UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Aprendizado de Máquina
Eficiência energética dos edifícios

Aluno: Guillaume Jeusel

Professor: Alexandre G. Evsukoff

Disciplina: Inteligência Computacional

11 de dezembro de 2016

Sumário

1 Introdução **3**
1.1 Problema . . . 3
1.2 Conhecimento Prévio . . . 3

2 Descrição dos dados - lembrete **4**
2.1 Dados . . . 4
2.2 Distribuições - Histogramas das variáveis não padronizadas . . . 5
2.3 Matriz de correlação . . . 6

3 Atividade preditiva: Regressões **8**
3.1 Metodologia seguida . . . 8
3.2 Modelo Linear . . . 8
3.2.1 Modelo Linear de primeira ordem . . . 9
3.2.2 Modelo Linear Polinomial de grau r . . . 9
3.2.3 Modelo Linear de primeira ordem com regularização de Tikhonov . . . 11
3.3 Random Forest Regressor . . . 13
3.3.1 Apresentação do Random Forest . . . 13
3.3.2 Estudo da influência do número de ávore escolhido . . . 13
3.3.3 Gráfico dos valores preditivas . . . 14
3.4 Conjunto de resultados e comparações . . . 15

4 Estudos de Regressões complementares **16**
4.1 Influência da variável X6 Orientation . . . 16
4.2 Influência da padronização dos dados . . . 16

1 Introdução

1.1 Problema

Com uma demanda de energia sempre crescente nosso mundo, o problema de economia de energia é colocado no centro das preocupações. O conceito de *négaWatt* [1] traduze uma economia de energia devido a uma mudança de comportamento ou de tecnologia usada, e veja essa economia como um ganho. Além disso, um campo cujo desperdício de energia fica ainda extremamente importante é o edifício.

Por conseguinte, as investigações na área do desempenho energético dos edifícios cresceu muito recentemente; uma ação prioritária que as sociedades deve ter em mente é a redução do consumo de energia dos novos edifícios, também como a renovação dos antigos. A propósito, a legislação sobre o desempenho energético dos edifícios é sempre mas exigente, especificamente nos países europeus com a directiva 2002/91/CE limitando o consumo de energia dos edifícios [2].

1.2 Conhecimento Prévio

Para o design desses edifícios, é necessário a computação dos termos chamados *"Heat Load" e "Cooling Load"* (que pode ser traduzido pelo "carga de aquecimento" e "carga de arrefecimento" respetivamente). Eles são diretamente ligados à especificação dos equipamentos responsáveis para manter uma temperatura confortável, e então ao consumo energético. Esses coeficientes são dependentes das características geométricas dos edifícios, como também do clima local e do uso deles (industrial, casal . . .).

Existem muitos diferentes software de simulação que são eficientes para prever o consumo energético dos edifícios em projeto com uma precisão aceitável. Eles resolvam as equações diferencias da termodinmica aplicada a uma geometria particular. No enquanto, essas simulações podem demorar muito tempo, sem mencionar que quando um parmetro é mudado, a simulação deve ser reinicializada desde ao início.

Desse fato, um interesso crescente sobre o uso das técnicas de aprendizado de máquinas nasci. A ideá é a seguinte: suponho que você tem um banco de dados recente com as características e cargas de um grande número de edifícios, o uso de estatísticas e aprendizado de máquinas pode reduzir o tempo de computação e facilitar o experimento de diversos parmetros. Nos podemos pensar até criar um banco de dados com os diferentes resultados de simulação, e depois prever o desempenho energético de um novo edifício com interpolação dos resultados que nos já temos.

Isto foi a ideá do engenheiro civil *Angeliki Xifara* e do matemático *Athanasios Tsanas* da universidade de Oxford. Usando o software Ecotect, um conjunto de dado foi criado da simulação do desempenho energético para 768 geometrias de edifícios, assumindo uma localização em Atena, Grécia e um uso residencial com sete pessoas. Nos vamos estudar esse banco de dados.

Para ter mais informações sobre as hipóteses de simulação, deve-se referir ao papel deles [3].

2 Descrição dos dados - lembrete

2.1 Dados

O dataset é tirado do web-site UCI – Machine Learning Repository [4]. A figura 2.1 contem um resumo geral desse conjunto de dados.

Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	2012-11-30
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	95751

Figura 2.1: Características dos dados

Para facilitar o estudo das regressões, nos vamos somar a carga de aquecimento e a carga de arrefecimento para ter uma única saída.

Ele é composto de 768 registros e tem 8 variáveis de entrada e 1 de saída que são as seguintes:

Tabela 2.1: Mathematical representation of the input and output variables

Mathematical representation	Input or output variable	Number of possible values	Unit
X1	Relative Compactness	12	None
X2	Surface Area	12	m²
X3	Wall Area	7	m²
X4	Roof Area	4	m²
X5	Overall Height	2	m
X6	Orientation	4	Unknown
X7	Glazing Area	4	m²
X8	Glazing Area Distribution	6	None
y	Heating Load + Cooling Load	636	Unknown

É importante de notar que as variáveis de entradas são descontinuidades.
Um estudo anterior foi realizada, concluindo que o conjunto de dados:

- não tinha valores ausentes
- não tinha valores aberrantes (outliers)

2.2 Distribuições - Histogramas das variáveis não padronizadas

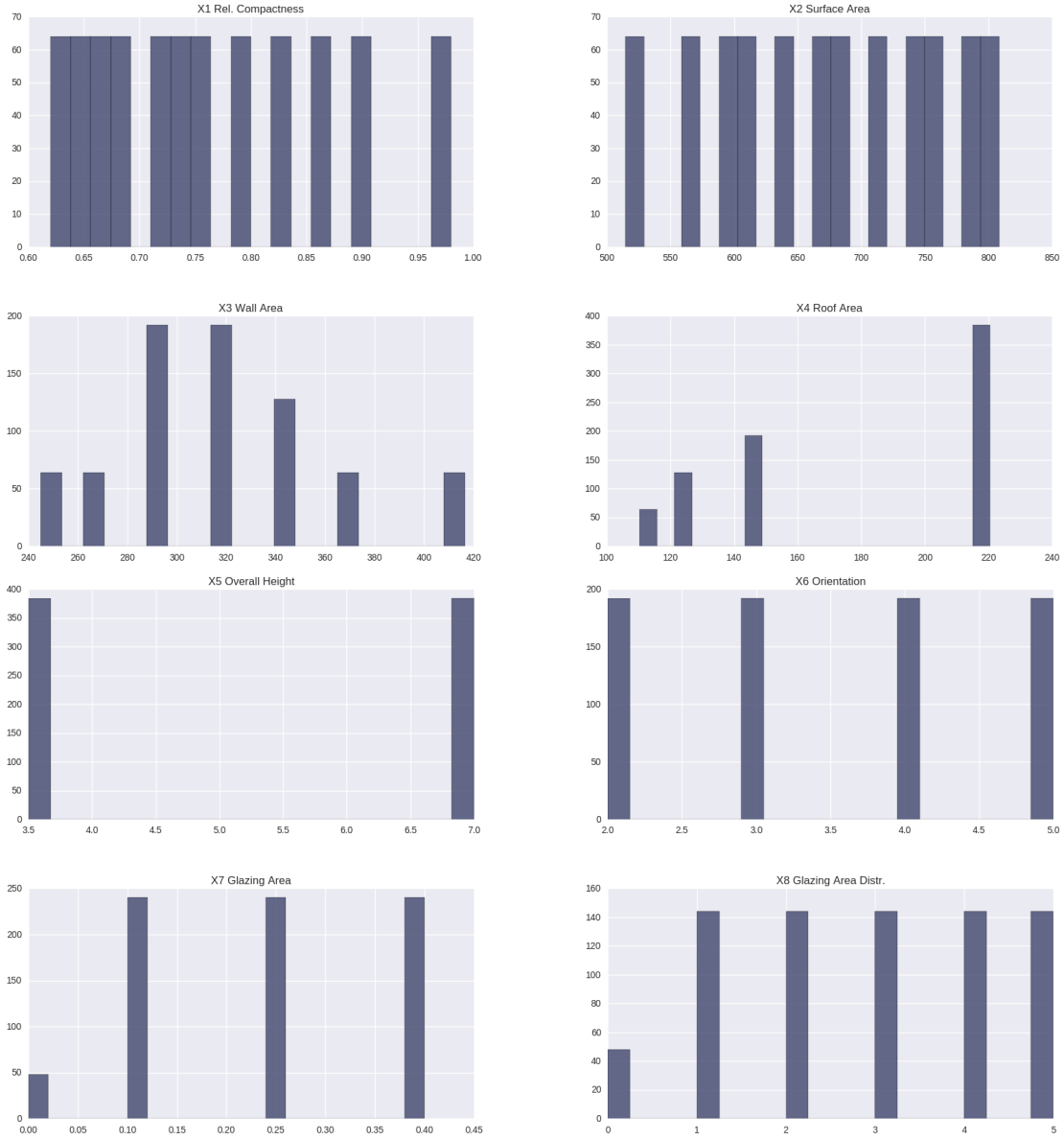

Figura 2.2: Histogramas das variáveis de entradas

Nos podemos comentar que as variáveis “X3 Wall Area”, “X4 Roof Area”, “X7 Glazing Area” e “X8 Glazing Area Distr.” não são bem centradas. Seja bem de processar com a metodologia de validação cruzada para ser robusto à escolha das partições de treinamento e validação.

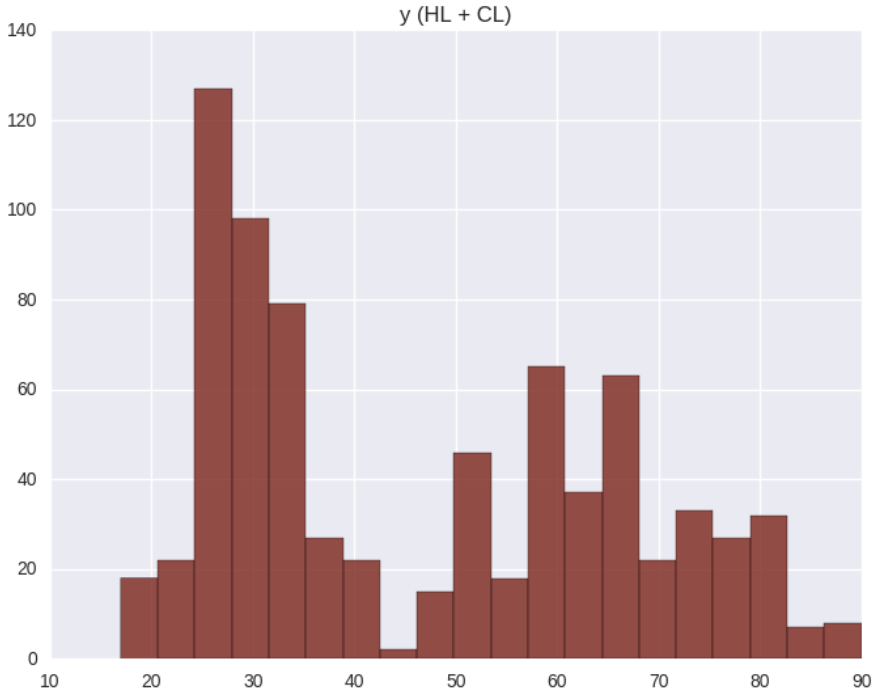

Figura 2.3: Histograma da variável de saída

Deve-se apontar a forma multimodal da variável de saída. Nos podemos já ter em mente que uma regressão linear não vai dar certo.

2.3 Matriz de correlação

A matriz de correlação é colocada na figura 2.4

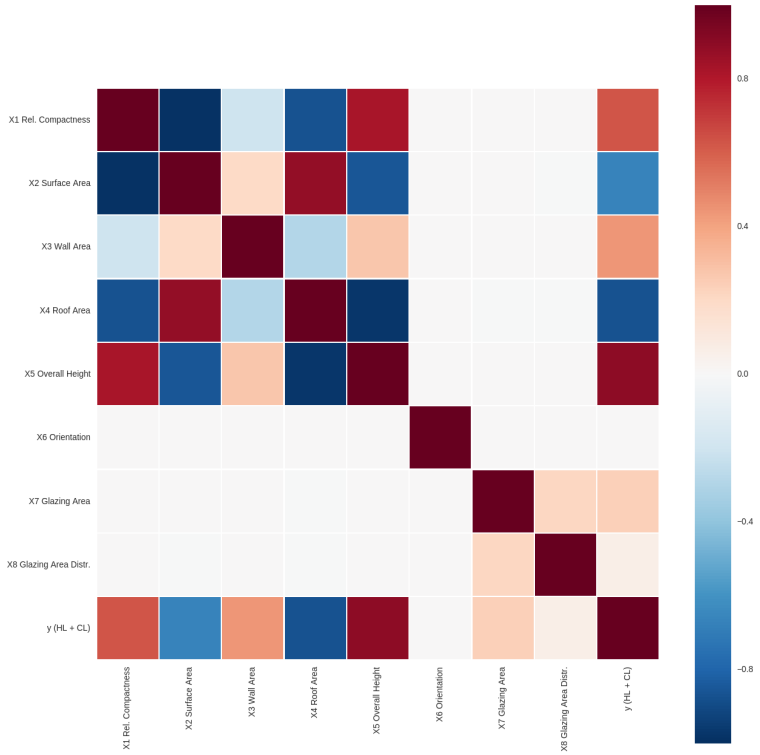

Figura 2.4: Matriz de correlação

As variáveis X1 (Relative Compactness) e X2 (Surface Area) são inversamente proporcional com um coeficiente de correlação igual a -1. Olhando no papel dos autores, nos podemos encontrar a explicação desse resultado: nos valores escolhidos para as simulações, eles fizeram a hipótese de um volume total dos edifícios constantes. Isto acarreta num relação analítica que liga X1 com X2. Observa-se o mesmo fenômeno com "X4 Roof Area"e "X5 Overall Height".

As variáveis de entradas “X4 Roof Area” e “X5 Overall Height” são variáveis altamente correlacionadas com a variável de saída. Elas vão ter um efeito importante na predição do y.

No entanto, nos vemos que a variavel "X5 Orientation"que pode ser retirada devido ao fato que ela não é correlacionada com nenhuma outra variavel: ela não da informações relevantes. Nós removemos essa variável para a continuação do estudo.

3 Atividade preditiva: Regressões

3.1 Metodologia seguida

Para cada modelo, será apresentado rapidamente o conceito matemático, e dado o gráfico (y medido, y predito) obtido.
A discussão sobre o desempenho de cada modelo será feita no final da secção, comparando todas as métricas de validação obtidas.

As métricas de validação usadas são:

- o *coeficiente de determinação* R2:

$$R^2 = \frac{\sum_{t=1}^{N}(\hat{y}(t) - \bar{y})^2}{\sum_{t=1}^{N}(y(t) - \bar{y})^2} \tag{3.1}$$

- *raiz quadrada do EMQ*, conhecida como RMS:

$$RMS = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{y}(t))^2} \tag{3.2}$$

- *erro absoluto médio percentual* MAPE:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y(t) - \hat{y}(t)}{y(t)}\right| \tag{3.3}$$

$\hat{y}(t)$ é a previsão de y calculada pelo modelo de regressão no ponto $x(t)$.

Finalmente, é importante de precisar que todos os $\hat{y}$ computados serão a união dos resultados de *predições cruzadas* de *10 ciclos*. Isto quer significar que o conjunto de dados vai ser dividido em 10 subconjuntos. Em cada ciclo (por um total de 10), o modelo é ajustado utilizando 9 subconjuntos e a saída é estimada por o subconjunto restante. No fim, todas as estimativas serão concatenadas de maneira que nos temos uma estimativa da saída para cada registros.
As estatísticas de validação serão calculadas com esse $\hat{y}$.

3.2 Modelo Linear

No modelo linear, a estimativa $\hat{y}$ da variavel de saída é procurado usando a forma seguinte:

$$\hat{y}(t) = f(x(t), \theta) = \hat{x}(t)\theta^T = \sum_{t=1}^{N}\hat{x}_i(t)\theta_i \tag{3.4}$$

com:

- $\hat{x}(t) = [1, h_1(x(t)), ..., h_N(x(t))]$ os regressores e $h_i(x(t))$ as funções de base
- $\theta = (\theta_1, ..., \theta_N)$ o vetor de parâmetros

Deve-se minimizar a função de custo, chamada de Erro Médio Quadrático para ajustar os parâmetros:

$$EMQ(\theta) = \frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{y}(t))^2 \tag{3.5}$$

3.2.1 Modelo Linear de primeira ordem

Nesse modelo, os regressores são as próprias variáveis de entrada: $\hat{x}(t) = [1, x(t)]$, i.e $h_i = Id$.
O gráfico dos valores preditivas é o seguinte:

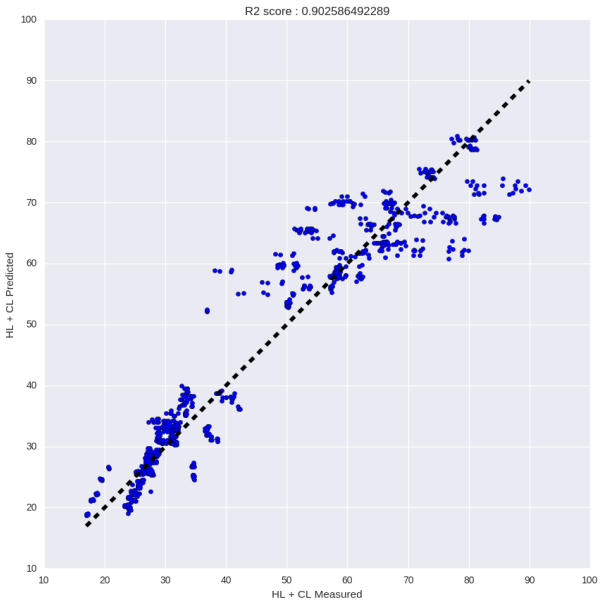

Figura 3.1: Predicted vs measured - Linear first order

3.2.2 Modelo Linear Polinomial de grau r

Nesse modelo, as funções de base são de forma polinômial: $h_i(x(t)) = x(t)^i$), com i variando de 0 até r o grau do polinômio.

O que dá: $\hat{x}(t) = [1, x(t), x(t)^2, ..., x(t)^r]$ como regressores.

Um estudo sobre a influência do grau escolhida do polinômio foi feita. As estatísticas de validação obtidas para cada grau é dado pela tabela 3.2 e plotada na figura 3.3.

	R2	RMS	MAPE
Polynomial deg 1	0.903231152681	36.6917578665	9.73994718219
Polynomial deg 2	0.982444549478	6.65648457785	4.43966732193
Polynomial deg 3	0.976892793022	8.7615391408	4.24348814077
Polynomial deg 4	-5.04146817577	2290.73812085	28.8408583508
Polynomial deg 5	-2.49924212741	1326.80452866	27.9708199736
Polynomial deg 6	0.886070306623	43.1986206204	8.12817149625
Polynomial deg 7	0.903003015841	36.7782603097	7.20115413124
Polynomial deg 8	0.916199561283	31.7745378986	7.30215428995
Polynomial deg 9	0.934045966841	25.0077321583	5.82373913709
Polynomial deg 10	0.942394503558	21.8422249081	6.43144195146
Polynomial deg 11	0.9078502191	34.9403505561	7.33365414704

Figura 3.2: Tabela influência grau - desempenho

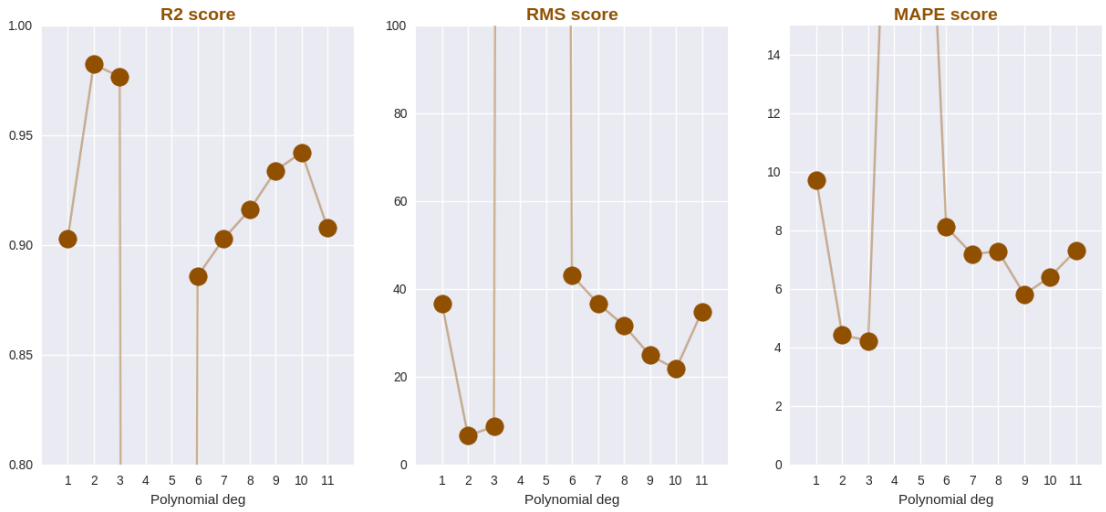

Figura 3.3: Gráfico influência grau - desempenho

Os scores R2 e RMS obtidos para os polinômios de grau 4 e 5 não faze nenhum sentido. Eu não consegui entender onde ficou o problema na hora da computação deles. Lendo a documentação scikit da função r2_score, um valor negativo significa que o modelo é *"arbitrarily worse"*.

No enquanto, nos podemos observar que os graus 2 e 3 são bem parecidos em termo de qualidade de modelagem. Além disso, para graus superiores nos podemos assumir uma situação de overfitting, com uma complexidade da modelagem superior ao que é preciso.
É interessante de notar que para o polinômio de grau 1, a solução do modelo linear simple é encontrada.

O gráfico dos valores preditivas para o Modelo Linear Polinomial de grau 3 é o seguinte:

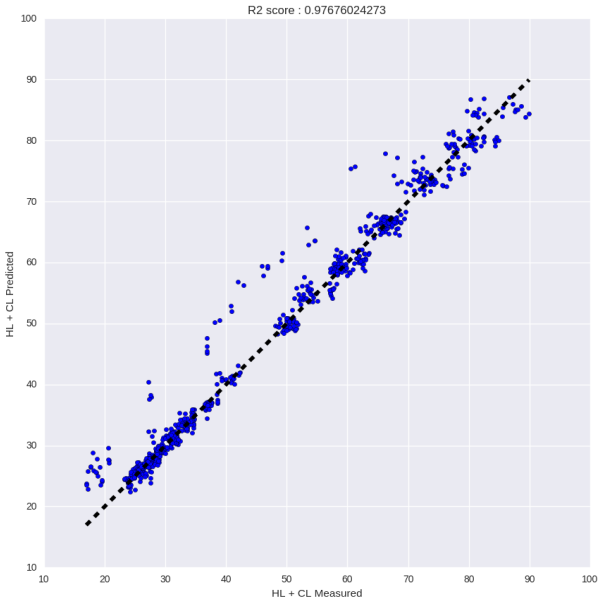

Figura 3.4: Predicted vs measured - Linear Polinomial deg 3

3.2.3 Modelo Linear de primeira ordem com regularização de Tikhonov

Usando a regularização de Tikhonov, chamada "Ridge regression"em inglês, a função de custo que tem que ser minimizada é da forma:

$$EMQR(\mu,\theta) = \frac{1}{N}\sum_{t=1}^{N}(y(t)-\hat{y}(t))^2 + \mu\|\theta\|^2 \tag{3.6}$$

Ela é chamada de Erro Médio Quadrático Regularizado. Isto é uma técnica de controle de complexidade do modelo através da aplicação de uma penalidade sobre o vetor de parametros.

A influência da penalidade escolhida (alpha) sobre as métricas de validação pode ser deduzido da figura 3.5.

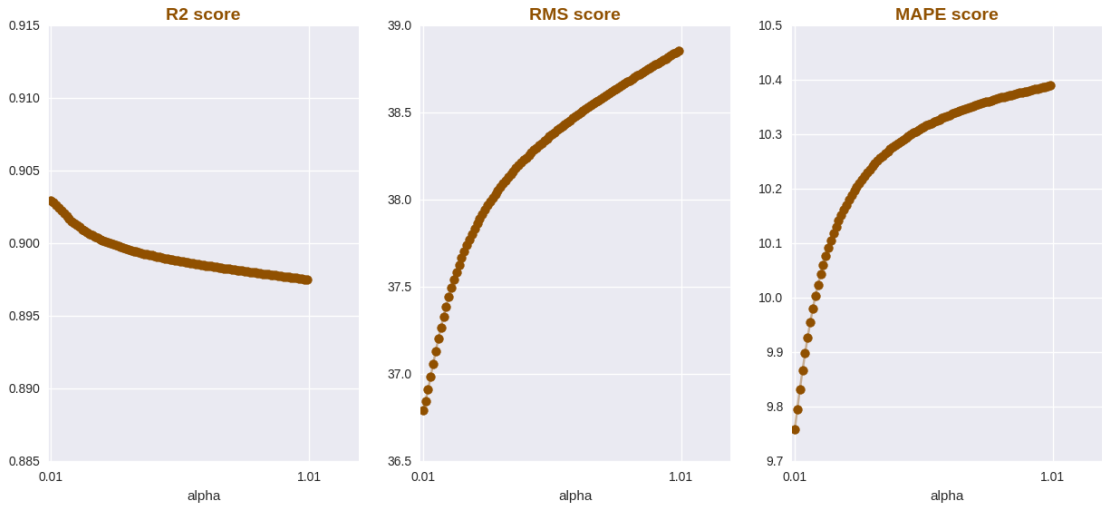

Figura 3.5: Gráfico influência alpha - desempenho

Dá para ver que os resultados são melhorados com um alpha pequeno, mas que esse ganho de desempenho é quase insignificante. Além disso, quando o alpha tende para 0, nos convergemos para a solução do problema linear de preimeira ordem sem regularização. Isto quer dizer que para esse conjunto de dados, nos não podemos esperar obter melhores resultados usando essa regularização.

O gráfico dos valores preditivas para o Modelo Linear de primeira ordem com regularização de Tikhonov é o seguinte:

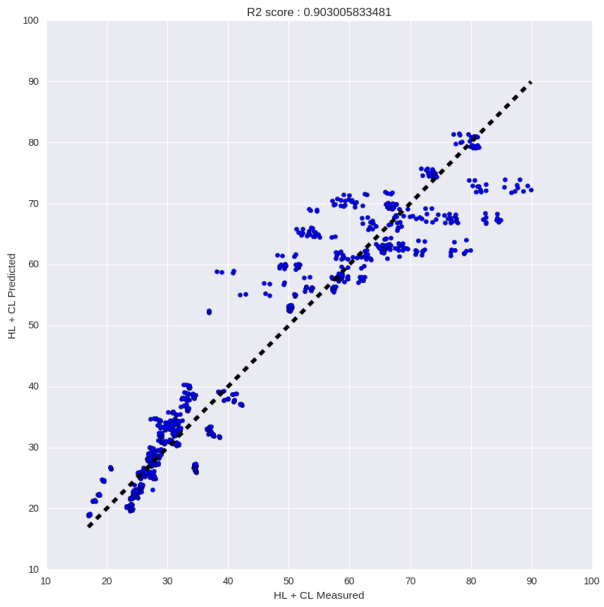

Figura 3.6: Predicted vs measured - Ridge alpha=0.001

3.3 Random Forest Regressor

3.3.1 Apresentação do Random Forest

A ideá seguida nesse modelo é simplesmente de treinar o modelo com um número de árvore de decisão grande com características aleatórias, e de pegar a media para melhorar a capacidade preditiva. Por lembrete, a árvore de decisão é uma estrutura de dados definida recursivamente como:

- Um nó folha que contém o valor de uma classe
- Um nó decisão que contém um teste sobre algum atributo.
- Para cada resultado do teste existe uma aresta para uma subárvore, que tem a mesma estrutura da árvore.

Inicialmente designado para problemas de classificação, nos podemos utilizar ele assumindo que as nó folhas contem os valores da variável de saída, o teste como uma verificação da distancia entre o valor previsto e o valor querido. A figura 3.7 do website scikit ilustra o algoritmo de ávore de decisão usada para aproximar uma curva de seno com ruído.

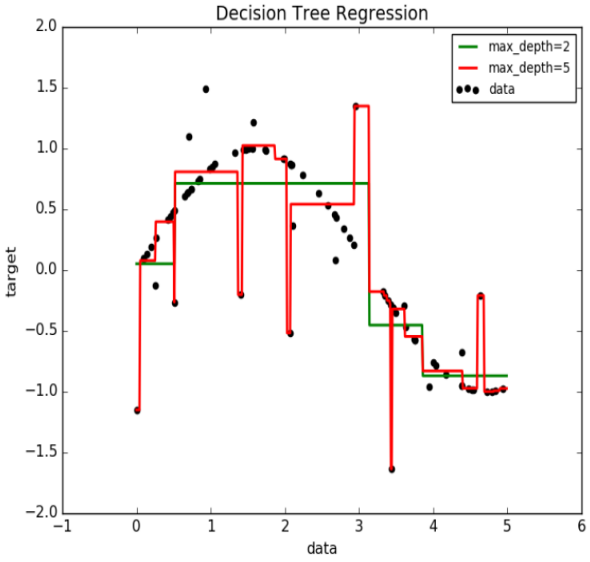

Figura 3.7: Exemple : decision tree used to estimate a sine curve with additional noisy observation

3.3.2 Estudo da influência do número de ávore escolhido

Sem limite de profundidade, as métricas de validação obtidas em função do número de árvore escolhido é colocado na figura 3.8.

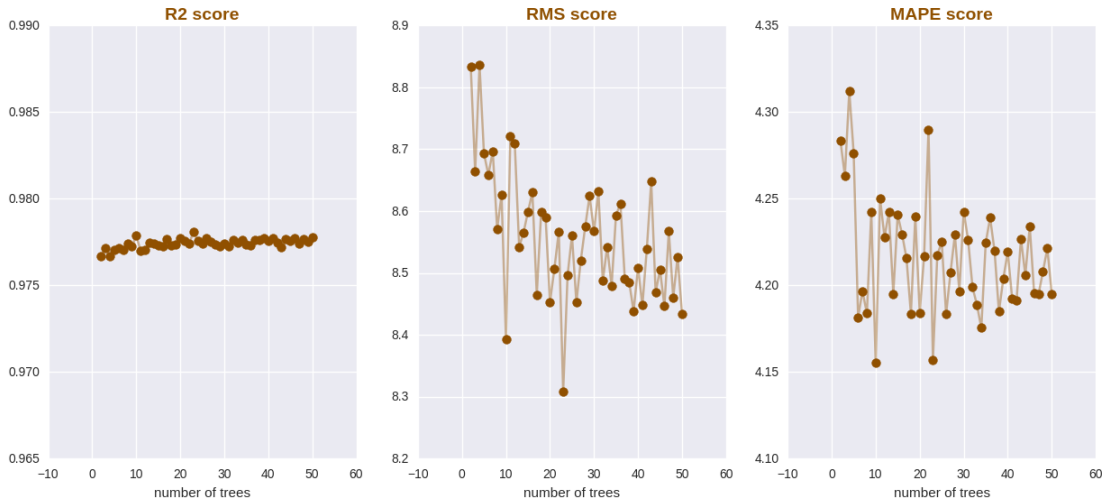

Figura 3.8: Gráfico influência número de árvore - desempenho

Nos podemos observar um variabilidade importante sendo as métricas RMS e MAPE, reflectindo o comportamento aleatório do algoritmo. No enquanto, a métrica R2 não muda muito em função do número de árvore escolhido. Isto é possivelmente devido ao fato que na formula do calculo do coeficiente de determinação, é pegado o valor média da saída $\bar{y}$, o que acarreta suavizar o comportamento aleatório.

3.3.3 Gráfico dos valores preditivas

O gráfico dos valores preditivas para o algoritmo de Random Forest com 10 árvores é o seguinte:

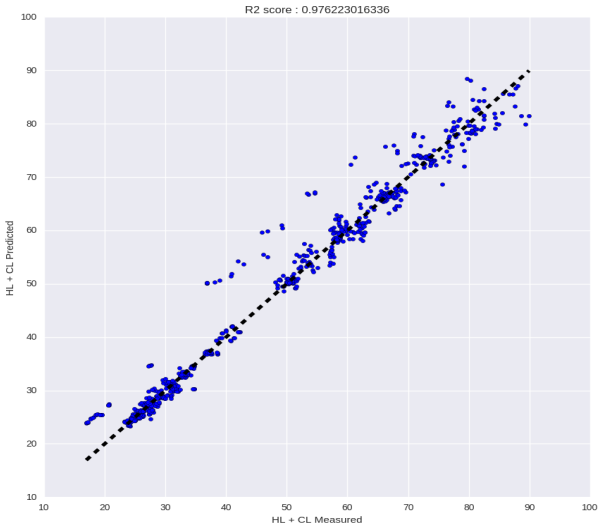

Figura 3.9: Predicted vs measured - RandomForest n_trees=10

3.4 Conjunto de resultados e comparações

As estatísticas de validação de todos os modelos anteriormente apresentados são colocado na tabela 3.10.

	R2	RMS	MAPE
Polynomial deg 3	0.976892793022	8.7615391408	4.24348814077
Random Forest 10 trees	0.977207734009	8.64212324667	4.28091530782
Linear + SVD regularization alpha=0.001	0.903008259797	36.7762719633	9.73658563994
Linear	0.90378456709	36.4819202168	9.48449357294

Figura 3.10: Métricas de validação em função do modelo

Nos podemos concluir que esse conjunto de dados não é bem modelado por modelos lineares simples. No enquanto o modelo linear polinômial de grau 3 tem aproximativamente o mesmo desempenho que o modelo de RandomForest. Seja bem de comparar esses modelos com modelos de redes neurais, mas eu não consegui instalar a última versão de scikit que tem esses algoritmos de rede neural.

4 Estudos de Regressões complementares

4.1 Influência da variável X6 Orientation

A mesma tabela de resultados pegando em conta a variável X6:

	R2	RMS	MAPE
Polynomial deg 3	0.97676024273	8.81179811753	4.33672910007
Random Forest 10 trees	0.976223016336	9.01549777214	4.24825810014
Linear + SVD regularization alpha=0.001	0.903005833481	36.7771919475	9.73732577983
Linear	0.902586492289	36.9361931748	9.71430145493

Figura 4.1: Métricas de validação em função do modelo - com a variável X6

Comparando com a tabela 3.10, nos podemos certificar que a variável X6 tem uma influencia desprezível.

4.2 Influência da padronização dos dados

Aplicando a padronização Z-score $\hat{X}_i(t) = \frac{X_i(t) - \bar{X}_i}{\hat{\sigma}_i}$ ao conjunto de dados, os resultados obtidos são os seguintes:

	R2	RMS	MAPE
Polynomial deg 3	-4.18966090619e+20	4.18420561856e+20	7.28874672447e+12
Random Forest 10 trees	0.977198092367	0.0227722176487	61.1605913264
Linear + SVD regularization alpha=0.001	0.90300466582	0.0968690381717	80.1875258347
Linear	0.903009427915	0.0968642822778	80.2462795794

Figura 4.2: Métricas de validação em função do modelo - com dados padronizadas

O Modelo Linear Polinomial de grau 3 não funciona mais.
Para os outros modelos:

- O coeficiente de determinação R2 obtido não é melhorado
- A métrica RMS parece justa, mas a sua leitura não dá mais para interpretar o erro da estimação sendo a escala mudada.
- A métrica MAPE acarreta ser totalmente errada, provavelmente devido à uma divisão de números pequenos na sua fórmula, que é mal administrado pelo computador.

Lista de Tabelas

2.1 Mathematical representation of the input and output variables 4

Lista de Figuras

2.1 Características dos dados 4
2.2 Histogramas das variáveis de entradas 5
2.3 Histograma da variável de saída 6
2.4 Matriz de correlação 7
3.1 Predicted vs measured - Linear first order 9
3.2 Tabela influência grau - desempenho 10
3.3 Gráfico influência grau - desempenho 10
3.4 Predicted vs measured - Linear Polinomial deg 3 11
3.5 Gráfico influência alpha - desempenho 12
3.6 Predicted vs measured - Ridge alpha=0.001 12
3.7 Exemple : decision tree used to estimate a sine curve with additional noisy observation . 13
3.8 Gráfico influência número de árvore - desempenho 14
3.9 Predicted vs measured - RandomForest n_trees=10 14
3.10 Métricas de validação em função do modelo 15
4.1 Métricas de validação em função do modelo - com a variável X6 16
4.2 Métricas de validação em função do modelo - com dados padronizadas 16

Referências

[1] Claude Crampes and Thomas Olivier Léautier. Pour une régulation intelligente de la demande d'électricité. *Les Echos*, 2010.

[2] Journal officiel des Communautés européennes. Directive 2002/91/ce du parlement européen et du conseil sur la performance énergétique des btiments, décembre 2002. http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/.

[3] Tsanas Athanasios and Xifara Angeliki. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, pages Vol. 49, pp. 560–567, 2012.

[4] A. Xifara A. Tsanas. Energy efficiency data set, 2012. https://archive.ics.uci.edu/ml/datasets/Energy+efficiency.