Tech Challenge - Fase 2: Modelo Preditivo para o IBOVESPA

1. Introdução

O documento visa apresentar a solução desenvolvida pelo grupo, para concluir o desafio proposto na segunda fase do *Tech Challenge*. A tarefa proposta foi de desenvolver um modelo de *machine learning* capaz de prever se o IBOVESPA (Índice Bovespa) iria subir ou descer no dia seguinte, com acurácia mínima de 75%.

O IBOVESPA é o principal indicador de desempenho do mercado de ações brasileiro. Ele representa uma carteira teórica composta pelas ações mais negociadas da B3 (Bolsa de Valores do Brasil) e serve como termômetro do comportamento do mercado. Dessa forma, prever sua variação diária é uma tarefa desafiadora, mas com grande valor estratégico, sobretudo para investidores e analistas financeiros.

2. Aquisição e Exploração dos Dados

A base de dados foi composta por registros históricos do IBOVESPA [1] , abrangendo um período de 10 anos (2015-2025), e esse intervalo foi escolhido para que houvesse uma grande quantidade de dados recentes, que representassem melhor o cenário atual da bolsa de valores brasileira.

Foram utilizadas variáveis como preços de abertura, fechamento, máxima, mínima, volume negociado e variação percentual diária. Durante o pré-processamento, os dados foram convertidos, normalizados e ordenados cronologicamente. Também foram criadas colunas derivadas como indicadores técnicos e *LAGs*.

Além do processo de tratamento dos dados, também foi realizado uma visualização de dados voltada para entender o funcionamento e a importância de cada variável, a correlação entre elas, e o impacto que cada uma poderia causar no modelo preditivo, seja de forma positiva ou negativa.

[1] https://br.investing.com/indices/bovespa-historical-data

Por se tratar de uma série temporal, a natureza sequencial dos dados foi respeitada em todas as etapas do *pipeline*. O *dataset* foi ordenado cronologicamente do mais antigo para o mais recente, garantindo que o modelo fosse treinado apenas com informações anteriores aos dados de teste. Além disso, foram criadas *features* derivadas baseadas em janelas temporais, como *LAGs* (valores defasados de um ou mais dias), *SMA* (médias móveis simples), *EMA* (médias móveis exponenciais), *Momentum* e *RSI*. Essas técnicas permitiram ao modelo captar padrões temporais recorrentes sem vazar informações futuras, mantendo a integridade do processo de modelagem preditiva.

3. Engenharia de Atributos

A engenharia de atributos foi essencial para extrair padrões temporais e comportamentais do mercado a partir das variáveis brutas. As técnicas adotadas foram escolhidas com base em sua ampla aplicação em análises financeiras e sua capacidade de capturar sinais úteis para previsão de tendências. As *features* escolhidas foram:

- *LAG* (defasagens): permite ao modelo considerar o comportamento do ativo nos dias anteriores, capturando possíveis efeitos de continuidade ou reversão de tendência.
- *Momentum*: avalia a força do movimento recente dos preços. Um valor positivo indica tendência de alta sustentada, enquanto um valor negativo aponta possível enfraquecimento.
- *SMA* (Média Móvel Simples): suaviza oscilações de curto prazo e ajuda a identificar tendências gerais de mercado em diferentes horizontes.
- *EMA* (Média Móvel Exponencial): semelhante à SMA, mas atribui maior peso aos dados recentes, tornando-a mais responsiva a mudanças rápidas no comportamento do mercado.
- *RSI* (Índice de Força Relativa): mede a velocidade e a mudança dos movimentos de preços, sendo útil para detectar condições de sobrecompra ou sobrevenda.
- Amplitude: reflete a volatilidade intradiária, capturando a diferença entre os valores máximos e mínimos do dia, o que pode indicar instabilidade ou oportunidades de reversão.

Essas features combinam informações de curto e médio prazo, enriquecendo a base de dados e aumentando a capacidade do modelo de identificar padrões preditivos com mais precisão.

4. Modelos Testados

Foram avaliados dois algoritmos principais: Regressão Logística e *Random Forest*. Ambos se mostraram adequados para o tipo de problema proposto de classificação binária, ou seja, distinguir se a bolsa subiu ou desceu no dia seguinte com base em dados históricos.

- Regressão Logística: é um modelo estatístico simples, de fácil implementação e rápida execução, ideal para problemas de classificação com variáveis contínuas e binárias. Foi escolhido inicialmente devido à sua praticidade de manipulação, sendo uma boa base de comparação para modelos mais complexos.
- *Random Forest*: é um modelo de ensemble baseado em múltiplas árvores de decisão. Sua principal vantagem está na robustez contra o overfitting e na capacidade de lidar com relações não lineares entre os atributos. Além disso, é menos sensível a outliers e permite avaliar a importância das variáveis no processo decisório.

A *Regressão Logística* foi o modelo final escolhido, pois obteve melhor desempenho ao lidar com os indicadores técnicos derivados. Após ajustes de hiperparâmetros e ampliação da base de dados para 10 anos de histórico, atingiu-se uma acurácia de 77,2%, superando a meta estabelecida.

5. Conclusão

O desenvolvimento deste projeto representou um desafio técnico e analítico importante, ao exigir a construção de um modelo de machine learning capaz de lidar com dados financeiros reais e históricos. Ao longo do processo, foi possível aplicar conceitos fundamentais de ciência de dados, como engenharia de atributos, tratamento de séries temporais e avaliação de modelos preditivos.

Por meio de um processo iterativo de testes e validações, a solução final baseada no algoritmo de *Regressão Logística* superou a acurácia mínima exigida, atingindo 77,2%. Esse resultado demonstra não apenas a viabilidade da abordagem adotada, mas também o potencial da aplicação de técnicas de aprendizado de máquina

na análise de mercado financeiro.

Além do resultado técnico, o projeto proporcionou um aprendizado sólido sobre modelagem preditiva, decisões de design de features e o impacto de diferentes abordagens sobre a performance. A experiência reforça a importância de respeitar o contexto dos dados, realizar escolhas embasadas e equilibrar desempenho com generalização.