Universidade Tecnológica Federal do Paraná - Campus Dois Vizinhos

Especialização em Ciência de Dados

Resolução de Problemas com Ciência de Dados

Estudo de caso

Mercado financeiro para compra e venda de ações

Grupo R

Dois Vizinhos - PR
2021

Identificação do Grupo R:

Ana Paula Fernandes Lucio Menezes

Gabriel Stankevix Soares

Heron Carlos Gonçalves

Isabela Fernanda Capetti

Contexto da área do problema/estudo de caso : Cada vez mais o mercado de ações brasileiro ganha público e importância para pequenos e médios investidores. Identificar a oportunidade de compra e venda de ações com objetivo de melhorar a rentabilidade é o desafio principal dos novos acionistas.

Definição do problema : Aplicar aprendizado de máquina para analisarmos a previsão de valores futuros e tendências de ações no mercado financeiro. Para este estudo de caso foram selecionadas as ações da Itaúsa, dados reais de mercado, durante o período de 2016 a 2021.

Justificativa/Importância: Entender os momentos mais adequados para fazer sua aplicação, garante maior segurança e robustez nas tomadas de decisão. Isso impacta diretamente no retorno de futuros aportes dos acionistas para maximização de ganhos e valorização do ativo.

Descrição dos dados :

Variável	Tipo	Descrição
Data	Date	Dados diários
Último	Float	Dado de fechamento
Abertura	Float	Dado de abertura
Máximo	Float	Valor máximo diário
Mínimo	Float	Valor mínimo diário
Volume	String	Volume total em reais negociados
Variação	Float %	Variação diária das negociações

Pré-seleção de variáveis (quais variáveis são mais importantes para solucionar o problema):
Inicialmente pensamos em considerar todas as variáveis disponíveis, porém efetuamos uma análise de correlação para definir quais seriam utilizadas.

Pré-seleção do método (qual o primeiro método (candidato) que será utilizado e por qual motivo): No primeiro momento, serão aplicados modelos da classe de autorregressão devido a característica dos dados e por serem os mais populares na literatura/academia, e estes foram utilizados considerando a presença ou não de variáveis exógenas para assim avaliar performance.

Estudo de caso – Mercado financeiro

Qual a fonte dos dados: Repositório Yahoo Finance disponível no link: https://finance.yahoo.com/

Como foi/será feita a coleta/geração dos dados?: Via download direto do site de acordo com as orientações contidas neste link https://help.yahoo.com/kb/download-historical-data-yahoo-finance-sln2311.html

- **Os dados serão/foram obtidos de terceiros? :** Dados obtidos de fonte pública disponibilizados pelo Yahoo Finance a partir do histórico de precificação da ação Itausa na B3 que é a empresa de infraestrutura de mercado financeiro que atua no Brasil em abinete de bolsa de valores.
- **Qual foi a metodologia para a geração/coleta destes dados ?:** Desde 1997 o Yahoo Finance fornece gratuitamente dados financeiros, notícias e cotações de diversas ações negociadas nas bolsas de valores do mundo todo com objetivo de apoiar no portfólios e gerenciamento das finanças dos usuários.

Quais as considerações éticas no uso destes dados? : As informações disponibilizadas são para fins acessos diários de usuários que podem realizar download para uso em análises financeiras.

Resultado preliminar da solução do problema

- Qual método foi aplicado e como (Metodologia)

Os dados coletados são caracterizados por uma sequência de valores que são dependentes das datas correspondentes a cada linha. Esse tipo de dado em que a ordem temporal é crucial para o entendimento das observações, além de possível presença de fatores como tendência e sazonalidade, caracterizam as chamadas séries temporais. Essa série temporal é da Itausa Investimentos Itaú SA, que é uma empresa sediada no Brasil com atividade principal no setor bancário. As atividades da Companhia estão divididas em dois segmentos de negócios: Financeiro e Industrial. A divisão Financeira concentra-se na gestão do Itaú Unibanco Holding SA, uma instituição bancária que oferece produtos e serviços financeiros, como empréstimos, cartões de crédito, contas correntes, apólices de seguros, ferramentas de investimento, corretagem de valores mobiliários, consultoria de tesouraria e investimentos para clientes individuais e empresas. A divisão Industrial é responsável pela operação da Itautec SA, que fabrica equipamentos de automação comercial e bancária, além de prestar serviços de tecnologia da informação (TI); Duratex SA, que produz painéis de madeira, louças sanitárias e metais sanitários, e Alpargatas, que produz calçados sob as marcas Juntas, Havaianas e Dupe, entre outros.
As etapas da modelagem foram as seguintes:

1. Identificação da variável target – "Último" que foi o valor de fechamento do dia
2. Identificação do modelo
3. Inclusão de novas variáveis
4. Estimação dos parâmetros
5. Análise da adequação do modelo
6. Inclusão das variáveis exógenas (que podem interferir em outra)
7. Previsão e validação

Considerando esse domínio temporal e os fatores citados, a modelagem da informação de fechamento do valor da ação Itausa foi avaliado pelo método Sarimax que explica a variável dependente pela combinação das variáveis exógenas e suas defasagens com defasagens da variável dependente, com os seguintes parâmetros :

- p é o número de defasagens da série (parte autorregressiva) não sazonal ou estacionária
- d é a ordem de diferenciação não sazonal para alcançar estacionariedade
- q é ordem não sazonal de médias móveis
- P é ordem da parte autorregressiva sazonal

- D é ordem da parte de diferenciação sazonal
- Q é ordem da parte sazonal de médias móveis

Além das variáveis contidas no conjunto de dados de origem, foram incluídas as seguintes variáveis exógenas:
Covid – data apartir de quando a OMS decretou pandemia mundial
CriticalCovid – datas de restrições determinadas pelo Governo de São Paulo – consideramos estado de SP como base
Mês – mês correspondente a data da informação de fechamento das ações
Quadrimestre – quadrimestre correspondente ao ano analisado
Dia_da_semana- dia da semana correspondente a data da informação de fechamento das ações

```
itausa
```

Data	Último	Abertura	Máxima	Mínima	Vol.	Var%	Covid	Mes	Quadrimestre	Dia_da_Semana	CriticalCovid
2016-04-01	4.220000	4.350000	4.370000	4.22	2.797000e+07	-0.040000	0.0	4.0	2.0	4.000000	0
2016-04-02	4.340000	4.320000	4.440000	4.21	4.178000e+07	0.020000	0.0	4.0	2.0	5.000000	0
2016-04-03	5.340000	5.440000	5.580000	5.21	8.398000e+07	0.040000	0.0	4.0	2.0	6.000000	0
2016-04-04	5.200000	5.320000	5.340000	5.16	3.445000e+07	-0.030000	0.0	4.0	2.0	0.000000	0
2016-04-05	5.360000	5.300000	5.380000	5.26	2.810000e+07	0.020000	0.0	4.0	2.0	1.000000	0
...	...	...	...	...	...	...	...	...	...	...	...
2021-03-25	10.260000	10.020000	10.260000	9.97	2.474000e+07	0.020000	1.0	3.0	1.0	3.000000	0
2021-03-26	10.360000	10.180000	10.380000	10.17	2.119000e+07	0.010000	1.0	3.0	1.0	4.000000	0
2021-03-27	10.346667	10.213333	10.376667	10.17	2.015667e+07	0.006667	1.0	3.0	1.0	2.666667	0
2021-03-28	10.333333	10.246667	10.373333	10.17	1.912333e+07	0.003333	1.0	3.0	1.0	1.333333	0
2021-03-29	10.320000	10.280000	10.370000	10.17	1.809000e+07	0.000000	1.0	3.0	1.0	0.000000	0

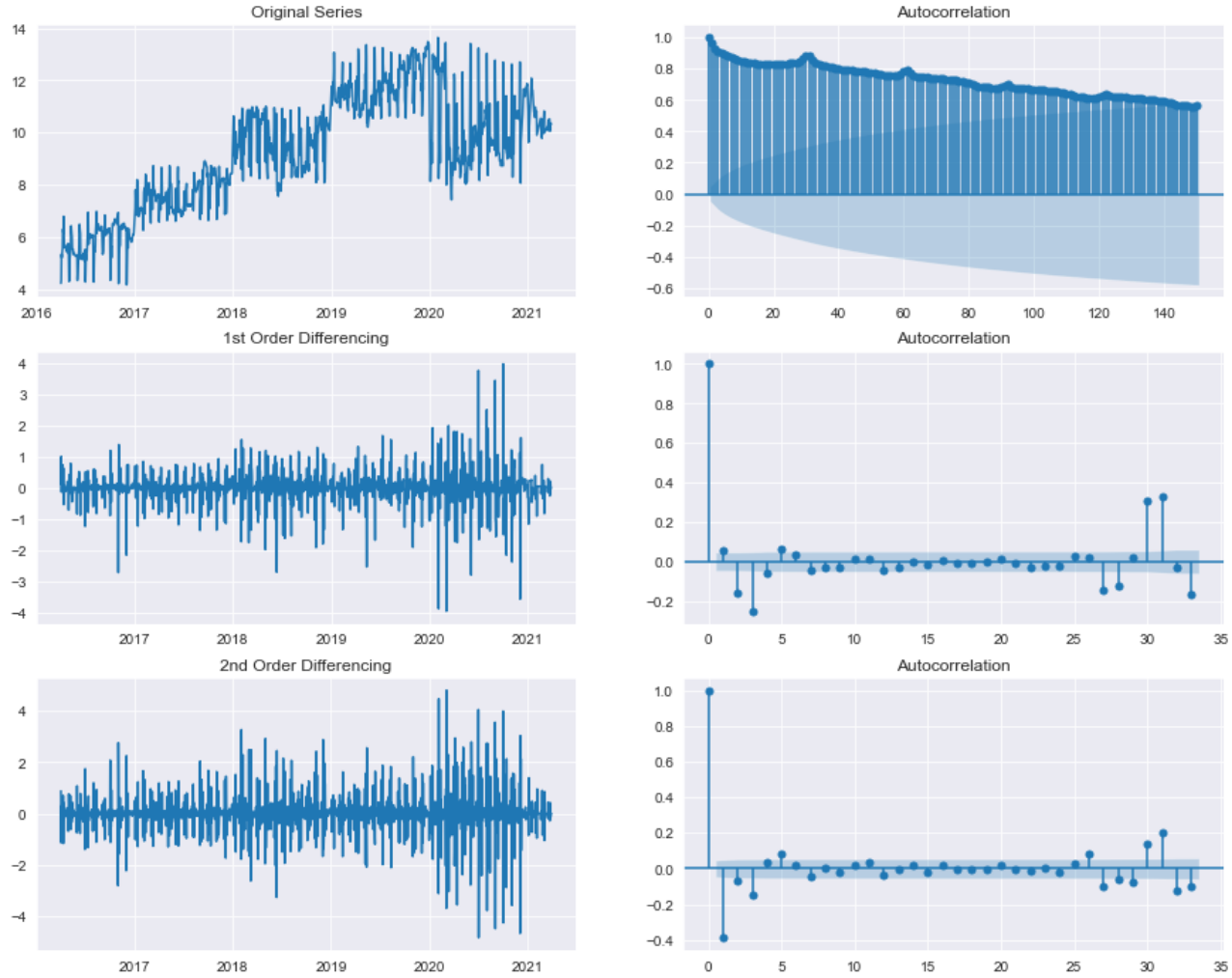
1824 rows × 11 columns

A aanlise de estacionariedade indica que a nossa série temporal pode ser predita a partir de um grau de diferenciação. Abaixo resultado dos testes aplicados:

Augmented Dickey-Fuller Test – 1 Grau de diferenciação
test statistic: -1.9936216495
p-value: 0.2893767676
critical values
1%: -3.4339921916
5%: -2.8631488249
10%: -2.5676264863
Aceitamos a Hipotese Nula

Augmented Dickey-Fuller Test – 2 grau de diferenciação
test statistic: -14.4013233828
p-value: 0.0000000000
critical values
1%: -3.4339942213
5%: -2.8631497210
10%: -2.5676269634
Rejeitamos a Hipotese Nula

Graficamente:



A correlação entre todas as variáveis é representada pela figura abaixo:

Index	Último	Abertura	Máxima	Mínima	Vol.	Var%	Covid	Mes	Quadrimestre	Dia_da_Semana	CriticalCovid
Último	1	0.997668	0.998876	0.999001	0.0412076	0.00549662	0.26207	-0.0830183	-0.0800458	-0.0109123	0.0724743
Abertura	0.997668	1	0.999031	0.998793	0.0440785	-0.0478193	0.267281	-0.0820336	-0.0787365	-0.00984082	0.0771106
Máxima	0.998876	0.999031	1	0.998608	0.0578134	-0.0240326	0.275555	-0.0864617	-0.0835825	-0.00975888	0.0838183
Mínima	0.999001	0.998793	0.998608	1	0.0281321	-0.0204547	0.256936	-0.0781333	-0.074927	-0.0104376	0.0682679
Vol.	0.0412076	0.0440785	0.0578134	0.0281321	1	-0.0431691	0.423381	-0.102493	-0.11088	0.0416267	0.285859
Var%	0.00549662	-0.0478193	-0.0240326	-0.0204547	-0.0431691	1	-0.0796852	-0.0198171	-0.015768	-0.00149262	-0.01842
Covid	0.26207	0.267281	0.275555	0.256936	0.423381	-0.0796852	1	-0.144426	-0.150169	-0.00265416	0.628258
Mes	-0.0830183	-0.0820336	-0.0864617	-0.0781333	-0.102493	-0.0198171	-0.144426	1	0.97156	0.00421013	-0.0617721
Quadrimestre	-0.0800458	-0.0787365	-0.0835825	-0.074927	-0.11088	-0.015768	-0.150169	0.97156	1	0.001974	-0.0783253
Dia_da_Semana	-0.0109123	-0.00984082	-0.00975888	-0.0104376	0.0416267	-0.00149262	-0.00265416	0.00421013	0.001974	1	0.0102792
CriticalCovid	0.0724743	0.0771106	0.0838183	0.0682679	0.285859	-0.01842	0.628258	-0.0617721	-0.0783253	0.0102792	1

Desta forma, utilizamos como variáveis exógenas Covid e CriticalCovid como parte do modelo de previsão.
Os critérios de avaliação do Sarimax foram:

- RMSE que é raiz do erro quadrático médio;
- MAPE é o erro de percentual médio absoluta;
- AIC – informação Akaike é um método matemático para avaliar o quão bem um modelo se ajusta aos dados

- BIC - é um critério para seleção de modelos entre um conjunto finito de modelos

Foi implementado em Python o Autosarimax para otimizar a busca dos melhores parâmetros (força bruta) utilizando variáveis exógenas e também sem variáveis exógenas, que retorna os 5 melhores resultados conforme os critérios de BIC, AIC e RMSE. O dataframe de treinamento utilizado dos anos de 2018, 2019 e parte de 2020. sendo o dataframe de teste 10 dias de 2020 durante período crítico da pandemia do Covid 19 no Brasil. Os resultados obtidos foram:

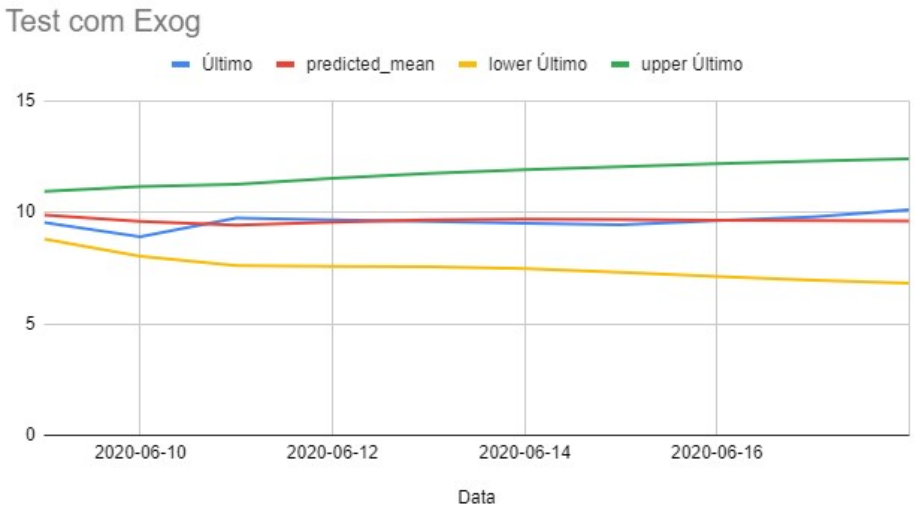
Com exog:

Neste caso, foram aplicadas as variáveis Covid e Critical Covid como parte do processo de aprendizagem do modelo.

pdq	pdqs	bic	aic	rmse
(2, 1, 2)	(2, 0, 1, 12)	1084.64	1040.2	0.326858
(2, 1, 2)	(1, 0, 2, 12)	1085.69	1041.25	0.330991
(2, 1, 2)	(2, 0, 0, 12)	1079.55	1039.55	0.331479
(2, 1, 2)	(0, 0, 2, 12)	1079.51	1039.51	0.333637
(0, 0, 0)	(0, 0, 0, 12)	4691.77	4678.44	0.340547

Resultado de Treinamento

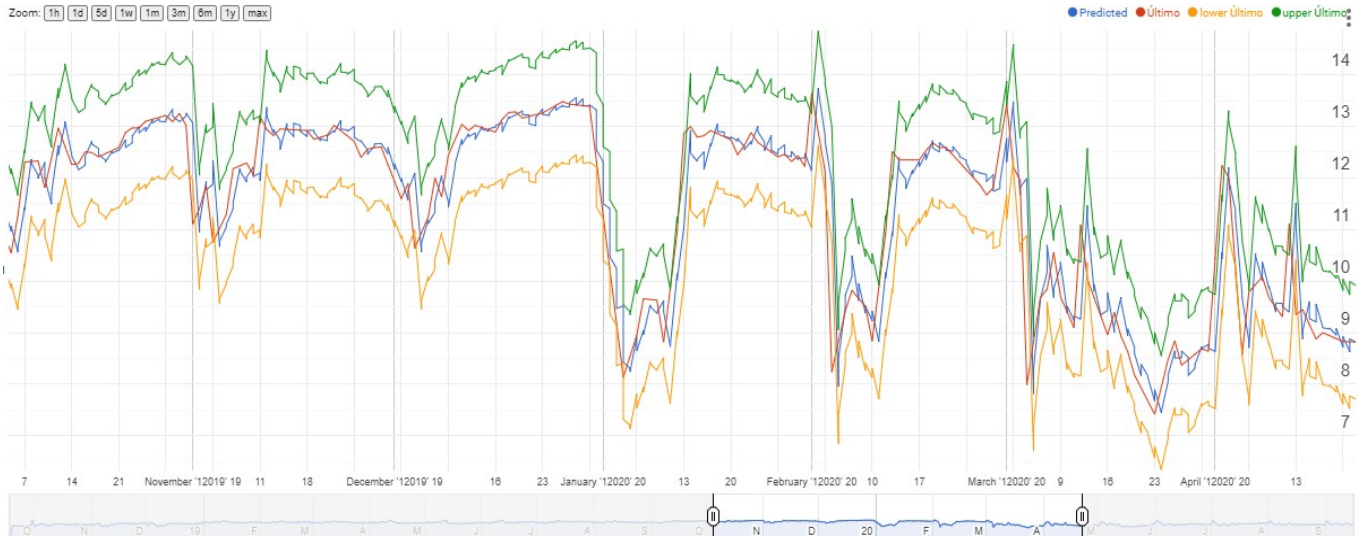
Resultado de Teste



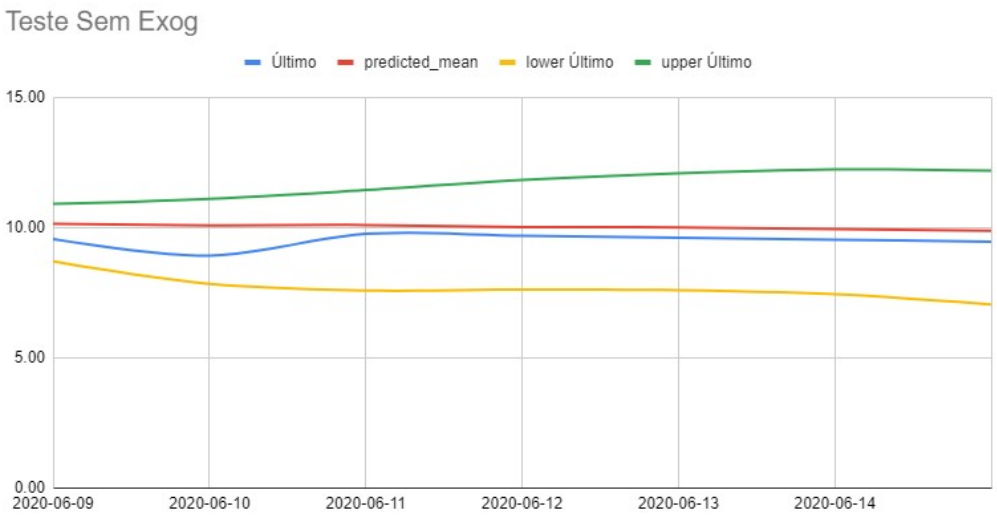
Sem exog:

pdq	pdqs	bic	aic	rmse
(2, 1, 2)	(2, 1, 1, 12)	1126.05	1090.65	0.275982
(2, 1, 2)	(2, 0, 2, 12)	1088.97	1048.98	0.292164
(2, 1, 2)	(1, 1, 1, 12)	1127.28	1096.31	0.294175
(2, 1, 2)	(1, 1, 2, 12)	1132.91	1097.51	0.301101
(2, 1, 2)	(2, 0, 1, 12)	1091.09	1055.53	0.313909

Resultado de Treinamento



Resultado de Teste



- **O problema proposto foi resolvido?**

Podemos notar que ambos os modelos do AutoSarimax retornam resultados razoavelmente satisfatórios na predição de novos valores, todavia, o grupo entende que a melhor combinação entre BIC, AIC e RMSE é o modelo que utilizou de variáveis exógenas no aprendizado e previsão. Desta forma, os parâmetros escolhidos foram de **order=(2,1,2) e seasonal_order=(2,0,1,12)** com as métricas de performance:

- Treinamento: RMSE 0,68
- Teste: RMSE de 0,32 e MAPE 98%.

- **Conclusão e/ou considerações**

Neste estudo para previsão de valores de ações, implementamos o autosarimax para otimização da força bruta dos melhores parâmetros de lags, diferenciação e média móvel sazonal. Apesar de conseguirmos encontrar os parâmetros ótimos com base no BIC, AIC e RMSE, este método demanda alto processamento e consumo de memória, demandando muito tempo de busca.

Importante citar que nem sempre a melhor RMSE representa o melhor modelo, deve-se encontrar um equilíbrio entre as métricas de performance do modelo de acordo com o nosso objetivo.

Como próxima etapa, o desenvolvimento de uma nova busca por otimização de parâmetros será necessária. Dois métodos iniciais foram identificados:

- Ant Colony Optimization
- Genetic Evolution

Desta forma será possível comparar os melhores parâmetros de ordem e ordem sazonal potencializando a otimização e custo computacional.

O link para o github onde contempla os dados e os comandos utilizados nesta análise segue:

https://github.com/Stankevix/StockPrices