Hackathon JPA Agro 2021

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

Identificação da equipe

Nome da equipe: Victor Cabral.
Integrante 1: Victor Gustavo Cabral Rodrigues.

Descrição da solução

Neste hackathon, foi proposto a realização de um modelo preditor para compreender o comportamento dos preços de venda de Polpa Cítrica. O pedido de obter os 30 valores para os próximos 30 dias foi entendido como um problema de extrema complexidade, pois por ser relacionado ao mercado de commodities como o Milho, o valor de venda da Polpa se mostra tão instável quanto o próprio mercado financeiro.

Para abraçar tal complexidade, foi proposto um modelo baseado a partir de Redes Neurais, utilizando a arquitetura LSTM (Long Short Term Memory), uma Rede Neural Recorrente (RNN), que carrega consigo dados recentes enquanto está prevendo novos dados. Tudo isso feito com o auxílio de médias móveis, para diminuir o ruído dos dados.

1. Entendimento do negócio

Com o auxílio de profissionais na área e pesquisa em sites do ramo, busquei entender melhor o mercado de commodities. A partir dessas buscas, procurei entender como o mercado futuro de Milho se comportava ao lado do preço de venda da Polpa Cítrica.

Assim como o preço da Polpa, a instabilidade está bem presente ao longo do tempo na cotação do Milho. Porém é visível que a commodity afeta consideravelmente o preço do produto. Sabendo disso, em um primeiro momento, busquei aplicar os dados lado a lado na arquitetura LSTM, mas sem sucesso, devido a um elevado RMSE na fase de testes.

Considerei utilizar os dados da cotação do milho dos 30 dias propostos (Agosto de 2019), mas acreditei que não seria coerente com a proposta, que é prever os dados em um período que ainda não há nenhum dado de mercado.

2. Pré-processamento dos dados

Assim que recebi os dados no dia 22 de fevereiro, comecei a analisar sua natureza. A princípio, nenhum padrão foi identificado, nada que pudesse ser previsível em suas séries temporais. Devido sua impresivibilidade, a quantidade de linhas disponíveis mostrou-se um tanto quanto pequena para uma análise mais aprofundada e uma previsão com elevada acurácia.

A imprevisibilidade dos dados é compreensível, pois tratam-se de valores baseados no mercado financeiro e de commodities. Porém, grandes mudanças nos preços tornaram essa previsão ainda mais complicada, é possível ver aumentos e diminuições extremas nos

valores de venda de um dia para o outro. Com esses "ruídos"nas séries temporais, viu-se necessário o uso das médias móveis para analisar o comportamento de maiores períodos de tempo e diminuir esse ruído.

3. Enriquecimento dos dados

Todos os atributos do banco de dados disponibilizado foram utlizados, tanto as datas quanto os preços de venda foram úteis na análise do problema. Além desses dados, busquei dados da cotação de cítricos e de milho, no qual apenas o segundo teve uma implementação e uma análise mais aprofundada.

Com a informação prévia de que os dados do mercado de milho podem influenciar no preço da Polpa Cítrica, busquei prever também os valores dos próximos 30 dias no mercado de milho. Os dados reais deste período estão disponíveis, porém baseado na busca da JPAgro em uma solução voltada para tomadas de decisão futuras, busquei tentar prever estes dados, baseados nos números a partir de 2004.

Utilizando a arquitetura LSTM para fazer as 2 previsões. vi o RMSE da minha solução aumentar cada vez mais. Fazer uma previsão de mercado futuro é uma tarefa árdua, mas nada comparado a fazer duas. Desta forma, utilizei os dados históricos do mercado de milho apenas para consultar a qualidade do modelo.

4. Modelos

Dada a complexidade da proposta, em um primeiro momento já me veio a ideia de utilizar de Deep Learning e redes neurais para o desenvolvimento do projeto. Buscando e analisando a utilidade dos modelos e algoritmos de Deep Learning, encontrei o conceito de Redes Neurais Recorrentes (RNN), e logo me deparei com a arquitetura LSTM(Long Short Term Memory).

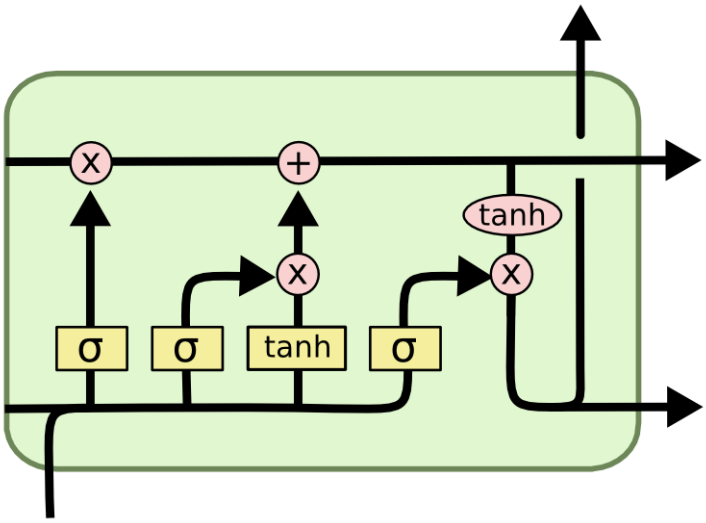
O diferencial dessa rede neural em relação ás tradicionais é a memória. Redes Neurais Recorrentes possuem loops que permitem que as informações persistam conforme o tempo avança. As RNNs podem ser imaginadas como várias cópias de uma mesma rede neural, passando mensagens ás redes sucessoras.

Por Estes fatores, LSTM é adequada para prever séries temporais com intervalos de tempo de duração não conhecidos, sua estrutura contém quatro redes neurais e blocos de memória chamado células, que retem a informação e as manipulações são feitas pelos gates : Forget Gate, Input Gate e Output Gate.

Forget Gate : Remove as informações que não são mais úteis para a célula, este gate é alimentado pela saída da célula anterior e a entrada no momento t. Estas são multiplicadas por suas matrizes de peso e adicionam o bias. O resultado passa por uma função de ativação que resulta em 0 para esquecer a informação e 1 para reter esta informação.

Input Gate : Este Gate contém as informações úteis para a célula. A informação que chega é regulada por uma função sigmoide, seguido da criação de um vetor usando a função tanh, retornando valores de -1 a 1. Estes valores são multiplicados pelos valores regulados para obter as informações necessárias do modelo.

Output Gate : Após receber as informações úteis do Input Gate, o Output Gate gera um vetor e aplica a função tanh na célula, seguido da função sigmoide que filtra os valores a ser lembrados. o vetor e os valores regulados se multiplicam e saem da célula, sendo entrada da célula seguinte.



5. Avaliação da solução

Após a implementação do modelo e o processamento dos dados, foi a hora de prever o conjunto de teste e avaliar a sua acurácia. A técnica de avaliação utilizada foi a mesma que será aplicada no resultado da proposta, o RMSE(Root Mean Square Error).

Buscando abordagens do LSTM em publicações de outros Cientistas de Dados, reparei que as mais comuns divisões dos conjuntos de treino e teste são : 67% treino e 33% teste ou 80% treino e 20% teste. Após fazer o fit do modelo com as 2 abordagens, a segunda aprsentou um RMSE menor.O resultado do fit foi um RMSE de aproximadamente 26.1.

Outros hiperparâmetros também foram essenciais para a eficiência do LSTM, as séries temporais foram particionadas de 90 em 90 para cada, partições utilizadas para prever o dado seguinte.

A seleção dos atributos para o treinamento do LSTM foram escolhidos respeitando a capacidade e o desempenho do computador utilizado. Para a previsão de cada dia futuro, foram utilizadas 100 epochs, particionando os dados em batches de 32 valores, levando cerca de 4 horas para realizar todas as iterações.

Aplicando uma algoritmo para descobrir os valores nas médias móveis, temos a previsão do modelo dos próximos 30 dias no preço da Polpa Cítrica.

332.67, 377.72 ,321.76 ,360.78 ,378.63 ,374.39 ,393.95 ,339.0 ,366.1 ,348.37 ,345.74 ,329.8 ,321.03 ,303.34 ,337.51 ,340.46 ,275.09 ,276.14 ,321.82 ,311.74 ,328.62 ,352.07 ,296.37 ,322.63 ,340.06 ,350.18 ,319.96 ,321.04 ,359.23 ,355.58

Referências

Indicador do Milho Esalq/BMFBovespa
`https://https://www.cepea.esalq.usp.br/br/indicador/milho.aspx`

Arquitetura de Redes Neurais Long Short Term Memory (LSTM).Deep Learning Book

http://deeplearningbook.com.br/
arquitetura-de-redes-neurais-long-short-term-memory

Machine Learning Mastery. How to Develop LSTM Models for Time Series Forecasting
https://machinelearningmastery.com/
how-to-develop-lstm-models-for-time-series-forecasting/

Multi-Step LSTM Time Series Forecasting Models for Power Usage
https://machinelearningmastery.com/
how-to-develop-lstm-models-for-multi-step-time-series-forecasting-of-household-power-