Sistemas Inteligentes

Relatório final

Predição de valor de Cotas de Fundos através de árvores de decisão

Luneque Del Rio de Souza e Silva Junior

Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas - UFABC

Beatriz Nogueira Costa 11058515

Igor Bandim de Oliveira 11064013

Gabriel Sena de Queiroz 11007815

Gabriel Mesquita de Souza 11057015

Dezembro de 2019

Santo André

1. Introdução

Neste projeto foi utilizada a biblioteca de software de código aberto *XGBoost* para Python para, através de árvores de decisão, prever o valor de cotas de fundos de investimento. Foi utilizado o Jupyter Notebook com Python 3, assim como utilizado para os laboratórios da disciplina.

O *dataset* utilizado possui as seguintes *features*:

- Valores de índices de mercado (IPCA, taxa Selic, Ibovespa, taxa de câmbio e dólar);
- Características de fundos investimentos;
- Características de gestoras responsáveis por estratégia do fundo;
- Valores da cota em dias anteriores.

Através do *XGBoost* foi utilizada a regressão obtendo, assim, o valor da cota para novas entradas. Além disso, foi utilizado um algoritmo de *shapley* para entender quais variáveis influenciam de forma mais significativa no modelo final.

2. Justificativa

Fundos de investimentos são uma categoria de produto de investimento voltados para diversificação de portfólio. Nessa classes de ativo os clientes podem investir em diversos outros produtos de forma indireta, sem de fato precisar acompanhar as oscilações de mercado; uma vez que a administração do fundo é realizado por uma gestora com experiência de mercado e estratégia de alocação bem definida.

Ao aplicar em um fundo, o cliente compra uma quantidade de cotas do mesmo. Seu valor é atualizado diariamente com base na performance do fundo e, dentre outros fatores, define a rentabilidade final do cliente.

Já as gestoras reúnem o capital de seus cotistas e realizam aplicações em diversas classes de ativos. Por exemplo, produtos de renda fixa, renda variável, operações estruturadas ou até mesmo em outros fundos de investimento que não são tão acessíveis ao público. Em geral, apesar dos fundos abrirem a sua estratégia

de alocação de forma mais ampla, é desconhecida a posição dos fundos em cada classe de ativo que os mesmos aplicam.

Existem áreas da economia que buscam realizar projeções de indicadores econômicos. Alinhando tais projeções ao modelo de predição, em um cenário ideal, seria possível propor estratégias de recomendações de fundos para clientes que se adaptassem a sua suscetibilidade à riscos, ao seu capital de aplicação e às suas necessidades de liquidez de forma mais sólida.

3. Objetivo

Utilizar indicadores de mercado, características dos fundos e o histórico dos valores de cota dos últimos dias para prever o valor da cota em uma nova data. A hipótese é que indicadores de mercado, alinhados à características macro dos fundo dão subsídio para prever oscilações bruscas nos valores das cotas.

4. Metodologia

O *XGBoost* é um algoritmo baseado em árvore de decisão e utiliza uma estrutura de *Gradient Boosting*. É utilizado quando deseja-se velocidade e performance. É recomendado para este projeto, uma vez que o *dataset* trata-se de dados estruturados.

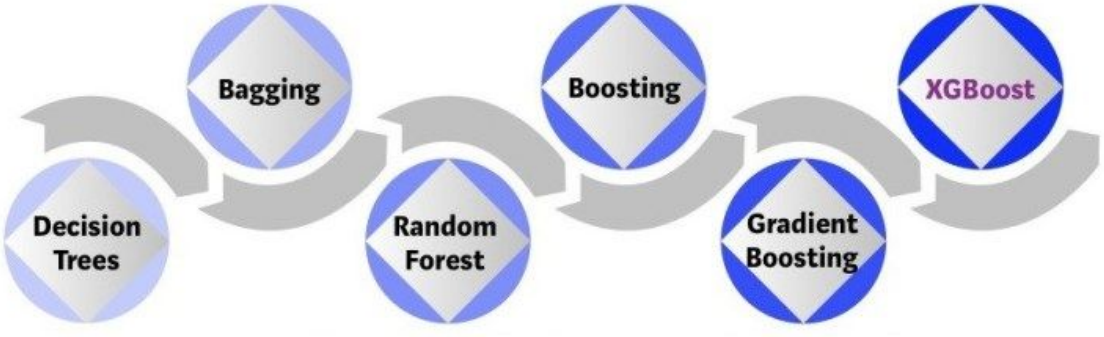


Figura 1 - Evolução dos algoritmos.

Há três formas de *Gradient Boosting* suportadas:

- *Gradient Boosting;*
- *Stochastic Gradient Boosting;*
- *Regularized Gradient Boosting.*

O algoritmo usa uma técnica de *ensemble* onde novos modelos são adicionados para corrigir erros realizados pelos modelos existentes. Isso é realizado de forma sequencial até não haver melhorias possíveis.

O algoritmo de *Shapley*, também *Gale-Shapley*, foi criado por David Gale e Lloyd Shapley para resolver o problema do emparelhamento estável.

Um exemplo em que o problema do emparelhamento estável surge é em uma rede de processos seletivos para universidades. Supõe-se que um determinado estudante foi aceito pela universidade A e encontra-se na lista de espera da universidade B (que ele prefere em relação à A). Caso, por algum motivo, o estudante venha a ser chamado pela universidade B, ele pode abandonar a universidade A para estudar na universidade B, e deixando A com uma vaga ociosa. Isso levaria A a chamar algum candidato de sua lista de espera, e deixaria uma vaga ociosa em alguma outra universidade. Essa situação pode fazer com que o processo seletivo seja finalizado com vagas ociosas em algumas universidades e alunos interessados nelas, mas que não foram aceitos em nenhuma universidade.

O principal objetivo do algoritmo é encontrar uma solução em que ambos os grupos dentro do emparelhamento sejam distribuídos da melhor forma conforme suas necessidades. Em suma, David Gale e Lloyd Shapley provaram que em casos em que dois conjuntos são iguais sempre há formas de criar um emparelhamento estável (onde as necessidades dos grupos são atendidas).

4.1 - Análise dos dados

Antes de partir para o treinamento do nosso principal objetivo, foi necessário fazer uma análise exploratória e gráfica sobre o dataset em si, e a correlação entre suas features.

Para tal análise foram criados seis notebooks, com seus respectivos objetivos:

1. Analising_funds_general: obter informações gerais acerca do dataset, e a correlação das features de todos os fundos com o valor da cota no dia atual.
2. Analising_funds_fix: Análise gráfica para os fundos de renda fixa.
3. Analising_funds_foreign: Análise gráfica para os fundos de investimento estrangeiro.
4. Analising_funds_forexInvesting: Análise gráfica para os fundos cambiais.
5. Analising_funds_stocks: Análise gráfica para os fundos de ações.
6. Analising_funds_multimarket: Análise gráfica para os fundos de multimercado.

A análise exploratória e gráfica nos trouxe informações importantes para o conhecimento geral do dataset, no entanto não foi tão clara no seu papel de mostrar a influência de cada variável no valor da cota para o dia atual, nos indicando a importância da utilização do Shapley em cumprir essa tarefa.

A exploração dos dados nos permitiu visualizar que se trata de um dataset com 136.924 linhas, que tratam das observações de cada cota em períodos variados, separados em 61 colunas, referindo-se ao target que pretendemos prever (valor da cota no dia atual), e outras 60 features separadas em:

Taxas: IPCA, Selic, Câmbio;

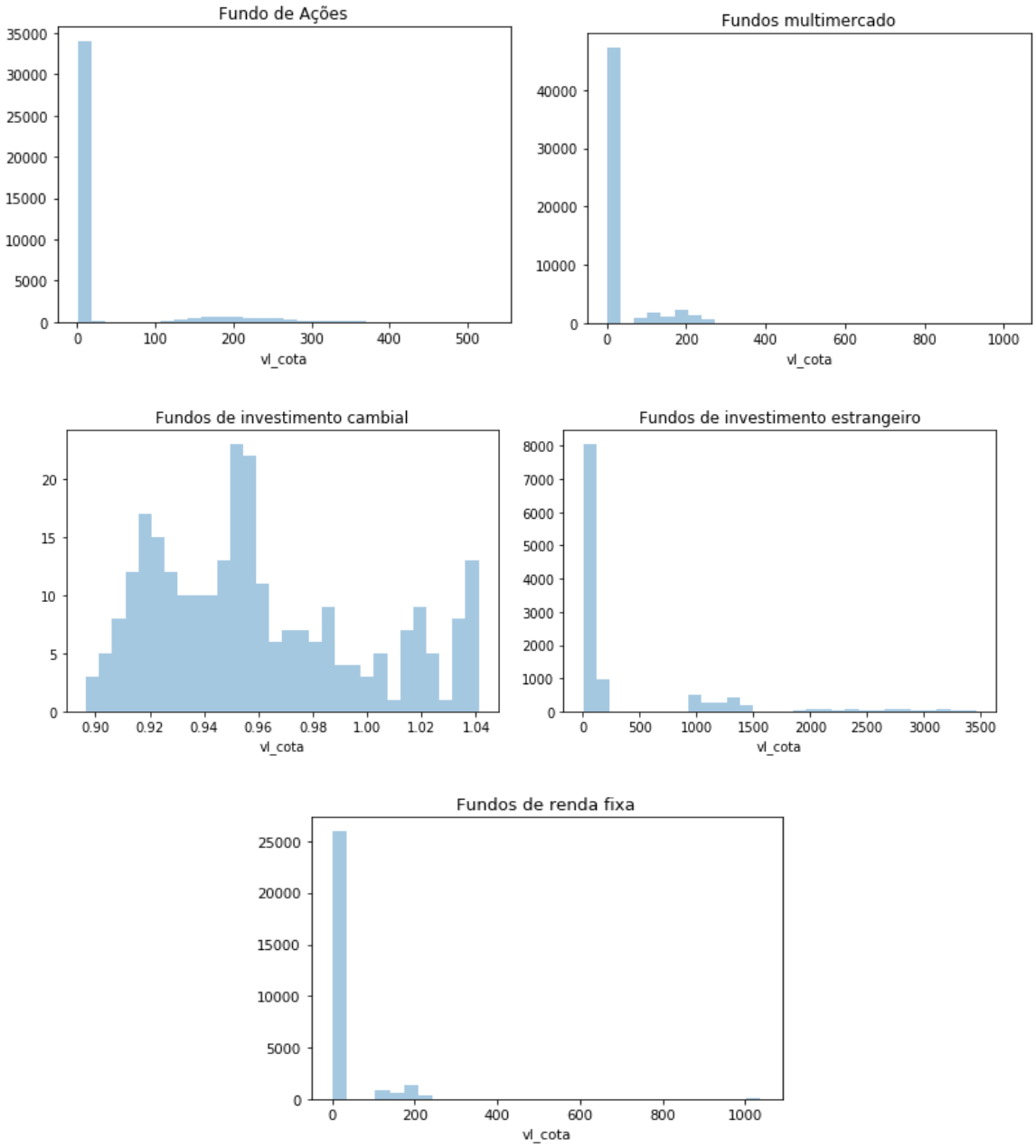
Bolsas: Índice Bovespa e Down Jones;

Métricas de avaliacao dos fundos;

Indicador binário de sugestao (indicado, não indicado);

Tipos de fundos (renda fixa, ações, multimercado, cambial e estrangeiro).

É importante ressaltar que o dataset continha os valores de até 5 dias anteriores para cada indicador (inclusive o valor da cota), sendo que cada dia estava em uma coluna. Tendo isso em mente, foi visto que a análise entre a variação das features com a variação do valor das cotas fez mais sentido, haja visto que os valores brutos se mostravam muito dispersos, e o principal influenciador eram os próprios valores da cota dos dias anteriores, algo já esperado.



Figuras 1: Histogramas dos valores da cota para cada tipo de fundo de investimento.

Informações importantes obtidas na análise exploratória podem ser vistas na tabela abaixo:

Parametros/ Fundos	Cambial	Ações	Renda fixa	Externo	Multimercado
Máximo	1.04098751	528.11445777	1037.0837469	3461.128376	1019.1525562
Mínimo	0.89652087	0.87522276	0.99869844	0.83707672	0.9458787

Média	0.959955063	33.904698648	26.429126907	354.58576703	28.2859596232841
Desvio padrão	0.03878380	83.47285464	88.32650211219	717.37019094	69.9430620410002

Tabela 1: Parâmetros importantes do dataset, dividido por tipos de fundos

A melhor distribuição dos dados encontra-se nos fundos de investimentos cambiais, no entanto isso é justificado pela pequena quantidade desse tipo de fundo no dataset, tal restrição fez com que esse grupo fosse retirado do treinamento do nosso modelo, apesar disso foi possível retirar um insight com relação a variação dos seus valores de cota: estão inversamente relacionados com a variação dos valores das bolsas, como pode ser visto no conjunto de figuras abaixo.

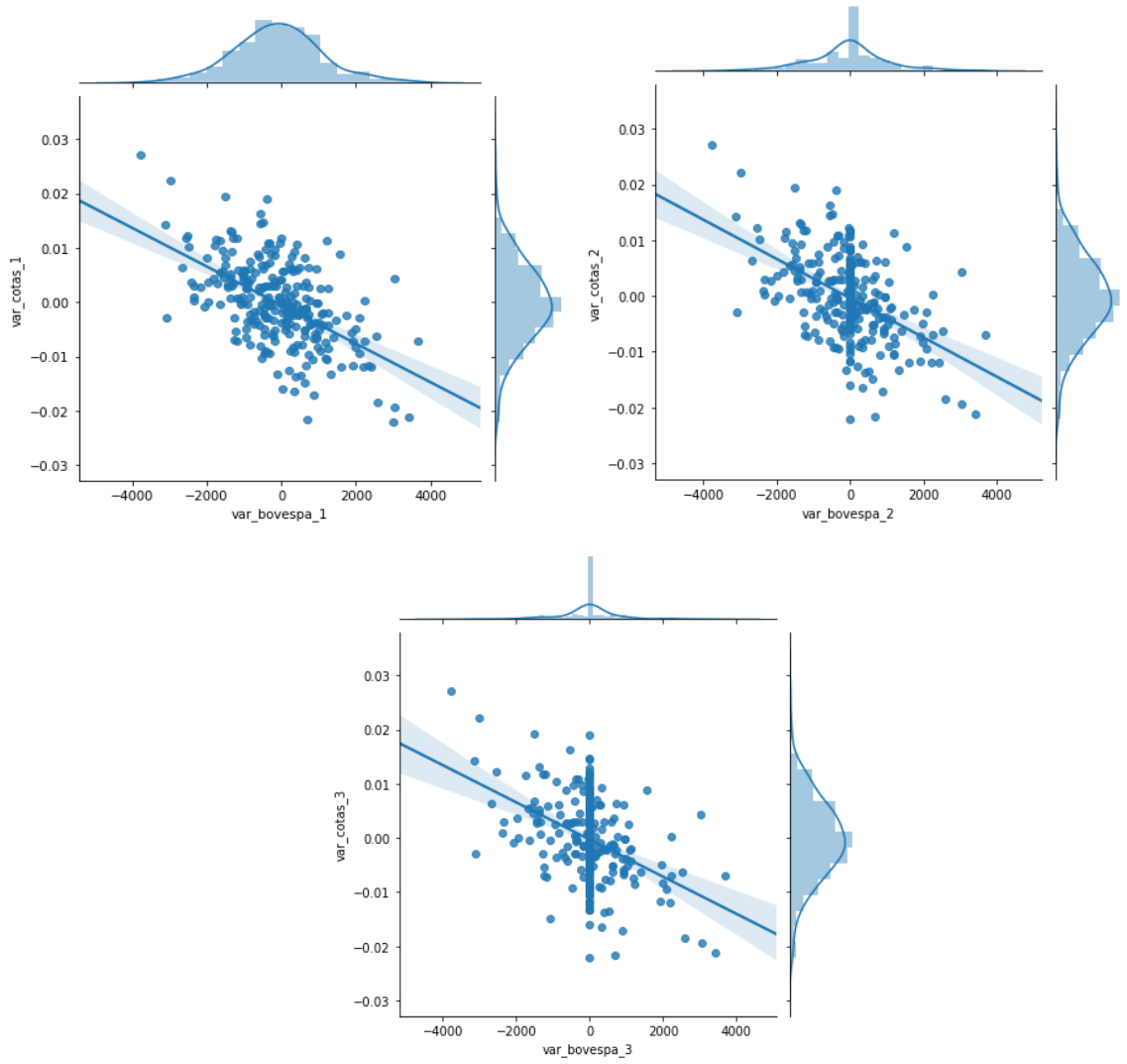


Figura 2: Pairplot entre as variações do índice Bovesp x Valores da cota para os fundos de investimentos cambiais.

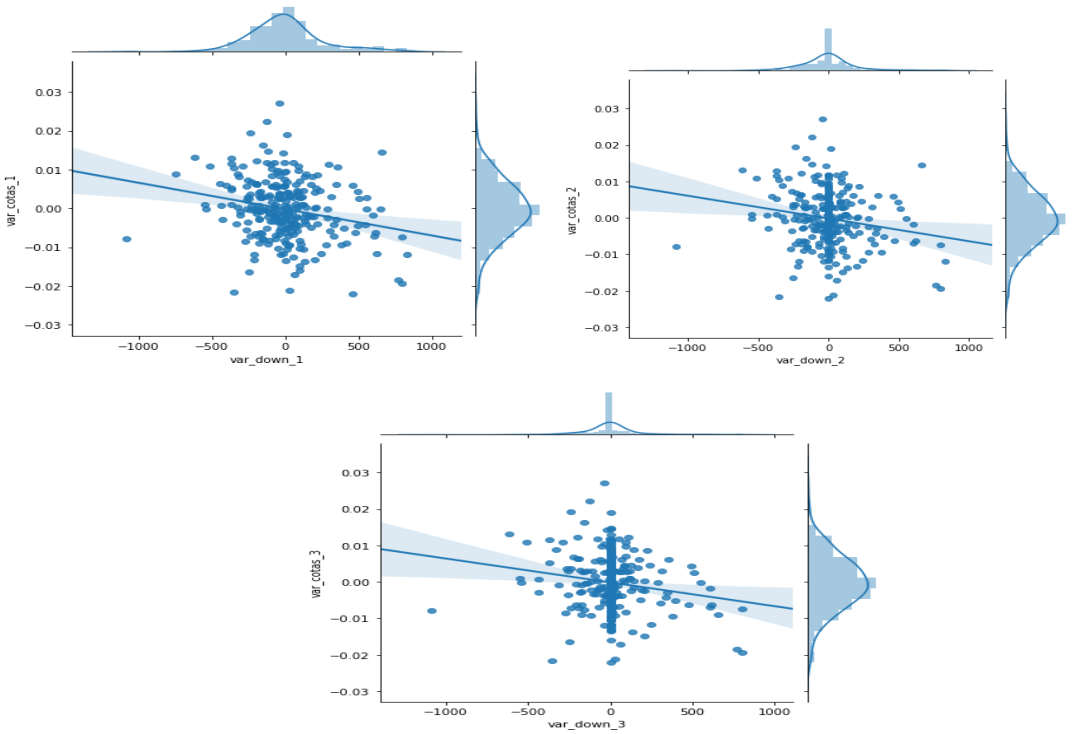


Figura 3: Pairplot entre as variações do Down Jones x Valores da cota para os fundos de investimentos cambiais.

4.2 - Treinamento do modelo

Para solução do problema proposto foram criados quatro *notebooks*, a saber:

1. 1_modelo_previsao_fundos_acoes.ipynb
2. 2_modelo_previsao_fundos_multi.ipynb
3. 3_modelo_previsao_fundos_extern.ipynb
4. 4_modelo_previsao_fundos_renda_fixa.ipynb

Além disso, foi utilizado o dataset *base_final_fundos.csv*.

Foi necessário fazer uso das bibliotecas pandas, sklearn, datetime, numpy, xgboost e shap.

Assim como na análise, o treinamento foi dividido para cada tipo de fundo, trabalhando para prever a variação percentual do valor da cota de fundos de ação

em relação ao valor de cinco dias anteriores foi utilizado a árvore de decisão *XGBoost.*

Com o intuito de analisar o modelo, e evitar overfitting de apenas algumas variáveis estarem predominando na predição dos valores, foram realizadas três formas de ajustar os dados para o treinamento do modelo:

1. Regressão utilizando os valores das variações percentuais dos dias anteriores (sempre com base no valor da cota há seis dias atrás);
2. Regressão anterior, excluindo os valores das variações percentuais dos dias anteriores (puramente com base no uso de indicadores);
3. Regressão anterior, excluindo os valores brutos dos índices de mercado.

Com base nisso, ao decorrer da descrição, os treinamentos serão citados pelos seus respectivos nomes de dataset utilizados: dataset 1 (dataset com valores das cotas anteriores), dataset 2 (dataset somente com os indicadores de mercado) e dataset 3 (exclusão dos valores brutos dos índices de mercado).

O dataset foi dividido em seis folds para realizar cada análise. Para verificar se não há variações abruptas do erro médio quadrático os cinco primeiros folds serão utilizados em um cross validation (Tabela 1). Uma vez validado, o sexto fold será utilizado para métrica de acurácia final do modelo.

Também foi utilizada a função *RandomizedSearchCV* para realizar a estimativa dos melhores parâmetros do *XGBoost* dentro uma lista de parâmetros específicos passados com base em um *score* pré definido.

Uma vez que o resultado obtido foi satisfatório para o *cross validation*, sem muita variação entre os *folds*, foi possível utilizar todos os folds para rodar um modelo final e verificar o seu erro médio quadrático.

Para verificar como cada variável explica o modelo foi utilizado o *features_importance* do *XGBoost.* Exemplificando, os resultados para o fundo de investimento em ações utilizando os valores das cotas dos dias anteriores (dataset 1), podem ser observados na Figura 1 abaixo.

	feature	importance
77	var_perc_vl_cota_d1	0.355258
76	var_perc_vl_cota_d2	0.0490165
52	var_perc_bovespa_d1	0.0365812
43	var_perc_selic_d5	0.0241613
13	dolar_d2	0.0206933
62	var_perc_dow_30_d1	0.0175119
75	var_perc_vl_cota_d3	0.0167299
22	dow_30_d5	0.0138078
32	tx_cambio_d3	0.0128718
71	var_perc_tx_cambio_d2	0.0123748

Figura 4 - Importância de cada variável para a predição utilizando *features_importance*.

Também foi utilizado o *shap* para visualizar a importância dos valores das features no modelo, tal agoritmo que foi detalhado anteriormente possibilitou uma visão ampla sobre a influência de cada feature no modelo de regressão. Como exemplo, o resultado para o fundo de mercado de ações com o dataset 1, pode ser observado na Figura 2 abaixo.

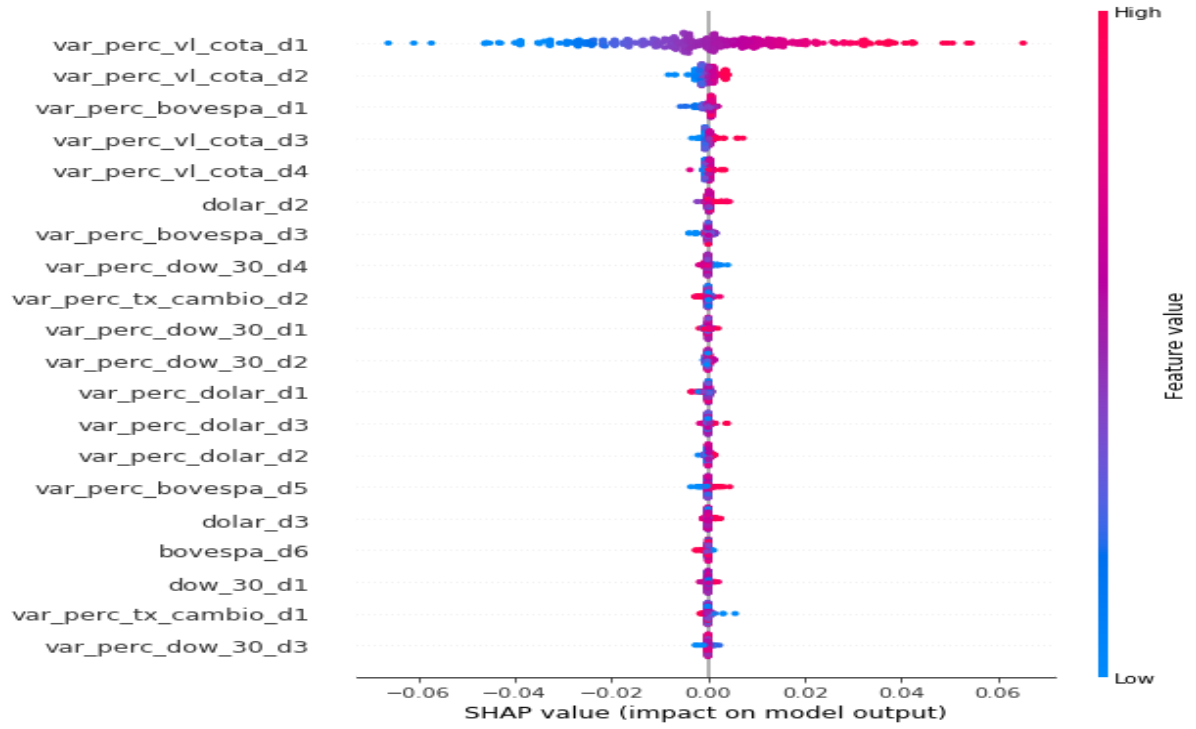


Figura 5 - Importância de cada variável para a predição utilizando *Shapley*.

Pode ser observado que o valor das cotas é um preditor de importância superior às demais *features*. Por isso, foi-se necessário realizar nova predição sem essa variável (e com as mesmas premissas) de forma a avaliar novamente os resultados.

Nas tabelas a seguir será possível detalhar os valores obtidos do mse final (erro médio quadrático final) para cada fundo, com seu respectivo dataset utilizado.

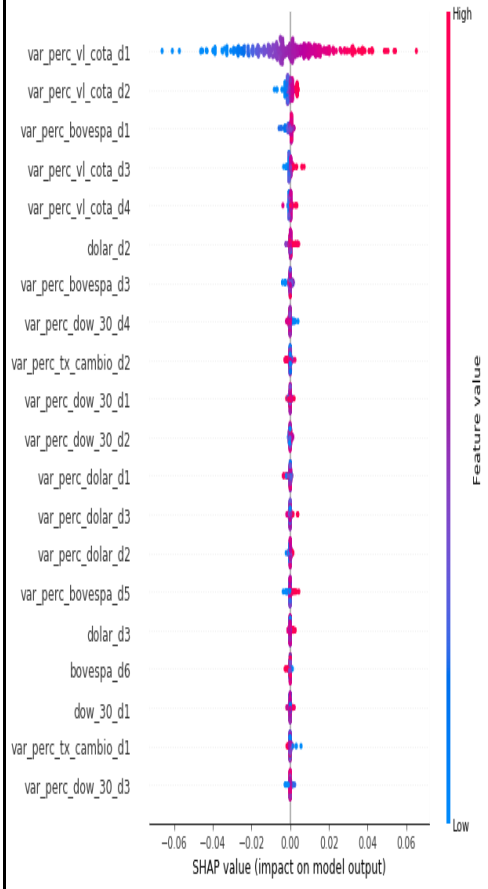
Fundo/Método	Dataset 1	Dataset 2	Dataset 3
Ações	0.000158823604767	0.000353050795060	0.000376661755808
Multimercado	0.000763720856724	0.000816474188802	0.000261101844510
Estrangeiro	0.000027549785668	0.000184171224019	0.000182470416107
Renda Fixa	0.000569405640306	0.000572013553859	0.000012178144461

Tabela 2 - MSE variando com a retirada de variáveis (diferentes datasets), para cada tipo de fundo.

Após verificar que o modelo está funcionando com ótimos valores percentuais de erro, a próxima análise a ser mostrada são a importância das features na regressão. O resultado pode ser visto com as imagens que a aplicação do Shapley nos fornece, além da função features_explorer, para cada tipo de fundo:

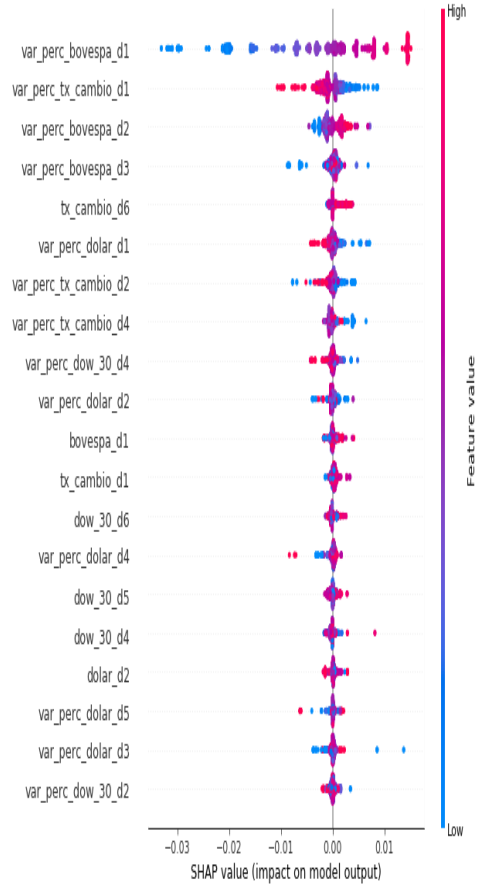
Fundos de ações

Dataset 1 (com valores das cotas anteriores)

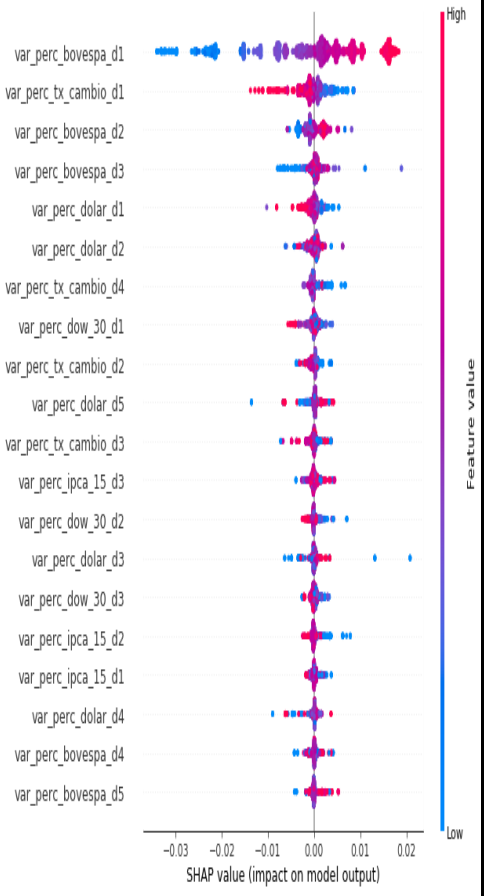


	feature	importance
77	var_perc_vl_cota_d1	0.355258
76	var_perc_vl_cota_d2	0.0490165
52	var_perc_bovespa_d1	0.0365812
43	var_perc_selic_d5	0.0241613
13	dolar_d2	0.0206933
62	var_perc_dow_30_d1	0.0175119
75	var_perc_vl_cota_d3	0.0167299
22	dow_30_d5	0.0138078
32	tx_cambio_d3	0.0128718
71	var_perc_tx_cambio_d2	0.0123748

Dataset 2 (sem valores das cotas)



Dataset 3 (Sem valores brutos)

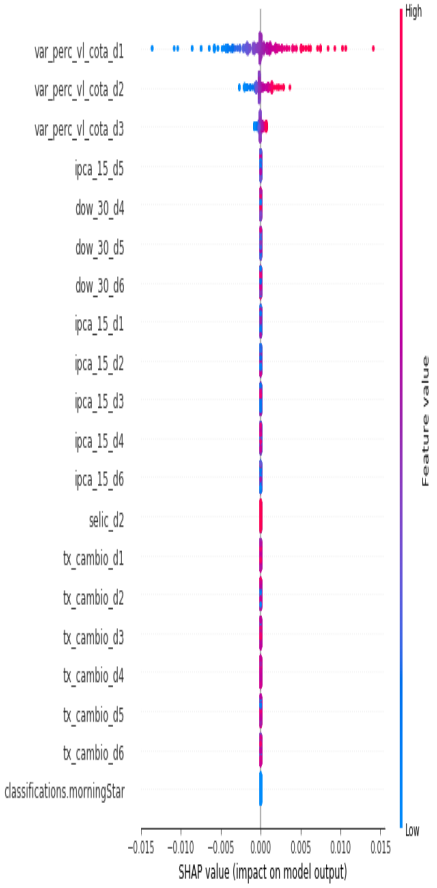


	feature	importance
16	var_perc_bovespa_d1	0.234549
36	var_perc_tx_cambio_d1	0.0611117
15	var_perc_bovespa_d2	0.0546684
34	var_perc_tx_cambio_d3	0.0374101
17	var_perc_dolar_d5	0.0364455
29	var_perc_ipca_15_d3	0.0326045
33	var_perc_tx_cambio_d4	0.0320974
31	var_perc_ipca_15_d1	0.0318745
21	var_perc_dolar_d1	0.0298807
30	var_perc_ipca_15_d2	0.029757

Tabela 3 - Importância das features para os fundos de ações..

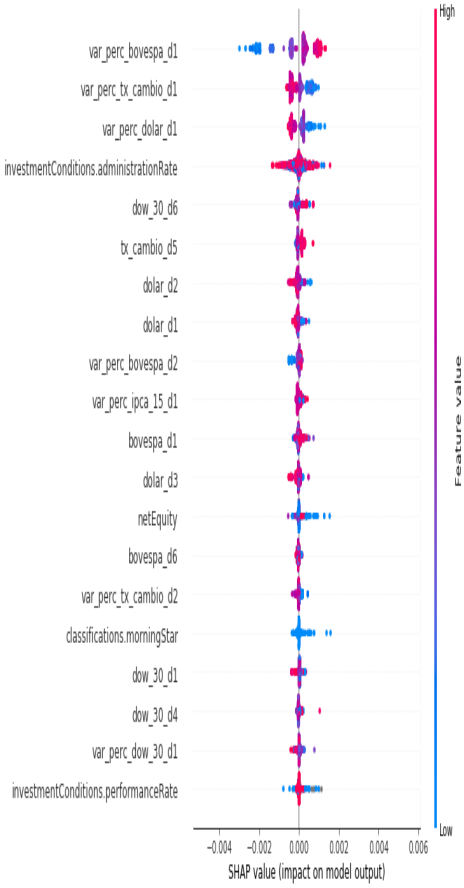
Fundos multimercado

Dataset 1



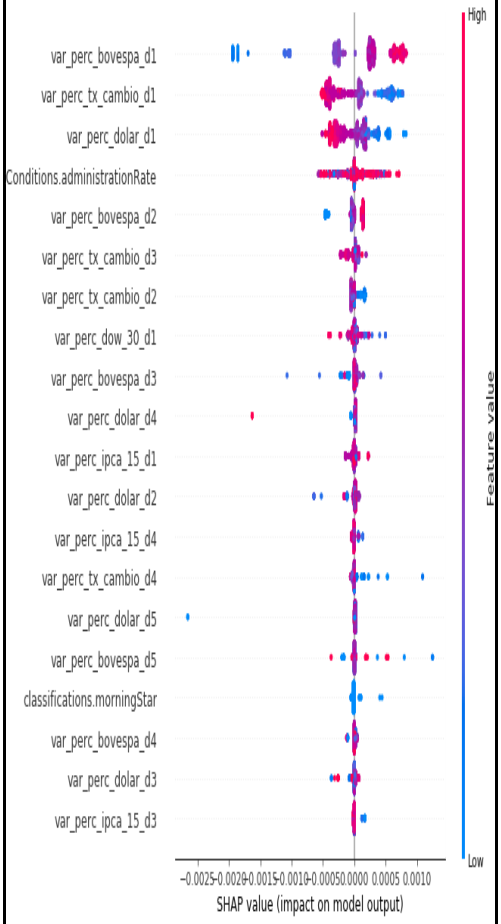
	feature	importance
77	var_perc_vl_cota_d1	0.223323
76	var_perc_vl_cota_d2	0.130544
62	var_perc_dow_30_d1	0.118541
15	dolar_d4	0.106165
13	dolar_d2	0.072879
75	var_perc_vl_cota_d3	0.0669159
8	bovespa_d3	0.0624641
16	dolar_d5	0.0493854
70	var_perc_tx_cambio_d3	0.0485368
52	var_perc_bovespa_d1	0.0467692

Dataset 2



	feature	importance
52	var_perc_bovespa_d1	0.0565997
1	selic_d2	0.0333581
53	var_perc_dolar_d5	0.0293815
18	dow_30_d1	0.0292591
57	var_perc_dolar_d1	0.0263926
30	tx_cambio_d1	0.0245619
72	var_perc_tx_cambio_d1	0.023431
32	tx_cambio_d3	0.0223542
21	dow_30_d4	0.0207706
49	var_perc_bovespa_d4	0.0187129

Dataset 3

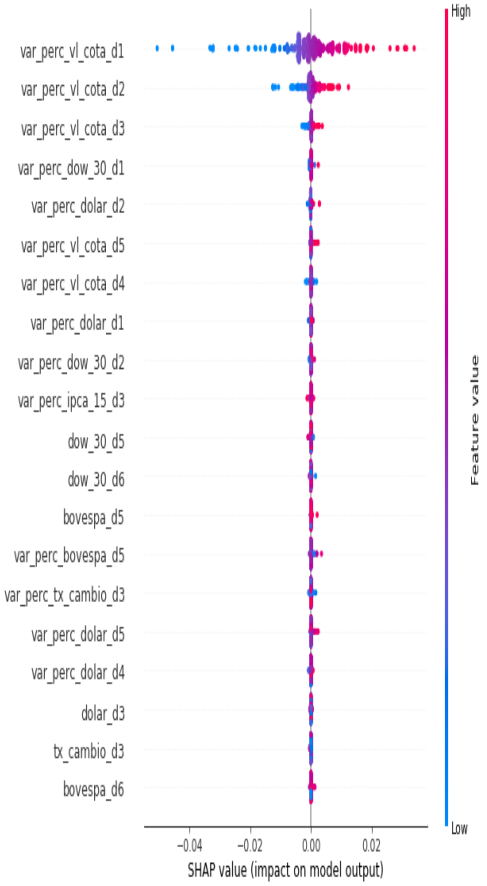


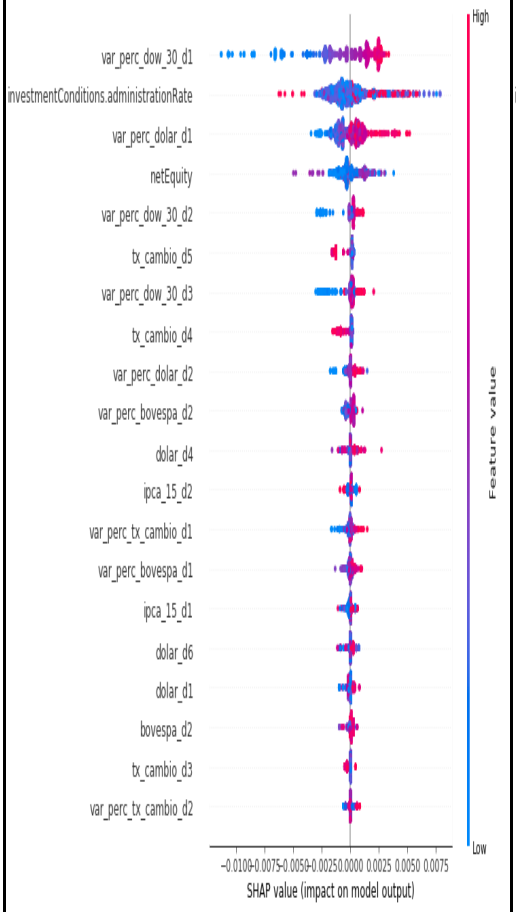
	feature	importance
16	var_perc_bovespa_d1	0.089997
36	var_perc_tx_cambio_d1	0.049432
21	var_perc_dolar_d1	0.0440546
18	var_perc_dolar_d4	0.041017
14	var_perc_bovespa_d3	0.0395092
26	var_perc_dow_30_d1	0.0391341
19	var_perc_dolar_d3	0.0378216
35	var_perc_tx_cambio_d2	0.0375832
15	var_perc_bovespa_d2	0.0354085
17	var_perc_dolar_d5	0.0348015

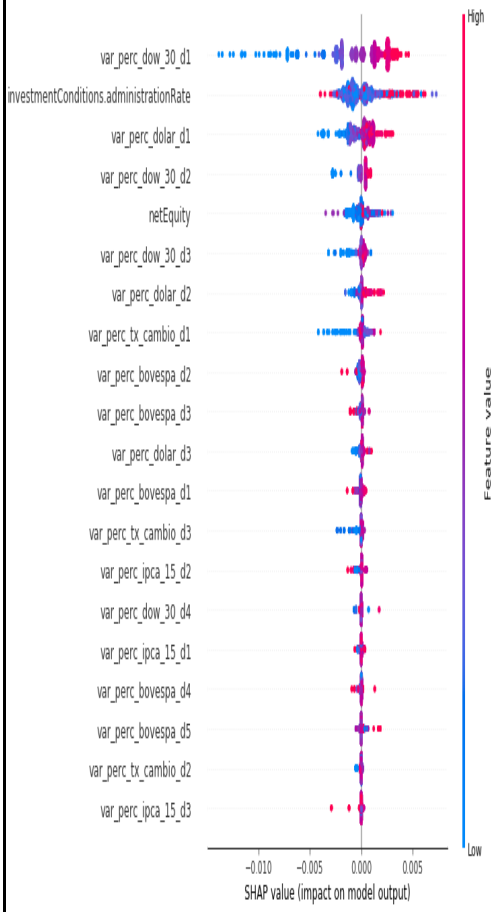
Tabela 4 - Importância das features para os fundos multimercado.

Fundos estrangeiros

Dataset 1 (com valores das cotas anteriores)	Dataset 2 (sem valores das cotas)	Dataset 3 (Sem valores brutos)







	feature	importance
77	var_perc_vl_cota_d1	0.264949
76	var_perc_vl_cota_d2	0.0879261
75	var_perc_vl_cota_d3	0.0265263
32	tx_cambio_d3	0.0208584
28	ipca_15_d5	0.0168889
62	var_perc_dow_30_d1	0.0166467
47	var_perc_selic_d1	0.0165679
65	var_perc_ipca_15_d3	0.0161254
22	dow_30_d5	0.0148404
63	var_perc_ipca_15_d5	0.0143745

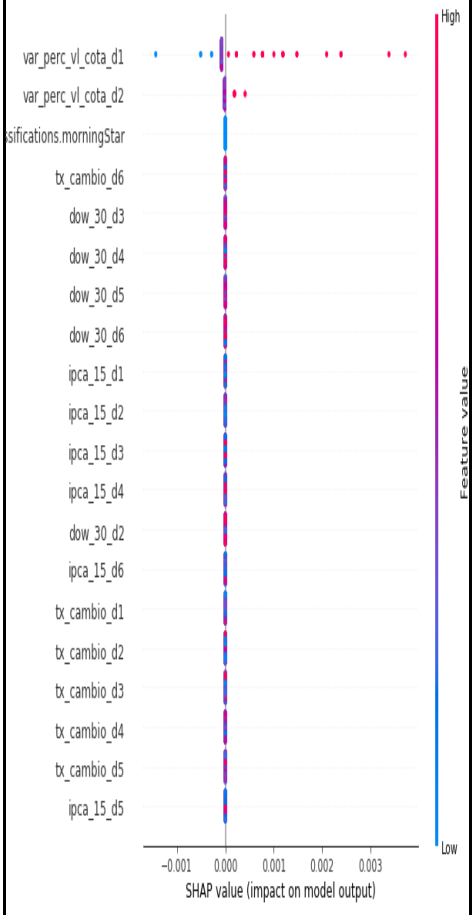
	feature	importance
62	var_perc_dow_30_d1	0.0545298
57	var_perc_dolar_d1	0.0347771
61	var_perc_dow_30_d2	0.0343655
34	tx_cambio_d5	0.03032
33	tx_cambio_d4	0.0253701
37	netEquity	0.0253437
60	var_perc_dow_30_d3	0.0224707
39	investmentConditions.administrationRate	0.0219317
26	ipca_15_d3	0.0210356
56	var_perc_dolar_d2	0.0202584

	feature	importance
26	var_perc_dow_30_d1	0.0905346
25	var_perc_dow_30_d2	0.0686995
21	var_perc_dolar_d1	0.0601399
3	investmentConditions.administrationRate	0.055415
1	netEquity	0.0508754
36	var_perc_tx_cambio_d1	0.0449471
24	var_perc_dow_30_d3	0.0447717
20	var_perc_dolar_d2	0.0395816
29	var_perc_ipca_15_d3	0.0370929
11	var_perc_selic_d1	0.0358636

Tabela 5 - Importância das features para os fundos estrangeiros.

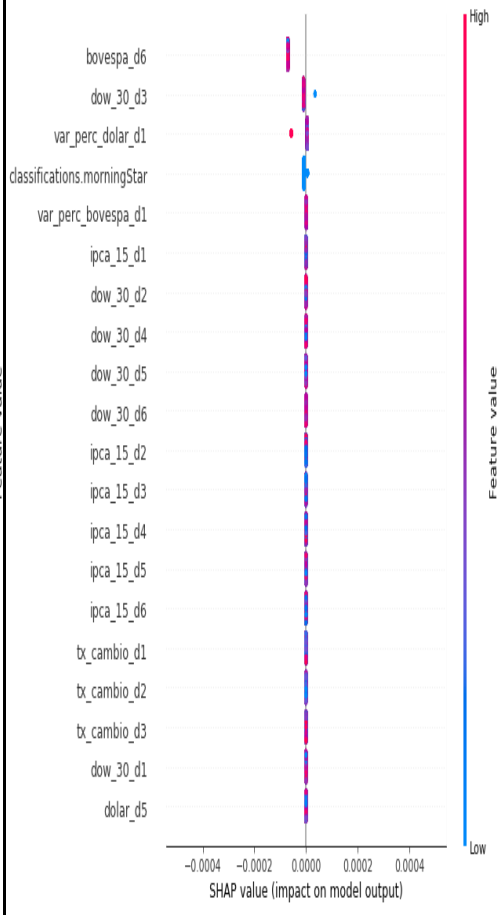
Fundos de renda fixa

Dataset 1 (com valores das cotas anteriores)



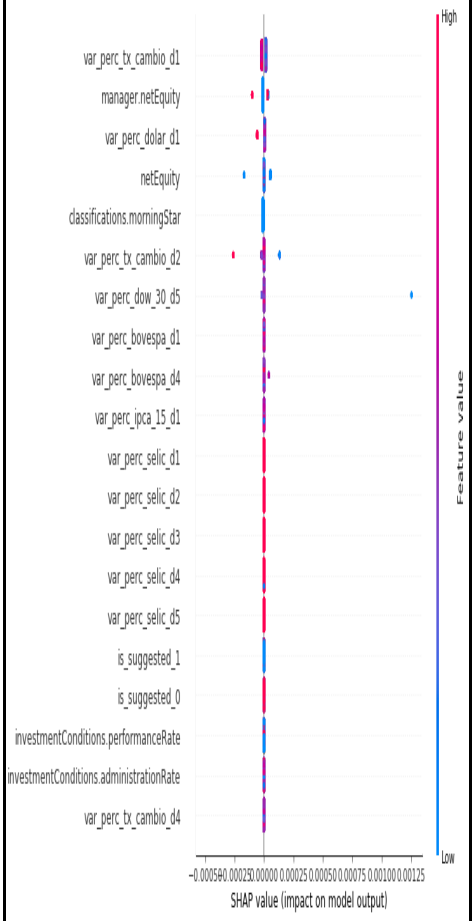
	feature	importance
77	var_perc_vl_cota_d1	0.218811
76	var_perc_vl_cota_d2	0.131812
18	dow_30_d1	0.107958
75	var_perc_vl_cota_d3	0.0865242
27	ipca_15_d4	0.0791974
14	dolar_d3	0.0722897
17	dolar_d6	0.0677383
0	selic_d1	0.0645637
54	var_perc_dolar_d4	0.0575203
13	dolar_d2	0.037274

Dataset 2 (sem valores das cotas)



	feature	importance
35	tx_cambio_d6	0.0399081
13	dolar_d2	0.0379251
37	netEquity	0.0299041
33	tx_cambio_d4	0.0295577
34	tx_cambio_d5	0.027557
71	var_perc_tx_cambio_d2	0.0264687
15	dolar_d4	0.0259877
19	dow_30_d2	0.0247351
58	var_perc_dow_30_d5	0.0246935
31	tx_cambio_d2	0.0241094

Dataset 3 (Sem valores brutos)



	feature	importance
4	investmentConditions.performanceRate	0.0675962
1	netEquity	0.0583645
21	var_perc_dolar_d1	0.0560108
22	var_perc_dow_30_d5	0.0554835
36	var_perc_tx_cambio_d1	0.0503712
16	var_perc_bovespa_d1	0.0494163
3	investmentConditions.administrationRate	0.0466081
19	var_perc_dolar_d3	0.0462651
0	classifications.morningStar	0.0430372
33	var_perc_tx_cambio_d4	0.0413763

Tabela 6 - Importância das features para os fundos de renda fixa, utilizando Shapley.

De uma forma geral podemos observar que utilizando apenas variações percentuais dos índices e *features* dos fundos é possível realizar uma predição mais acurada da variação da cota e que mesmo eliminando os valores das cotas de dias anteriores (maiores preditores do valor atual), ainda assim é obtido um modelo com baixos valores de erro.

Após eliminar os valores das cotas dos dias anteriores, os índices de mercado em relação às bolsas e às taxas de câmbio são os fatores que mais acarretam na variação percentual da cota, mas um fato curioso é a importância da métrica de avaliação referente a administração da bolsa, principalmente para os fundos de investimentos estrangeiros.

Considerando os fundos de ações, temos a variação do valor de bovespa em D-1 e dos valores da taxa de câmbio como grandes preditores do modelo. No caso da taxa de câmbio valores negativos do mesmo diminuem o valor da cota dos fundos, ao passo que no caso do bovespa variações positivas no bovespa aumentam o valor da variação percentual da cota.

Vale a pena ressaltar que a variação dos índices é um preditor melhor que o valor do índice propriamente dito. Isso faz sentido, pois o valor da variação dá uma ideia de queda/aumento do indicador, algo que o seu valor absoluto não consegue promover. Hoje 125 mil pontos no ibovespa corresponderia à um record histórico. Daqui a 5 anos, esse valor numa mesma data pode representar uma queda

Já para os fundos em multimercados, comparando-se com os fundos de ações podemos observar que o valor da predição é ligeiramente melhor. Isso se dá pois os fundos multimercados têm estratégias diversificadas sendo, portanto, o valor da sua cota é menos volátil.

Ainda sobre os fundos multimercados, além dos valores da taxa de câmbio e do dólar, temos também a selic e dow 30 como uma das variáveis mais correlacionadas com o valor da cota. Isto faz sentido, partindo do princípio que esses fundos investem também em produtos de renda fixa, e em investimentos exteriores.

Um grande indicador foram as métricas de administração dos fundos, com base em uma análise via Shapley, tais métricas aparecem em diversos casos. Em geral, fundos com taxas de administração mais alta costumam ter uma boa gestão, o que pode torná-los um pouco mais estáveis e melhor a predição do valor da sua cota. Abaixo podemos ver com qual feature ele mais se relaciona e como é essa relação:

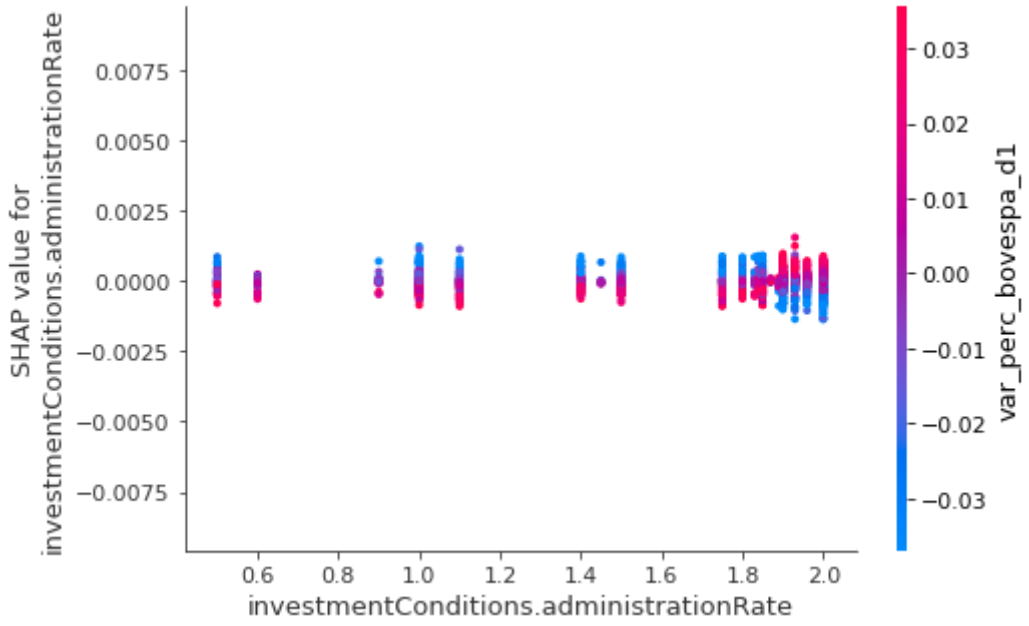


Figura 6 - Correlação entre a administração do fundo com a variação do índice Bovespa

Fundos com taxas de administração mais altas passam a ter aumentos mais significativos em períodos em que a bovespa está em alta e quedas superiores quando a mesma está em queda.

Para os fundos estrangeiros, temos que o valor do dólar e do índice Dow Jones são as features que mais explicam o modelo, visto que esses fundos investem em ativos internacionais.

As taxas de administração mostraram-se como a segunda principal feature na variação da predição, sendo esse tipo de fundo em que tal métrica obteve melhor posição como preditora.

Os fundos de renda fixa foram os que tiveram o melhor valor da predição. Em geral esses fundos são os menos voláteis, dado que os menos investem em tesouro e produtos com garantias do FGC, dentre outros ativos.

Após o treinamento com os datasets 2 e 3, em geral as variáveis de maior importância não possuem muito poder explicativo. Os fundos de renda fixa tendem a ter uma baixa oscilação ao longo do tempo já que normalmente são investimentos de médio a longo prazo, correspondendo às ínfimas variações do valor da sua cota em um curto intervalo, como o utilizado para predição nesse projeto.

5. Conclusão

Após todos os processos de pré processamento, análise e treinamento dos dados, é possível concluir que, foi obtido a criação de um modelo de regressão aparentemente satisfatório e com alta precisão para prever as variações percentuais dos valores de cotas de fundos de investimento para um curto período de tempo, bem como a utilização desse modelo para sugerir ou não a aplicação financeira em um determinado fundo, com base nos indicadores de mercado, sendo que tal sucesso foi adquirido por diversos motivos.

O primeiro motivo para o bom resultado apresentado passa pelo database obtido, trata-se de uma base com uma enorme quantidade de dados (mais de 123 mil registros de fundos), e que conta com uma grande quantidade de features. Os resultados final das métricas usadas para teste são totalmente influenciadas pela quantidade de dados que foi utilizado, pois com um acervo maior de dados, é facilitado e apurado o treinamento de qualquer modelo de aprendizado.

A escolha do algoritmo de treinamento também tem fator crucial no desempenho do mesmo, o XGBoost tem sido eleito o algoritmo campeão em maior número de competições do Kaggle, além disso, analisando o campo de machine learning de uma forma geral, quando se trata de dados estruturados/tabulares, algoritmos baseados em árvore de decisão são considerados os melhores da sua classe no momento, principalmente em modelos de regressão, como o realizado no projeto. Outro ponto favorável do XGBoost é sua flexibilidade em se adequar ao problema, dado que possui um grande número de hiperparâmetros que podem ser ajustados afim de obter o melhor resultado possível.

A escolha dos melhores parâmetros para o algoritmo de aprendizado só foi possível graças a utilização da função RandomizedSearchCV, permitindo que o eficiente XGBoost fosse utilizado da melhor forma para o dataset de trabalho.

Outro fator importante de abordar é a presença ou não de overfitting no modelo, já que não foram testados dados futuros para prever as variações das cotas. No entanto, é possível avaliar que há baixas chances de ocorrer um overfitting devido ao processo rigoroso de cross validation a que o modelo foi submetido, além da enorme quantidade de dados disponível no dataset, dificultando que o modelo "decore" os dados utilizados para treino, junto com a aprimorada abordagem para obtenção dos melhores hiperparâmetros do XGBoost.

O modelo de Shapley tornou possível a abstração e desencapsulamento das correlações entre as features e os valores a serem previstos, tal abordagem tornou o modelo útil não somente para predição de valores, mas como ferramenta para sugestão de fundos com maior esperança de rentabilidade, conhecendo apenas as variações dos índices de mercado que mais influenciam cada tipo de fundo.

Um fator negativo do projeto é o fato de que não foi possível ao curto prazo de tempo realizado, aprimorar o modelo para testá-lo e ajustá-lo para a predição de valores em maior prazo, o que torna-o temporariamente inútil para grande parte de investidores, no entanto é possível aprimorá-lo estipulando os valores de treino de forma sequencial até se chegar no prazo estipulado para o valor da cota.

6. Referências

- Conheça o algoritmo XGBoost, disponível em https://www.datageeks.com.br/xgboost/

- A Gentle Introduction to XGBoost for Applied Machine Learning, disponível em https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

- Gale–Shapley algorithm simply explained, disponível em https://towardsdatascience.com/gale-shapley-algorithm-simply-explained-caa344e643c2

- College Admissions and the Stability of Marriage, disponível em http://www.eecs.harvard.edu/cs286r/courses/fall09/papers/galeshapley.pdf