Entrega 2ª Fase

GVCode

Bruno Ting, Ricardo Semião e Victor Dutra

1. O Modelo

O modelo utilizado para realizar a predição proposta na primeira fase foi feito por Python, através do treinamento de uma rede neural, denominada Perceptron Multicamadas (MLP), retirada da biblioteca sklearn ensembles. O modelo foi treinado em um período de 3 anos (2016 até 2018) e tem como variáveis o índice imobiliário, financeiro, energia elétrica, dólar, S&P 500 e Nasdaq 100. Além disso, a rede neural foi parametrizada utilizando o mecanismo de GridSearch, que buscou os fatores ótimos de acordo com o RMSE. A predição do modelo apresentou resultado com 658.082 de RMSE no período de teste da primeira fase, o que representa uma diminuição de aproximadamente 4,12% em relação ao modelo anterior (686.363 de RMSE).

2. Análise da Série

2.1. Outliers e valores NA

Como procedimento padrão, o primeiro passo foi analisar graficamente o fluxo, o que permitiu identificar a presença de muitos outliers, principalmente durante os anos de 2020 e 2021. Dado suas altas magnitudes, iriam impactar negativamente o modelo e, portanto, decidimos tratá-los. Testamos dois métodos: transformar todas as observações fora dos limites superior e inferior de 5 intervalos interquartis (25% a 75%) nos próprios limites; e realizar uma decomposição STL, identificando os outliers no resíduo da decomposição e os substituindo pelo valor da tendência. O primeiro método gerou um modelo final mais preciso. Além disso, como o RMSE pune muito erros grandes, buscamos modelos que evitassem errar muito na presença de outliers, mesmo que errassem um pouco mais na média. Por fim, algumas séries apresentavam valores NA, mas sem nenhum padrão aparente e em pouca frequência, de modo que preenchemos esses valores com os aqueles válidos mais próximos.

2.2. Tendência e sazonalidade

Em seguida, foi analisado a sazonalidade e tendência na série. Em relação a sazonalidade, testamos se alguns fatos estilizados de outras variáveis financeiras estavam presentes (como efeito dos dias da semana, trimestres do ano etc.), mas em sua maioria não foram significantes. Além disso, modelos ARMA com parte sazonal não se mostraram melhor que os sem sazonalidade em explicar o fluxo da bolsa. Por fim, limpar a série da sazonalidade não melhorou o desempenho de nosso modelo final.

Já sobre a tendência, a série era centrada no zero, e como esperado, o teste de Dikey-Fuller indicou ausência de tendências estocásticas, o que nos fez descartar modelos de cointegração. Ao redor de 2019 a tendência parece mudar, indicando uma possível quebra estrutural.

2.3. Quebras estruturais

Seguindo o estudo da tendência, foram analisadas as quebras estruturais com regressões segmentadas e identificamos diferenças significantes na tendência a partir de 2019. Isso pode indicar uma mudança no processo gerador dos dados ou um período de maior instabilidade. Assim, foram testados três métodos diferentes: treinar o modelo apenas com os dados "limpos" das tendências identificadas; usar apenas dados pré 2019; e apenas dados pós 2019. Dentre as várias opções testadas, o modelo treinado no período de 2016 a 2018 foi o que apresentou melhores resultados. Além disso, foi utilizado somente três anos pois, quando se estendia o treinamento para períodos mais antigos, a predição apresentava um erro maior, dado que estava se baseando em tendências diferentes das atuais.

3. Seleção de variáveis

Em seguida, analisamos a correlação com o fluxo de mercado para determinar quais variáveis seriam utilizadas no modelo multivariável por rede neural. Para isso, foram feitos 3 testes diferentes, sendo eles a correlação simples de Pearson, testes estatísticos univariáveis (SelectKBest) e uma permutação de features por árvores de decisão randomizadas (ExtraTrees). Os resultados demonstraram que, dentre as variáveis fornecidas na primeira etapa do desafio, aquela que mais se correlacionava com o fluxo de mercado no período de 2012 a 2021 foi o índice Ibovespa. Contudo, buscando outros indicadores representativos do mercado de investimento, o grupo descobriu que, ao decompor o Ibovespa por segmento e setor, foi possível obter uma correlação elevada e melhores resultados da predição nos índices imobiliário (IMOB), financeiro (IFNC) e energia elétrica (IEE), com destaque no setor de Real Estate. Os dados foram facilmente coletados através da API e site do Yahoo Finance.

Como a teoria econômica indica que os agentes devem responder a mudanças nos mercados e moedas externas alterando a quantidade de dinheiro investida em cada bolsa, buscamos incluir índices da bolsa e câmbio da zona do euro, Inglaterra, Australia, China, Japão e Estados Unidos. Dentre eles, os que obtiveram melhores resultados na predição foram os índices americanos, sendo eles o dólar, S&P 500 e Nasdaq 100. Outras variáveis foram testadas no modelo, porém não foram consideradas por conta da baixa relevância estatística ou por impactarem negativamente na acurácia do modelo, sendo elas a taxa de desemprego, IPCA, SELIC, SELIC descontada, PIB, Fed Funds Rate e Dow Jones Industrial Average. O índice Ibovespa também foi descartado para simplificar o modelo pois tinha alta covariância com os demais índices da bolsa.

4. Seleção do modelo

4.1. Modelos univariados

Para ajudar a entender a dinâmica própria da série, analisamos os modelos univariados SARIMA, GARCH, e rede neural autorregressiva. Todos fitaram bem os dados, o GARCH lidou bem com a mudança de volatilidade na série, a rede neural se destacou - indicando que a dinâmica da série possa ser não linear – mas nenhum deles foi útil para fazer previsões tão longas (300 passos), já que acumulam os erros de previsão. Deste modo, foi preferível dar foco nos modelos multivariados.

4.2. Modelos multivariados

A fim de determinar qual o melhor modelo a ser utilizado na previsão da série temporal, foi realizada uma validação cruzada por K-Fold de diversos modelos de regressão multivariados, sendo eles: ARDL, VAR, Support Vector (SVR), k-Nearest Neighbors (KNN), Decision Tree, Extreme Gradient Boosting (XGBoost), Multiayer Perceptron (MLP) e Gradient Boosting. Além disso, também foi montado um modelo de ensemble por stacking, utilizando alguns dos modelos citados como base e uma regressão linear como meta learner. Após realizado os testes de validação, foi possível determinar que a rede neural MLP apresenta o menor RMSE entre os modelos testados e, portanto, é a mais adequada para a predição. Além disso, a rede neural foi parametrizada utilizando GridSearch, uma ferramente da biblioteca sklearn que realiza diversos testes para cada combinação de parâmetros fornecidos. Nessa busca, foi modificado os números de neurônios nas camadas escondidas, função de ativação, solver para otimização de pesos, parâmetro de regularização (alpha) e a taxa de aprendizagem. Dessa forma, foi possível construir um modelo muito bem adaptado para realizar a predição do fluxo de mercado.

Com isso, pode-se concluir que uma rede neural bem treinada e parametrizada geralmente traz um maior desempenho do que até mesmo diversos modelos juntos (modelo de stacking), quando se trata de bases de dados mais complexas.

5. Melhorias no modelo

Como foi dito na seção 1 do relatório, o modelo apresentou uma diminuição de aproximadamente 4,12% do RMSE em relação ao modelo passado, entregue na primeira etapa do desafio. Isso pode ser

explicado principalmente pelas alterações no período de treino, seleção de variáveis e parametrização.

Primeiramente, o grupo descobriu que o período utilizado para o treinamento do modelo - 2018 à 2021 - não era adequado para a predição, uma vez que estava treinando sobre uma tendência anormal relacionada à pandemia do Covid-19. Durante essa crise, o valor do fluxo de mercado apresentou um desvio padrão muito elevado, o que possivelmente impactou negativamente na predição do período pós pandêmico. Assim, foi preferível treinar no período de 2016 à 2018, que apresenta dados relativamente recentes, porém antes do início da pandemia.

Em seguida, foi percebido que a seleção das variáveis não estava levando em conta o mercado externo. Portanto, o grupo expandiu a análise de correlação para diversos mercados estrangeiros e determinou que a utilização dos índices americanos S&P 500, dólar e Nasdaq 100 ocasionavam melhores resultados no modelo.

Por fim, a rede neural MLP do primeiro modelo não estava parametrizada de acordo com os fatores ótimos para a predição do fluxo de mercado. Para resolver esse problema, foi realizado uma busca pelos melhores parâmetros através do GridSearch, que proporcionou um modelo mais bem adaptado para realizar a predição.

6. Considerações finais

O principal objetivo desse trabalho foi desenvolver um modelo de predição que pudesse, através de variáveis explicativas, predizer o fluxo de dinheiro investido dentro dos fundos de ações para um longo período e realizar uma análise sobre o impacto que os índices têm sobre o mercado de ações, a fim de proporcionar uma visão mais clara ao investidor**.**

Com isso em mente, o grupo conseguiu desenvolver um modelo preditivo com um bom desempenho através de um mecanismo complexo de rede neural que, mesmo assim, é fácil de ser utilizado em qualquer máquina em poucos minutos. Além disso, os dados fornecidos para o treinamento do modelo são de rápido acesso, por meio do Yahoo Finance, que oferece dados atualizados para todos os índices utilizados. O modelo desenvolvido pode ser utilizado para aprofundar o entendimento sobre as interrelações entre as diversas variáveis que influenciam o mercado de investimentos, auxiliando o profissional experiente a desenvolver cenários futuros baseados em suas predições, como a construção de cenários pessimistas, realistas ou otimistas. Contudo, em um cenário real de previsão, as covariadas também teriam que ser previstas e a inclusão dessa incerteza teria que ser levada em conta ao se tomar decisões com base no modelo. Ademais, seria interessante recalibrar o modelo de acordo com o período de previsão exigido.

Em relação aos insights extraídos do projeto, pode-se destacar a dificuldade e complexidade de se prever o fluxo de mercado, visto que existem inúmeras variáveis que afetam o seu valor que muitas vezes não podem ser previstas, tais como a pandemia de Covid-19. Isso é evidenciado no alto número de outliers, que possivelmente representam ocasiões em que fatores anormais agiram para gerar uma grande variação no fluxo. Adicionalmente, as moedas e índices de bolsa de outros países não melhoraram o modelo final - com exceção do dólar, NASDAQ e S&P500 - indicando que, no âmbito da previsão, a maior parte da dinâmica entre os investidores brasileiros e o cenário internacional foi captada pelos índices norte-americanos. Outro ponto que chamou a atenção do grupo foi o impacto que a pandemia causou no cenário macroeconômico. Durante o teste de correlação, percebeu-se que certas variáveis que explicavam bem o fluxo do mercado no período pré-pandêmico impactam negativamente a predição do período pós-pandêmico, o que demonstra uma mudança na dinâmica do mercado de investimentos.

Por fim, analisando o produto final feito pelo grupo, pensamos em otimizar o código até o final do desafio, buscando trazer menor uso de processamento e tempo, e deixá-lo mais limpo, a fim de que possa ser facilmente entendido e adaptado por outras pessoas que desejem utilizar o programa. Além disso, o modelo pode ser aperfeiçoado a partir da introdução de variáveis mais correlacionadas com o fluxo de mercado no período pós-pandêmico pois, analisando o gráfico da base de teste, algumas tendências futuras não foram identificadas pelo modelo, evidenciando a existência de variáveis inexploradas.