Hackathon JPA Agro 2021

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

Identificação da equipe

Nome da equipe: RHandOn
Integrante 1: Ana Claudia Festucci de Herval
Integrante 2: Renata Aparecida Cintra

Descrição da solução

O problema proposto baseia-se no cálculo da previsão de 30 valores referentes aos preços da polpa cítrica a partir de um banco de dados fornecido com 1637 observações, datados de 07/01/2014 a 31/07/2019.

Inicialmente, a partir de uma análise exploratória dos dados, com tabelas e gráficos descritivos, foi possível notar que a amostragem de dados foi dada de maneira irregular, contando dias úteis e não úteis, apontando que haviam dados faltantes (*missing data*).

Foi realizada uma interpolação dos dados faltantes e optou-se por seguir com o ajuste dos modelos considerando ambos os conjuntos de dados: o original e o completo, com dados interpolados.

A partir de algumas pesquisas, foi identificado que alguns estudos apresentavam uma forte influência do preço do milho no preço da polpa cítrica, bem como o desempenho da safra da laranja, além de questões climáticas e choques exógenos. Houve certa dificuldade em se obter informações relativas a essas possíveis covariáveis, por exemplo, preço do milho apenas para determinada região e/ou numa frequência mais baixa, como mensal, ao invés de diária. Por fim, optou-se por utilizar preço das *commodities* do milho e preço do suco de laranja da negociados na Bolsa NY, obtidos no site Investing.com, pois as mesmas são possíveis variáveis que influenciam no preço da polpa cítrica.

A exploração dos modelos foi feita segundo as ferramentas com as quais os participantes tem familiaridade. Foi o caso com modelos da classe ARIMA e modelos bayesianos, conhecidos como modelos latentes gaussianos.

Também conhecidos como modelos paramétricos de Box e Jenkins, os modelos ARIMA (autorregressivos integrados de médias móveis) são frequentemente utilizados devido aos bons resultados em previsões com ajuste sazonal. No entanto, nestes modelos a ordem é dada principalmente pelas funções de autocorrelação e de autocorrelação parcial, e esta mostrou-se difícil de captar a dinâmica da correlação serial em um modelo parcimonioso, apresentando valores significativos em *lags* de ordem mais alta, como 26 e 31. Além disso, é sabido que modelos com maior número de parâmetros influenciam negativamente nas obtenções de previsões (MORETTIN & TOLOI, 2006).

Já os modelos latentes gaussianos (LGM), são uma classe de modelos bayesianos que englobam a grande maioria dos modelos mais utilizados em estatística. Para citar alguns, temos os modelos aditivos mistos generalizados, os modelos lineares generalizados, modelos espaciais para dados de área, superfícies contínuas (geoestatísticos) e de processos

pontuais, modelos temporais, espaço-temporais e entre outros. Os LGM são descritos em uma estrutura hierárquica de 3 estágios. A primeira é descrita pela verossimilhança dos dados e esta pode descrever dados de diferentes formas: discretos, contínuos, categóricos, etc. O segundo estágio é dado pelo campo latente, que por sua vez é descrito por variáveis que temos o interesse em realizar alguma inferência, porém, tais variáveis não são diretamente observáveis. Estas variáveis do campo latente, assumem prioris gaussianas. O terceiro estágio é descrito pelos hiperparâmetros, que podem assumir qualquer forma de densidade e se referem, na maioria das vezes, aos parâmetros de precisões, correlações das prioris do campo latente e também das precisões da verossimilhança (JAMES et. al, 2013, MARTINO & RUE, 2009; RUE, 2017).

Uma grande vantagem dos LGM é que sua estimação é realizada através da metodologia INLA (do inglês *Integrated Nested Laplace Approximation* que é um algoritmo determinístico para realizar inferência nos modelos LGM. A metodologia INLA é uma alternativa superior aos métodos de Monte Carlo via Cadeia de Markov (MCMC) quando utilizada denro da classe de LGM, pois fornece estimativas mais precisas e com um tempo computacional bem inferior.

Para o problema em questão, ajustou-se um modelo com verossimilhança gamma, sendo utilizado um modelo *random walk* de ordem 1 para modelar a tendência temporal.

Por fim, optou-se pela utilização de um *Random Forest* (RF) (BREIMAN, 2001) para dados de séries temporais. Para isto, uma transformação na variável foi realizada: logaritmo para estabilizar a variância e uma diferença para eliminar a tendência e obter estacionariedade. A partir destas transformações também fica mais claro avaliar qual a estrutura de dependência no processo de preço em relação à ordem do *lag* a ser considerada (analogamente ao que é feito no ajuste do modelo ARIMA).

É preciso considerar uma matriz que incorpora as defasagens no tempo e que vai viabilizar a utilização de uma técnica de *Machine Learning* (ML). Isto porque técnicas de ML, em geral, não tem consciência sobre o tempo, ao invés disso, consideram que as observações são independentes e identicamente distribuídas, suposição claramente violada em dados de séries temporais, caracterizados pela dependência serial.

O *Random Forest* é um algoritmo de aprendizagem de máquina simples, flexível e poderoso, que produz excelentes resultados na maioria das vezes, mesmo sem ajuste de hiperparâmetros. O algoritmo cria várias árvores de decisão e as combina para obter uma predição com maior acurácia e mais estável.

Como a maioria dos métodos de ML, o RF não têm consciência do tempo. Pelo contrário, consideram que as observações são independentes e distribuídas de forma idêntica. Esta suposição é obviamente violada em dados de séries temporais que são caracterizados pela dependência serial. Contudo, com alguns truques, pode-se fazer previsão de séries temporais com *Random Forest*. Basta um pouco de pré e pós processamento e é possível calcular predições beneficiadas pela robustez do método.

Além disso, RF ou métodos baseados em árvore de decisão são incapazes de prever uma tendência, ou seja, eles não extrapolam. Para entender o porquê, lembre-se de que as árvores operam por regras *if-then* que dividem recursivamente o espaço de entrada. Portanto, eles são incapazes de prever valores que estão fora da faixa de valores da meta no conjunto de treinamento.

Dentre as especificações do RF, como o horizonte de previsão definido era 30, iniciamos a investigação utilizando *hold-out*, com amostra teste e treinamento, o treinamento contendo os exatos 30 últimos valores do conjunto de dados. Além disso, foi realizado um estudo para identificar a melhor defasagem a ser considerada na matriz de defasagens.

Com o auxílio da fac da variável transformada, identificamos que os valores poderiam variar de 2 a 8 considerando as ordens mais baixas, até 31, ordem mais alta que apontou a dependência. A partir do valor obtido de RMSE (*Root Mean Square Error*) e MAPE (*Mean Absolute Percent Error*), concluiu-se por utilizar k=2.

De maneira semelhante, as mesmas métricas foram utilizadas ao se comparar as previsões para horizonte de previsão 30 e ordem de defasagens de 2 a 8, considerando dados originais e completados pela interpolação. Por fim, optou-se por utilizar o conjunto de dados da maneira como foi fornecido também por demonstrarem ser mais robusto à previsão dos valores subsequentes à amostra treinamento utilizada.

A utilização das covariáveis citadas (preço do milho e preço do suco de laranja) não apresentou melhora na acurácia do modelo e, portanto, foram retiradas do modelo final de predição.

Todas as análises foram feitas através do *software* R (R CORE TEAM, 2020).

Referências

BREIMAN, L. Random forests. *Machine learning*, 45(1), 5-32, 2001.

JAMES, Gareth, WITTEN, Daniela; HASTIE, Trevor and TIBSHIRANI, Robert. An introduction to statistical learning. Vol. 112. New York: Springer, 2013.

MARTINO, Sara, RUE, Håvard. *Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program.* Department of Mathematical Sciences, NTNU, Norway, 2009.

MORETTIN, P.A.; TOLOI, C.M.C. *Análise de Séries Temporais* 2 ed. São Paulo: Blucher, 2006.

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/`.

RUE, Håvard et al. *Bayesian computing with INLA: a review.* Annual Review of Statistics and Its Application 4: 395-421, 2017.