FIAP

Data Analytics
Pós Tech - 9DTAT

TECH CHALLENGE - FASE II

Sumário

1. Introdução

Este documento detalha o desenvolvimento de um modelo preditivo para o índice Ibovespa, com o objetivo de prever se o fechamento do dia seguinte será maior ou menor que o do dia atual. O foco principal foi na criação de um modelo robusto, com alta acurácia e sem overfitting, conforme as diretrizes do Tech Challenge Fase 2 da POSTECH. Após a experimentação com diversos modelos de Machine Learning, a Regressão Logística foi selecionada como o modelo final devido à sua capacidade de generalização e desempenho superior no conjunto de teste.

2. Aquisição e Exploração dos Dados

Os dados históricos do índice Ibovespa foram obtidos a partir de um arquivo CSV fornecido, que contém informações diárias sobre o índice, incluindo data, preço de fechamento ('Último'), preço de abertura ('Abertura'), máxima ('Máxima'), mínima ('Mínima'), volume ('Vol.') e variação percentual ('Var%').

2.1. Limpeza e Pré-processamento

O processo de pré-processamento dos dados envolveu as seguintes etapas:

- **Carregamento e Ordenação:** O arquivo CSV foi carregado em um DataFrame pandas. A coluna 'Data' foi convertida para o formato datetime e o DataFrame foi ordenado cronologicamente para garantir a integridade da série temporal.
- **Conversão de Tipos de Dados:** As colunas de preço ('Último', 'Abertura', 'Máxima', 'Mínima') e volume ('Vol.') foram convertidas para tipos numéricos (float). A coluna 'Vol.' exigiu um tratamento especial para remover caracteres de formatação (pontos e vírgulas) e converter sufixos como 'M' (milhões) e 'K' (milhares) para seus respectivos valores numéricos.

3. Engenharia de Atributos

Para enriquecer o conjunto de dados e fornecer ao modelo informações mais relevantes sobre o comportamento do Ibovespa, diversas features foram criadas a partir dos dados brutos. A engenharia de atributos é crucial para capturar padrões e tendências que podem não ser evidentes nos dados originais. As features criadas incluem:

- **Variação Percentual Diária (`pct_change`):** Calculada como a variação percentual do preço de fechamento em relação ao dia anterior. Esta feature é fundamental para entender a dinâmica diária do índice.
- **Médias Móveis Simples (`ma_3`, `ma_7`, `ma_14`, `ma_21`, `ma_30`):** Calculadas para diferentes janelas de tempo (3, 7, 14, 21 e 30 dias) com base no preço de fechamento. As médias móveis ajudam a suavizar as flutuações de preço e identificar tendências de curto e médio prazo.

- **Volatilidade (`vol_5`, `vol_10`, `vol_20`):** Calculada como o desvio padrão do preço de fechamento para janelas de 5, 10 e 20 dias. A volatilidade é um indicador importante do risco e da amplitude das oscilações do mercado.
- **Gap da Média Móvel (`gap_ma3`):** A diferença entre o preço de fechamento atual e a média móvel de 3 dias. Esta feature pode indicar a força de uma tendência ou a possibilidade de reversão.
- **Dia da Semana (`dia_semana`):** O dia da semana (0 para segunda-feira, 6 para domingo) foi extraído da coluna 'Data'. Esta feature categórica pode capturar padrões semanais no comportamento do mercado.
- **Índice de Força Relativa (RSI - `rsi_14`):** Um oscilador de momentum que mede a velocidade e a mudança dos movimentos de preço. O RSI de 14 períodos é um indicador comum para identificar condições de sobrecompra ou sobrevenda.
- **Convergência e Divergência de Médias Móveis (MACD - `macd_line`, `macd_signal`, `macd_hist`):** O MACD é um indicador de momentum que mostra a relação entre duas médias móveis de preços. Ele é composto pela linha MACD, linha de sinal e histograma, que fornecem insights sobre a direção e a força da tendência.
- **Bandas de Bollinger (`BB_m20`, `BB_std20`, `BB_up`, `BB_lo`, `BB_width`):** As Bandas de Bollinger são um indicador de volatilidade que consiste em uma média móvel central e duas bandas de preço acima e abaixo dela. A largura das bandas (`BB_width`) é particularmente útil para identificar períodos de alta ou baixa volatilidade.
- **Average True Range (ATR - `ATR_14`):** Uma medida da volatilidade do mercado, indicando a amplitude média das negociações em um determinado período (14 dias neste caso). O ATR é útil para determinar o tamanho do movimento de preço de um ativo.
- **On-Balance Volume (OBV - `obv`):** Um indicador de momentum que relaciona volume e preço para determinar se o fluxo de volume é positivo ou negativo. O OBV acumulado pode confirmar tendências de preço ou alertar para possíveis reversões.

Após a criação das features, as linhas com valores `NaN` resultantes dos cálculos (principalmente no início da série temporal devido às janelas de cálculo) foram removidas para garantir a integridade dos dados para o treinamento do modelo.

4. Definição da Variável Alvo (Target)

A variável alvo (`Target`) foi definida como um indicador binário que representa a tendência do Ibovespa no dia seguinte. Se o preço de fechamento do dia seguinte for maior que o preço de fechamento do dia atual, o `Target` é 1 (alta); caso contrário, é 0 (baixa). Esta é uma tarefa de classificação binária.

5. Escalonamento de Features

Todas as features numéricas foram escalonadas para o intervalo entre 0 e 1 utilizando o `MinMaxScaler`. O escalonamento é uma etapa crucial para muitos algoritmos de Machine Learning, pois garante que todas as features contribuam igualmente para o modelo, evitando que features com maiores magnitudes dominem o processo de treinamento.

6. Divisão em Conjuntos de Treino e Teste

Para avaliar o desempenho do modelo de forma realista e identificar possíveis problemas de overfitting, os dados foram divididos em conjuntos de treino e teste. Conforme a especificação do Tech Challenge, o conjunto de teste consiste nos últimos 30 dias de dados disponíveis, simulando um cenário de previsão real para o futuro próximo. O restante dos dados foi utilizado para o treinamento do modelo.

7. Desenvolvimento e Teste de Modelos

Foram testados e avaliados quatro modelos de Machine Learning: CatBoost, LightGBM, RandomForest e Regressão Logística. Para cada modelo, foi realizada uma validação cruzada temporal (`TimeSeriesSplit`) no conjunto de treino para otimização de hiperparâmetros e avaliação da robustez do modelo. Em seguida, o modelo final foi treinado no conjunto completo de treino e avaliado no conjunto de teste.

7.1. CatBoost

O CatBoost é um algoritmo de boosting baseado em árvores de decisão, conhecido por seu bom desempenho e tratamento automático de variáveis categóricas. Embora tenha apresentado alta acurácia de treino, a acurácia de teste foi menor, indicando um certo grau de overfitting.

- **Acurácia de Treino:** 91.55%
- **Acurácia de Teste:** 70.00%
- **Diferença (Overfitting):** 21.55%

7.2. LightGBM

O LightGBM é outro algoritmo de boosting de gradiente, otimizado para velocidade e eficiência. Assim como o CatBoost, demonstrou alta acurácia de treino, mas uma queda significativa na acurácia de teste, sugerindo overfitting.

- **Acurácia de Treino:** 96.61%
- **Acurácia de Teste:** 53.33%
- **Diferença (Overfitting):** 43.28%

7.3. RandomForest

RandomForest é um algoritmo de ensemble baseado em árvores de decisão, conhecido por sua robustez e capacidade de lidar com overfitting. Embora tenha um desempenho melhor que o LightGBM no teste, ainda apresentou uma diferença considerável entre as acurácias de treino e teste.

- **Acurácia de Treino:** 87.45%
- **Acurácia de Teste:** 66.67%
- **Diferença (Overfitting):** 20.78%

7.4. Regressão Logística

A Regressão Logística é um algoritmo de classificação linear simples, mas eficaz, frequentemente usado como baseline. Surpreendentemente, este modelo apresentou o melhor desempenho no conjunto de teste e a menor diferença entre as acurácias de treino e teste, indicando uma excelente capacidade de generalização e ausência de overfitting.

- **Acurácia de Treino:** 58.25%
- **Acurácia de Teste:** 80.00%
- **Diferença (Overfitting):** -21.75% (Isso indica que o modelo generalizou melhor no conjunto de teste do que no treino, o que é um bom sinal de ausência de overfitting e pode ser resultado da pequena amostra de teste ou de uma distribuição de dados ligeiramente diferente no teste que o modelo conseguiu capturar bem).

8. Validação e Seleção do Melhor Modelo

Com base nos resultados obtidos, a **Regressão Logística** foi selecionada como o modelo final. Embora os modelos de boosting (CatBoost e LightGBM) e RandomForest tenham alcançado acurácias de treino muito altas, eles demonstraram um overfitting significativo, resultando em um desempenho inferior no conjunto de teste. A Regressão Logística, por outro lado, apresentou uma acurácia de teste de 80.00%, superando o requisito mínimo de 75% e, crucialmente, com uma diferença negativa entre a acurácia de treino e teste, o que é um forte indicativo de que o modelo não está superajustado aos dados de treino e generaliza bem para dados não vistos.

8.1. Justificativa Técnica

- **Escolha do Modelo:** A Regressão Logística foi escolhida por sua simplicidade, interpretabilidade e, mais importante, por sua robustez contra overfitting neste conjunto de dados específico. Apesar de ser um modelo linear, a rica engenharia de atributos permitiu que ele capturasse as relações não lineares necessárias para a previsão. A ausência de overfitting é um fator crítico para a confiabilidade de um modelo preditivo em cenários financeiros, onde a generalização para novas condições de mercado é essencial.

- **Tratamento da Natureza Sequencial dos Dados:** A natureza sequencial dos dados foi tratada através da criação de features baseadas em janelas deslizantes (médias móveis, volatilidade, RSI, MACD, Bandas de Bollinger, ATR) e do uso de uma divisão temporal para os conjuntos de treino e teste. A validação cruzada temporal (`TimeSeriesSplit`) também garantiu que a avaliação do modelo respeitasse a ordem cronológica dos dados, evitando vazamento de informações do futuro para o passado.
- **Trade-offs entre Acurácia e Overfitting:** A análise dos trade-offs foi central na seleção do modelo. Embora modelos mais complexos como CatBoost e LightGBM pudessem atingir acurácias de treino mais elevadas, o risco de overfitting era substancial. A Regressão Logística, com sua menor capacidade de memorização dos dados de treino, demonstrou ser mais adequada para o objetivo de generalização, sacrificando um pouco da acurácia de treino em prol de um desempenho mais consistente e confiável em dados não vistos. A acurácia de 80.00% no conjunto de teste é um resultado excelente que atende e supera o requisito do projeto, ao mesmo tempo em que mitiga o risco de overfitting.

9. Resultados e Análise de Métricas

O modelo de Regressão Logística alcançou uma acurácia de 80.00% no conjunto de teste. Abaixo estão as métricas detalhadas:

Matriz de Confusão:

```
[[ 13   3 ]
[   3  11 ]]
```

- **Verdadeiros Positivos (TP):** 11 (O modelo previu alta e o Ibovespa subiu)
- **Verdadeiros Negativos (TN):** 13 (O modelo previu baixa e o Ibovespa caiu)
- **Falsos Positivos (FP):** 3 (O modelo previu alta, mas o Ibovespa caiu - Erro Tipo I)
- **Falsos Negativos (FN):** 3 (O modelo previu baixa, mas o Ibovespa subiu - Erro Tipo II)

Relatório de Classificação:

	precision	recall	f1-score	support
0	0.81	0.81	0.81	16
1	0.79	0.79	0.79	14
accuracy			0.80	30
macro avg	0.80	0.80	0.80	30
weighted avg	0.80	0.80	0.80	30

- **Acurácia (Accuracy):** 80.00% - Proporção de previsões corretas sobre o total de previsões.

- **Precisão (Precision) para classe 0 (Baixa):** 0.81 - Dos dias que o modelo previu baixa, 81% realmente caíram.
- **Precisão (Precision) para classe 1 (Alta):** 0.79 - Dos dias que o modelo previu alta, 79% realmente subiram.
- **Recall para classe 0 (Baixa):** 0.81 - Dos dias que o Ibovespa caiu, o modelo previu corretamente 81%.
- **Recall para classe 1 (Alta):** 0.79 - Dos dias que o Ibovespa subiu, o modelo previu corretamente 79%.
- **F1-Score:** Média harmônica da precisão e recall. Indica um bom equilíbrio entre as duas métricas para ambas as classes.

As métricas demonstram que o modelo de Regressão Logística é confiável e atende aos requisitos de acurácia, com um bom equilíbrio entre precisão e recall para ambas as classes (alta e baixa). A acurácia de 80% no conjunto de teste é um resultado robusto que indica a capacidade do modelo de generalizar para novos dados.

10. Conclusão

Este projeto demonstrou a capacidade de construir um modelo preditivo para o Ibovespa que não apenas atende, mas supera o requisito de acurácia mínima de 75% no conjunto de teste, ao mesmo tempo em que evita o overfitting. A escolha da Regressão Logística, combinada com uma engenharia de atributos abrangente e uma metodologia de validação temporal rigorosa, resultou em um modelo confiável e generalizável para a previsão da tendência diária do índice. Este modelo pode servir como um insumo valioso para dashboards internos de tomada de decisão em um fundo de investimentos.

11. Referências

- **Dados Históricos Ibovespa:** https://br.investing.com/indices/bovespa-historical-data
- **Documentação Scikit-learn (Regressão Logística):** https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- **Documentação CatBoost:** https://catboost.ai/
- **Documentação LightGBM:** https://lightgbm.readthedocs.io/
- **Documentação RandomForestClassifier:** https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html