MECAI 2024 III Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Avaliação de Técnicas de Aprendizado Estatístico no Contexto de Previsão de Índices de Preços de Commodities

Larissa Carolina Correa Neves [1] Adriano Kamimura Suzuki[2]

1 Introdução

A Ciência de Dados é, em resumo, um campo multidisciplinar amparado por conceitos estatísticos, computacionais e matemáticos que busca a extração de conhecimento e entendimentos de padrões a partir dos dados, sendo estes das mais diversas fontes e áreas do conhecimento. Nos tempos atuais, a capacidade dos computadores transformou a Ciência de Dados em uma das mais promissoras em muitos setores de negócio [2]. Neste sentido, recentes avanços tecnológicos têm permitido uma verdadeira revolução na forma como áreas do conhecimento e empresas evoluem seus produtos. No caso do setor financeiro, o valor desse avanço está, não só na melhora do processo de tomada de decisão, como também no controle do risco envolvendo essas escolhas.

A predição de índices de preço no mercado de commodities é um importante item não só no planejamento das empresas produtoras da matéria ou que fabricam produtos a partir desse bem, mas também nas projeções macroeconomicas de uma nação, impactando intituições financeiras governamentais e do setor privado. Sua importância permeia todas as áreas do conhecimento, pois as previsões podem ser usadas como base para a revisão/implementação de políticas públicas, desenvolvimento de planejamento estratégico nas empresas e decisões no mundo corporativo [4].

Tradicionalmente, métodos estatísticos paramétricos são utilizados para tal tarefa, considerando suposições sobre a distribuição dos dados e relacionamento dos mesmos. Utilizar tais métodos nesse processo de previsão se mostrou elevante ao longo dos anos, contudo, atualmente com o avanço das técnicas de aprendizado novas estratégias podem enriquecer o conjunto de técnicas a serem pensadas para uma tarefa desse tipo.

Diante do cenário por ora construído, este projeto de pesquisa propõe a investigação por meio da experimentação de técnicas tradicionais estatístico em séries temporais como modelos autoregressivos e técnicas de aprendizado de máquinas (métodos ensemble - como *Bagging* e *Gradient Boosting*) para o problema em questão.

[1] larissaneves@usp.br
[2] suzuki@icmc.usp.br

2 Referencial Teórico

Tradicionalmente, modelos estatísticos são utilizados para predições em dados de característica temporal, para além disso o modelo ARIMA (modelo autoregressivo de médidas móveis) é frequentemente aplicado como referência em tarefas de previsão de preço [3]. Apesar de úteis, tais modelos carregam uma série de suposições sobre os dados, como estacionariedade, distribuição paramétrica dos resíduos e relacionamento linear. Em muitos casos reais, principalmente se tratando de séries de preços de commodities que apresentam características cíclicas, tais suposições podem ser inalcansáveis [3].

Nos últimos anos, pesquisas envolvendo aplicação de técnicas de aprendizado têm recebido destaque nesse cenário, permitindo incorporar covariáveis associadas a variável alvo, bem como informações econômicas que possam impactar no índice de preço estudado. Nesse sentido, surgem como candidatos uma variedade de modelos capazes de lidar com o problema em questão, podendo esses modelos serem paramétricos ou não paramétricos. A escolha sobre qual desses modelos utilizar passa pela adequação dos dados às suposições estabelecidas por cada um deles. De forma geral, técnicas não paramétricas oferecem maior liberdade e têm o potencial de apresentar melhor desempenho, caso os dados não se adéquem às suposições do modelo paramétrico.

Em [3] um estudo que tinha por objetivo a comparação de métodos ensembles e tradicionais para a predição de preços de petróleo bruto no setor de energia foi consuzido. Os resultados mostraram que tais métodos apresentaram melhor performance quando comparados a metodologias autoregressivas.

Outro estudo [4] focou na avaliação da performance de métodos ensemble como Floresta aleatória e *Gradient Boosting* e o modelo *Support Vector Regression* para a predição de preço das commodities agrícolas de soja e milho. O estudo foi feito considerando diferentes horizontes de previsão. De forma geral, os resultados se mostraram favoráveis a utilização de tais métodos quando comparados a performance de módelos únicos.

As vantagens da combinação de modelos para a previsão de séries temporais podem ser creditadas a três fatores:(i) aumenta em grande medida a acurácia global das previsões através da utilização de técnicas de agregação adequadas, (ii) existe frequentemente uma grande incerteza sobre o modelo de previsão ótimo e, em e, em tais situações, as estratégias de combinação são as alternativas mais adequadas mais adequadas, e (iii) a combinação de múltiplas previsões pode reduzir de forma eficaz os erros [5].

3 Método

O processo de trabalho irá compreender cinco etapas de forma geral, sendo elas: Obtenção dos dados, Entendimento dos dados, Preparação dos dados, Modelagem e Avaliação.

3.1 Geração/Obtenção dos dados

No momento, os dados a serem trabalhados ao longo do desenvolvimento deste projeto não está definido. Tem sido feito uma pesquisa entre portais públicos de dados para definição do setor de commoditie a ser estudado, podendo ser Energia, Metais, Agricultura ou Pecuária. O portal do CEPEA/Esalq-USP (Centro de Estudos Avançados em Economia Aplicada - Escola Superior de

Agricultura Luiz de Queiroz/Universidade de São Paulo), disponibiliza dados de índices de preço no setor agrícula que podem ser consultados e considerados para esse projeto.

3.2 Entendimento dos dados

Essa etapa compreende o contato inicial com a matéria prima da solução, os dados. Aqui serão feitas análises exploratórias, a fim de obter insights e formular hipóteses, bem como a análise da qualidade dos dados, com o objetivo de levantar as etapas de transformações necessárias para tratamento dos mesmos.

Outro aspecto importante dessa etapa é a observação de características dos dados que permitam direcionamento para a escolhas das técnicas a serem utilizadas, como a dimensionalidade e complexidade dos dados.

3.3 Preparação dos dados

A etapa de preparação dos dados irá compreender os tratamentos necessários para adequação dos dados às técnicas de modelagem a serem testadas, pode incluir imputação de dados faltantes, tratamento de escala para variáveis assimétricas, adequação de variáveis categóricas para *input* do modelo, geração de novos atributos a partir dos já existentes e etc.

Um aspecto importante dessa etapa é a adequação do dado temporal ao método de modelagem que não visualiza o dado como uma série no tempo. Dessa forma, a criação de lags e features temporais são necessárias.

3.4 Modelagem

A etapa de modelagem, irá compreender o ajuste das técnicas aos dados em questão. Dentro do escopo deste projeto, as primeiras técnicas a serem estudadas são modelos estatísticos tradicionais como ARIMA (descrito em (1)) e a evolução do estudo consiste em comparar os desempenho dessas técnicas com métodos atuais de aprendizado de máquinas, como métodos ensembles.

$$X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \ldots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \qquad (1)$$

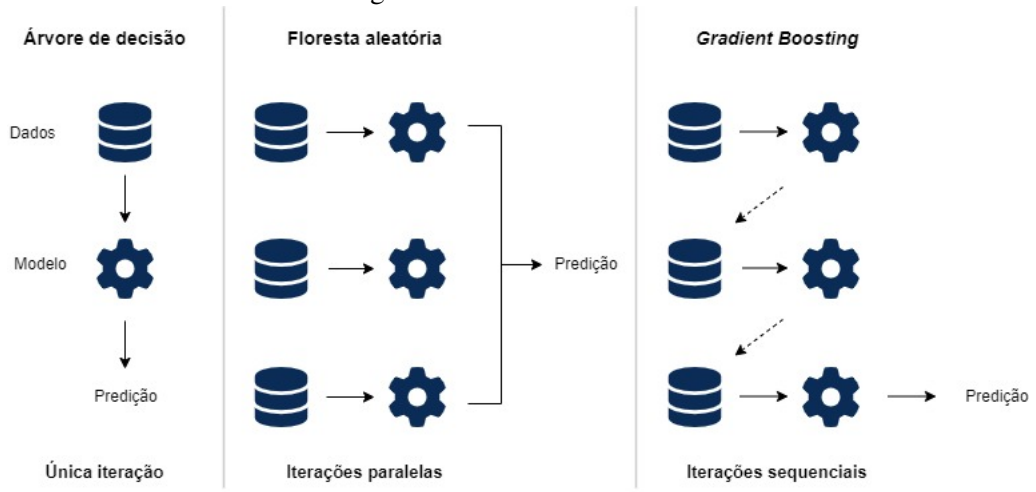
3.5 Métodos Ensemble

Os métodos ensemble são abordagens que combinam vários modelos ”simples”, muitas vezes chamados de modelos fracos, de forma a obter um único modelo com maior poder de predição [1].

Na **Floresta aleatória** modelos de árvore de decisão são combinados de forma não correlacionada e paralela. A cada iteração do algoritmo, uma amostra com reposição é retirada do conjunto de dados, bem como uma amostra de covariáveis e então uma árvore de decisão é ajustada a partir das amostras selecionadas. Ao final, no caso de resposta numérica, o valor médio dos valores preditos de cada árvore ajustada é considerada como a predição final.

O ponto em considerar apenas uma subamostra das variáveis preditoras está em superar o problema de ajustar vários modelos de árvores altamente correlacionados, uma vez que em situações que alguns preditores tenham forte relacionamento com a resposta, o fato de ajustar modelos de

Figura 1: Métodos *Ensemble*.



Fonte: Imagem criada pela autora.

árvores utilizando o mesmo conjunto de covariáveis levaria a pouca ou nenhuma mudança entre as predições de cada um, já que esses preditores altamente relacionados a resposta teriam uma importância maior em todos os modelos ajustados [1].

Nos algoritmos **Gradient *Boosting***, modelos com menor poder de predição (frequentemente árvore de decisão) são combinados em série de maneira que o modelo aprenda de forma gradual. A cada iteração do algoritmo, os resíduos das predições anteriores são computados de forma a agregar informação ao próximo modelo ajustado, melhorando de forma sucessiva as predições até a predição final. Considerando o resíduo do modelo anterior como *input* para o próximo melhoramos o desempenho em áreas que o modelo não teve bom desempenho [1]. A figura 1 mostra como os métodos de floresta aleatória e *gradient boosting* diferem na forma de treinamento.

3.6 Avaliação

Uma vez ajustados os diferentes modelos, métricas de performance baseadas na distância entre o valor predito pelo modelo ($\hat{y}_i$) e o valor real (y_i) observado serão utilizadas, sendo duas principais o Erro percentual absoluto médio (2) e o Erro quadrático médio (3).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{2}$$

$$RSME = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Além da qualidade do ajuste através das métricas de desempenho, outros aspectos podem ser avaliados como critérios para a escolha do melhor modelo, entre eles a sensibilidade do modelo aos dados de treinamento, viés e a dimensão do conjunto de treinamento. No primeiro aspecto,

essa avaliação passa por experimentações a partir de indução de perturbações no conjunto de treinamento e a avaliação do impacto desses ruídos na variabilidade dos coeficientes e predições do modelo. Além disso métricas de negócio podem ser consideradas no contexto empresarial.

Referências

[1] G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer New York, 2014.

[2] H. Albrecher, A. Bommier, D. Filipovi´c, P. Koch-Medina, S. Loisel and H. Schmeise. Insurance: Models, digitalization, and data science, European Actuarial Journal, 2019.

[3] YU, L.; DAI, W.; TANG, L. A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. Engineering Applications of Artificial Intelligence, v. 47, p. 110–121, 1 jan. 2016.

[4] Ribeiro M. H. D. M.; Dos Santos Coelho, L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing Journal, v. 86, 1 jan. 2020.

[5] Allende, H. e Valle, C. Ensemble Methods for Time Series Forecasting. In: Seising, R., Allende-Cid, H. (eds) Claudio Moraga: A Passion for Multi-Valued Logic and Soft Computing. Studies in Fuzziness and Soft Computing, vol 349. Springer, Cham. 2017. https://doi.org/10.1007/978-3-319-48317-7-13