Predicting the Volume of Solids Combining Density Functional Theory and Machine Learning

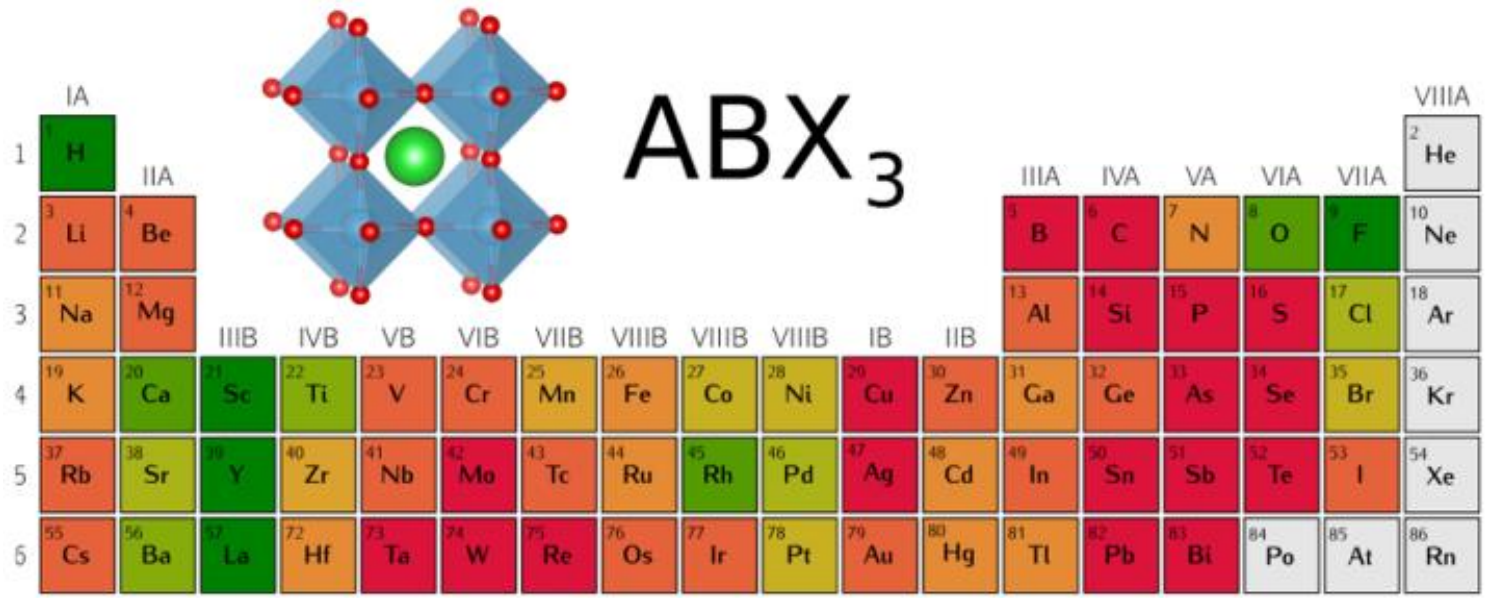
JONATHAN SCHMIDT, JINGMING SHI, PEDRO BORLIDO, LIMING CHEN, SILVANA BOTTI & MIGUEL A. L. MARQUES

Objetivo

Vamos tentar computar a estabilidade termodinâmica de sólidos

Partimos de um dataset de cálculos de teoria de densidade funcional baseado em uma estrutura cúbica (ABX_3)

Todas a espécies atómicas menos gases nobres e lantanídeos



Introdução

Neste trabalho vamos testar para um mesmo dataset vários modelos para calcular o Volume de cada composto

Comparar a performance de cada modelo, com analise gráfica e comparação de erros

Metodologia

Vamos usar uma estequiometria ternária relativamente simples (ABX_3) e uma estrutura cristalina fixa (cubic perovskite)

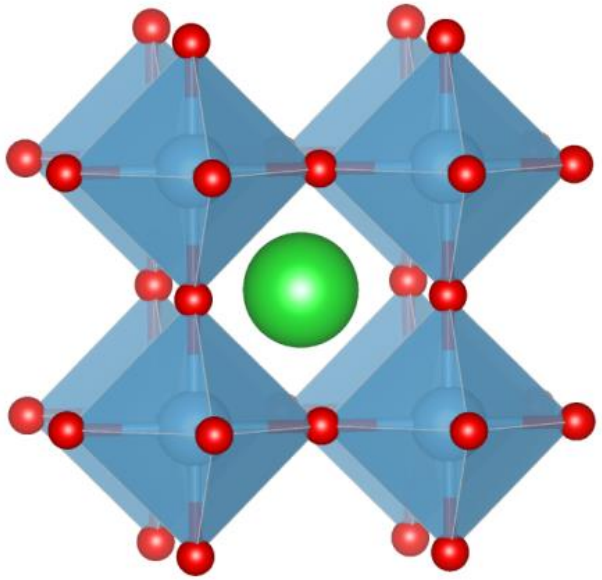
Variando A, B e X, e fazendo cálculos de density-functional theory (DFT) para cada combinação vamos obter um set convergente de dados

	A	B	X	Ehull
0	W	Ir	Pt	2210.302

	A	B	X	volume
0	Mn	Bi	Se	162.56

Vamos usar depois métodos de regressão para tentar prever valores para cada combinação

Usou se modelo linear, regressão Lasso, e Random Forest



Modelos de Regressão

Os modelos de regressão seguem todos um certo procedimento:

- Carregar os dados e separar em features e target (volume)
- Separar os dados em treino e teste (90:10)
- Criar o modelo
- Fazer fit do modelo aos dados (fazer regressão)
- Prever novos valores Target com features_treino
- Análise

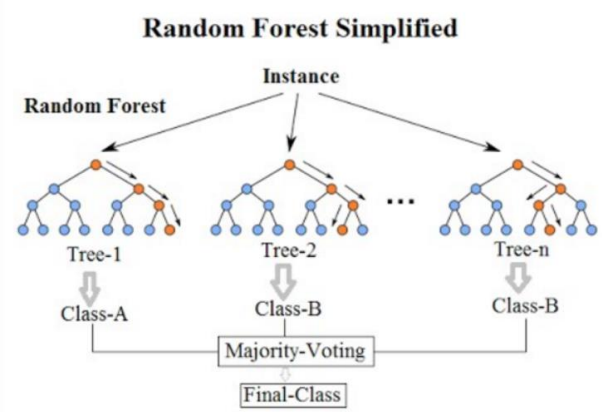
Modelos de Regressão (descrição)

Modelo linear:

- Necessita da existência de correlação linear entre dados
- Ótimo para modelos mais simples

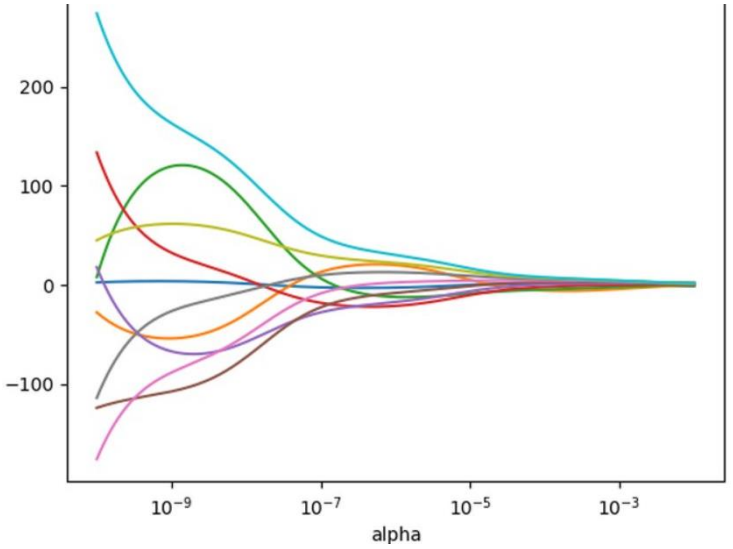
Random Forest:

- Método de aprendizagem de “orquestra”
- Ótima performance, mas menor precisão que alguns modelos

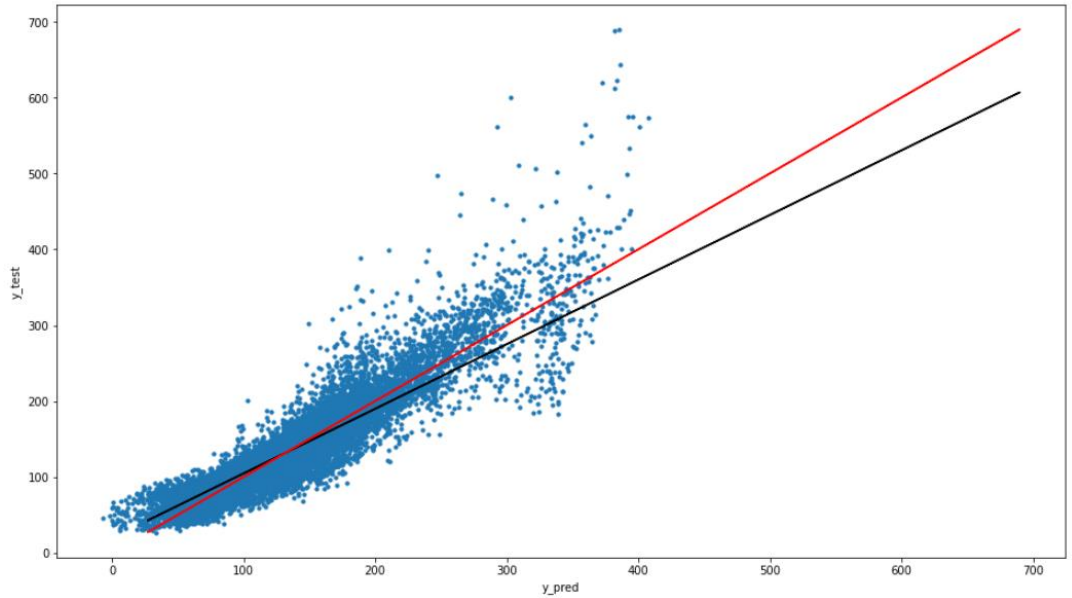


Lasso regression:

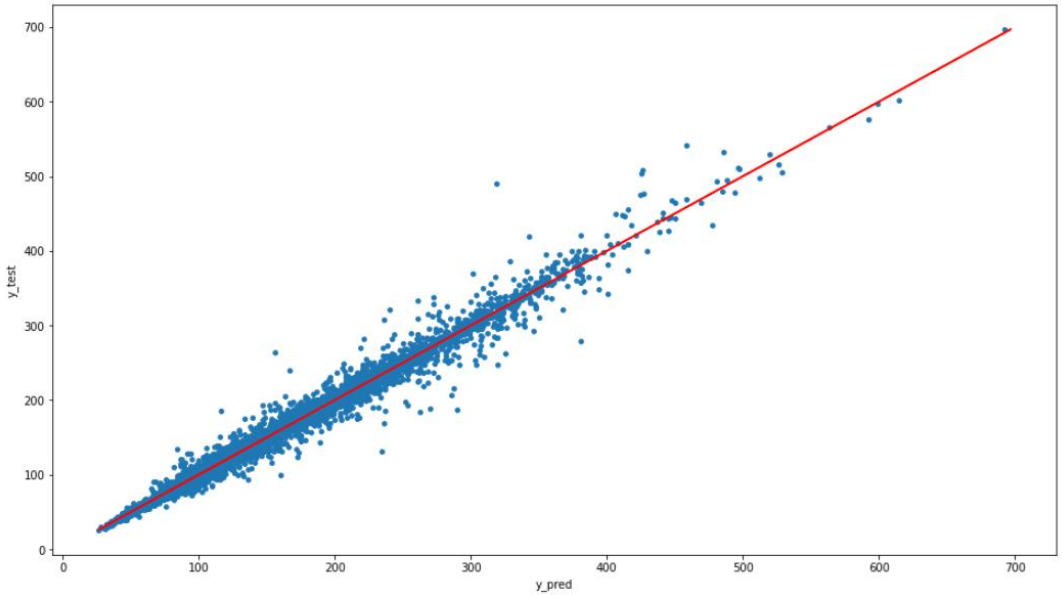
- Modelo baseado em “encolher” os dados, ou seja, aproxima os valores da media pretendida
- Encoraja o uso de modelos com menos features, e de maior importância



Comparação de modelos

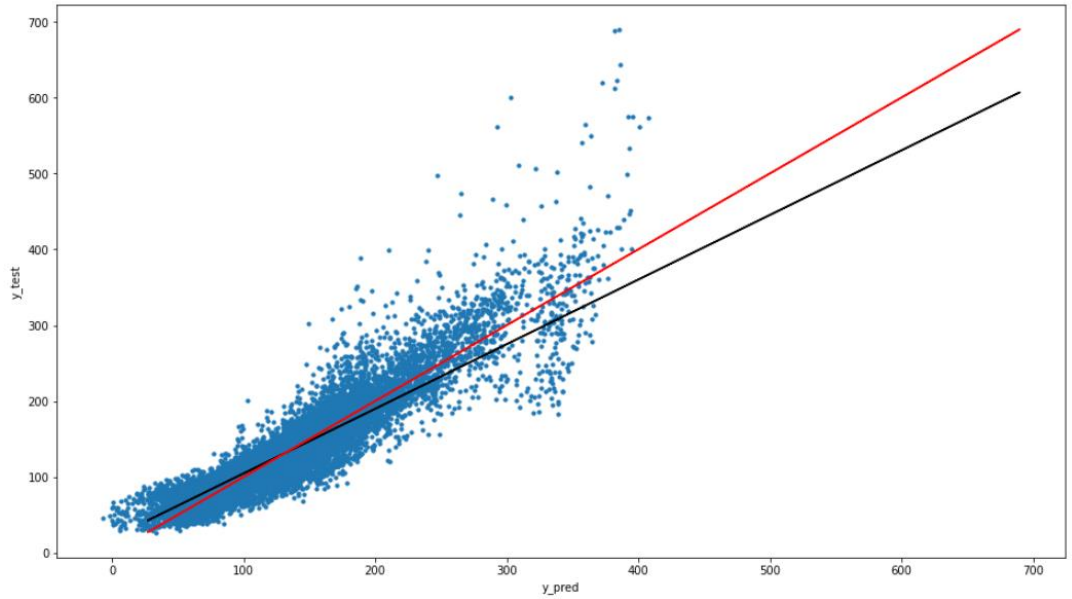


Modelo linear

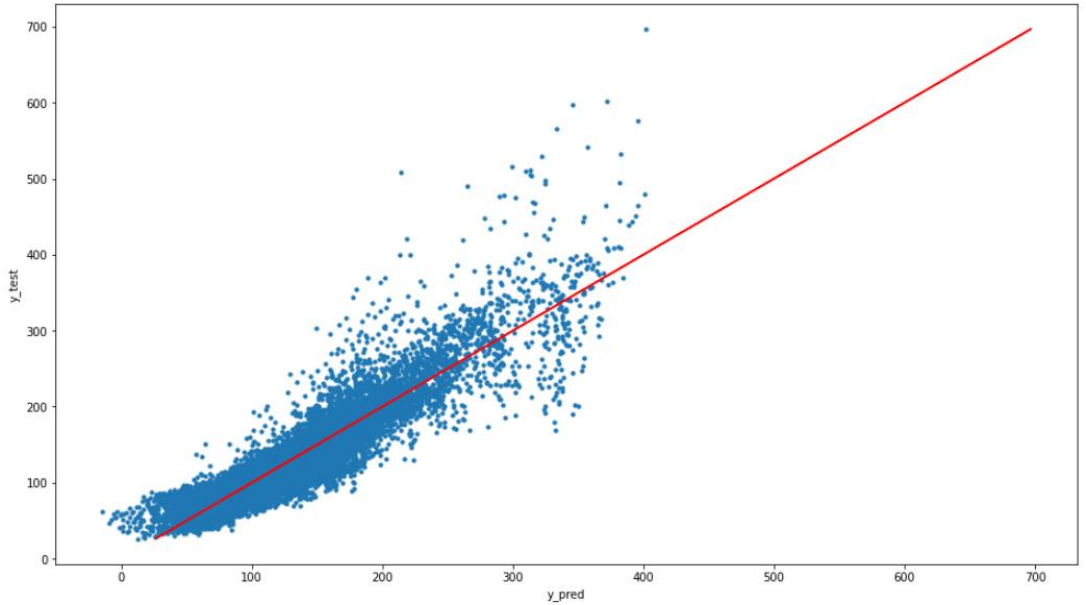


Random Forest

Comparação de modelos

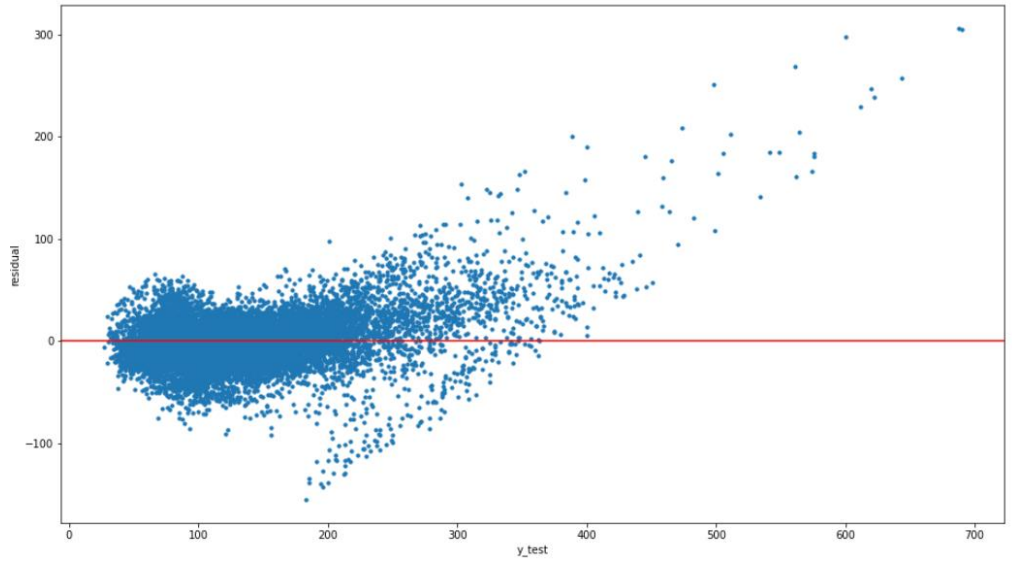


Modelo linear

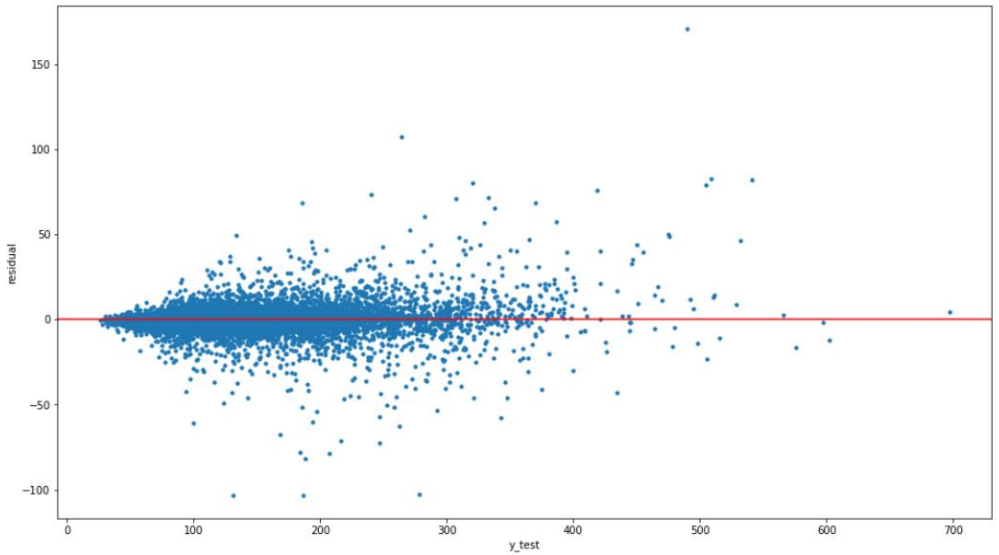


Lasso regression

Comparação de modelos (resíduo)

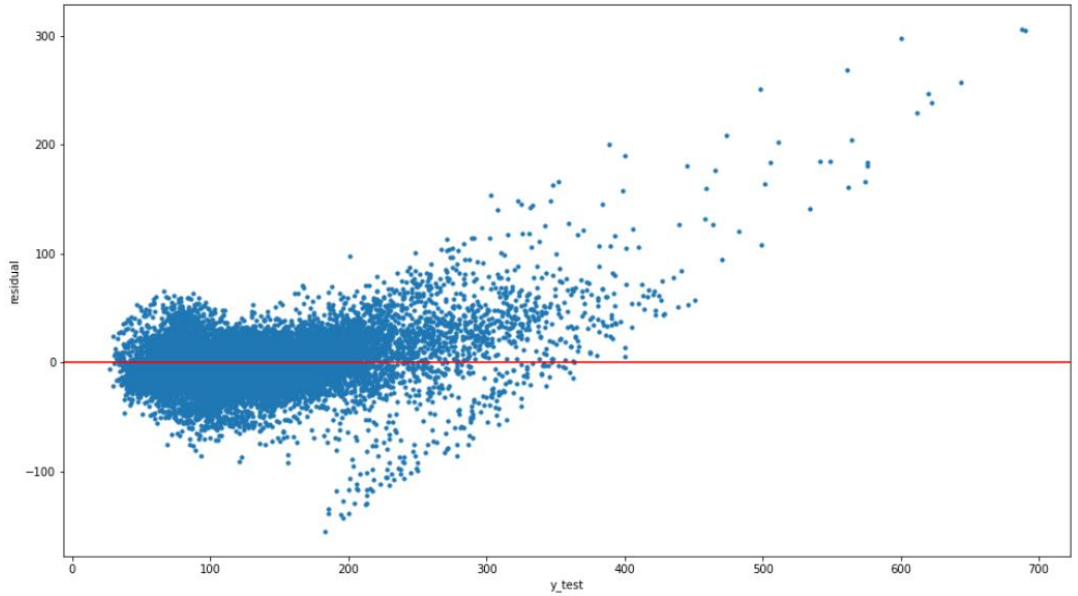


Modelo linear
$\bar{x}$ = 0,114

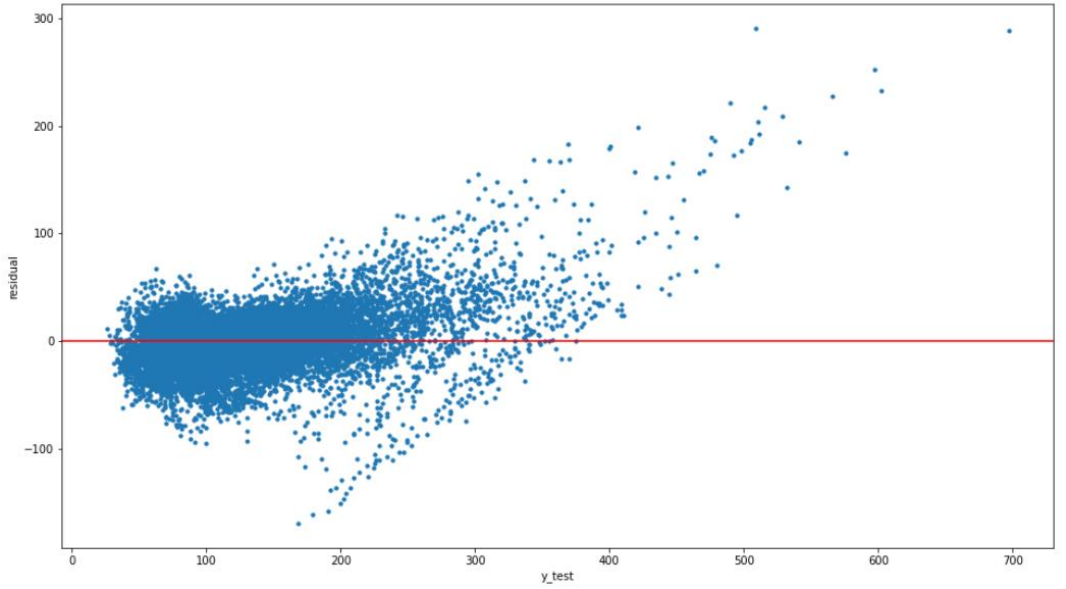


Random Forest
$\bar{x}$ = 0,028

Comparação de modelos (resíduo)

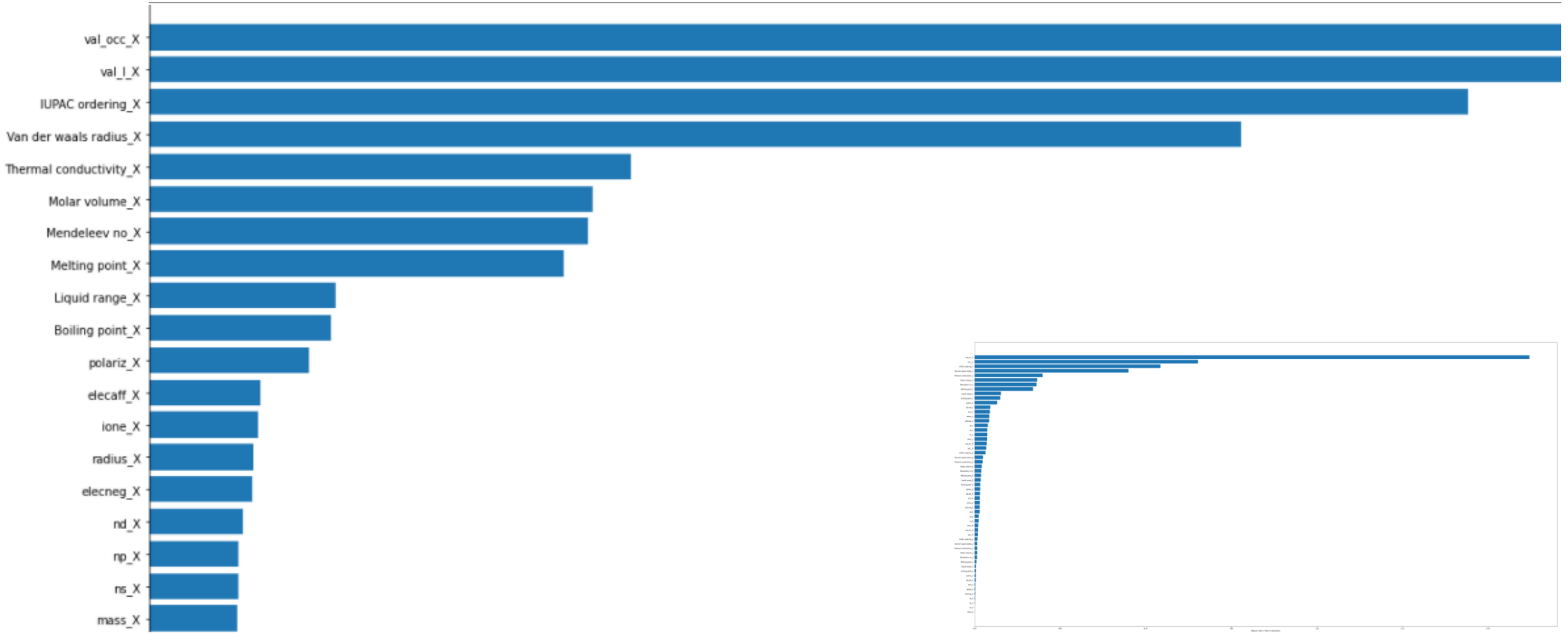


Modelo linear
$\bar{x}$ = 0,114



Lasso regression
$\bar{x}$ = -0,097

Peso de features



Comparação de modelos (erros)

Modelo Linear:

- RMSE: 23,739
- MAE: 3,986
- RMSE(%): 17,98

Lasso regression:

- RMSE: 23,83
- MAE: 4
- RMSE(%): 18,06

Random forest:

- RMSE: 6,668
- MAE: 1,797
- RMSE(%): 5,05

Conclusões

Os modelos baseados em regressão linear vão apresentar resultados basicamente iguais com no caso do Modelo linear e na regressão de Lasso

O Modelo Random Forest é significamente melhor, não tendo um erro maior do que 10%

Ensembles de modelos de previsão fracos vão ter um desempenho superior aos modelos de regressão linear