Hackathon JPA Agro 2021

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

Identificação da equipe

Nome da equipe: Sirius.
Integrante 1: Thiago Mantuani de Souza.
Integrante 2: Cecília Ramos de Oliveira.

Descrição da solução

1. Entendimento do negócio

1.1 A empresa

A empresa JPA AGRO está presente no mercado há mais de 10 anos, e possui parcerias com empresas do segmento de granjas, fábricas de ração, cooperativas, entre outros. A empresa definida como holding busca inovar desde a fabricação, comercialização a conexão de produtos para clientes e parceiros. Ao todo a empresa trabalha com mais de 30 produtos, divididos entre farelos, caroços, cascas, óleos e polpa cítrica, sendo está última o que mais se destaca.

propondo soluções inovadoras na conexão, comercialização e fabricação de produtos para os clientes e parceiros

1.2 Polpa cítrica

É um produto derivado da laranja e utilizado como alimento na dieta de ruminantes, como um substituto do milho.

1.3 Problema

A polpa cítrica e um produto que possui bastante varição de preço, devido a ser influenciada pela cotação do milho. Saber um provável preço auxilia a empresa nas suas tomadas de decisão.

1.4 Avaliação

Para avaliar as previsões será utilizado a métrica RMSE (*Root Mean Square Error*).

1.5 Entendimento

Foi necessário buscar informações sobre o produto em questão e também os fatores que o influenciavam.

2. Pré-processamento dos dados

2.1 Base de dados

A base de dados fornecida corresponde a uma série de preços da polpa cítrica do período de 2014 á julho de 2019. É composta por 3 atributos:

- **product**: produto comercializado (valor constante igual a 'Polpa Cítrica'.
- **negotiation_date**: data do faturamento do produto.
- **sold_price**: preço de venda do produto (variável a ser predita).

Como o atributo *product* possui apenas um valor, o mesmo foi removido do conjunto de dados, pois não irá agregar nenhum valor aos modelos de aprendizado de máquinas.

2.2 Estatística descritiva

Conforme Tabela 1 podemos observar que o menor preço foi de 145,00 e o maior durante o período de 2014 a julho de 2019 de 845,00. A média e a mediana são próximas que a distribuição se aproxima de uma normal, que pode ser confirmado através dos valores do coeficiente de assimetria e curtose.

Tabela 1: Estatística descritiva

	Média	**Mediana**	**Desvio**	**Mín**	**Máx**	**Assimetria**	**Curtose**
sold_price	379,73	343,60	130,37	145,00	845,00	0.74	0.04

2.3 Falha na sequência temporal

Analisando a série foi possível identificar falhas temporais, ou seja, a série não é regular. A Tabela 2 nos mostra uma quebra de sequência após o dia 12/01/2014, no qual há um salto para 15/01/2014, e após 15/01/2014 a próxima data é 21/01/2014.

Tabela 2: Falha na sequência temporal

negotiation_date	**sold_price**
09/01/2014	295,00
10/01/2014	324,00
11/01/2014	250,71
12/01/2014	250,00
15/01/2014	305,00
21/01/2014	335,00

A sequência com falhas foi corrigida através de um preenchimento dos dias faltantes, e para os registros do atributo sold_price os mesmos foram preenchidos com o mesmo valor do dia anterior (Tabela 2).

Tabela 3: Amostra de sequência corrigida

negotiation_date	sold_price
12/01/2014	250,00
13/01/2014	250,00
14/01/2014	250,00
15/01/2014	305,00
16/01/2014	305,00
17/01/2014	305,00
...	...
21/01/2014	335,00

2.4 Análise exploratória

Através da Figura 1 pode-se observar que a série possui uma distribuição assimétrica positiva.

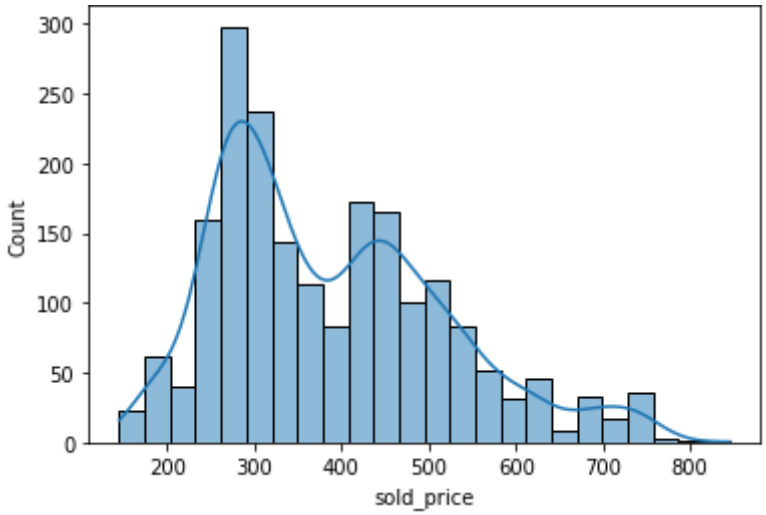


Figura 1: Distribuição da sold_price

A Figura 2 nos mostra que o preço possui uma tendência crescente até determinado ponto e após isso há um decaimento e novamente um crescimento, através desse comportamento podemos observar uma certa sazonalidade.

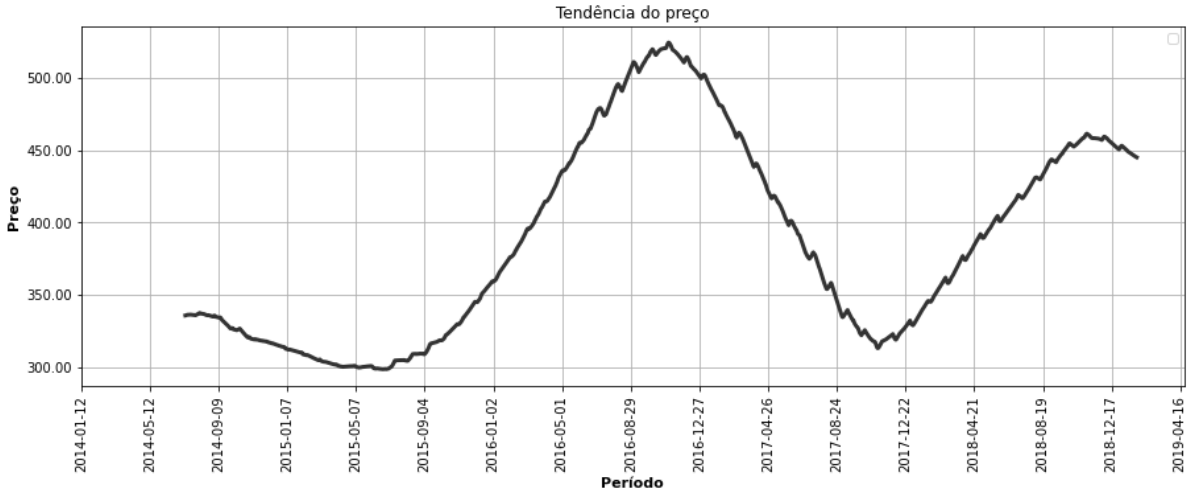


Figura 2: Tendência do preço

3. Enriquecimento dos dados

3.1 Engenharia de recursos

Através do atributo data foram criadas novas variáveis, como ano, mês, dia, dia da semana, semana do ano. Já com o atributo sold_price foram criadas variáveis com preços anteriores (lags) e atributos com diferenciação de valores de preços com lags anteriores. Para esse processo foi definido um lag de 5. Novos atributos criados:

- **day**: dia.
- **month**: mês.
- **year_week**: semana do ano.
- **day_week**: dia da semana.
- **sold_price_lag_1**: preço do dia anterior.
- **sold_price_lag_2**: preço de 2 dias anteriores.
- **sold_price_diff_1**: diferença do preço atual com do dia anterior.
- **sold_price_diff_2**: diferença do preço atual com do dois dias anteriores

Obs: foram criadas variáveis sold_price_lag e sold_price_diff até 5 dias anteriores.

4. Modelos

4.1 Transformação dos dados

Os atributos sold_price_lag_X foram normalizadas através da equação da Figura 3. Já os atributos sold_price_diff_X foram transformadas através da equação da Figura 4 e para os atributos que possuem uma natureza cíclica como dia, mês, ano e semana do ano foram colocadas como arcos (seno e cosseno). E na variável resposta sold_price foi feito uma transformação logarítmica afim de torná-la uma distribuição normal.

Figura 3: Normalização

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Figura 4: Robust Scaler

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

4.2 Seleção de atributos

Nessa etapa foi utilizado o algoritmo Boruta, afim de remover variáveis colineares. os atributos selecionados foram: sold_price_lag_1, sold_price_lag_2, sold_price_diff_1, sold_price_diff_2, sold_price_diff_3, sold_price_diff_4, sold_price_diff_5 e day_sen.

4.3 Modelos utilizados

Os modelos utilizados para previsão, foram os baseados em árvores, como o LightGBM e o XGBoost. Ambos os modelos foram selecionados, criando-se uma combinação de ambos, com uma média ponderada.

5. Avaliação da solução

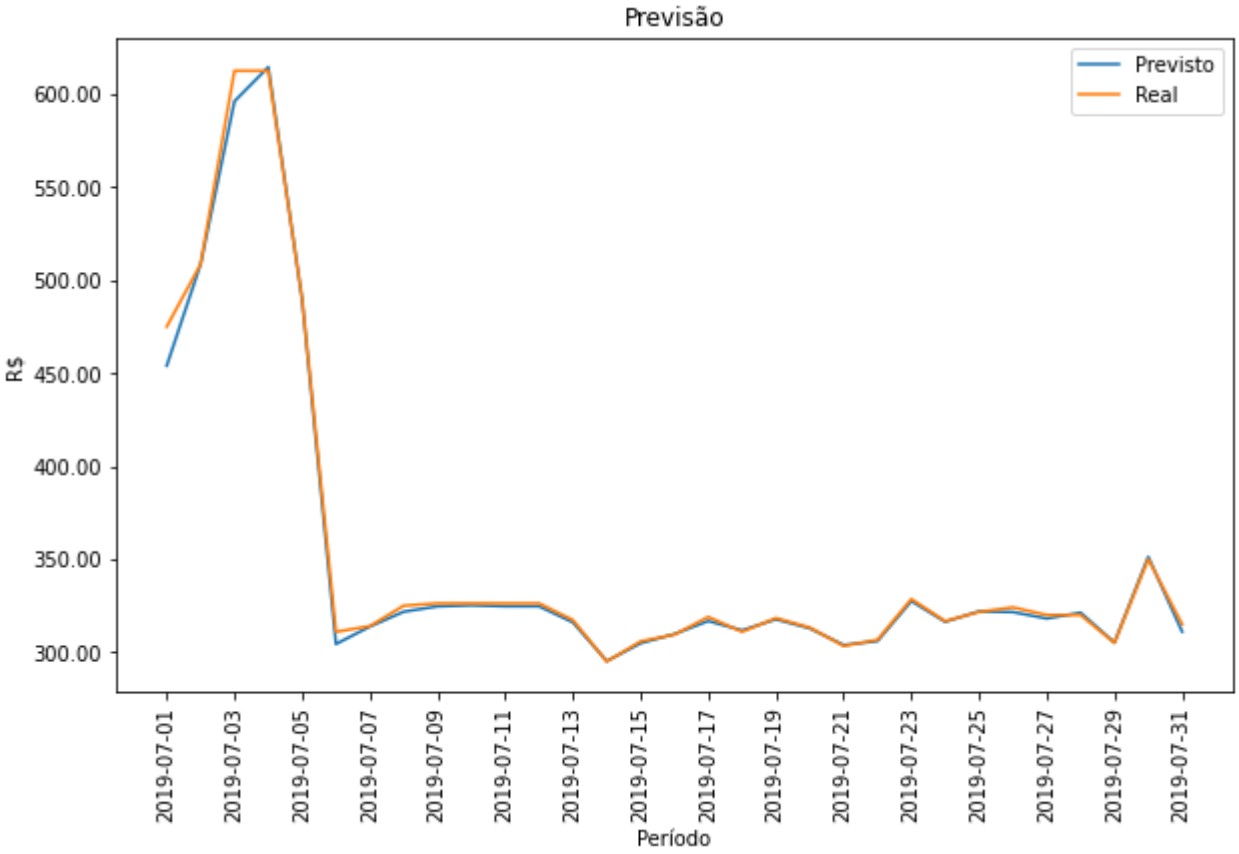
5.1 Divisão entre treino e teste

Para avaliar o modelo, o conjunto de dados foi dividido entre treino e teste, sendo o conjunto de treino contemplando um período de 2014 á junho de 2019, e para teste de 01/07/2019 á 31/07/2019.

5.2 Resultados

O modelo LightGBM obteve um RMSE de 4.37 e o XGBoost um RMSE de 4.60 no conjunto de testes. A figura 5 nos mostra as previsões para o mês de julho/2019 do conjunto de testes. Para um melhor resultado, foi combinado os 2 modelos através de uma média ponderada, sendo 0.6 para o LightGBM e 0.4 para o XGBoost.

Figura 5: Previsão do conjunto de testes



Referências

XGBoost Documentation. Acesso em 23 de fevereiro de 2020. URL `https://xgboost.readthedocs.io/en/latest/`.
LightGBM Documentation. Acesso em 23 de fevereiro de 2020. URL `https://lightgbm.readthedocs.io/en/latest/`.
Boruta Explained. Acesso em 23 de fevereiro de 2020. URL `https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a`.
Indicador do Milho. Acesso em 23 de fevereiro de 2020. URL `https://www.cepea.esalq.usp.br/br/indicador/milho.aspx`.
Uso de polpa cítrica na alimentação de ruminantes. Acesso em 23 de fevereiro de 2020. URL `https://www.educapoint.com.br/blog/pecuaria-geral/polpa-citrica-alimentacao-ruminantes/`.
Hackathon JPA Agro 2021. Acesso em 23 de fevereiro de 2020. URL `https://github.com/dsrg-icet/hackathon_JPAAgro`.