Hackathon JPA Agro 2021

Data Science Research Group - DSRG
Universidade Federal de Lavras - UFLA

Identificação da equipe

Nome da equipe: Stats Group
Integrante 1: Alice Silva Duarte
Integrante 2: Leonardo Biazoli
Integrante 3: Matheus Saraiva Alcino

Descrição da solução

1. Entendimento do negócio

A negociação de um dos principais produtos da JPA Agro, polpa cítrica, pode estar relacionada a outros produtos do setor agropecuário e do mercado financeiro, como, por exemplo, a laranja, o milho e a cotação do dólar.

Nesse sentido, buscou-se estudar modelos utilizados na literatura para previsões de dados financeiros. Conforme Ceretta, Righi e Schlender (2010), as redes neurais artificiais podem ser mais eficientes na previsão do comportamento do mercado financeiro e, assim, uma alternativa em relação aos modelos tradicionais de séries temporais (ARIMA, SARIMA, ARCH e GARCH). Diante disso, realizou-se a previsão dos dados da JPA Agro considerando modelo de rede neural artificial incorporado com modelo ARIMA.

2. Pré-processamento dos dados

O conjunto de dados consiste em dados diários do preço de venda de Polpa Cítrica fornecida pela JPA Agro no período de janeiro de 2014 a julho de 2019. Os dados iniciam no dia 07/01/2014 e encerram no dia 31/07/2019. Verificou-se que não ocorreu uma frequência periódica nos dados, tendo em vista que em alguns dias não tinham observações do preço da polpa cítrica. Com o objetivo de solucionar esse obstáculo as datas com informações faltantes foram preenchidas com o ultimo valor observado, ou seja, se a linha i apresentasse valor ausente, o valor a ser substituído refere-se ao valor da linha $i - 1$.

O resumo estatístico dos dados é mostrado na Tabela 1.

Tabela 1: Resumo Estatístico dos dados

Média	Máximo	Mínimo	Mediana	1º Quartil	3º Quartil
388.3	845.0	145.0	360.0	280.0	470.0

Com a finalidade de capturar a variabilidade anual do preço da polpa cítrica a Figura 1 exibe um diagrama de caixa (*boxplot*) e um diagrama de dispersão, referentes a cada

ano presente nos dados. Na Figura 1 é possível observar também que o ano de 2017 apresentou a maior variabilidade dentre os anos do estudo.

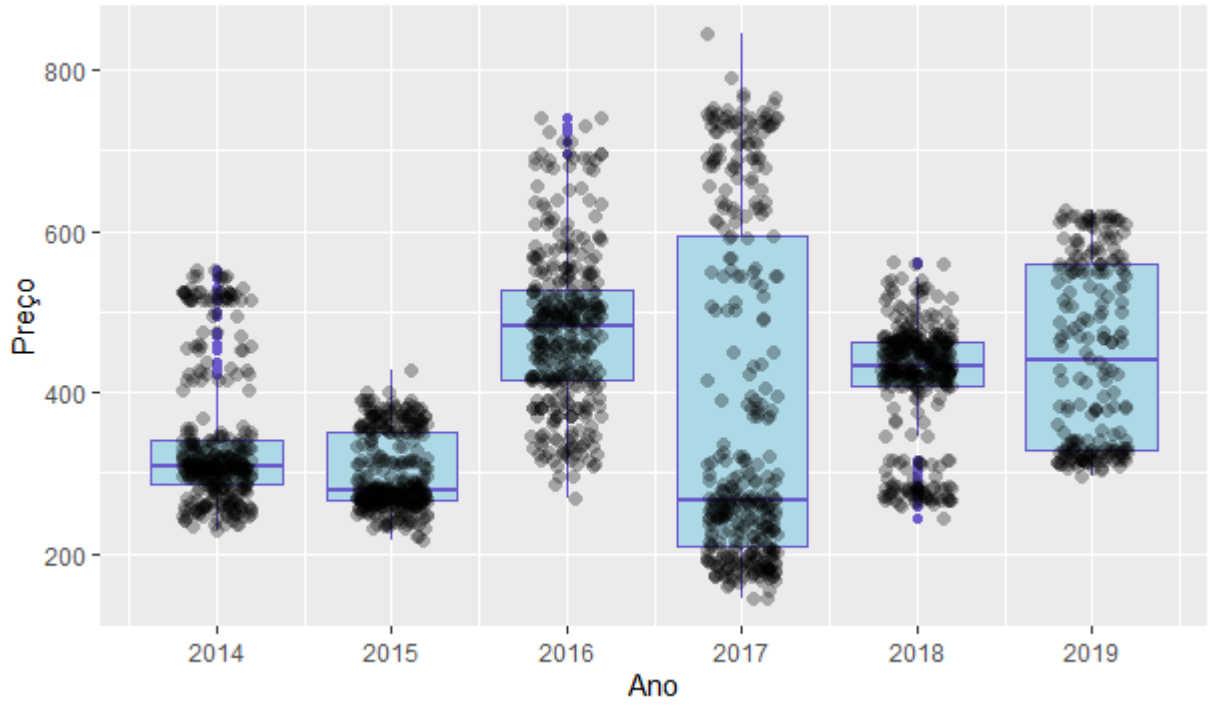

Figura 1: Variabilidade anual da série de preços.

3. Enriquecimento dos dados

Para uma melhor modelagem os dados faltantes foram preenchidos com o valor anterior a ele, ou seja, nas observações em que não há a informação do preço da polpa da laranja, estas foram preenchidos com a informação do dia anterior com a finalidade de melhorar a qualidade dos dados.

4. Modelos

Um estudo descritivo da série apontou que há dependência entre as observações e, dessa forma, é conveniente a utilização de modelos de séries temporais. Com base nas informações de estatísticas descritivas e de autocorrelação, inicialmente buscou-se encontrar modelos da classe ARIMA. O modelo ARIMA *(Autoregressive Integrated Moving Average)*, criado por Box e Jenkins (1970), é um modelo usado na estatística e econometria para lidar com séries temporais. Trata-se de um caso geral do modelo auto-regressivo de médias móveis. Os modelos ARIMA (p, d, q) são, de forma geral, a diferenciação dos processos não estacionários ARMA (Auto regressivos de Médias Móveis). A ideia é que a diferenciação (d) torna o modelo estacionário, ou seja, processo este transforma séries não estacionárias em estacionárias através de suas diferenças (TSAY, 2005; CERETTA et al, 2010).

Apesar destes modelos apresentarem resultados promissores, os modelos convencionais da classe ARIMA utilizados não atingiram poder de previsão considerável. A alta variabilidade presente na série de dados pode ser um dos motivos que justificam tal fato.

Por causa disto, foi utilizado um modelo de séries temporais mais robusto capaz de captar tamanha volatilidade. Existem algumas alternativas na literatura que poderiam ser utilizadas, tais como modelos da classe GARCH (e todas as suas variações) que modelam a volatilidade condicional de uma série temporal (MORETTIN, 2006).

Ainda que modelos da classe GARCH atinjam resultados melhores do que os modelos da classe ARIMA, em termos de erro quadrático médio, optou-se por utilizar modelos de redes neurais para a modelagem proposta. A rede construída leva em conta a estrutura de um modelo ARIMA, mas também a estrutura de uma rede neural para que o modelo "aprenda"com o conjunto de dados e atinja melhores resultados de previsão.

O modelo utilizado foi, portanto, uma rede neural com 5 neurônios (ou nós) que leva em conta a estrutura de um modelo ARIMA(32, 1, 3) sem considerar qualquer variável exógena como regressora. Tal parametrização foi dada pela observação de resultados de modelos anteriores e, com relação a quantidade de camadas que compõem a rede, é preciso uma certa calibragem para que o modelo não decore os dados, mas aprenda com eles e consiga realizar previsões mais precisas. A implementação do modelo foi dada através da linguagem de programação R (R CORE TEAM, 2020), através do pacote *forecast*.

Desta forma, a Figura 2 exibe a serie trabalhada até o dia 31/07/2019 em preto e a previsão realizada para o mês de agosto de 2019 em vermelho.

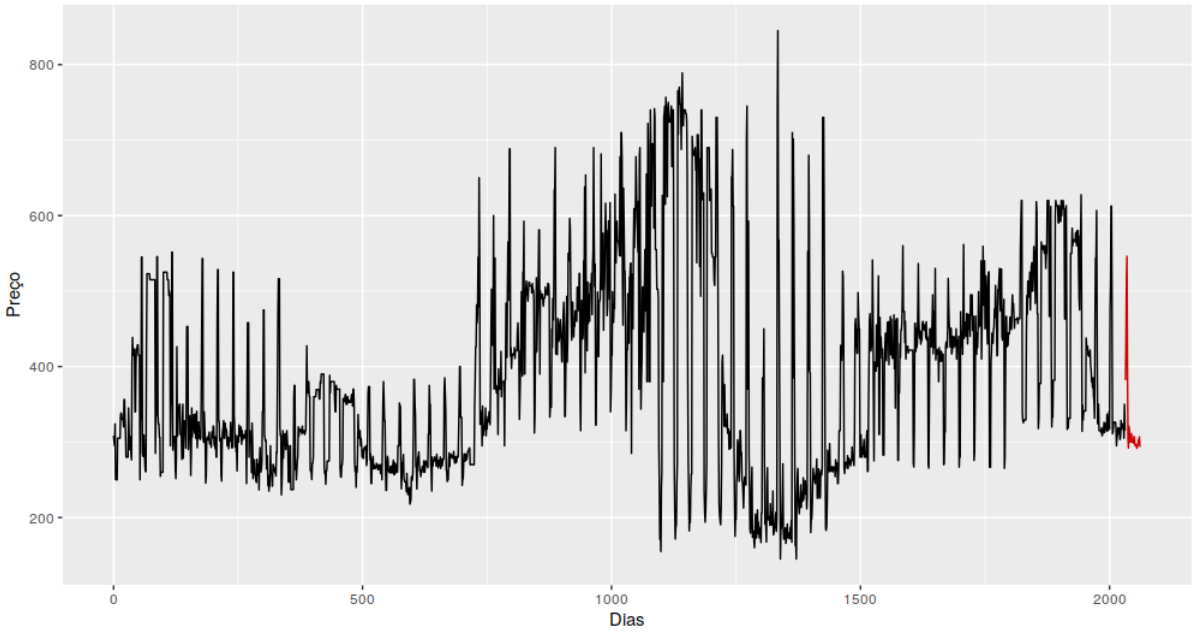

Figura 2: Série com a previsão para os próximos 30 dias.

5. Avaliação da solução

Para avaliar o modelo foi separado uma parte da amostra correspondente a 90% para treinamento da rede neural e o restante (10%) para teste da mesma. Os resultados da previsão podem ser observados através da Figura 3.

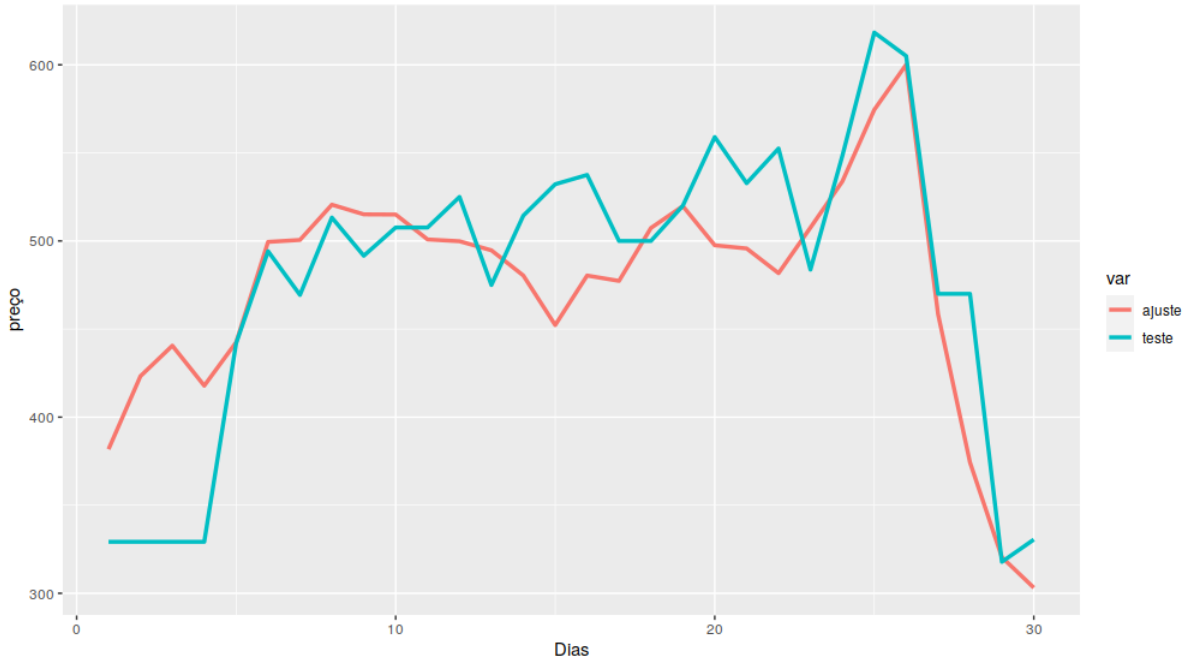

Figura 3: Treinamento versus teste.

O erro quadrático médio obtido na previsão utilizando os dados de treinamento e comparados com os dados de teste foi de 48,98.

Referências

BOX, G. E; JENKINS, G. M; REINSEL, G. C & LJUNG, G. M. Time series analysis: forecasting and control **Holden-day San Francisco**. BoxTime Series Analysis: Forecasting and Control Holden Day, 1970.

CERETTA, P. S; RIGHI, P. B; SCHLENDER, S. G. Previsão do preço da soja: uma comparação entre os modelos ARIMA e redes neurais artificiais. **Informações Econômicas**, v. 40, n. 9, p. 15-27, 2010.

DE ARCE, R; MAHÍA, R. Modelos Arima. Programa CITUS: Técnicas de Variables Financieras, 2003.

HYNDMAN R, ATHANASOPOULOS G, BERGMEIR C, CACERES G, CHHAY L, OHARA-WILD M, PETROPOULOS F, RAZBASH S, WANG E, YASMEEN F. forecast: Forecasting functions for time series and linear models. R package version 8.13, 2020 `https://pkg.robjhyndman.com/forecast/`.

MORETTIN, P. A.; TOLOI, C. **Análise de séries temporais**. In: Análise de séries temporais. 2006. p. 538-538.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. URL `http://www.R-project.org/`.

TSAY, R. S. **Analysis of financial time series.** John wiley & sons, 2005.