Previsão de vendas numa empresa de TI utilizando time series

Sales forecast in an IT company using time series

Pedro Sobreiro
ESDRM, Instituto Politécnico de Santarém; ISLA Santarém; CEPESE
Santarém, Portugal; Porto, Portugal
sobreiro@esdrm.ipsantarem.pt

Domingos Martinho
ISLA Santarém; CEPESE
Santarém, Portugal; Porto, Portugal
domingos.martinho@islasantarem.pt

António Pratas
ISLA Santarém
Santarém, Portugal
antonio.pratas@islasantarem.pt

Resumo — **A previsão de vendas constitui um fator fundamental para o planeamento da atividade das empresas fornecendo indicadores importantes para o suporte das decisões dos gestores. Este estudo visa explorar o potencial dos algoritmos para previsões baseado em time series, numa empresa da área das TI. A previsão foi realizada com base nos dados de faturação da empresa durante 192 meses de atividade. A análise dos dados baseou-se na abordagem CRoss Industry Standard Process for Data Mining e para o tratamento recorreu-se ao Anaconda IPython e Pandas. A previsão foi realizada em R com três modelos: *Exponencial Smoothing* (Holt-Winters), *autoregressive integrated moving average* (ARIMA) e *artificial neural networks* (ANN). A comparação do desempenho de cada um dos métodos mostra que o modelo baseado em artificial *neural networks* tem uma maior exatidão na previsão. Estes resultados carecem de um aprofundamento do estudo alargando o universo dos contextos estudados, contudo a simplicidade na aplicação do modelo baseado artificial neural networks, possibilita a sua utilização em aplicações informáticas utilizadas pelas empresas sem necessidade conhecimentos específicos, fornecendo um instrumento fiável que possibilite o apoio à tomada de decisão por parte dos gestores.**

Palavras Chave – algoritmos; redes neuronais; suporte à decisão; time series.

Abstract - **The sales forecast is fundamental for the planning of the activity of the companies providing, important indicators for the support of the decisions of the managers. This study aims to explore the potential of time series prediction algorithms in an IT company. The forecast was based on the company's billing data for 192 months of activity. The analysis of the data was based on the Cross Industry Standard Process for Data Mining approach and for the treatment; we used the Anaconda IPython and Pandas. We developed the prediction with three models using R: Exponential Smoothing (Holt-Winters), autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN). The comparison of the performance of each of the methods shows that the model based on artificial neural networks has a greater accuracy in the prediction. These results need deepening the study to broaden the universe of the studied contexts. However, the simplicity in the application of the artificial neural networks model makes possible its use in computer applications without specific knowledge, giving a reliable instrument that allows the supporting decision-making by managers.**

Keywords - algorithms; decision support; neural networks; time series.

I. INTRODUÇÃO

A previsão de vendas é uma atividade da maior importância para as empresas, que despendem recursos humanos e financeiros significativos para obter informação fiável sobre esta componente essencial para o desenvolvimento do negócio. É assumido que esta tarefa se não for realizada corretamente, provoca perturbações na gestão dos stocks com aumentos desnecessários ou ruturas que não respondem às necessidades dos clientes, sendo muito importante desenvolver previsões eficazes [1]. Apesar desta importância, muitas empresas não utilizam mecanismos adequados para prever as vendas e utilizam normalmente aproximações que passam por calcular a média de vendas mensais ou a média de vendas daquele mês num historial de alguns anos [2].

Para obviar a estes inconvenientes podem ser utilizadas para a previsão das vendas algoritmos baseados em *time series* que podem ser classificados em lineares ou não-lineares, dependendo do tipo do modelo em que são baseados [3]. Os métodos a aplicar na previsão de vendas podem ser tradicionais como o *Exponencial Smoothing* de Holt e Winters [4] – [5] ou o *Autoregressive Integrated Moving Average* (ARIMA). Contudo estas aproximações não são adequadas se se verificar a não-linearidade nos dados, o que normalmente se verifica [6]. Mais recentemente surgiram previsões que recorrem a *Artificial Neural Networks* (ANN) que podem ser aplicadas em previsões.

O *Exponential Smoothing* proposto por Holt [4] e Winters [5] tornou-se em um dos métodos com mais sucesso na previsão recorrendo a séries. As previsões são realizadas com médias ponderadas com um peso menor das observações passadas, permitindo criar rapidamente previsões em séries não estacionárias [7].

O *auto-regressive integrated moving average* (ARIMA) combina três elementos distintos, uma função *auto-regressive* (AR) nos valores passados, função regressiva *moving average* (MA), uma parte integrada (I) para tornar as séries estacionárias através da diferenciação [8].

As *artificial neural networks* (ANN) inspiram-se em estudos do cérebro e sistema nervoso, têm poucos pressupostos à priori e ao contrário do ARIMA e *Exponencial Smoothing*. fornecem uma alternativa aos métodos tradicionais, sendo orientadas aos dados, com métodos adaptativos, aprendem com os dados e são adequadas para problemas em que as soluções requerem conhecimento que é difícil de especificar mas que têm dados suficientes [6]. As ANN podem ser entendidas como um método estatístico não-paramétrico multivariado não-linear [9].

O estudo foi desenvolvido com o intuito de verificar o desempenho na previsão de vendas baseado em *time series* dos modelos: *Exponencial Smoothing*; ARIMA e ANN, para avaliarmos qual apresenta melhores resultados. Com o objetivo de incorporar as técnicas selecionadas em aplicações informáticas para disponibilizarmos um modelo de previsão na empresa onde desenvolvemos o estudo e para prestação de serviços aos seus clientes.

Este artigo está organizado em cinco seções. A seção 2 explica metodologia utilizada para o desenvolvimento do estudo, a abordagem utilizada para a análise dos dados, as técnicas utilizadas para realizarmos as previsões e por último descrevemos como é que identificamos o modelo com melhor desempenho. Na seção 3 analisam-se os resultados obtidos na aplicação de cada um dos modelos. Na seção 4 interpreta-se o desempenho de cada um dos modelos utilizados. Por último, na seção 5 apresentam-se as conclusões do estudo, bem como propostas de trabalho futuro.

II. METODOLOGIA

O trabalho foi desenvolvido numa empresa da área das tecnologias de informação (TI) sedeada na região centro de Portugal utilizando os dados obtidos através do ERP PHC CS Desktop Advanced versão 22, suportado no SQL Server 2016, onde se encontram registados todos os dados acumulados na atividade da empresa, desde 2002, numa base de dados com cerca de 5GB. Os dados analisados neste projeto foram disponibilizados pelos responsáveis da empresa e são referentes aos registos da faturação da empresa envolvendo cerca de 24000 registos obtidos nos últimos 192 meses (janeiro de 2002 a dezembro de 2017).

Atendendo aos objetivos e ao problema que se pretende resolver, optou-se por utilizar a metodologia CRISP-DM (CRoss Industry Standard Process for Data Mining) [10], que consiste no standard mais utilizado devido à flexibilidade da sua implementação em qualquer domínio e à compatibilidade com qualquer ferramenta de data mining. A metodologia CRISP-DM comtempla seis fases flexíveis, tendo sido desenvolvidos os seguintes passos [10]: compreensão do negócio; compreensão dos dados; preparação dos dados, modelação, avaliação e implementação.

A fase compreensão do negócio permitiu alinhar os objetivos do projeto com a necessidade de dados, alinhando a definição do problema com a mineração dos dados.

A compreensão dos dados tem por objetivo efetuar a coleta inicial dos dados e familiarização, possibilitando identificar problemas de qualidade de dados obtendo-se os resultados óbvios. Para concretizar esta fase recorreu-se à análise ao dicionário de dados e da exploração dos dados diretamente sobre o SQL Server.

A preparação dos dados consistiu no registro e seleção de atributos e limpeza dos dados, seguindo-se a modelação dos dados através da utilização das ferramentas de data mining. O resultado destas fases consistiu na criação de uma Data Mart, ou seja, de uma base de dados de pequena dimensão onde foram carregados os dados a utilizar neste projeto [11]. As fases de preparação e modelação foram realizadas com recursos aos softwares Anaconda I Python [12] e Pandas [13].

A avaliação tem por objetivo determinar se os resultados obtidos atendem aos objetivos de negócios, identificando-se os problemas que foram abordados anteriormente. Considerando a importância da previsão das vendas para a organização e a sua integração nos serviços disponibilizados aos clientes, avaliaram-se os dados e a facilidade da sua obtenção para que o processo possa ser facilmente aplicado na empresa e nas organizações clientes.

A fase de implementação será realizada no final do projeto com a disponilibilização na empresa.

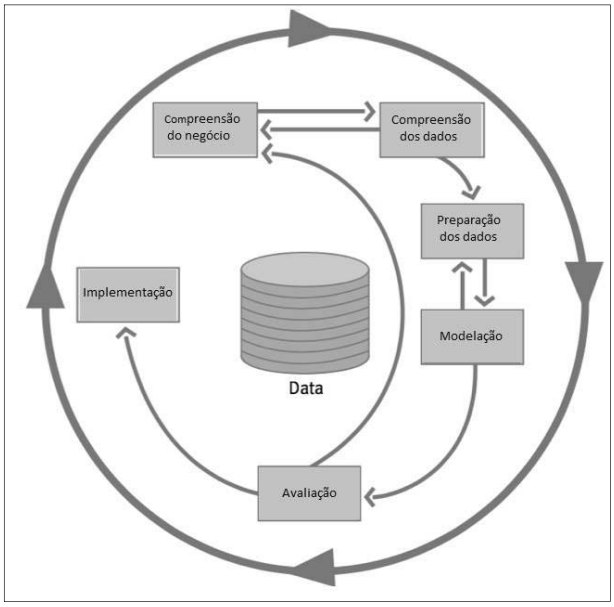

Figura 1. Modelo de referencia CRISP-DM, adaptado [10]

A análise e previsão de *time series* foi desenvolvida no R [14], através da biblioteca *forecast* [15]. Realizou-se a previsão aplicando primeiro *Exponencial Smoothing* (HoltWinters),

depois *autoregressive integrated moving average* (ARIMA) e, por último, *artificial neural networks* (ANN). No final comparou-se o desempenho que cada um dos métodos.

Para avaliar o desempenho da previsão existem vários indicadores que podem ser utilizados [7] para avaliar a exatidão da previsão, tais como *mean error* (ME); *root mean squared error* (RMSE); *mean absolute error* (MAE); *mean percentage error* (MPE); *mean absolute percentage error* (MAPE); *mean absolute scaled error* (MASE) e *first-order autocorrelation coefficient* (ACF1). Segundo Davydenko e Fildes [16] o MAE tem mais vantagens enquanto o MAPE é referido como um dos indicadores mais utilizados na literatura publicada [17], contudo o MAE é simples de interpretar e fácil de calcular [18].

III. ANÁLISE

A análise inicial dos dados, cujos resultados são apresentados na figura 2, mostra que se verifica alguma sazonalidade na evolução das vendas durante o período analisado.

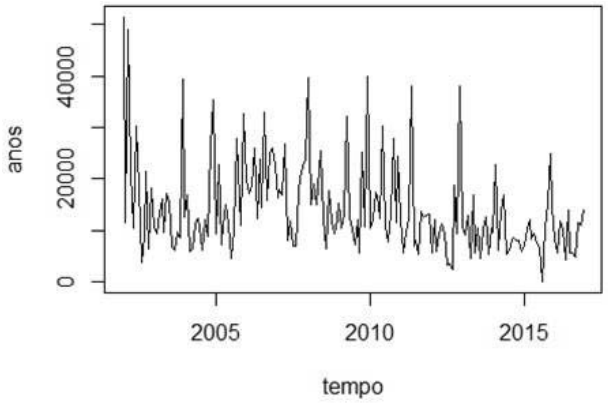

Figura 2. Evolução das vendas

Em seguida fez-se a aplicação de um modelo aditivo realizando a decomposição da tendência, sazonalidade e residual (Figura 3), onde o primeiro gráfico representa o valor original, o segundo a estimação da tendência, o terceiro sazonalidade e no último o residual.

Na tendência pode-se observar um decréscimo seguindo de uma tendência de crescimento depois de 2003 e subidas e descidas entre 2007 e 2010 com uma queda acentuada a partir de 2010.

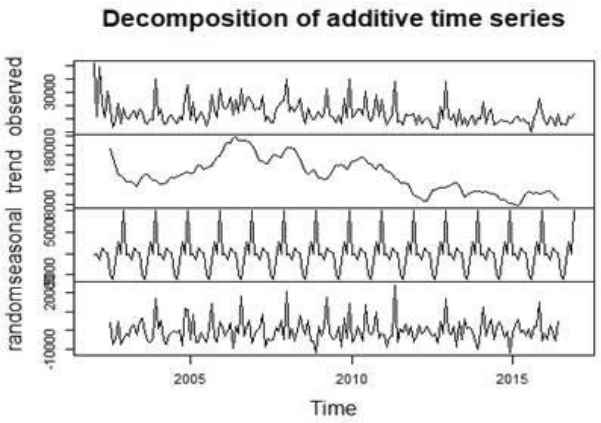

Figura 3. Decomposição da time serie

A. *Exponencial Smoothing*

Ajustou-se sazonalidade retirando a componente sazonal da série original e procedeu-se à previsão recorrendo à função HoltWinters [18] com parâmetros beta=FALSE e gamma=FALSE para suavizar os dados exponencialmente. A previsão foi realizada para 12 meses (Figura 4).

A previsão baseada no HoltWinters é apresentada a azul, 80% de confiança no intervalo a azul escuro e 95% confiança no intervalo mais claro. Os erros de previsão são calculados a partir da diferença entre a previsão e o valor observado para cada valor.

Os erros de previsão são guardados em *residuals*. Estes elementos também foram obtidos nos métodos seguintes. Para realizar a avaliação de eventual melhoria do modelo recorreu-se à avaliação das correlações entre os erros de previsão, através do ACF (*auto correlation function*). Testou-se ainda a significância através do Ljung-Box test não se identificando correlações entre os valores residuais (p-value> 0.05).

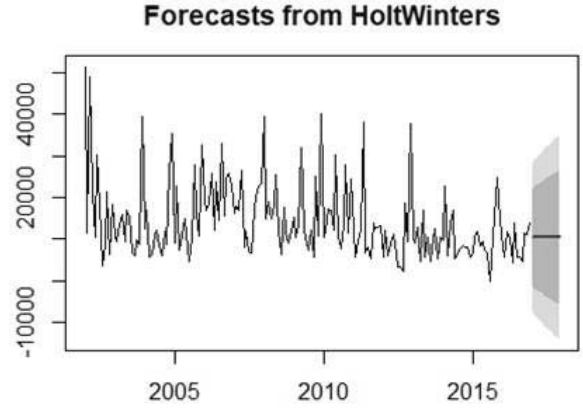

Figura 4. Previsão HoltWinters a 12 meses

B. *ARIMA*

A utilização do método ARIMA requer que as previsões sejam realizadas sem a existência de correlações entre os valores e que as séries sejam estacionárias, isto implica iniciar-se a aplicação com uma série não estacionária, removendo as

diferenças até ser obtida uma série estacionária. Na aplicação do modelo ARIMA utilizou-se o método de Box-Jenkins [8] com as seguintes fases:

1. Escolher o modelo ARMA se for estacionária ARIMA se não for;
2. Identificar e utilizar os dados e a informação relacionada para selecionar um modelo que melhor sumarize os dados:
   a. Verificar se a série é estacionária ou não e quantas diferenciações são necessárias até ser estacionária.
   b. Identificar os parâmetros do modelo ARMA através da *Auto Correlation Function* (ACF) e *Partial Autocorrelation Function* (PACF)
3. Estimar através da utilização dos dados;
4. Avaliar o modelo e verificar onde pode ser melhorado. Proceder à verificação com o ACF e PACF, onde não devem existir correlações ou correlações parciais nos erros.

Na diferenciação (d) verificou-se que com uma diferenciação (d=1) existiam dados aparentemente estacionários. Na avaliação do ACF e PACF obteve-se um p=1 e q=2, o que permite utilizar um ARIMA (1,1,2). Com estes parâmetros realizou-se uma previsão a 12 meses (Figura 5). Onde obtivemos também o intervalo de confiança representado a 80% a azul escuro e o de 95% confiança a azul mais claro. Os erros de previsão são calculados a partir da diferença entre a previsão e o valor observado para cada valor.

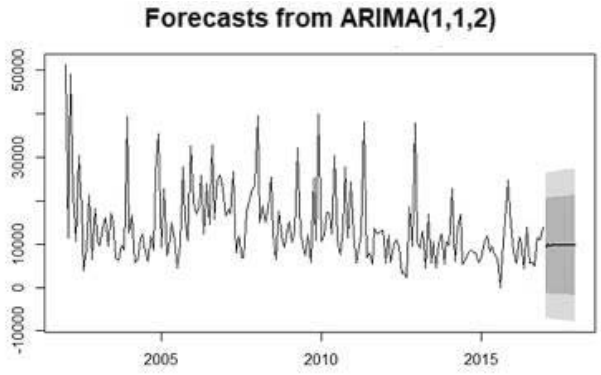

Figura 5. Previsão ARIMA a 12 meses

A biblioteca forecast possuí o auto ARIMA que retorna o melhor modelo para a aplicação do ARIMA, e que permite a identificação dos parâmetros, identificando um modelo sazonal do tipo ARIMA (p, d, q) × (P, D, Q) S, em que p = ordem do AR não sazonal, d = diferenciação do não-sazonal, q = ordem não sazonal do MA, P = ordem sazonal do AR, D = diferencial sazonal, Q = ordem sazonal do MA e S = período de tempo para repetir o padrão sazonal, onde obtivemos um ARIMA (1,1,2) (2,0,2) [13] (Figura 6).

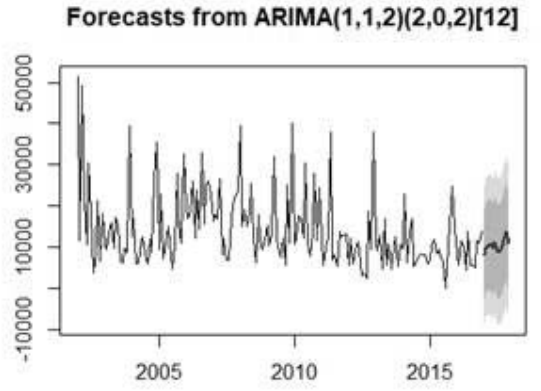

Figura 6. Previsão com auto ARIMA a 12 meses

C. *Artificial Neural Networks*

A realização da previsão com a função Neural Network Time Series Forecasts (NNETAR), disponível na biblioteca forecast do R, recorrendo a um modelo NNAR(p,P,k), onde p corresponde ao número de observações como input, P se é sazonal ou não, onde P=1 indica sazonalidade e k o número de neurons no layer escondido. Obteve-se um modelo NNAR(15,1,8). Os resultados obtidos indicam que o modelo utilizou 15 observações como input e 8 neurons num *layer* não visível. Como existe sazonalidade nos dados obteve-se um p=1 (Figura 7).

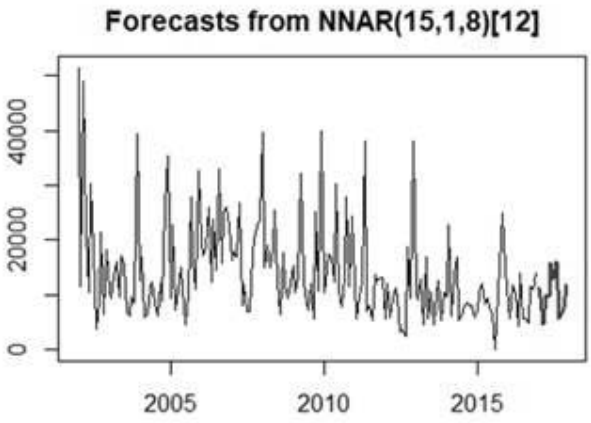

Figura 7. Previsão NNAR a 12 meses

IV. RESULTADOS

O desempenho dos modelos utilizados está representado no Quadro 1, onde podemos comprovar as capacidades preditivas do *Exponencial Smoothing*, ARIMA, AUTO.ARIMA e ANN. Para a comparação do desempenho das várias técnicas utilizou-se MAE e MAPE para avaliar o desempenho.

O MAE (*Mean Absolute Error*) e o MAPE (*Mean Absolute Percentage Error*) são utilizados como um indicador para avaliar qual é o modelo com melhor desempenho. Os resultados demonstram que as redes neuronais têm uma maior exatidão na previsão no MAE e MAPE uma vez que apresenta

os valores mais baixos em ambos indicadores. O valor foi obtido através da utilização da função *accuracy*.

Quadro 1. Comparação da performance dos modelos

Modelo	MAE	MAPE
Exponencial Smoothing	6363.39	58.2596
ARIMA	6169.044	59.9269
AUTO.ARIMA	6122.355	59.4057
ANNs	506.170	4.968

V. CONCLUSÕES E TRABALHO FUTURO

A utilização das técnicas de previsão de vendas recorrendo a time series baseadas nos modelos: *Exponencial Smoothing*; ARIMA; Auto_ARIMA e ANNs mostrou que o melhor desempenho foi obtido com a aplicação do modelo baseado em ANN.

Estes resultados também foram obtidos por Hann e Steurer [19], onde o ANN ultrapassa o desempenho de modelos lineares, na utilização de dados semanais ou mensais. Khashei e Bijari [20] também obtiveram resultados superiores, mas ao aplicarem um modelo hibrido combinando ARIMA numa primeira fase com uma segunda fase aplicando ANN sobre os valores obtidos na primeira fase, obtiveram melhores resultados. O ANN torna-se útil para a resolução de problemas onde existem dados e conhecimentos sobre pressupostos teóricos subjacentes ao problema que estamos a analisar [6], o que permite que o modelo seja aplicado facilmente.

Contudo, os resultados obtidos carecem de desenvolvimento e aprofundamento do trabalho realizado para que se possa determinar qual das técnicas se revela mais fiável no contexto da aplicação generalizada a este tipo de problemas. Como trabalho futuro a exploração de técnicas hibridas poderá interessante para a avaliação do seu desempenho, apesar da simplicidade do ANN.

O trabalho futuro passa pela continuação do teste das técnicas de previsão baseadas em time series de modo a desenvolver um modelo de previsão fiável e pela incorporação desse modelo nas aplicações informáticas atualmente utilizadas o que constituirá uma ferramenta muito importante para o suporte da decisão dos gestores.

REFERÊNCIAS BIBLIOGRÁFICAS

1 Nunnari, G., & Nunnari, V. (2017). Forecasting Monthly Sales Retail Time Series: A Case Study. Em *2017 IEEE 19th Conference on Business Informatics (CBI)* (Vol. 01, pp. 1–6).

2 Ribeiro, A., Seruca, I., & Durão, N. (2016). Sales prediction for a pharmaceutical distribution company: A data mining based approach. Em *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–7).

3 Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering*, *75*(2), 196–204.

4 Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10.

5 Winters, P. R. (1976). Forecasting Sales by Exponentially Weighted Moving Averages. Em *Mathematical Models in Marketing* (pp. 384–386). Springer, Berlin, Heidelberg.

6 Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

7 Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

8 Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5 edition). Hoboken, New Jersey: Wiley.

9 Cheng, B., & Titterington, D. M. (1994). Neural Networks: A Review from a Statistical Perspective. *Statistical Science*, *9*(1), 2–30.

10 Wirth, R & Jochen I. (2000). CRISP-DM: Towards a standard process model for data mining. Em Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (pp. 29–39).

11 Khan, R., & Quadri, S. (2012). Business Intelligence: An Integrated Approach. *Business Intelligence Journal, Vol 5. n.º 1*, 64-70.

12 Continuum Analytics. (2016). *Anaconda Software Distribution*. Obtido de https://continuum.io

13 McKinney, W., & others. (2010). Data structures for statistical computing in python. Em *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). SciPy Austin, TX. Obtido de https://pdfs.semanticscholar.org/f6da/c1c52d3b07c993fe52513b8964f86e8fe381.pdf

14 R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Obtido de http://www.R-project.org

15 Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22.

16 Davydenko, A., & Fildes, R. (2016). Forecast error measures: critical review and practical recommendations. *Business Forecasting: Practical Problems and Solutions. Wiley*, *34*. https://doi.org/10.13140/RG.2.1.4539.5281

17 Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, *8*(3), 147–156. https://doi.org/10.1016/S0969-6989(00)00011-4

18 Hyndman, R. J., & Athanasopoulos, G. (2013). *Forecasting: principles and practice*. S.l.: OTexts.

19 Hann, T. H., & Steurer, E. (1996). Much ado about nothing? Exchange rate forecasting: Neural networks vs. linear models using monthly and weekly data. *Neurocomputing*, 10(4), 323–339. https://doi.org/10.1016/0925-2312(95)00137-9

20 Khashei, M., & Bijari, M. (2010). An artificial neural network (p,d,q) model for timeseries forecasting. *Expert Systems with Applications*, 37(1), 479–489. https://doi.org/10.1016/j.eswa.2009.05.044