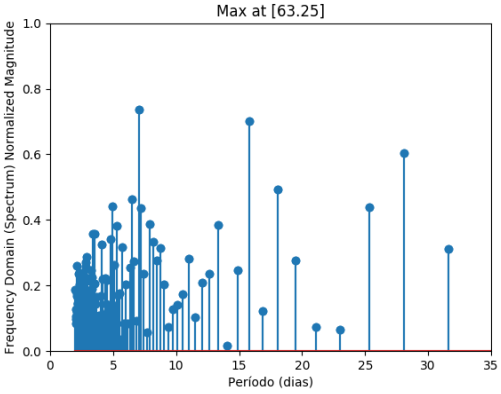

Figura 4. *Zoom* da Transformada de Fourier para P1.

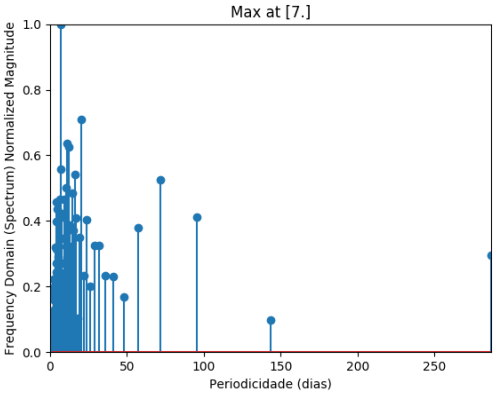

Figura 5. Transformada de Fourier para P7

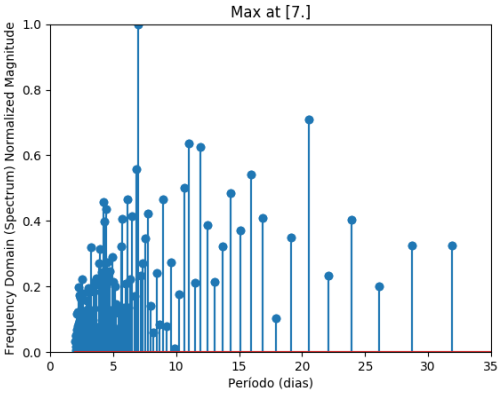

Figura 6. *Zoom* da Transformada de Fourier para P7.

Após a análise do espectro dos dados, a etapa seguinte foi pesquisar um modelo que se adequasse ao problema. O modelo escolhido foi o Long Short-Term Memory Network [Murphy 2013, Brownlee] (LSTM), uma variação de Recurrent Neural Network (RNN). Nunca tinha utilizado, apenas ouvido falar e aproveitei a oportunidade para conhecer melhor.

Os atributos utilizados no modelo foram o dia da semana $w(t)$ e a média de preços do dia $p(t)$. O objetivo é encontrar o número de itens vendidos $q(t)$. De forma geral:

$$q(t) = f_{LSTM}(w(t), p(t)) \tag{1}$$

A implementação da LSTM utilizada foi a da biblioteca Keras [Chollet et al. 2015] com TensorFlow de *backend*. O modelo precisa dos parâmetros `epochs`, `batch_size`, `units`, `past_sight` e `future_sight`. O parâmetro `epochs` é o número de iterações para treinar o modelo, `batch_size` é o tamanho do subconjunto utilizados para o treinamento e `units` é o "tamanho" do modelo. Os parâmetros `past_sight` e `future_sight` indicam quantos dias anteriores serão usados e quantos dias serão previstos por vez, respectivamente.

Para encontrar um bom conjunto de parâmetros, foram testados os seguintes parâmetros:

- `epochs` $\in \{50, 500, 1000, 5000\}$
- `batch_size` $\in \{15, 32\}$
- `units` $\in \{5, 10, 50, 100, 200\}$
- `past_sight` $\in \{1, 7, 15\}$
- `future_sight` $\in \{1\}$

Para treinar e testar o modelo, o conjunto de dados foi dividido em dois. O conjunto de teste com os últimos 15 dias e o conjunto de treinamento com o restante. O melhor conjunto é dado pelo que tem o menor erro médio quadrático no conjunto de teste. As figuras a seguir ilustram os resultados, o título das figuras estão codificadas nos parâmetros do modelo.

Título
`{P#}_{past_sight}_{future_sight}_{epochs}_{batch_size}_{units}`

O pontos em azul são os valores reais, verdadeiros. Os pontos em laranja são as previsões do modelo para o conjunto de teste. Os pontos mais importantes estão em verde, que são as previsões para o conjunto de teste. Os melhores parâmetros encontrados foram: `epochs` $= 5000$, `batch_size` $= 32$, `units` $= 5$ e `past_sight` $= 1$. `future_sight` é sempre 1. Alguns dos erros correspondentes aos modelos estão em `data/models_rmse.csv`. Durante os testes, os menores erros estavam por volta de 75.

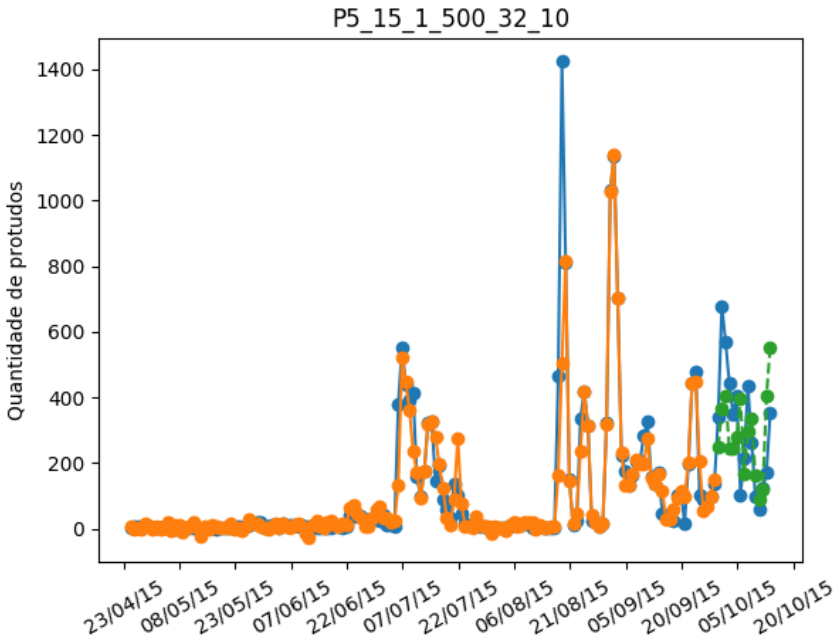

Figura 7. Ajuste e previsão do modelo LSTM para os parâmetros do título

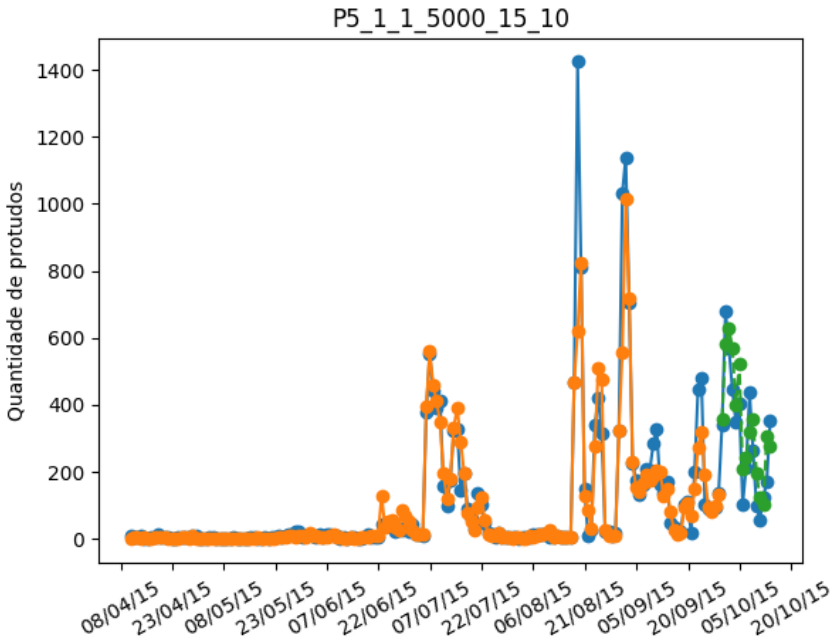

Figura 8. Ajuste e previsão do modelo LSTM para os parâmetros do título

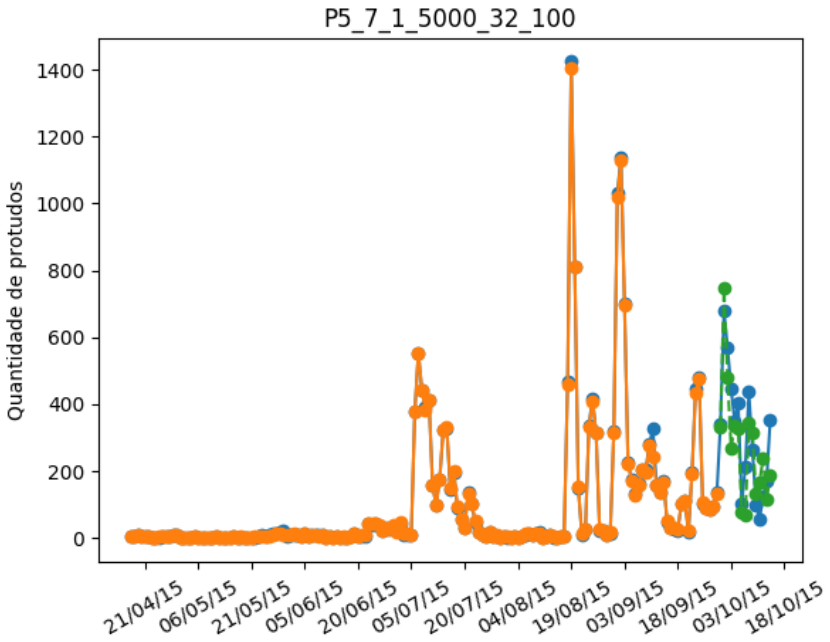

Figura 9. Ajuste e previsão do modelo LSTM para os parâmetros do título

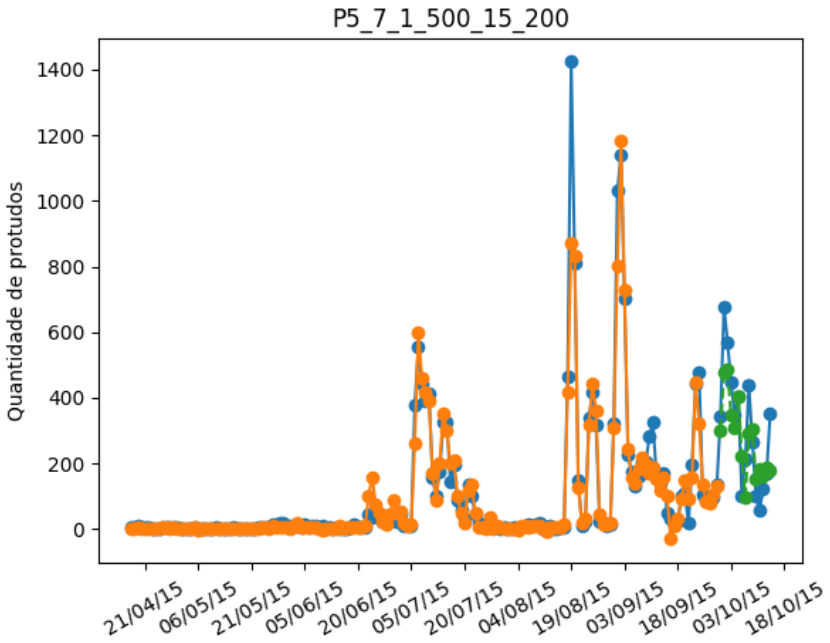

Figura 10. Ajuste e previsão do modelo LSTM para os parâmetros do título

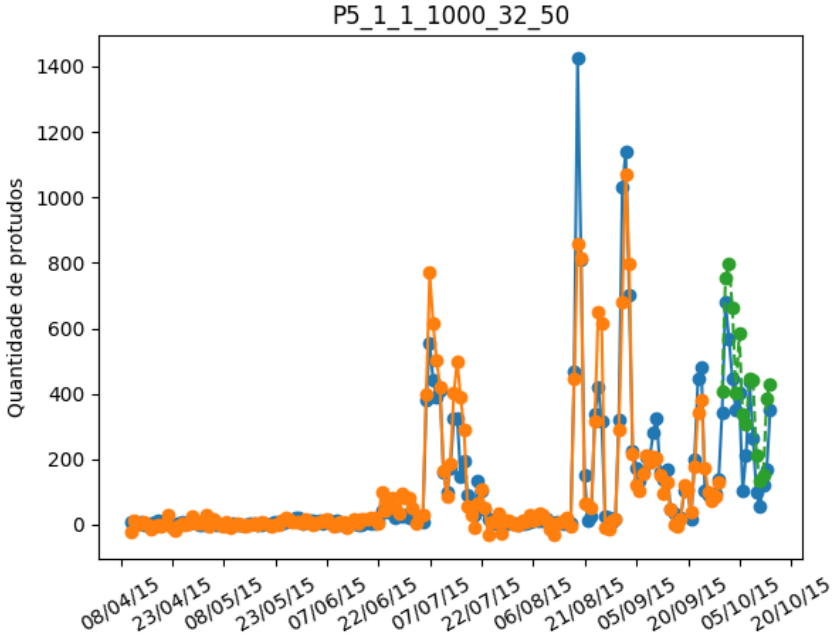

Figura 11. Ajuste e previsão do modelo LSTM para os parâmetros do título

2. Descrição dos dados

A Tabela 1 apresenta informações da quantidade média de vendas diárias, o preço médio e a média da quantidade por venda para cada produto. Foram calculados os coeficientes de correlação linear [Stewart 2009] entre os seguintes conjuntos:

- Quantidade vendida (Qt) e preço (Pc) para cada produto (P)
- Preço de cada concorrente (C) para cada produto (P) para cada tipo de pagamento (t)

As tabelas a seguir estão organizadas com o conjunto e o produto (P) indicados na primeira linha. Na segunda linha, estão indicados quais colunas se referem aos conjuntos (Relação) ou ao coeficiente de correlação (Fator). O fator é o coeficiente de correlação linear entre o conjunto da coluna Relação e o conjunto da primeiro linha.

Importante ressaltar que o coeficiente **não implica em causalidade**.

As tabelas 2 e 3 apresentam os 2 maiores (excluindo a auto correlação linear) e os 2 menores coeficientes de correlação para a quantidade vendida e o preço para cada produto, respectivamente. Um aspecto é interessante, mas não surpreendente: para cada produto, a quantidade vendida é inversamente proporcional ao seu preço, sendo esse coeficiente o menor entre todos calculados.

Também é interessante a relação de P8 e P9, que tem a média de preço muito similar e suas correlações lineares são muito próximas de 1 tanto para a quantidade quanto para o preço.

Todos os preços diminuem com o tempo (inversamente proporcional a data), exceto o produto P1. O preço do produto P7 do concorrente C6 apresenta alto coeficiente de correlação com o preço de venda do produto P7. O preço de P5 apresenta alto coeficiente com o preço de vários outros produtos (P2, P3, P4, P6, P8).

Tabela 1. Médias dos preços, quantidades diárias e por venda dos produtos

Média	Vendas diárias	Preço	Quantidade por venda
P1	16,559	1459,13	1,01284
P2	236,390	731,08	1,03696
P3	11,769	1289,03	1,01325
P4	86,979	525,13	1,16044
P5	115,054	869,80	1,03928
P6	14,814	1775,58	1,02483
P7	742,884	779,50	1,06640
P8	142,000	466,18	1,04636
P9	93,623	465,54	1,03821

Tabela 2. Correlação linear das quantidades vendidas por produto.

Quantidade P1		Quantidade P2		Quantidade P3		Quantidade P4		Quantidade P5	
Relação	Fator	Relação	Fator	Relação	Fator	Relação	Fator	Relação	Fator
Qt P3	0,29	Qt P7	0,60	Qt P1	0,29	Qt P6	0,44	Qt P9	0,64
Qt P7	0,15	Qt P8	0,45	Qt P5	0,13	Qt P7	0,24	Qt P8	0,68
Pc P1	-0,56	Pc P2	-0,41	Pc P3	-0,37	Pc P4	-0,57	Pc P5	-0,60
C4 P6 t1	-0,20	Pc P7	-0,21	Pc P1	-0,24	Pc P7	-0,24	Pc P7	-0,51

Quantidade P6		Quantidade P7		Quantidade P8		Quantidade P9	
Relação	Fator	Relação	Fator	Relação	Fator	Relação	Fator
Qt P4	0,44	Qt P2	0,60	Qt P9	0,93	Qt P8	0,93
Qt P7	0,12	Qt P5	0,46	Qt P5	0,62	Qt P5	0,64
Pc P6	-0,19	Pc P7	-0,38	Pc P8	-0,62	Pc P9	-0,61
Pc P4	-0,10	Pc P2	-0,15	Pc P9	-0,59	Pc P8	-0,61

Tabela 3. Correlação linear dos preços por produto.

Preço P1		Preço P2		Preço P3		Preço P4		Preço P5	
Relação	Fator	Relação	Fator	Relação	Fator	Relação	Fator	Relação	Fator
Data	0,18	Pc P5	0,80	Pc P5	0,81	Pc P5	0,64	Pc P8	0,84
C6 P2 t1	0,07	Pc P8	0,74	C6 P7 t1	0,78	Pc P7	0,63	C6 P7 t1	0,84
Qt P1	-0,56	Data	-0,85	Data	-0,83	Data	-0,65	Data	-0,91
Qt P3	-0,24	Qt P5	-0,48	Qt P5	-0,38	Qt P4	-0,57	Qt P5	-0,60

Preço P6		Preço P7		Preço P8		Preço P9	
Relação	Fator	Relação	Fator	Relação	Fator	Relação	Fator
Pc P2	0,67	C6 P7 t1	0,82	Pc P9	0,97	Pc P8	0,97
Pc P5	065	C6 P7 t2	0,82	Pc P5	0,84	Pc P6	0,82
Data	-0,72	Data	-0,82	Data	-0,87	Data	-0,87
Qt P9	-0,37	Qt P9	-0,53	Qt P8	-0,62	Qt P9	-0,61

Outra informação extraída foi a decomposição dos dados $\mathcal{D}(t)$ em tendência $\mathcal{T}(t)$, sazonalidade $\mathcal{S}(t)$ e o resíduo $r(t)$, de modo que

$$\mathcal{D}(t) = \mathcal{T}(t) + \mathcal{S}(t) + r(t) \tag{2}$$

Para tal foi utilizado a biblioteca Statsmodels [Seabold and Perktold 2010]. O resultado para P5 está na Figura 12, os demais resultados estão em `/app/`. Não encontrei compreensão adicional nos dados a partir da decomposição.

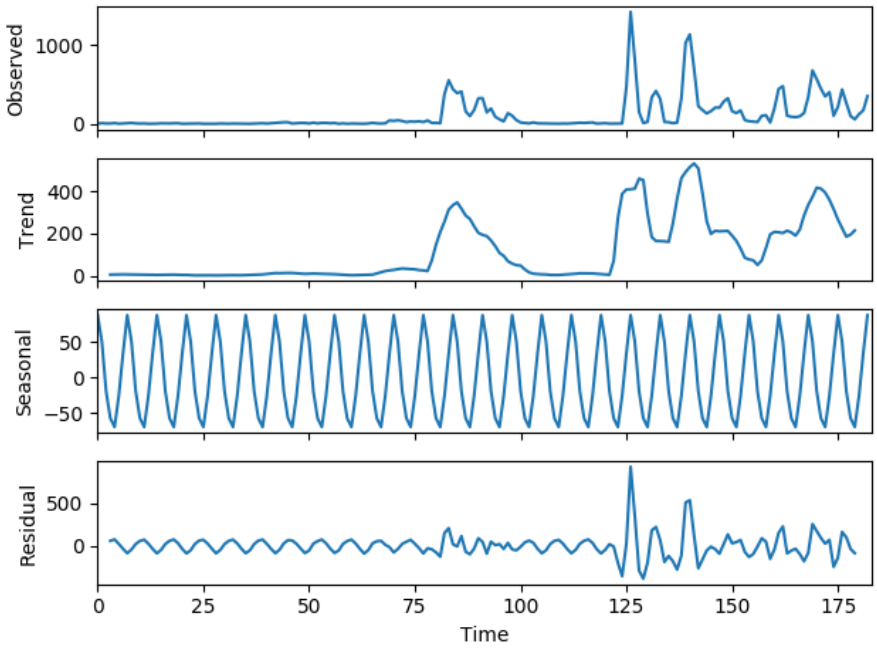

Figura 12. Decomposição de P5 (Eq 2)

3. Conclusão

A análise espectral (Figura 2) e decomposição (Figura 12) levam a crer que o o dia da semana é um fator relevante para o problema. Os produtos P8 e P9 parecem se tratar do mesmo produto, levando em conta as informações. Talvez algo como marcas diferentes de um mesmo produto. P7 e P2 apresentam características parecidas, mas em menor grau.

Na correlação de preços, apenas P7 apresenta um alto coeficiente com algum concorrente. O preço do produto P5 é relacionado ao preço de muitos produtos, como se P5 fosse utilizado na produção dos demais. Se P5 fosse o preço da farinha, os demais produtos poderiam ser os pães e bolos de uma padaria.

Referências

Brownlee, J. Time series forecasting with the long short-term memory network in python. https://machinelearningmastery.com/time-series-forecasting-long-short-term-memory-network-python/. Accessed: 2019-06-13.

Chollet, F. et al. (2015). Keras. https://keras.io.

de Figueiredo, D. (1977). *Análise de Fourier e equações diferenciais parciais*. Projeto Euclides. Instituto de Matemática Pura e Aplicada, CNPq.

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. Princeton University Press, Princeton, NJ, USA.