

Desvendando o iBovespa: Modelagem Preditiva de Dados

**Uma Jornada Através da Análise Preditiva na
Bolsa de Valores de SP**

Para profissionais de dados e líderes de negócio, esta apresentação aborda um projeto de modelagem preditiva, desde a concepção até a interpretação dos seus resultados.

Olá, prazer! Sou o Breno

Com uma formação interdisciplinar em Economia (FEA) e Jornalismo (ECA) pela Universidade de São Paulo (USP), atuo atualmente como Assistente de Negócios no Banco do Brasil.

Sou apaixonado por tecnologia e por desvendar padrões e histórias por trás dos números, especialmente no dinâmico mercado financeiro. Este projeto reverbera com meu interesse e compromisso em aplicar a análise de dados para gerar insights valiosos e previsões estratégicas.

1. Introdução ao Projeto

Contexto do Problema

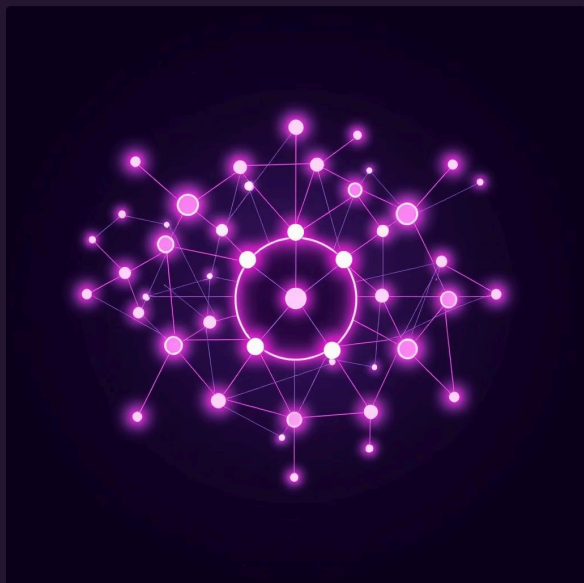
Em um mercado volátil, como é o Mercado de Ações no Brasil, a previsão da variação do iBovespa é crucial para otimizar resultados. Variações inesperadas geram perdas. Assim, a um analista de dados é passada a Tarefa fundamental de predição do índice Bovespa diário.

Objetivo do Projeto

Desenvolver um modelo preditivo para prever os movimentos do iBovespa, com **75% de acurácia**, tendo uma base de dados de pelo menos dois anos, e uma base de teste de 30 dias.

- Predição de alta (↑) ou de baixa (↓) no iBovespa

2. Aquisição e Exploração dos Dados



Fontes de Dados e Variáveis Chave

Os dados foram coletados do site Investing, integrando informações históricas do iBovespa.

- Requisito mínimo: dois anos de dados
- Serie de tempo escolhida a partir de 2022, pós pandemia

Variáveis contidas no Banco de Dados

- **Data:** Data referenciada dos dados da linha
- **Ultimo:** Valor de fechamento iBovespa daquele dia.
- **Abertura:** Valor de abertura iBovespa daquele dia.
- **Volume:** volume de ações transacionadas naquele dia.
- **Maxima e Mínima:** valores de max e mín daquele dia.
- **Var%:** percentual de variação entre a abertura e o fechamento.

3. Engenharia de Atributos



Lagged - D1 e D2

As colunas `Último_Lag1` e `Último_Lag2` foram adicionadas para incluir os preços de fechamento dos dois dias anteriores como preditores. Isso permite que os modelos considerem o histórico recente do preço ao fazer previsões.



Janela Deslizante

Este é outro indicador comum em análise de séries temporais que suaviza as flutuações de preço e ajuda a identificar tendências, incorporando informações de uma "janela deslizante" de 5 dias



Divisão Temporal

Garante que o modelo seja avaliado em dados futuros que ele não "viu" durante o treinamento, simulando um cenário de previsão realista e evitando *look-ahead bias*. Features baseadas em tempo como `Year`, `Month`, `DayOfWeek`, e `DayOfYear` também ajudam os modelos a capturar padrões sazonais ou cíclicos nos dados.

4. Preparação da Base para Previsão

Estratégia de Divisão




A serie temporal foi dividida ente antes de depois de aprox 30 dias. A data estabelecida foi **26 de Julho de 2025**.

- **Target:** Variação Diária
- **Janela de Tempo Móvel:** Treino com dados a partir de 22 - pós pandemia e teste com o mês subsequente. Essa abordagem simula o cenário real de previsão contínua.

Tratamento de Dados e Codificação

- **Padrão pt-br:** Adaptação da base de dados original em .CSV para leitura python - valores B e M /
- **Dados Faltantes:** Tratamento dos dados faltantes com o comando `dropna`.
- **Treino / Validação / Teste**

5. Escolha do Modelo e Justificativa Técnica

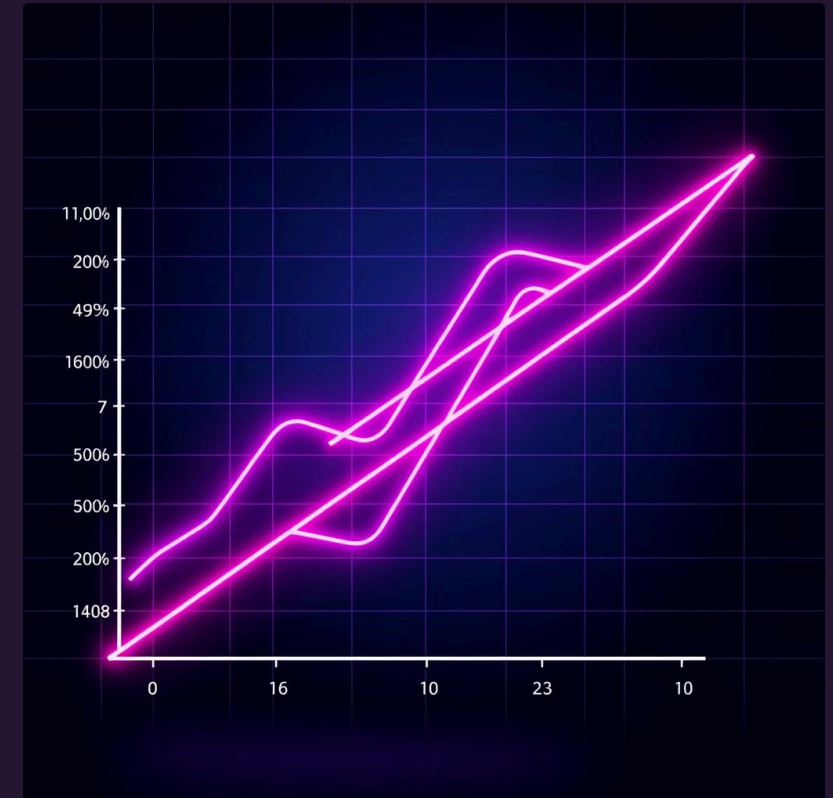
 98	 63	 0
XGBoost (Escolha) <p>É um modelo baseado em árvores que combina previsões de múltiplas árvores de decisão sequencialmente para melhorar a acurácia. Escolhido pela performance em tarefas de regressão, como foi o caso, demonstrou capacidade de lidar com dados tabulares e robustez a outliers</p>	Random Forest <p>Também um modelo baseado em árvores, escolhido como um bom ponto de comparação para o XGBoost. Ele constrói múltiplas árvores de decisão independentemente e agrega suas previsões, o que ajuda a reduzir o overfitting em comparação com árvores de decisão individuais</p>	Regressão Linear <p>Incluído como um modelo de linha de base simples para comparar a complexidade e performance dos modelos baseados em árvores. Embora menos capaz de capturar relações não-lineares, é rápido de treinar e interpretar</p>

6. Resultados e Métricas

O modelo XGBoost demonstrou a melhor performance comparada.

RMSE (Erro Quadrático Médio da Raiz)	3577.70
MSE (Erro Absoluto Médio)	12799934.10
R ² (Coeficiente de Determinação)	0.98

O **R² de 0.98** indica que o modelo explica 98% da variância na demanda, superando o benchmark de 0.75 definido como target de objetivo para o trabalho.



Matriz de Confusão e Desempenho do Modelo

Avaliamos o desempenho do modelo no período de teste utilizando a matriz de confusão, uma ferramenta essencial para entender a acurácia das classificações binárias (alta ou baixa do iBovespa).

A matriz revela que o modelo obteve um total de **16 acertos (7 + 9)** nas previsões para o período de teste. Isso inclui 7 previsões corretas de alta (verdadeiros positivos) e 9 previsões corretas de baixa (verdadeiros negativos). Em contrapartida, houve apenas **3 erros (2 + 1)**, demonstrando a capacidade do modelo em capturar a direção do mercado com alta precisão.

Real: Alta	7	2
Real: Baixa	1	9

Estes resultados confirmam que o modelo está desempenhando **boas previsões**, validando sua aplicação para análise de tendências do iBovespa.

Conclusões e Próximos Passos

Confiabilidade & Trade-offs

O modelo preditivo de demanda baseado em XGBoost demonstrou **alta confiabilidade e acurácia** (R^2 de 0.98), superando os objetivos iniciais do projeto. É um tipo de modelo poderoso, e pode capturar relações complexas. No entanto, podem ser mais propensos ao overfitting e seus hiper-parâmetros precisam de ajustes com frequência. Um modelo com alta acuracidade no treinamento mas baixa acuracidade no teste é um sinal clássico de overfitting

Próximas Etapas

- **Implementação:** Integrar o modelo aos sistemas de gestão de investimentos e BI para uso em tempo real e otimização do resultado.
- **Monitoramento:** Estabelecer um dashboard de acompanhamento de performance do modelo, com alertas para desvios.
- **Melhorias:** Explorar a inclusão de dados de redes sociais e clima, além de técnicas de Deep Learning para casos mais específicos e movimentos de manada.



O Futuro já é Agora.

**Decisões Orientadas por Dados,
Resultados Extraordinários.**

Agradecemos sua atenção. Perguntas?