Relatório:

Previsão da série temporal do preço do ouro (2000-2021)

Alunos: Andreza (RA: 164213), Gil (RA: 225323) e Yan (RA: 118982)

1. Introdução

Na literatura, encontramos diversos trabalhos para a predição do preço do ouro. Muitos deles utilizam outras variáveis independentes para a previsão, como por exemplo, índices da bolsa de valores, preço de outros metais preciosos, ou então índices econômicos de expectativa dos agentes. Historicamente, o ouro apresenta-se como um ativo de reserva de valor, utilizado em momentos de incerteza nos mercados internacionais. Entretanto, esse ativo também é objeto de especulação de curto prazo no mercado internacional. Dessa forma, seu comportamento é volátil e influenciado por variáveis econômicas de curto e longo prazo.

Mesmo assim, parte da literatura busca prever o preço do ouro apenas com os valores passados, utilizando janelas de previsão para o valor da próxima semana. Este trabalho se encaixa nessa vertente de predição do preço do ouro.

2. Análise exploratória e Pré Processamento

Há uma tendência de crescimento do preço do ouro ao longo do tempo. Na figura 1, podemos verificar o preço desse ativo ao longo dos anos. Esta série claramente não é estacionária, e para análises clássicas de série temporais, necessitamos transformar esta série para uma estacionária. Muitos exemplos de análise de ativos financeiros utilizam o retorno dos ativos para a predição, já que

muitas vezes estas séries são estacionárias. Na figura 2 mostramos a variação do retorno do ouro entre 2000 e 2021. O retorno nada mais é do que a variação percentual do instante atual (t) em relação ao instante anterior (t-1).

Fig 1. Preço do ouro ao longo do tempo (2000-2021)

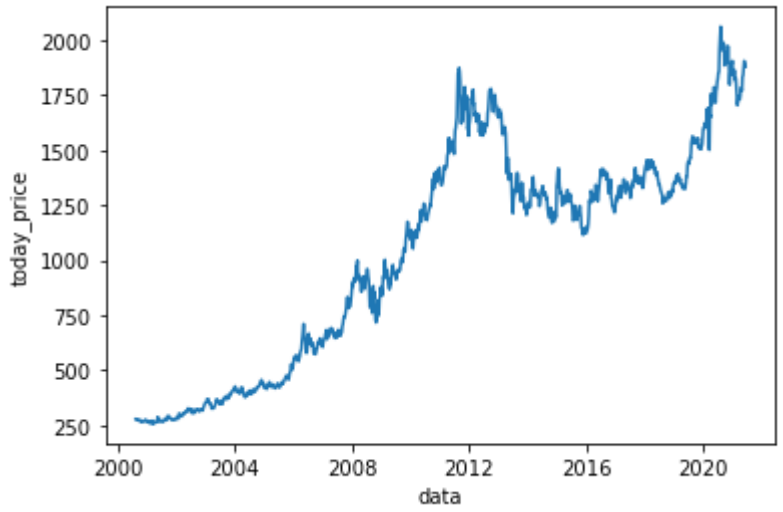
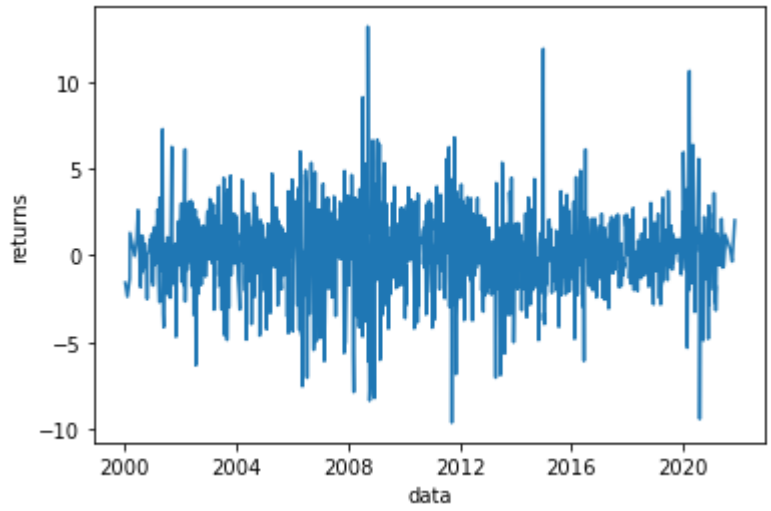


Fig 2. Variação do ouro ao longo do tempo (2000-2021)



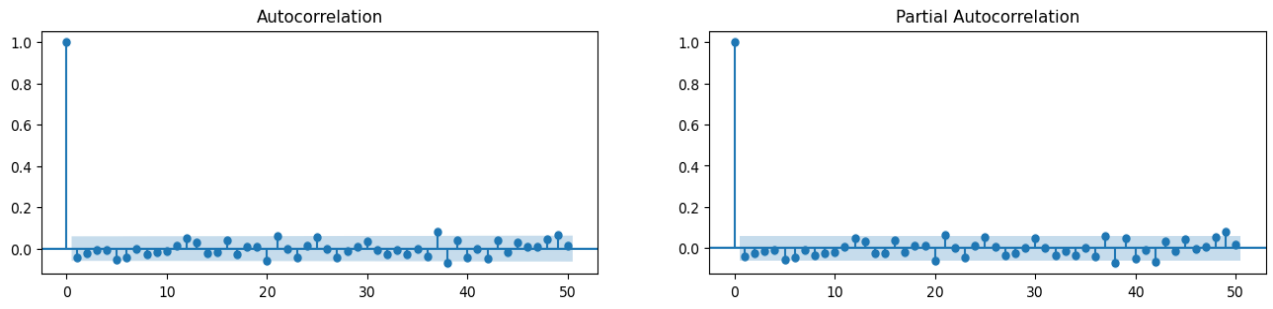
Para checarmos a estacionariedade da série do ouro, aplicamos o teste Aumentado de Dickey–Fuller (ADF). A hipótese nula do teste é que a série não é estacionária, e a hipótese alternativa é que a série é estacionária. Como encontramos um p-valor abaixo da região crítica de 1%, rejeitamos a hipótese nula, confirmando que a série é estacionária.

Fig 3. Estatística do teste Dick-Fuller aumentado.

```
ADF Statistic: -34.47391268771333
p-value: 0.0
Critial Values:
   1%, -3.436341508283391
Critial Values:
   5%, -2.864185524365606
Critial Values:
   10%, -2.5681785627437677
```

A seguir, checamos a autocorrelação e a autocorrelação parcial do retorno dos preços do ouro. Como vemos na figura 4, somente o valor 1 está fora do intervalo de confiança, tanto para a autocorrelação quanto para a autocorrelação parcial. Isso nos indica que o valor do preço do ouro nessa semana apenas está correlacionado com o preço da semana passada. Isso é uma indicação de que a melhor janela de valores para a modelagem provavelmente é apenas 1.

Fig 4.Autocorrelação e Autocorrelação Parcial dos retornos



3. Objetivo e Linha de Base

O objetivo do presente trabalho é treinar modelos e testar suas performances para a predição do preço do ouro na próxima semana, bem como se o preço da próxima semana será maior que o da anterior, caracterizando uma subida do preço, ou não (estabilidade ou queda).

Foi disponibilizado o arquivo **ouro2.csv**, com os preços semanais de fechamento do ouro, iniciando em 18/06/2000, terminando em 13/06/2021, num total de 1096 dados.

Os últimos 100 dados (mais recentes) da série deverão ser separados e usados apenas para teste dos modelos já treinados.

Para a predição do preço do ouro na semana seguinte, o parâmetro para definir qual o melhor modelo, será o RMSE. Consideramos como linha de base, chutar que o valor da semana seguinte será o valor atual, e calcular o RMSE sobre essa predição. Esse cálculo sobre os últimos 100 dados da série, geraram um RMSE de 1925.8675.

Para a predição de subir ou não o preço do ouro na semana seguinte, utilizaremos a acurácia para definir o melhor modelo. Consideramos como linha de base, um classificador “burro”, que apenas considere que o ouro subirá na semana seguinte, já que existe essa tendência do preço do ouro. Dos últimos 100 dados da série, 58 são de subida de preço, e 42, de não subida, gerando uma acurácia de 58% para este classificador “burro”.

4. Construção do modelo de Regressão

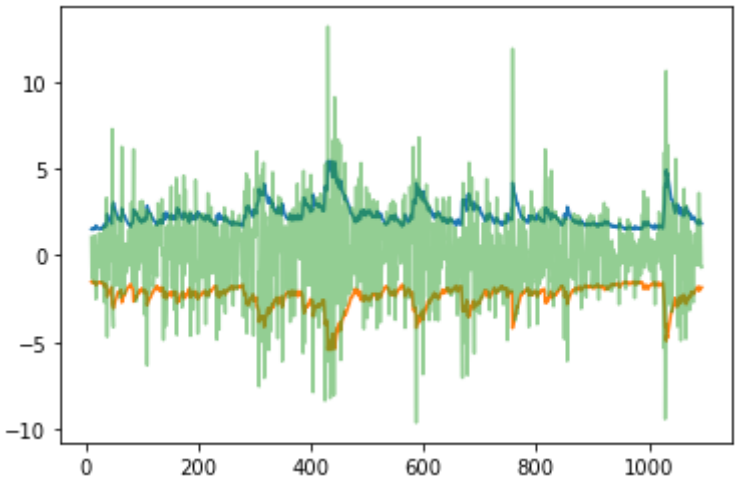
a. ARIMA(p,d,q)

Um modelo tradicional de séries temporais é o modelo ARIMA(p,d,q), que contém componentes autorregressivos (AR), de média-móveis (MA) e de diferenciação (I) . Entretanto, para utilizarmos o modelo ARIMA, pressupõe-se que a série é homocedástica, ou seja, que a variância no instante t-1 não afeta a variância em t. Entretanto, a série do ouro é heterocedástica, ou seja, a variância ao longo do tempo não é constante. Para isso, necessitamos utilizar o modelo GARCH, que é um modelo de heterocedasticidade condicional auto-regressiva generalizada.

b. GARCH(p,q)

O modelo GARCH(p,q) é utilizado na literatura para calcular a volatilidade de uma série heterocedástica. O modelo mais frequente na literatura aplica um modelo ARIMA para previsão da média condicional e um modelo GARCH(1,1) para a variância.

Fig 5. Previsão volatilidade GARCH(1,1) (azul e laranja) e retorno do preço do ouro (verde) para o preço do ouro (2019-2021).



Para nosso trabalho, utilizamos os valores da volatilidade para um regressor Random Forest, para classificarmos a subida ou descida do ouro. Mas antes, testamos alguns modelos para predição.

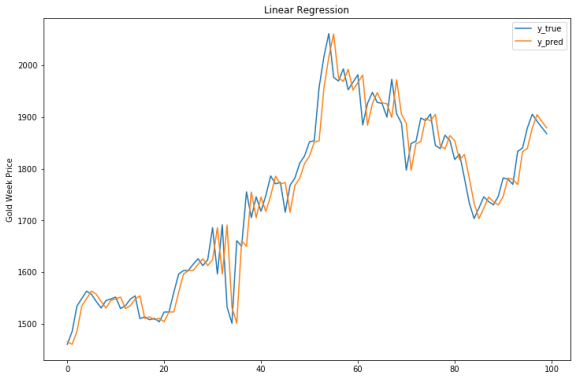
c. Regressão Linear

O modelo de regressão linear foi testado com algumas variações no conjunto de dados. Partimos da forma mais simples, enviando apenas os preços atuais como entrada e pedindo como saída o preço da semana seguinte (#1). A seguir, agregamos aos dados de entrada outras medições, como delta de uma semana para a anterior, médias móveis para janelas de 3, 5, 7 e 9 semanas (m3, m5, m9).

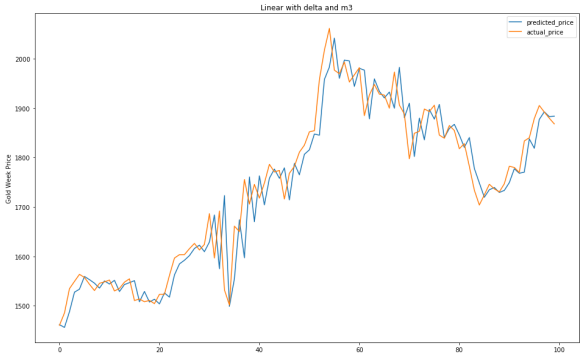
Tab.1 Especificação dos modelos #1 e #2

	entradas	**saída**	**RMSE**	**Acurácia sobre subida da taxa**
#1	preço semana atual	preço semana futura	1923.716	49%
#2	delta e m3	preço semana futura	2304.516	49%

Fig 6. Previsão (linha laranja) e Valor Atual (linha azul) para os dois modelos.



Experimento Regressão Linear #1



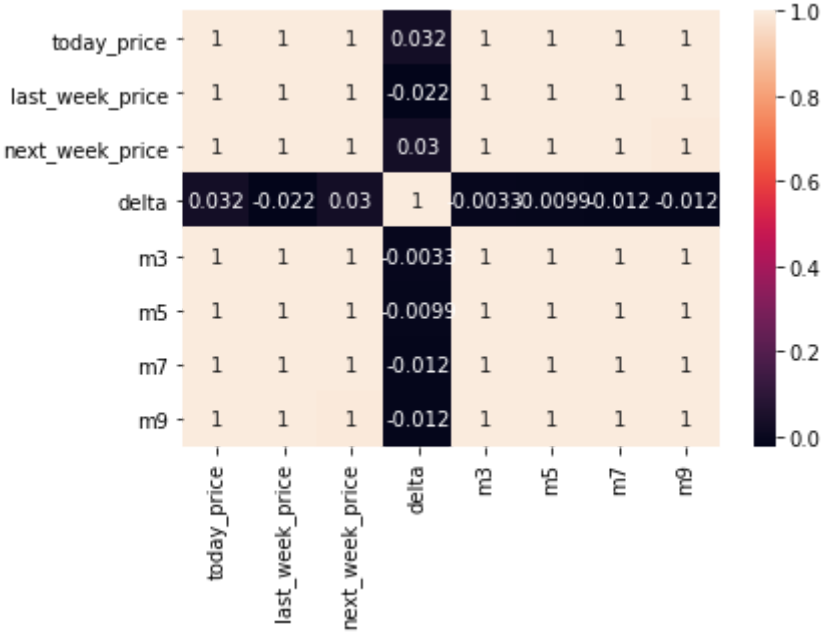
Experimento Regressão Linear #2

Observamos que, ao tentarmos predizer o preço da semana seguinte com essas informações, a predição se comporta com uma semana de atraso.

As penalizações com regularizações Ridge (L2) e Lasso (L1) indicavam, assim como os coeficientes da Regressão Linear sem penalização, que os dados de de longo prazo possuem menor peso, sendo inclusive zerados na regularização L1.

Ao observarmos a matriz de correlação de nossas variáveis, percebemos que as independentes estão completamente correlacionadas. Isso significa que o coeficiente de uma dessas variáveis não capta todo o impacto dessa variável, mas é dividido com o coeficiente da outra variável independente.

Fig 7. Matriz de correlação de Pearson das variáveis.



d. Ridge e Lasso

Para o modelo de regressão com penalidade L2 (Ridge) e L1 (Lasso), foram realizados dois experimentos, sendo o primeiro com o conjunto de entrada composto pelos deltas e o m3 e conjunto de saída o preço da próxima semana e o segundo enviando como entrada os deltas e o m7 e o mesmo conjunto de saída. Para todos os modelos testados utilizamos foi utilizado a classe TimeSeriesSplit do sklearn para gerar os conjuntos de treino e validação.

Tab 2. Resultados dos experimentos com os regressores Ridge e Lasso.

	entradas	**saída**	**RMSE**	**Acurácia sobre subida da taxa**
Ridge	delta e m3	preço semana futura	2293.5397	48%
Ridge	delta e m7	preço semana futura	3599.1081	43%
Lasso	delta e m3	preço semana futura	2293.5397	48%
Lasso	delta e m7	preço semana futura	3599.1081	43%

Como podemos observar na tabela acima e comparando o treinamento de todos os regressores lineares entre si, podemos concluir que eles não são capazes de realizar esse tipo de predição.

e. Random Forest

Utilizamos também um modelo regressor Random Forest para prever o preço do ouro, usando variações semelhantes às da Regressão Linear, aumentando inclusive as janelas ou médias móveis disponíveis para ter mais dados e gerar maior variedade de árvores, e buscando hiperparâmetros que melhor atendessem o caso.

Independente das tentativas, o modelo não logrou êxito sobre o de Regressão Linear, tanto para prever o valor como para a subida ou não do valor, chegando a no máximo 52% de acurácia mas um RMSE de 25413.9876.

f. SVR

O SVR Linear apresentou comportamento semelhante ao da Regressão Linear. Mesmo com a busca por hiperparâmetros, e variações nos dados de entrada, não houve melhora. A adição do Kernel RBF piorou muito a previsão de dos preços, porém, se este regressor for usado para a classificação, gera uma leve melhora, mas sem efeito real sobre um classificador que apenas chute a subida, dentro do conjunto de testes.

5. Construção dos modelos de Classificação

a. Melhor modelo: Random Forest

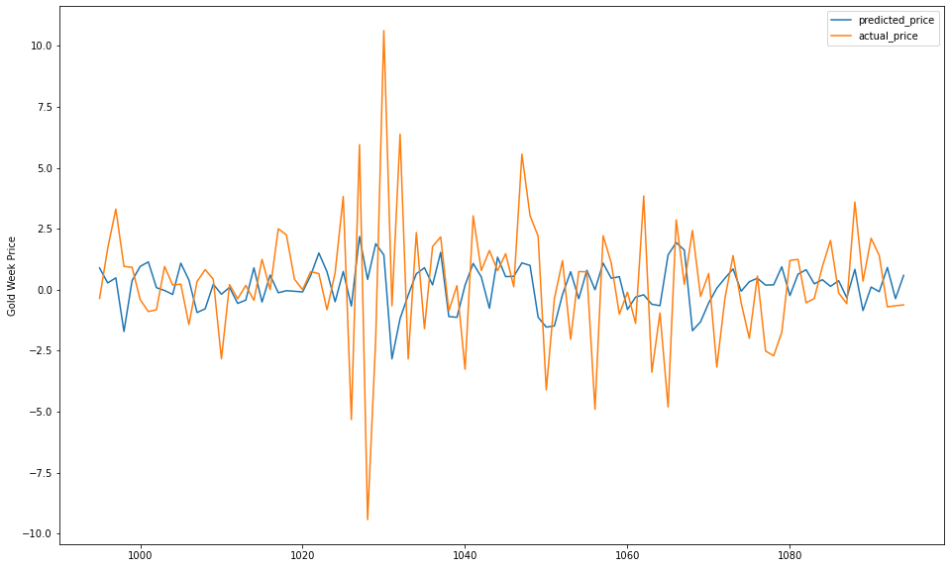
O melhor modelo para a classificação foi, na verdade, um regressor Random Forest com as seguintes variáveis independentes: a variação percentual do ouro em relação ao preço do instante anterior, a média móvel de 3 lags desses retornos e a variância dos retornos calculada pelo modelo GARCH(1,1).

O número ótimo de árvores encontradas foi de 75, apresentando um R^2 baixo: 0.04, o que é um problema pro modelo.

Para os últimos 100 valores, 58 deles eram de subida e 42 de descida. Dessa forma, se o nosso modelo chutasse todos os valores do teste como de subida, obteríamos uma acurácia de 58%. O nosso melhor modelo conseguiu superar essa acurácia, porém seu acerto foi de 60%. Outros modelos que testamos, como o SVC, indicou todos como a classe de subida, sendo irrelevante para nossa análise.

A acurácia do melhor modelo, de 60%, é pouco melhor do que jogar uma moeda ao acaso para decidir se o preço do ouro vai subir ou descer. Todavia, chegamos neste modelo buscando prever a variação do preço do ouro e, se essa variação fosse positiva, a próxima semana seria de alta, e se a variação fosse negativa, aí seria de queda. Abaixo mostramos o gráfico dos nossos valores de predição com os valores atuais para as 100 últimas semanas.

Fig 8. Gráfico do valor predito pelo modelo, em azul, e o valor real, em laranja.



b. Outros: SVR, SVC, Random Forest Classifier, LR

Como já dito, experimentamos outros classificadores, como SVC e Regressão Logística e até outros regressores, como o SVR. Mas, ou eles estimavam tudo como classificação de subida ou eles apresentavam resultados abaixo do obtido pelo Random Forest Regressor, com acurácia abaixo de 50%.

6. Resultados e conclusões

Abaixo, trazemos uma tabela com todos os modelos testados. É importante notar que alguns modelos tentam predizer o preço do ouro na seguinte semana, e apresentam erros RMSE na ordem dos milhares. E há outros que apresentaram um erro RMSE entre 6 e 7, pois estes tentam predizer a variação do preço, e não o preço em si. Embora não seja possível a comparação do RMSE entre modelos com estas duas diferentes saídas, podemos comparar a acurácia em relação à classificação de o preço do ouro subir ou descer. O que percebemos é que os classificadores que mediram a variação do preço, e que incorporaram a variância estimada pelo modelo Garch(1,1) se saíram melhores do que os demais, com uma predição de 60% - acima inclusive do que chutar que o preço sobe toda semana, que daria uma acurácia de 58%.

Tab 3. Especificação de alguns dos modelos testados

	Modelo	**Entradas**	**Saída**	**RMSE**	**Acurácia**
#1	Linear Regression	preço atual	preço futuro	1923.7162	49%
#2	Linear Regression	delta e m3	preço futuro	2293.5397	48%
#3	Random Forest Regressor	variação percentual do ouro, média móvel de 3 dias dessa variação, e variância estimada por GARCH(1,1)	variação percentual do preço do ouro na semana futura	6.1947	60%
#4	Gradient Boosting Regressor	variação percentual do ouro, média móvel de 3 dias dessa variação, e variância estimada por GARCH(1,1)	variação percentual do preço do ouro na semana futura	6.1815	60%
#5	SVR Linear	variação percentual do ouro, média móvel de 3 dias dessa variação, e variância estimada por GARCH(1,1)	variação percentual do preço do ouro na semana futura	6.4249	58%, mas estima todos como subida
#6	Random Forest Classifier	variação percentual do ouro, média móvel de 3 dias dessa variação, e variância estimada por GARCH(1,1)	Target binário de subida ou descida	--	49%
#7	Lasso	preço atual	preço futuro	1923.7163	49%
#8	Random Forest Regressor	preço atual	preço futuro	5770.6604	49%
#9	Decision Tree	delta e m3	Target binário	--	54%

	Classifier		de subida ou descida		
#10	Random Forest	delta e m7	preço futuro	25413.9876	52%

Como conclusão do trabalho, apesar dos esforços de testar diferentes modelos, com diferentes valores de entradas e saídas, não conseguimos um modelo razoável de predição e classificação. Gerar “novas features” com o dado disponibilizado em si, não parece melhorar o resultado dos modelos testados. Como trabalho futuro, seria interessante incorporar outras variáveis econômicas para a construção dos modelos, como valores de ações e outros metais preciosos, bem como explorar novos modelos e suas combinações.