Modelos aditivos generalizados aplicados na eficiência energética de processadores

Universidade Federal da Paraíba - CCEN

Paulo Ricardo Seganfredo Campana

29 de abril de 2024

Índice

Resumo

texto

Introdução

texto

Metodologia

Dados

O conjunto de dados provêm do site de entusiastas de hardware *TechPowerUp* e está disponível no pacote `openintro` do software R pelo nome `cpu`, ele contêm informações sobre 875 processadores lançados entre 2010 e 2020, é composto por 12 colunas, nas quais 4 são nominais, 7 são numéricas e uma temporal: a data do lançamento, as variáveis são as seguintes:

Empresa: Responsável pela tecnologia, design e lançamento do produto, são apenas duas empresas que nas últimas décadas competem na venda de processadores para o consumidor comum, Intel e AMD.

Nome: Nome do processador, é unico para cada observação do banco e diz a respeito sobre o uso recomendado, a classe de performance e geração do processador.

Codinome: Nome de uso interno, único para cada geração de processadores de cada empresa.

Cores: Núcleos físicos de processamento contídos na CPU.

Threads: Núcleos virtuais de processamento, é comum ser o dobro do número de cores pela técnica de "hyper-threading", aumentando a capacidade de processamento paralelo.

Base Clock: Frequência de operação padrão do processador, medida em gigahertz (GHz), todas as instruções do processador seguem esse rítmo para sincronização de operações.

Boost Clock: Frequência de operação máxima do processador, é atingida temporariamente durante operações intensivas, está apenas disponível em alguns processadores que contém essa tecnologia.

Encaixe: Tipo de conexão física do processador, indica quais placas-mães são compatíveis com o produto.

Processo: Termo de marketing usado para designar a geração da tecnologia de fabricação do chip, é medido em nanometros (nm) porém essa quantidade não possui relação certeira com as dimensões físicas das menores partes do chip.

Cache L3: Tamanho de armazenamento da terceira camada de cache do processador, medido em megabytes (Mb), é usado para acelerar a transferência de dados entre o processador e a memória RAM.

TDP: Do inglês "Thermal Design Power" é a quantidade máxima de energia dissipada pelo processador como calor, medida em Watts (W), é usado como indicação do tipo de sistema de refrigeramento recomendado.

Lançamento: Data de lançamento do processador.

A seguir está a estrutura das primeiras 8 observações do conjunto de dados em ordem decrescente de lançamento:

Lançamento	Empresa	Nome	Codinome	Encaixe	Processo
2020-05-27	Intel	Core i3-10100	Comet Lake	1200	14
2020-05-27	Intel	Core i3-10300	Comet Lake	1200	14
2020-05-27	Intel	Core i3-10320	Comet Lake	1200	14
2020-05-27	Intel	Core i3-10350K	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10400	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10400F	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10500	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10600	Comet Lake	1200	14

Cores	Threads	Base Clock	Boost Clock	Cache L3	TDP
4	8	3.6	4.3	6	65
4	8	3.7	4.4	8	62
4	8	3.8	4.6	8	91
4	8	4.1	4.8	8	91
6	12	2.9	4.0	12	65
6	12	2.9	4.3	12	65
6	12	3.1	4.6	12	65
6	12	3.3	4.8	12	65

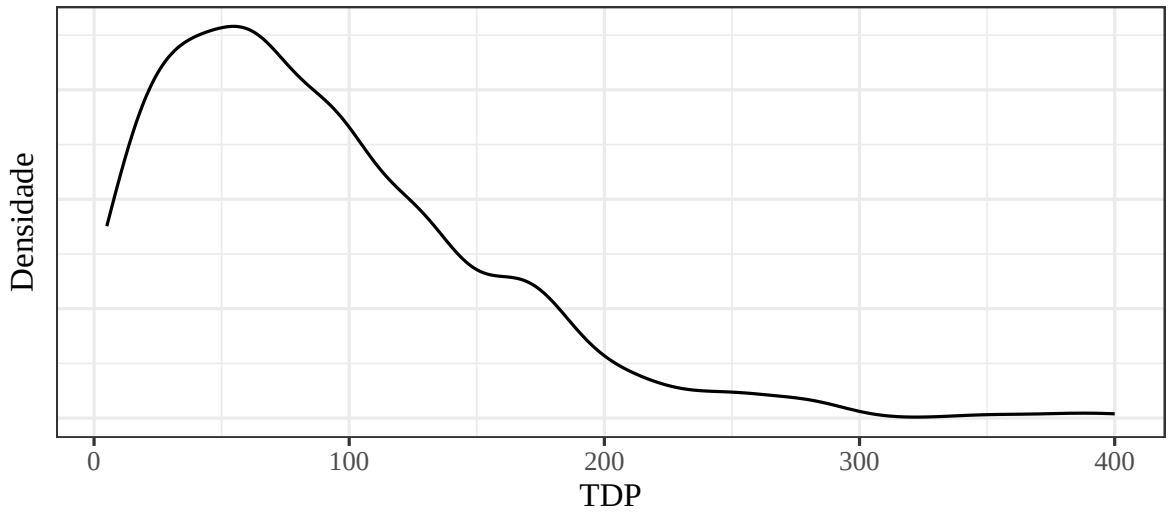
Todas as variáveis foram testadas para inclusão no modelo final, exceto o nome, codinome, encaixe e data de lançamento do processados pois além de não exercerem influência direta na eficiência energética do processador, são variáveis nominais que possuem muitos valores únicos, inviabilizando a utilização de métodos como variáveis dummy para serem inclusas no modelo de regressão.

Além disso, alguns processadores mais antigos não possuem as tecnologias de Boost Clock e terceira camada de cache, para o uso destes regressores no modelo, foram excluidos do banco de dados os processadores sem estas informações.

Modelo aditivo generalizado

Quando a variável resposta do modelo de regressão linear não tem distribuição normal, podemos partir para modelos lineares generalizados, que produzem melhores inferencias doque transformações na variável resposta. A eficiência energética dos processadores não possui distribuição normal como pode ser visto no gráfico de densidade:

Figura 1: Densidade da distribuição da variável resposta do modelo, a eficiência energética do processador.



Porém quando a contribuição de cada regressor na variável resposta não tem relação linear, este modelo não é valido, uma boa alternativa é o uso dos modelos aditivos generalizados, principalmente quando não conhecemos a formula da relação exata para o uso de um modelo de regressão não linear.

O modelo aditivo generalizado tem a seguinte forma:

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + \varepsilon$$

Onde $g(\mu)$ representa a função de ligação entre a média da variável resposta e os preditores, β_0 representa um intercepto e as funções $f_1, f_2, \cdots$ são funções não paramétricas que podem ser estimadas ou especificadas como lineares, e atuam em sua variável regressora correspondente.

Na prática, estas funções são suavizações baseadas nos dados que podem assumir formas como suavização por regressão local, *b-splines*, *splines* cúbicas ou polinômios ortogonais.

Validação do modelo

Foram usados três métricas para avaliação da performance do modelo: raiz do erro quadrático médio ($\hat{\sigma}$), erro médio absoluto (EMA) e coeficiente de determinação (R^2) com suas formas dadas a seguir, onde $\hat{y}$ são as previsões do modelo e $\hat{\rho}$ é o coeficiente de correlação de Pearson.

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{EMA} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad R^2 = \hat{\rho}^2_{y\hat{y}}$$

As previsões usadas no cálculo das métricas provem da separação do conjunto de dados em uma parte usada para ajuste do modelo e outra para estimação das métricas, utilizando a abordagem de validação cruzada com 10 *folds* e 10 repetições, assim obtemos não só uma estimativa pontual para as métricas como também erro padrão e intervalo de confiança.

Também é necessário que o modelo cumpra algumas suposições como a normalidade dos resíduos, ausência de auto correlação nos resíduos, além de que a contribuição de cada regressor seja significante para obtermos um modelo válido. Tudo isso será verificado com testes de hipótese.

Os resultados a seguir foram obtidos no software estatístico e linguagem de programação R (R Core Team 2024) fazendo uso dos pacotes `tidymodels` (Kuhn e Wickham 2020) para validação cruzada e estimação das métricas de performance, e `mgcv` (S. N. Wood 2011) que implementa os modelos aditivos generalizados.

Resultados

Pelo gráfico de densidade da variável resposta na Figura 1, vemos que está é uma variável com distribuição positiva e assimétrica, portanto a distribuição Gama é uma boa escolha para a parte generalizada do modelo aditivo, a função de ligação logarítimica apresentou melhores resultados de performance do modelo portanto foi escolhida, assim temos um modelo em que o logarítimo da esperança condicional da eficiência energética é explicada por uma combinação aditiva entre regressores, o intercepto e um erro com distribuição normal.

$$\log \mu = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + \varepsilon$$

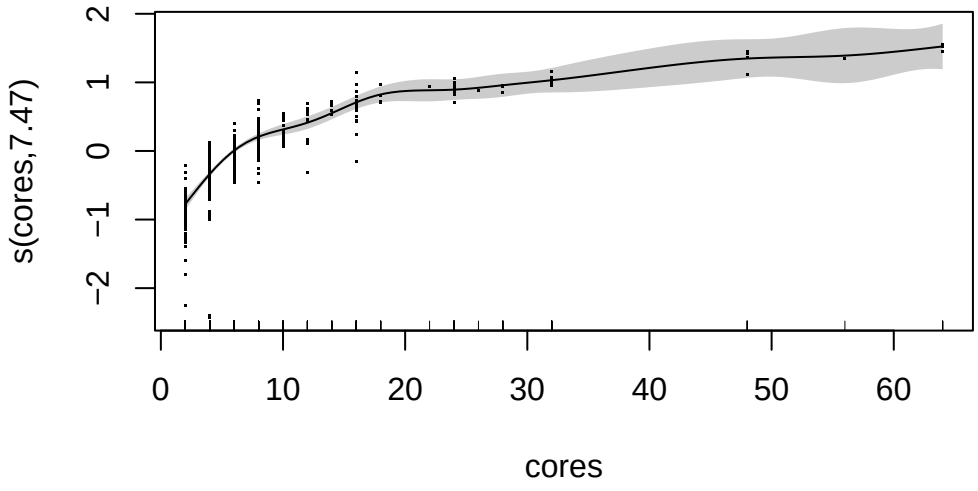
O tipo de função suavizadora escolhida foi *splines* isotrópicas de baixo posto, o padrão da função `s()` do pacote `mgcv` pois de certa forma é a melhor suavização segundo Simon N. Wood (2003). Essa suavização é aplicada nos regressores *cores*, *base clock*, *boost clock* e *cache L3*, não foi aplicada no regressor *processo* pois esta apresenta uma relação com a variável resposta que reduz a uma estimação linear nas *splines*.

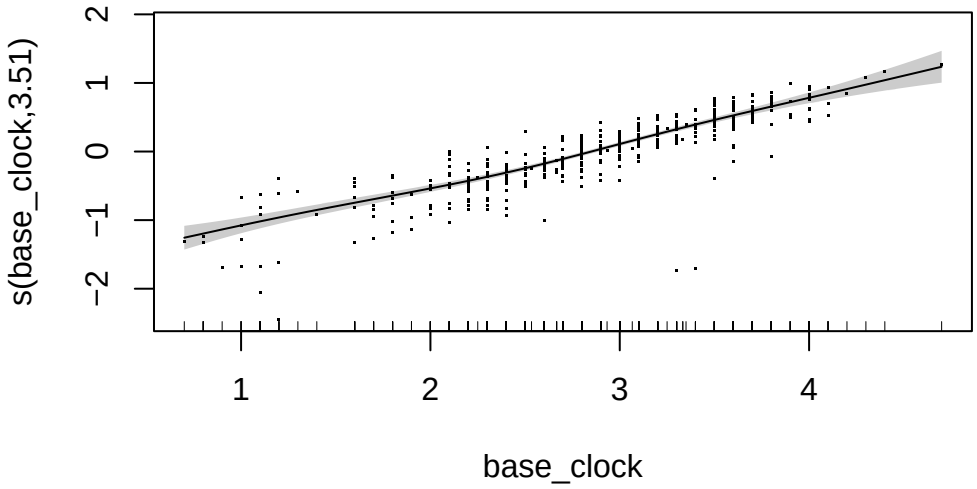
O regressor *empresa* foi transformado em variável dummy sem interação para uso no modelo, pois há diferença significativa na eficiência energética de processadores Intel e AMD nesta década.

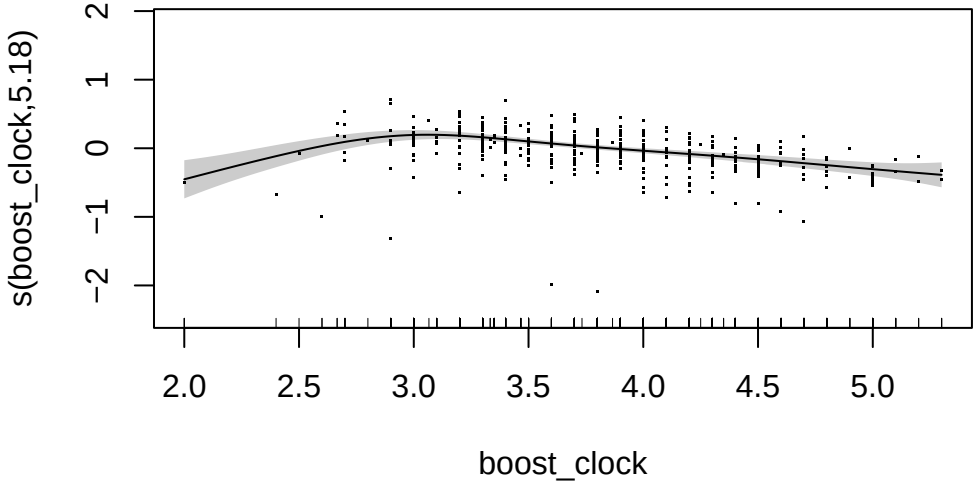
```
# A tibble: 2 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 rmse    standard   18.8      20 0.953   Preprocessor1_Model1
2 rsq     standard    0.914    20 0.00633 Preprocessor1_Model1
```

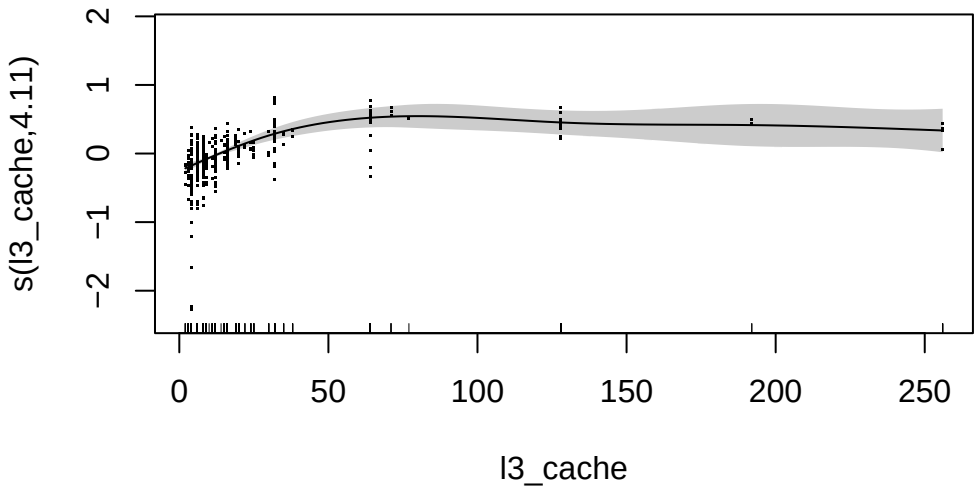
```
Family: Gamma
Link function: log

Formula:
tdp ~ company + process + s(cores) + s(base_clock) + s(boost_clock) +
    s(l3_cache)

Parametric coefficients:
```

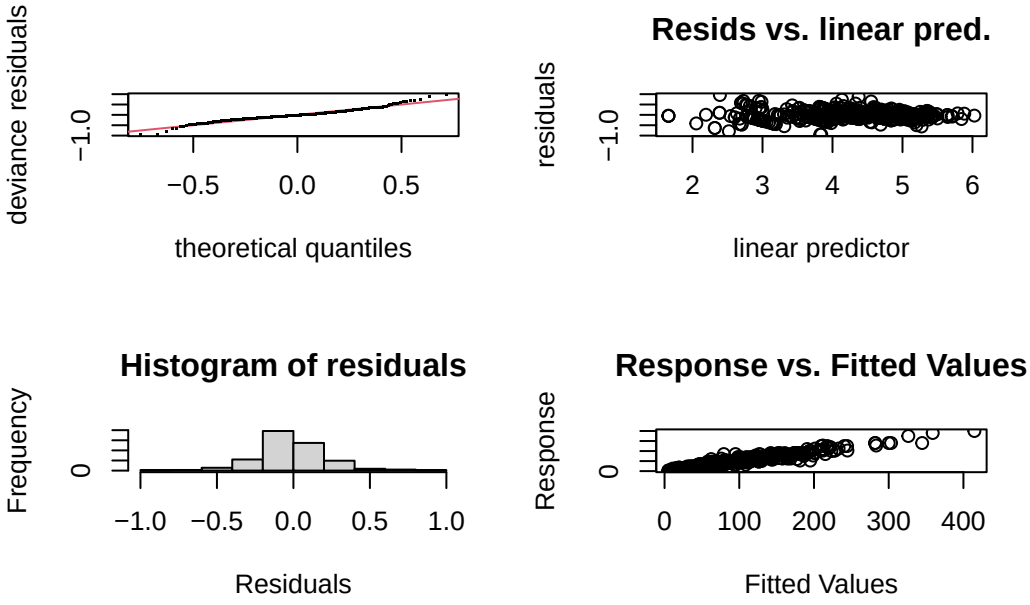
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.65268    0.03486  104.79   <2e-16 ***
companyIntel  0.31189    0.02776   11.24   <2e-16 ***
process       0.02090    0.00161   12.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                edf Ref.df      F p-value
s(cores)      7.470  8.286  52.84  <2e-16 ***
s(base_clock) 3.515  4.408 208.99  <2e-16 ***
s(boost_clock) 5.181  6.298  13.06  <2e-16 ***
s(l3_cache)   4.113  4.883  13.25  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.925   Deviance explained = 91.3%
GCV = 0.056991  Scale est. = 0.056956  n = 472
```




s(base_clock,3.51)
2
1
0
−1
−2
1
2
3
4
base_clock


s(boost_clock,5.18)
2
1
0
−1
−2
2.0
2.5
3.0
3.5
4.0
4.5
5.0
boost_clock






```
Method: GCV   Optimizer: outer newton
full convergence after 6 iterations.
```

```
Gradient range [-3.654351e-10,1.270523e-08]
(score 0.05699129 & scale 0.05695618).
Hessian positive definite, eigenvalue range [8.304573e-05,0.000295325].
Model rank =  39 / 39

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

                 k'  edf k-index p-value
s(cores)       9.00 7.47    0.57  <2e-16 ***
s(base_clock)  9.00 3.51    0.92   0.045 *
s(boost_clock) 9.00 5.18    0.82  <2e-16 ***
s(l3_cache)    9.00 4.11    0.59  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Kuhn, Max, e Hadley Wickham. 2020. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wood, S. N. 2011. «Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models». *Journal of the Royal Statistical Society (B)* 73 (1): 3–36.

Wood, Simon N. 2003. «Thin Plate Regression Splines». *Journal of the Royal Statistical Society (B)* 65 (1): 95–114. https://doi.org/10.1111/1467-9868.00374.