Hackathon JPA Agro 2021

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

Identificação da equipe

Nome da equipe: TOP AGRO AI
Integrante 1: Nélio Lemos Freire Junior.
Integrante 2: Evandro Nunes Miranda.
Integrante 3: Luis Otávio Santos.
Integrante 4: Henrique Geraldo Guimarães Silva.

1. Entendimento do negócio

A **polpa cítrica** é um coproduto muito utilizado na nutrição de vacas leiteiras, possuindo aproximadamente 87% da energia do milho (ESALQLab). Nos últimos anos vem se tornando uma excelente fonte de alimentação, com isso, há grande variação no preço nos últimos anos. Dada a complexidade de se prever preços dentro de uma série temporal, utilizamos diferentes estratégias que captassem o comportamento do preço ao longo do tempo. Os algoritmos de série temporal são usados extensivamente para analisar e prever dados baseados no tempo. No entanto, dada a complexidade de outros fatores além do tempo, o aprendizado de máquina surge como uma alternativa poderosa para entender complexidades ocultas em dados de séries temporais e gerar boas previsões.

2. Pré-processamento dos dados

Como se trata de uma série temporal, é importante captar toda a variabilidade dos preços da polpa cítrica, até para aumentar a capacidade do modelo de aprendizado máquina em captar corretamente o preço futuro do produto. Deste modo, não foi efetuado um pré-processamento dos dados, sem a necessidade de limpeza de dados e inspeção. Utilizamos como modelo para avaliar a série temporal o Random Forest, um Ensemble de modelos(árvores de decisão), e isso também impacta na melhora de suas aproximações quando comparado a modelos avaliados individualmente, como o LSTM, ARIMA, etc, uma vez que métodos Ensemble são capazes de mitigar efeitos atrelados ao dilema entre bias e variância.

3. Enriquecimento dos dados

Quando utilizamos um algoritmo de aprendizado de máquina é necessário converter os dados de séries temporais em variáveis de tempo. Deste modo, através das datas fornecidas pelo problema usamos o pacote “lubridate” do R para criar mais variáveis de tempo. Neste problema adicionamos ao banco de dados as variáveis ano, dia no ano, trimestre, mês, dia e dia da semana.

4. Modelos

Foram testados diferentes algoritmos clássicos de série temporal como ARIMA, SARIMA, PROPHET e LSTM, e algoritmos de aprendizado de máquina para definir a melhor estratégia, como Random Forest (RF), Rede Neural Artificial e ANFIS. Com base nestes algoritmos, utilizamos o root-mean-square deviation (RMSE) como métrica avaliadora, e também observamos o comportamento gráfico dos dados analisados e estimados. Com isso definimos o RF como a melhor estratégia para predizer os dados da série temporal.

O RF é um algoritmo de aprendizado de máquina muito utilizado para problemas de regressão e classificação, foi implementado pelo matemático Breiman (2001). O RF é um método flexível o suficiente para modelar interações em dimensões altas, criando muitas árvores de regressão e calculando a média de suas previsões. As árvores são criadas desenhando um subconjunto de amostras de treinamento através da substituição (uma abordagem de ensacamento), ou seja, algumas podem ser selecionadas novamente, enquanto outras não (Belgiu and Drăgu, 2016).

Para o ajuste do modelo, o algoritmo requer três parâmetros a serem definidos para produzir as árvores de decisão: ntree (número de árvores treinadas na floresta), nodesize (tamanho do nó do terminal de destino) e mtry (número de recursos aleatórios usados para dividir um nó da árvore) (O'Brien and Ishwaran, 2019). Ele se divide em dois níveis de randomização, o primeiro é a agregação de bootstrap ("ensacamento"), onde um subconjunto aleatório de dois terços das observações são usadas para treinar as árvores, e o terço restante dos dados (observações "fora do saco") são excluídos para validação (Woznicki et al., 2019). O segundo nível é referente a cada nó das árvores de decisão, onde de forma randômica é selecionado um número de variáveis, e a variável que apresentar a melhor divisão é selecionada para aumentar a árvore nesse nó (Woznicki et al., 2019). Para a avaliação de cada nó, é utilizado métricas. Para problemas de regressão a métrica de avaliação é o mean squared error (MSE, em português erro quadrático médio). A árvore de decisão cresce até onde o parâmetro de decisão de término é decidido pelo usuário (nodesize). A árvore de decisão que recebeu o menor valor de MSE ou que recebeu mais votos na classificação é a resposta final do algoritmo (Belgiu and Drăgu, 2016).

5. Avaliação da solução

Como entrada do modelo de aprendizado de máquina utilizamos 6 entradas, elas sendo: ano, dia no ano, trimestre, mês, dia e dia da semana. O algoritmo apresenta dois parâmetros que influenciam seu desempenho, mtry e ntree. O primeiro se refere ao número de variáveis utilizado em cada árvore de decisão, o segundo a quantidade de árvores que a floresta contém. Neste problema, definimos mtry e ntree como 6 e 10000, respectivamente.

Como a estratégia do problema é definir a predição da polpa cítrica para 30 dias além dos dados fornecidos, optamos por dividir os dados em conjunto de treinamento e de teste, com uma maior quantidade de amostras para treino, com 1637 e 30 para teste. Optamos por essa opção por ser mais próximo da data final, treinando assim possíveis intervenções ou acontecimentos climáticos e mercantil próximo a data que se pretende prever.

E para avaliar o modelo de aprendizado de máquina utilizamos o mean-square

deviation (MSE) como métrica avaliadora, posteriormente calculamos root-mean-square deviation (RMSE) e também vimos o comportamento gráfico dos dados observados e estimados.

6. RESULTADOS

O modelo de Random Forest obteve uma ótima aproximação dos dados observados para polpa cítrica nos dados de treino, com valor do RMSE de R$ 19.40 de erro. Espera-se que a estimativa para os 30 dias futuro em relação ao conjunto de dados tenha um baixo de RMSE. Neste caso, o Random Forest por ser um método Ensemble de modelos (árvores de decisão), obteve aproximações melhores quando comparado a modelos avaliados individualmente, como o LSTM, ARIMA, etc, uma vez que métodos Ensemble são capazes de mitigar efeitos atrelados ao dilema entre bias e variância.

Referências

ESALQLab, 2020. Polpa cítrica: qual a qualidade nutricional do co-produto que utilizo na minha propriedade? Acessado dia 23 de fevereiro de 2021. Disponível em: https://www.milkpoint.com.br/colunas/esalqlab/polpa-citrica-qual-a-qualidade-nutricional-do-coproduto-que-utilizo-na-minha-propriedade-220726/

Belgiu, M., Drăgu, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

O'Brien, R., Ishwaran, H., 2019. A random forests quantile classifier for class imbalanced data. Pattern Recognit. 90, 232–249. https://doi.org/10.1016/j.patcog.2019.01.036

Woznicki, S.A., Baynes, J., Panlasigui, S., Mehaffey, M., Neale, A., 2019. Development of a spatially complete floodplain map of the conterminous United States using random forest. Sci. Total Environ. 647, 942–953. https://doi.org/10.1016/j.scitotenv.2018.07.353