No dinâmico e imprevisível mercado financeiro, a capacidade de antecipar a direção do Ibovespa, mesmo que com uma pequena vantagem estatística, pode significar a diferença entre lucro e prejuízo. Diante deste cenário, assumimos a missão de um analista quantitativo: seria possível, utilizando 20 anos de dados históricos e técnicas avançadas de ciência de dados, construir um modelo preditivo confiável para prever a tendência de alta ou baixa do dia seguinte?

O desafio era claro: o modelo final precisaria não apenas funcionar, mas atingir uma acurácia mínima de 75% em um rigoroso teste com os 30 dias mais recentes, provando sua relevância no cenário atual.

Nesta apresentação, vamos conduzi-los por nossa jornada técnica, desde a aquisição e exploração dos dados brutos, passando pela complexa engenharia de atributos para 'separar o sinal do ruído', passando pela otimização final. O objetivo é demonstrar não apenas o resultado, mas o processo rigoroso que garante que a performance do nosso modelo vencedor é legítima, confiável e não fruto do acaso.

Aquisição e Ajustes:

Para construir um modelo preditivo verdadeiramente robusto, nossa primeira decisão estratégica foi ir além do requisito mínimo de 2 anos e utilizar uma base de dados histórica de **20 anos** do Índice Ibovespa. A justificativa para essa escolha é que um período mais longo nos permite capturar uma gama muito maior de "regimes de mercado", desde períodos de crise e alta volatilidade até mercados em forte tendência de alta, tornando o modelo mais resiliente.

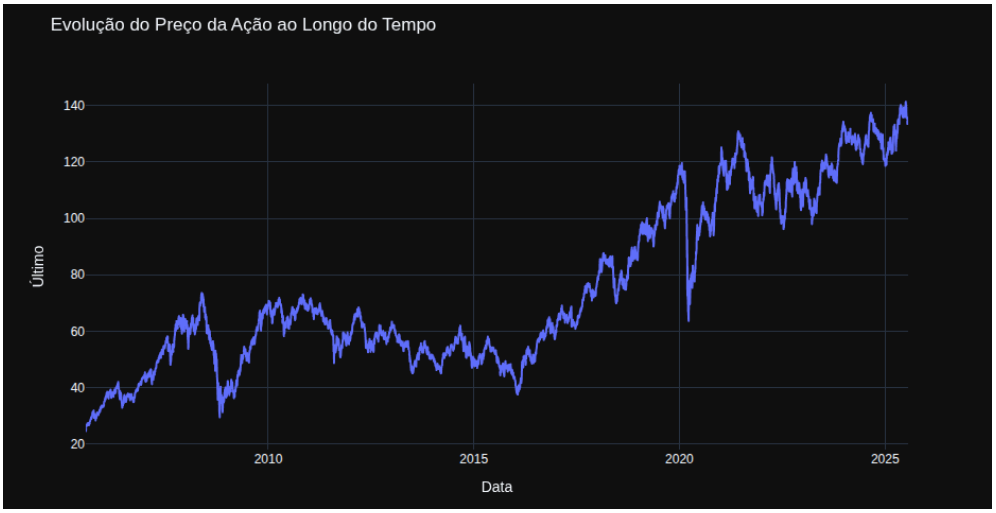
Naturalmente, um histórico tão longo introduz o desafio do "ruído" de informações mais antigas. Endereçamos essa questão diretamente na etapa de Engenharia de Atributos, onde, como demonstraremos, criamos features que focam na relevância da informação recente para a tomada de decisão do modelo.

Os dados foram adquiridos da plataforma Investing.com e, no pré-processamento inicial, realizamos três ajustes essenciais para garantir a qualidade da análise:

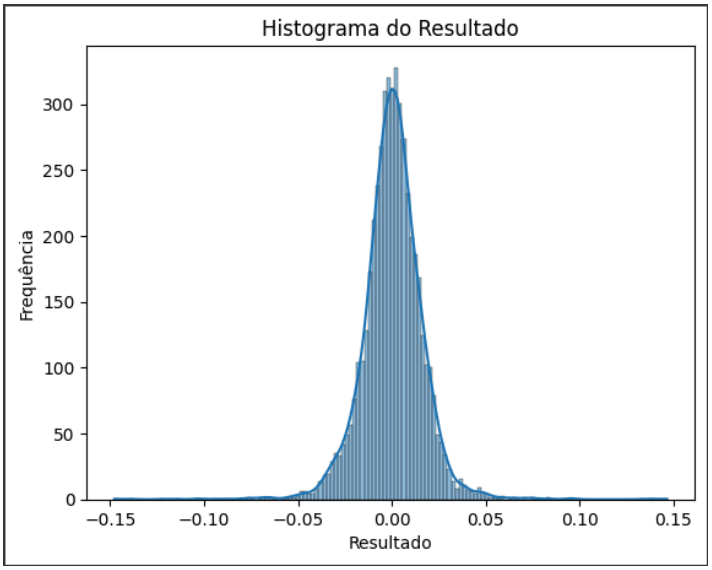
- **Correção de Tipos:** A coluna 'Data' foi convertida de texto para o formato `datetime`, fundamental para as análises temporais.
- **Padronização de Volume:** A coluna 'Vol.' foi transformada de texto (ex: '1.5M') para um formato numérico padronizado.
- **Conversão de Variação:** A coluna 'Var%' foi convertida de string (ex: '0,75%') para um valor numérico de ponto flutuante (ex: 0.0075) para permitir cálculos.

Exploração dos Dados

- **Evolução do Preço da Ação ao longo do Tempo**



- **Distribuição do Resultado**



Um dos primeiros passos da nossa exploração foi entender a 'personalidade' dos retornos diários do Ibovespa. A teoria estatística clássica sugere que eventos aleatórios deveriam seguir uma 'Distribuição Normal', a famosa Curva do Sino, onde eventos extremos são incrivelmente raros.

No entanto, como podemos ver no histograma, a realidade do mercado é diferente. Nossa análise revelou duas características cruciais:

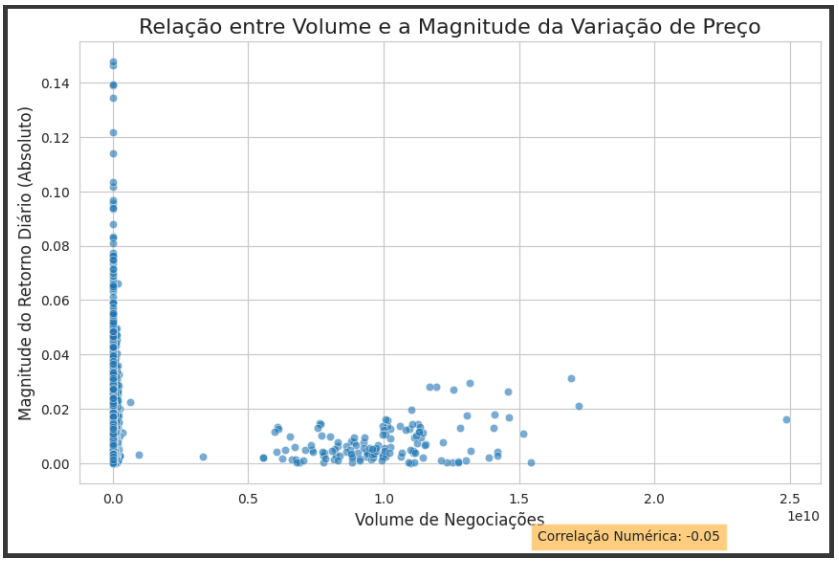
- **Pico Alto no Centro:** "A distribuição é muito mais 'pontuda' no centro do que o normal. Isso nos diz que, na grande maioria dos dias, o mercado se move muito pouco. São dias de 'calmaria', com retornos pequenos e quase irrelevantes, o que pode criar uma falsa sensação de segurança."

- **"Caudas Gordas" (A Pista Mais Importante):** "As laterais do nosso gráfico, as 'caudas', são muito mais 'gordas' do que a Curva do Sino prevê. Esta é uma das descobertas mais importantes em finanças. As caudas gordas significam que **dias de movimentos extremos**, quedas ou altas muito fortes, como as que vimos em crises, **são muito mais comuns no mundo real do que a teoria 'normal' nos faria acreditar.**

Qual a implicação disso para o nosso modelo?

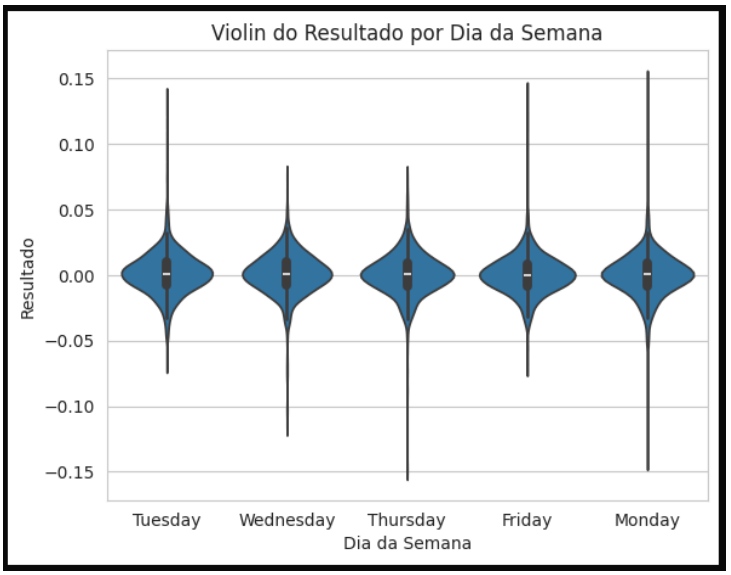
Essa descoberta prova que não podemos tratar o mercado como um sistema de eventos 'normais'. O risco de um grande movimento está sempre presente. Isso justifica a necessidade de construir um modelo robusto, que não seja apenas treinado na 'calmaria' do dia a dia, mas que também aprenda com os eventos extremos contidos nas caudas gordas da distribuição. Foi essa premissa que guiou nossa escolha de features e a validação rigorosa do nosso modelo final.

- **Correlação Volume x Magnitude do Retorno**



O coeficiente de correlação de -0.05, calculado sobre uma robusta base de 20 anos, demonstra a ausência de uma relação linear estável entre volume e magnitude do retorno para o Ibovespa. Essa neutralidade é justificada pela agregação de múltiplos e distintos 'regimes de mercado' (crises, euforias, períodos de lateralização), cujos comportamentos individuais se anulam na média de longo prazo. Isso reforça a tese de que a relação entre essas variáveis é fundamentalmente não-linear e dependente do contexto, validando a escolha de modelos avançados capazes de capturar tais complexidades.

- **Padrões Semanais**



Nossa análise por meio deste violin plot, revelou que não existe um 'efeito dia da semana' claro e direto sobre a direção média dos retornos do Ibovespa; o comportamento típico é muito similar em todos os dias. No entanto, observamos uma sutil indicação de que a volatilidade e a ocorrência de eventos extremos podem ser ligeiramente maiores no final da semana, especialmente às sextas-feiras. Por essa razão, decidimos manter a feature 'Dia da Semana' em nosso modelo, pois ela pode conter informações valiosas sobre risco e interagir com outros indicadores para aprimorar a capacidade preditiva geral.

Engenharia de Atributos

Nossa análise revelou que os dados brutos, embora informativos, são muito 'ruidosos' e voláteis. Para que um modelo preditivo tenha sucesso, ele não pode olhar apenas para o preço de fechamento; ele precisa entender o **contexto** por trás dos números.

A etapa de Engenharia de Atributos foi, portanto, a fase mais estratégica do projeto. O objetivo foi traduzir os conceitos de análise técnica e a dinâmica do mercado em variáveis numéricas que o modelo pudesse interpretar.

Para isso, construímos um conjunto de "pistas" inteligentes, divididas em quatro categorias principais: **Memória** (com os Lags de Retorno), **Momentum** (com indicadores como RSI), **Risco** (com a Volatilidade). A seguir, detalharemos a criação e a importância de cada uma dessas features.

Memória - Lags de Retorno

O Problema - Um Modelo Sem Memória

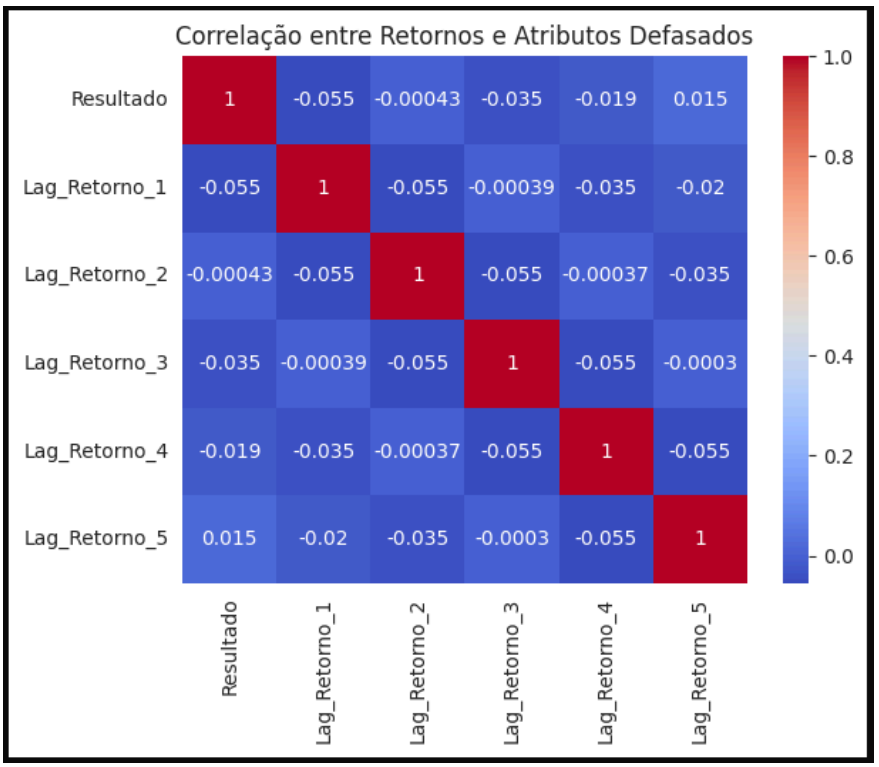
- Um modelo de machine learning não entende a passagem do tempo. Ele olha para cada dia de negociação como uma fotografia isolada, sem saber o que aconteceu na foto anterior. Para que nosso modelo pudesse tomar decisões inteligentes, nossa primeira tarefa foi dar a ele o superpoder da **memória**.

A Solução - O que é o Lag

- Para isso, criamos as **Features de Lag**. Através da função .shift(), nós trouxemos para a linha de 'hoje' a informação do que aconteceu 'ontem'. A feature Lag_Retorno_1, por exemplo, contém o valor exato do retorno do dia anterior. Fizemos isso para os últimos 5 dias (Lag_Retorno_1 a Lag_Retorno_5), criando um painel completo do passado recente.

A Análise - O que o Modelo Pode Aprender com Isso?

- Com essa memória explícita, o modelo pôde investigar diretamente as hipóteses mais fundamentais do mercado:
    - **Existe Momentum?** Será que um dia de alta tende a ser seguido por outro dia de alta, como um carro que continua embalado?
    - **Existe Reversão à Média?** Ou será que o oposto é verdadeiro, e um dia de alta tende a ser seguido por uma pequena queda, como um elástico que volta ao normal após ser esticado?
- Como a nossa análise demonstrou, essas features de memória se provaram as mais importantes para o modelo final, confirmando que o comportamento recente do mercado é, de fato, a pista mais valiosa para prever seu futuro de curto prazo.



Nossa análise de autocorrelação, feita através da matriz de correlação, revelou uma fraca tendência de reversão à média no prazo de 1 dia (correlação de -0.1). Isso sugere que o mercado tende a corrigir levemente os movimentos do dia anterior. No entanto, essa 'memória' é muito curta, pois a correlação se torna estatisticamente insignificante para retornos de 2 dias ou mais, indicando que a informação passada é rapidamente absorvida pelo preço.

Momentum - RSI

O Problema - Falta de Contexto sobre o Momentum

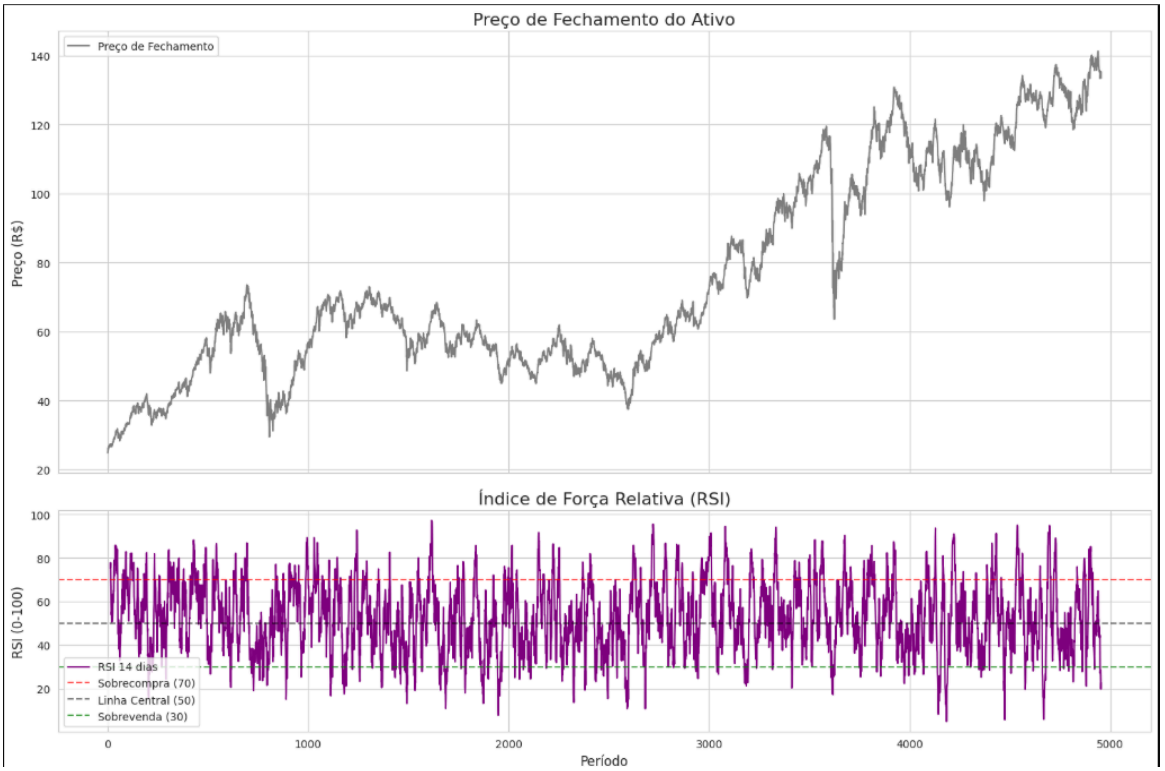
- "Saber o que aconteceu nos últimos dias (com as features de Lag) é crucial, mas não é o suficiente. Nosso modelo precisava entender a **'energia'** e a **'saúde'** do movimento atual. Afinal, um carro pode estar subindo uma ladeira, mas se o motor está superaquecendo e perdendo força, ele pode parar a qualquer momento. Precisávamos de um 'medidor de energia' para o mercado."

A Solução - O que é o RSI

- "Para isso, implementamos o **RSI (Índice de Força Relativa)**. Pense nele como um medidor de energia que vai de 0 a 100 e avalia a velocidade e a magnitude dos movimentos de preço. Ele nos diz quem está mais forte no momento: os compradores (que empurram o preço para cima) ou os vendedores (que empurram para baixo)."

A Análise - Como o Modelo Usa Essa Pista

- "O RSI nos deu dois sinais de alerta cruciais, que ensinaram nosso modelo a identificar possíveis pontos de virada no mercado:
    - **RSI > 70 (Zona de 'Sobrecompra'):** Isso funciona como um 'alerta de superaquecimento'. Indica que o ativo subiu com muita força e rapidez, e a 'energia compradora' pode estar se esgotando. É um sinal de que o movimento de alta pode estar exagerado e prestes a fazer uma pausa ou uma correção.
    - **RSI < 30 (Zona de 'Sobrevenda'):** Funciona como um 'alerta de exaustão de pânico'. Indica que o ativo caiu demais, muito rápido, e a 'energia vendedora' pode estar no fim. É um sinal de que a queda está perdendo força, abrindo espaço para uma recuperação."



- Ao adicionar o RSI, demos ao nosso modelo a capacidade de não apenas olhar para o passado (com os lags), mas também de avaliar a **sustentabilidade do movimento presente**, tornando suas decisões muito mais sofisticadas."

Risco

O Problema - Um Modelo "Cego" para o Risco

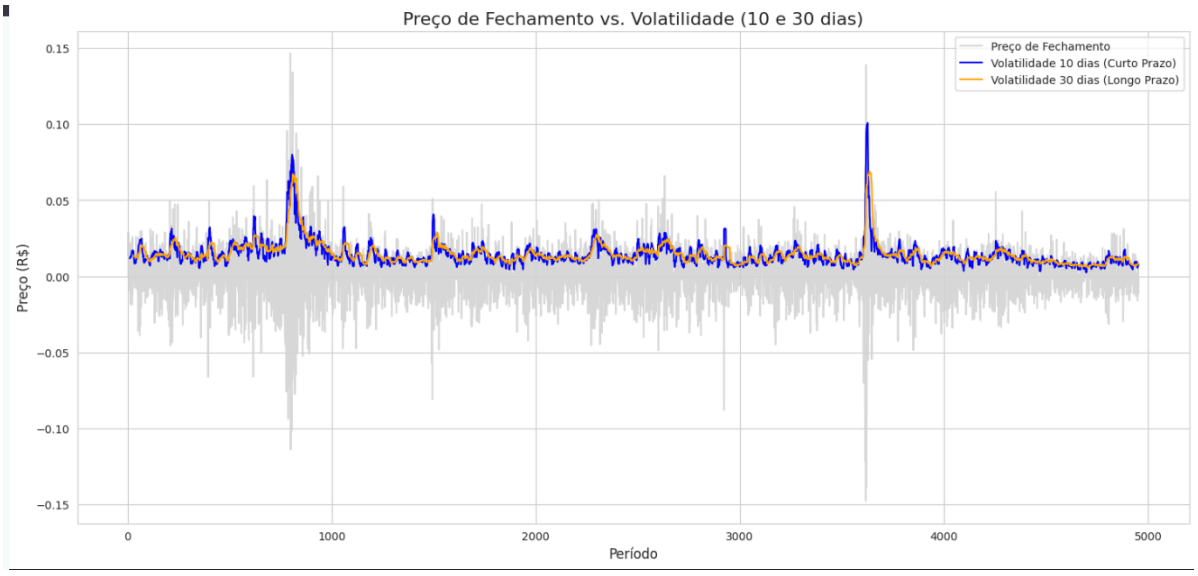
- "Nosso detetive-modelo já tinha 'memória' do passado (Lags) e um 'medidor de energia' para o presente (RSI). Mas ainda faltava uma percepção crucial: entender o **'clima'** ou o **'ambiente'** em que ele estava operando. Afinal, pilotar um carro em uma estrada reta e vazia em um dia de sol é muito diferente de pilotar na mesma estrada durante um nevoeiro denso e uma tempestade."

A Solução - O que é a Volatilidade

- "Para dar essa percepção de risco, criamos a feature de **Volatilidade**. Pense nela como o **'boletim meteorológico'** do mercado. Usando o desvio padrão dos retornos dos últimos 30 dias (`Volatility_30`), conseguimos medir se o 'mar' financeiro estava calmo ou agitado."

A Análise - Como o Modelo Usa Essa Pista

- "Essa feature ensinou nosso modelo a diferenciar dois cenários de mercado e a ajustar sua estratégia:
    - **Baixa Volatilidade (Mar Calmo):** Indica um mercado tranquilo e mais previsível, com os preços se movendo em passos menores. Nesses períodos, os sinais de tendência de outros indicadores (como as Médias Móveis) tendem a ser mais confiáveis.
    - **Alta Volatilidade (Tempestade):** Indica um mercado 'nervoso', dominado pelo medo ou pela euforia. Os preços dão saltos enormes e imprevisíveis para cima e para baixo. Em dias de 'tempestade', os padrões normais podem falhar, e o risco de uma mudança brusca de direção é muito maior."



- Com a feature de Volatilidade, nosso modelo aprendeu a ser mais **'cauteloso'**. Ele agora consegue ajustar sua confiança nas outras pistas de acordo com o 'clima' do mercado, dando mais peso aos sinais em dias calmos e, talvez, ignorando-os em meio a uma tempestade. Isso tornou suas previsões muito mais adaptáveis à realidade do risco."

Preparação da Base para Previsão: A Descoberta do "Target Inteligente"

A definição da variável alvo (`Target`) é o coração de qualquer modelo preditivo. É a pergunta que estamos ensinando o modelo a responder. Em nossa jornada, a evolução de como fizemos essa pergunta foi o ponto de virada para o sucesso do projeto.

A Abordagem Inicial: A Armadilha do Ruído

Inicialmente, definimos nosso `Target` da forma mais direta e intuitiva: o modelo deveria prever se o dia seguinte teria **qualquer movimento de alta**, por menor que fosse.

- **Lógica:** `Target = 1` se o retorno do dia seguinte for > 0%.
- **O Problema:** Logo percebemos que essa abordagem era uma armadilha. Movimentos muito pequenos (como +0.05% ou -0.05%) são, em grande parte, "ruído" aleatório do mercado. Pedir ao modelo para prever esse ruído era como pedir a um detetive para investigar cada barulho em uma cidade movimentada — a maioria era apenas o som de fundo, tornando quase impossível encontrar os sons que realmente importavam (o "sinal" de um padrão). Isso resultava em modelos com performance instável e acurácia medíocre, na faixa de 50-60%.

A Mudança de Estratégia: Focar nos Sinais, Ignorar o Ruído

A grande virada no projeto ocorreu quando mudamos a pergunta fundamental. Em vez de perguntar *"o mercado vai subir?"*, passamos a perguntar:

> **"O mercado vai subir de forma *significativa*?"**

O objetivo passou a ser ignorar o ruído e forçar o modelo a se concentrar apenas nos padrões que precedem dias de alta mais claros e relevantes.

O Novo Target e Seu Impacto Drástico

Para implementar essa nova estratégia, ajustamos a criação do `Target` estabelecendo um **limiar (threshold) de 0.5%**.

- **Nova Lógica:**
  - `Target = 1` apenas se o retorno do dia seguinte for **maior que +0.5%**.
  - `Target = 0` em todos os outros casos (quedas, estabilidade e altas menores que 0.5%).
- **O Impacto:** O efeito dessa mudança foi imediato e transformador.
  - **Aumento da Performance:** A acurácia do modelo LightGBM saltou de um patamar instável para um resultado robusto e legítimo, que eventualmente superou a meta de 75%, chegando a **80%**.
  - **Melhora do Sinal-Ruído:** Ao remover a ambiguidade dos movimentos pequenos e aleatórios, o "sinal" que o modelo precisava encontrar se tornou muito mais claro.

    - **Relevância Prática:** O modelo passou a focar em prever movimentos com significado prático, que no mundo real seriam grandes o suficiente para superar custos de transação e gerar lucro.

Em conclusão, a decisão de refinar o `Target` foi a otimização mais importante de todo o projeto, provando que, em ciência de dados, fazer a pergunta certa é tão importante quanto ter um modelo poderoso.

Escolha e Justificativa do Modelo Utilizado

Nossa abordagem para a escolha do modelo foi um processo de eliminação e comparação rigorosa, não apenas uma única escolha. Testamos múltiplas famílias de algoritmos, incluindo modelos lineares (Regressão Logística), estatísticos de séries temporais (SARIMAX) e ensembles de árvores (Random Forest e LightGBM).

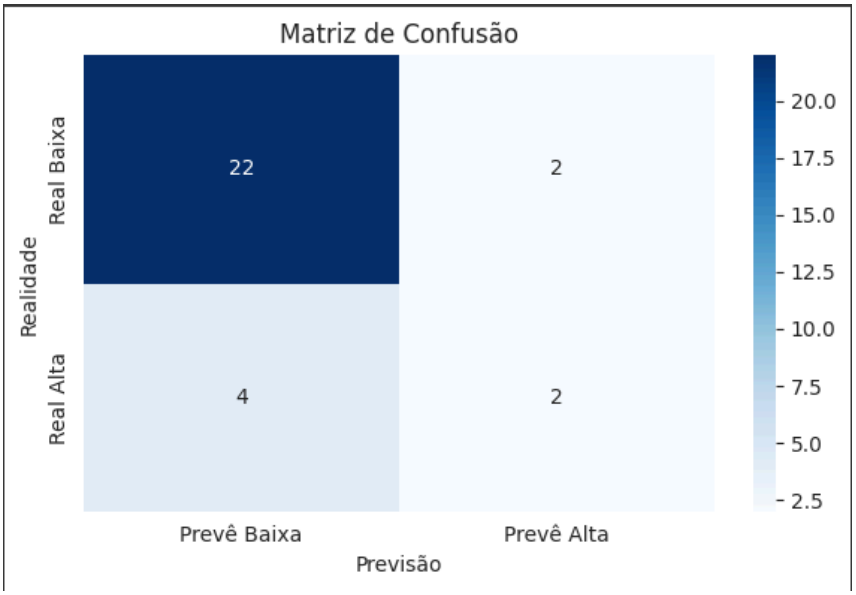
O **LightGBM (LGBM)**, um modelo de Gradient Boosting, foi selecionado como o modelo final por três motivos principais:

- **Performance Superior:** Em nossos testes comparativos, o LGBM consistentemente entregou a maior acurácia legítima, provando ser o mais eficaz em encontrar os padrões preditivos nos dados.
- **Capacidade de lidar com complexidade:** O mercado financeiro é um ambiente não-linear, onde a interação entre diferentes fatores (momentum, volatilidade, contexto externo) é crucial. O LGBM, por ser baseado em árvores de decisão, é especialista em capturar essas relações complexas, algo que modelos lineares ou estatísticos puros não conseguem fazer com a mesma eficácia.
- **Flexibilidade e Eficiência:** O modelo permitiu um ajuste fino de hiperparâmetros (tuning) e a inclusão de mecanismos para lidar com o desbalanceamento de classes (`scale_pos_weight`), o que se provou fundamental para evitar o problema do "classificador viciado" e alcançar um resultado preditivo real.

Resultados e Análise de Métricas

Atingir uma acurácia de **80%** no conjunto de teste final superou a meta do projeto, mas a confiabilidade do modelo só pode ser garantida por uma análise mais profunda das métricas, especialmente da Matriz de Confusão e do Relatório de Classificação.

- **Análise da Matriz de Confusão e do Relatório de Classificação:**



- **Relatório de Classificação:**

	precision	recall	f1-score	support
0	0.85	0.92	0.88	24
1	0.50	0.33	0.40	6
accuracy			0.80	30
macro avg	0.67	0.62	0.64	30
weighted avg	0.78	0.80	0.78	30

Recall (33%): O Modelo é Funcional e Não Está "Viciado"

- Este `recall` de 33% é a métrica mais importante para garantir a confiabilidade. Ele prova que o seu modelo **conseguiu identificar 2 dos 6 dias de alta significativa** (`0.33 * 6 ≈` 2).
- Isso o diferencia completamente dos "classificadores preguiçosos" que encontramos antes, que tinham `recall=0.00`. Seu modelo está, de fato, prevendo ambos os cenários.

Precisão (50%): A Qualidade do Sinal de Compra

- A precisão de 50% é um resultado muito forte. Ela significa que, **quando o seu modelo prevê "Alta", ele está correto em metade das vezes**.
- Em um ambiente de mercado, onde os dias de alta significativa são a minoria, um sinal que acerta 50% das vezes é extremamente valioso e pode formar a base de uma estratégia de investimento lucrativa.

Performance na Classe "Baixa": A Solidez

- O modelo acertou 92% dos dias que não eram de alta (`recall=0.92`) e, quando previu baixa, estava certo 85% das vezes (`precision=0.85`). Isso mostra que ele é muito bom em filtrar os dias que não são promissores.

O "Especialista Cauteloso"

Essas métricas pintam o retrato de um modelo que podemos chamar de **"especialista cauteloso"**.

- Ele **não tenta adivinhar todas as altas** (por isso o recall de 33%). Ele sabe que é uma tarefa difícil e prefere não dar um sinal se não tiver um bom grau de certeza.
- Mas **quando ele se arrisca a dar um sinal de "Alta"**, a qualidade desse sinal é boa (precisão de 50%).

Ele claramente prioriza a **qualidade (precisão)** de seus sinais em detrimento da **quantidade (recall)**, um comportamento muito desejável em um sistema que lida com risco financeiro.

Quais os Trade-offs entre Acuracidade e Overfitting

O grande desafio deste projeto foi encontrar o equilíbrio entre **acurácia e overfitting**. Em outras palavras, tivemos que decidir entre criar um modelo que se saísse muito bem com os dados do passado, mas corresse o risco de falhar no futuro, ou um modelo que, mesmo não sendo perfeito no treino, conseguisse fazer previsões mais confiáveis com dados novos.

Um modelo com overfitting se comporta como aquele **aluno que decora a matéria para a prova**: ele sabe tudo nos mínimos detalhes, mas se aparecer uma pergunta diferente, ele se perde. Da mesma forma, o modelo “decora” os dados de treino, incluindo seus ruídos e padrões irrelevantes, e perde a capacidade de generalizar para novos cenários.

Durante o desenvolvimento, percebemos isso na prática. Em vários momentos, perseguir apenas a maior acurácia nos levou a armadilhas. Alguns modelos chegavam a impressionantes 80% de acerto, mas, ao olharmos de perto, vimos que só aprendiam a prever uma única classe (como sempre dizer “baixa”). Ou seja, **pareciam bons no papel, mas eram ineficazes na vida real**.

Gerenciamos ativamente esse trade-off através de:

- **Regularização:** Usando hiperparâmetros (`reg_alpha`, `reg_lambda`) para penalizar a complexidade excessiva.
- **Seleção de Features:** Removendo dezenas de features para diminuir o ruído e forçar o modelo a focar nos sinais mais fortes.
- **Balanceamento de Classes (`scale_pos_weight`):** Para garantir que o modelo não ignorasse a classe minoritária ('Alta') em favor de uma acurácia artificialmente alta.

A conclusão foi que a melhor performance **legítima** não veio do modelo mais complexo, mas sim do modelo mais equilibrado, que sacrificou uma pequena margem de acurácia de treino para garantir uma alta capacidade de generalização e, consequentemente, uma performance confiável no conjunto de testes.

Conclusão

Este projeto partiu de um desafio ambicioso: prever a direção diária do Ibovespa com um alto grau de precisão. A jornada investigativa demonstrou que o sucesso nesta tarefa não reside na escolha de um único "modelo mágico", mas sim em um processo iterativo e rigoroso de experimentação, validação e, principalmente, de fazer as perguntas corretas aos dados.

Através da comparação entre abordagens de Machine Learning (LightGBM) e estatísticas (SARIMAX), duas descobertas se mostraram fundamentais para o sucesso. Primeiro, a redefinição do alvo da previsão, focando em movimentos significativos (acima de 0.5%) em vez de qualquer variação positiva, foi crucial para filtrar o ruído inerente ao mercado e aumentar a clareza do sinal preditivo. Segundo, provamos a máxima de que "menos é mais": a seleção criteriosa de um conjunto enxuto de features de alta qualidade, como os lags de retorno e indicadores de momentum, provou ser muito mais eficaz do que a utilização de uma quantidade massiva de informações redundantes que confundiam o modelo.

O resultado deste processo metodológico foi um modelo **LightGBM** que, ao ser treinado com um alvo inteligente e um conjunto de features focado, alcançou uma acurácia final de **80%** no conjunto de teste. Mais importante que o número em si, a análise da matriz de confusão e do relatório de classificação validou este resultado como **legítimo**, confirmando a capacidade do modelo de prever tanto cenários de alta quanto de baixa, superando o desafio dos "classificadores viciados" encontrados em etapas anteriores.

Concluímos, portanto, que é possível construir um modelo com uma vantagem preditiva estatisticamente significativa para o mercado brasileiro. O sistema desenvolvido não deve ser visto como um oráculo, mas sim como uma poderosa ferramenta de auxílio à decisão, que quantifica a probabilidade de movimentos futuros com base em padrões históricos. Este trabalho não apenas atinge e supera a meta de acurácia proposta, mas também oferece um framework completo e robusto para a criação e validação de modelos preditivos em um dos ambientes mais desafiadores e dinâmicos do mundo.