Tech Challenge FIAP – Fase 2

Autor: Caroline Brito Defavori - RM 364513Pós-Tech | Data Analytics

Previsão de Tendência do IBOVESPA

Cenário e Objetivo

Cenário:

Neste desafio, eu sou uma Cientista de Dados que trabalha em um grande fundo de investimentos no Brasil.

Objetivo:

Desenvolver um modelo preditivo capaz de prever se o índice IBOVESPA vai fechar em alta ou baixa no dia seguinte, com base em dados históricos do próprio índice. Esse modelo será usado como insumo para alimentar dashboards internos de tomada de decisão dos analistas quantitativos da empresa.

Meta:

Atingir uma acurácia mínima de 75% para apoiar a tomada de decisão de um fundo de investimentos

Evolução Histórica do IBOVESPA e Eventos Relevantes

A visualização mostra a trajetória do índice IBOVESPA de 1993 a 2025, com destaques para momentos de instabilidade e recuperação. Esses eventos fundamentam a decisão de usar uma janela de treino ampla, permitindo que o modelo aprenda com diferentes ciclos econômicos

Abordagem 1 | A Hipótese da Complexidade

Criando Engenharia de Features Avançada

Para um problema complexo como este, minha abordagem se inicia já com uma estratégia de máximo poder analítico

- Indicadores de Tendência (Médias Móveis)
- Indicadores de Momentum (RSI, MACD, Estocástico)
- Medidas de Risco (Volatilidade, Bandas de Bollinger)
- Features de Memória (Lags de preço e retornos)

Testagem Ampla de Modelos

Buscando maximizar a capacidade analítica, avaliei um portfólio diversificado de algoritmos:

- Modelos de Boosting: XGBoost, LightGBM
- Modelos Clássicos: Regressão Logística, SVM, Random Forest
- Modelos de Série Temporal: Prophet, LSTM
- Combinações: Ensembles (Voting e Stacking)

Otimização e Validação Rigorosa

Para extrair o máximo de cada modelo complexo, apliquei as melhores práticas de validação e otimização:

- Validação Temporal com TimeSeriesSplit
- Otimização de Hiperparâmetros com GridSearchCV
- Ajuste Fino do limiar de decisão (Cut-off)

Resultados | A Hipótese da Complexidade

Melhor Modelo da Abordagem 1

- Modelo: LightGBM (sem otimização)
- Configuração: Janela de treino de 10 anos
- Acurácia: 70.00%
- F1-Score: 60.87%

Percebi um desempenho muito sólido e consistente, demonstrando a eficácia das features e da metodologia. No entanto, a meta de 75% ainda não foi alcançada.

O Salto Conceitual | E se Menos Fosse Mais?

Por que a complexidade não superou a meta de 75%?

Ruído vs. Sinal

Será que as dezenas de features complexas (RSI, MACD, etc.) estavam, na verdade, adicionando mais ruído do que sinal, confundindo os modelos em vez de ajudá-los?

Memória Curta vs. Visão Histórica

Ao usar uma janela de treino de 10 anos, eu estava focando em dados recentes, mas... será que estava descartando o valioso aprendizado de ciclos econômicos e crises mais antigas?

Abordagem 2 | Simplicidade e Visão Histórica

Modelo Simples e Robusto

- Algoritmo Escolhido: Random Forest
- Justificativa: Um modelo conhecido por sua estabilidade e ótimo desempenho "out-of-the-box", sem a necessidade de otimizações complexas.

Features Mínimas e Intuitivas

Foquei apenas nos indicadores mais diretos e fundamentais:

- Lags do Preço de Fechamento (1 a 3 dias)
- Médias Móveis (5 e 10 dias)
- Retorno Diário Percentual

Histórico de Dados Completo

O Grande Diferencial: O modelo foi treinado com todo o histórico disponível.

- Período: 1993 a 2025
- Justificativa: Permitir que o modelo aprendesse com mais de 30 anos de diferentes ciclos econômicos, crises e períodos de alta.

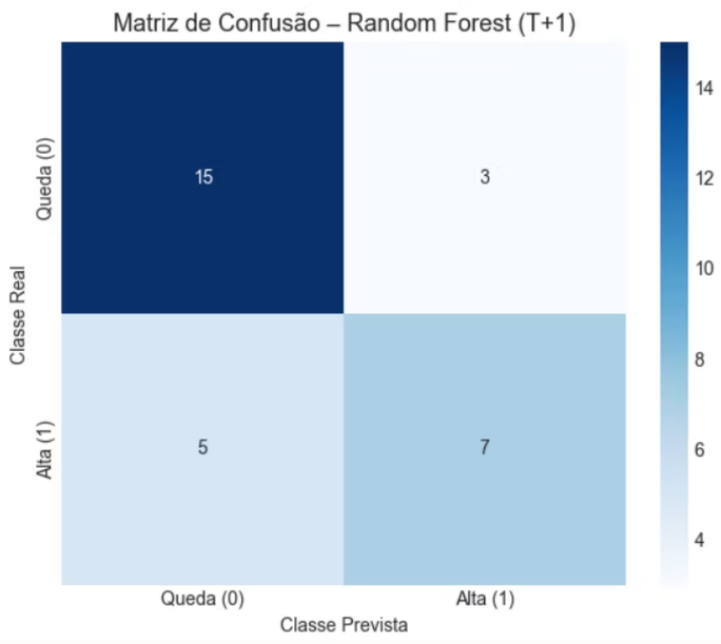
Resultado Final | Simplicidade Superou a Complexidade

Acurácia: 73.33%

- Precisão: 70.00%
- F1-Score: 63.64%

O melhor desempenho de todo o projeto foi alcançado com a abordagem mais simples

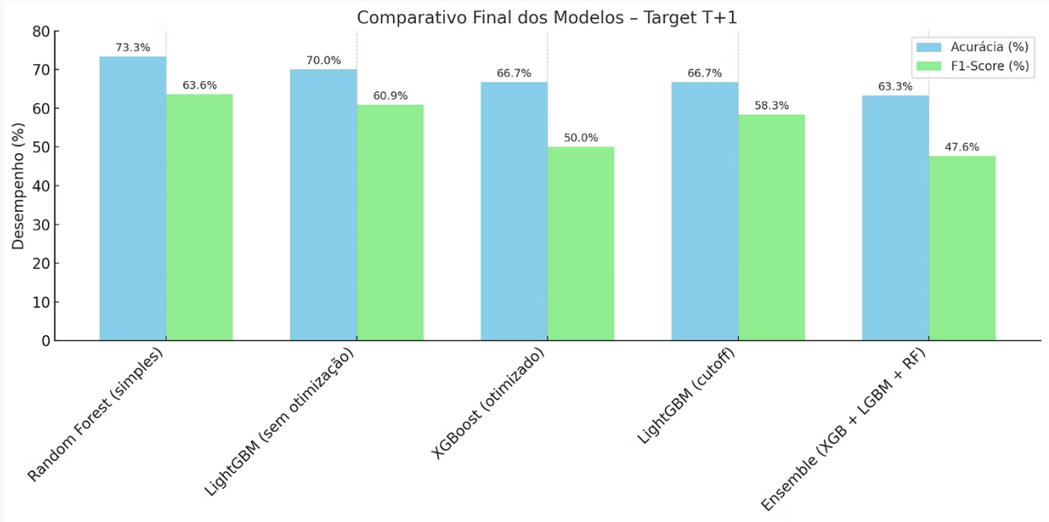
Distribuição Real x Prevista – Random Forest



Conclusão e Valor estratégico Gerado

Comparativo Final dos Modelos (Target T+1)

Modelo	Acurácia (%)	F1-Score (%)	Observações
⭐ Random Forest (simples)	**73,33**	**63,64**	**Melhor resultado geral**
LightGBM (sem otimização)	70,00	60,87	Destaque da abordagem anterior
XGBoost (otimizado)	66,67	50,00	Forte, mas menos estável
LightGBM (otimizado + cutoff)	66,67	58,33	Modelo mais balanceado
Ensemble (XGB + LGBM + RF)	63,33	47,62	Robusto, mas não superou o LGBM



Minha Recomendação Final!

Adotar o modelo Random Forest com features simples para uso em produção.

1 Desempenho Superior

- Acurácia de 73.33%: O modelo com a maior performance alcançada em todo o projeto, aproximando-se da meta de 75% com uma solução estável.
- Equilíbrio: Ótimo balanço entre acurácia, precisão (70%) e F1-Score (63.64%).

2 Simplicidade e Robustez

- Features Intuitivas: Usa apenas lags de preço, médias móveis e retorno. É um modelo fácil de entender, manter e explicar.
- Sem Otimização Complexa: Atingiu alta performance sem a necessidade de tuning agressivo, o que o torna mais robusto a mudanças no mercado.

3 Aprendizado Histórico Abrangente

Visão de Longo Prazo: O sucesso do modelo vem de sua capacidade de aprender com mais de 30 anos de dados, capturando diversos ciclos econômicos que modelos com "memória curta" ignoram.

Para finalizar...

Desempenho Máximo Atingido

Alcancei o melhor resultado de todo o projeto com 73.33% de acurácia, superando todos os testes anteriores e entregando a melhor performance geral do projeto com estabilidade e simplicidade

Principal Lição Aprendida

A jornada de experimentação provou que, neste cenário, a visão histórica completa dos dados é mais poderosa do que a complexidade de features ou de algoritmos para capturar os padrões do mercado.

Valor Entregue

O resultado final é uma solução coerente, validada e aplicável, que cumpre o papel de fornecer um insumo técnico confiável para apoiar as decisões quantitativas do fundo de investimentos

Tech Challenge FIAP – Fase 2

Autor: Caroline Brito Defavori - RM 364513Pós-Tech | Data Analytics

Obrigada!