Hackathon JPA Agro 2021

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

Identificação da equipe

Nome da equipe: Lorem Ipsum.
Integrante 1: Chrystian Arriel Amaral.
Integrante 2: Fernando Elias de Melo Borges.
Integrante 3: Gabriel Aparecido Fonseca.
Integrante 4: Jorge Sassaki Resende Silva.

Descrição da solução

1. Entendimento do negócio

O mercado de polpa cítrica para nutrição animal no ano de 2020 sofreu um aumento significativo nos preços, uma vez que outros produtos utilizados para este fim mantiveram os preços em alta no mesmo ano [1][2].

Tendo em vista esta alteração no mercado de polpa cítrica, fazer predições dos preços de venda em um futuro de curto prazo pode ser interessante para compradores, negociadores e vendedores do produto, de maneira que se possa aproveitar melhor as baixas do preço e se preparar em eventuais subidas da mesma variável.

Baseando-se nessa problemática, esta competição visa a predição dos preços de venda da polpa cítrica por meio de uma série temporal disponibilizada.

2. Pré-processamento dos dados

Os dados da série temporal consistem em 3 variáveis, sendo: a data da venda, disponível no formato $dd-mm-aaaa$, o valor da venda da polpa cítrica e o tipo do produto. Esta última variável possui valor constante e não foi utilizada como preditora.

Com relação às duas demais variáveis, ambas não possuíam valores inconsistentes e/ou ausentes, com base nesta informação, não foram necessários procedimentos de remoção ou imputação de valores nestas observações.

Os dados foram reorganizados de acordo com a data, alocando a mesma como índice para uso no modelo preditivo, logo, a base de dados ficou indexada pela data, de maneira que possa ser apresentada ao modelo preditivo como fator de tempo para predição de série temporal.

3. Enriquecimento dos dados

Foram utilizados as datas e os preços de venda da polpa cítrica. Durante os testes chegou a ser incluído o preço do milho no atacado, dado púbico disponibilizado pelo Conab (Companhia Nacional de Abastecimento), contudo, tal dado não contribuiu para melhoria

do modelo preditivo, portanto, os dados novos não foram incluídos restando somente os dois primeiros atributos do banco de dados disponibilizado.

4. Modelos

O modelo final para predição utilizado foi o *Prophet*, obtido por meio da biblioteca de mesmo nome disponível na linguagem Python, o código-fonte da biblioteca é de licença livre (*open-source*). O algoritmo foi desenvolvido pelo Facebook e se baseia modelos logísticos de crescimento por partes [3].

5. Avaliação da solução

O conjunto de dados disponibilizado foi dividido em 80% para treino e 20% para teste de maneira que o conjunto de treinamento ficasse com os dados mais antigos da série enquanto o modelo foi testado com os 20% dados mais recentes.

Como métrica de avaliação, foi utilizado o RMSE (*root-mean-square error*) para comparar os dados de teste preditos pelo modelo com os dados reais disponibilizados. O RMSE obtido durante a execução do modelo foi de, aproximadamente, 100. Para a execução do *forecast* para as 30 amostras futuras não contidas no banco de dados foi utilizada toda a série temporal para a predição destes novos eventos.

Referências

Referências

[1] "Polpa cítrica está custando 43,4% mais em 2020", site Polpa DBO . URL `https://www.portaldbo.com.br/polpa-citrica-esta-custando-434-mais-em-2020/`.

[2] "O mercado de polpa cítrica em 2019"site Canal Rural. URL `https://blogs.canalrural.com.br/blogdoscot/2019/07/01/o-mercado-de-polpa-citrica-em-2019/`

[3] TAYLOR, Sean J.; LETHAM, Benjamin. Forecasting at scale. The American Statistician, v. 72, n. 1, p. 37-45, 2018.