House Prices
(Advanced Regression Techniques)

Igor Martinelli[1], **Zoltán H. Jetsmen**[2]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC) - Universidade de São Paulo (USP)
13566-590 - São Carlos - SP

igor.martinelli@usp.br, zoltan.jetsmen@usp.br

Abstract. *This article describes the methods used and the learning obtained in solving the problem selected by the class of the discipline SCC0277 - Competições de Ciências de Dados, located in the competitions of the site Kaggle and titled House Prices: Advanced Regression Techniques.*

Resumo. *Este artigo descreve os métodos utilizados e o aprendizado obtido na resolução do problema selecionado pela turma da disciplina SCC0277 - Competições de Ciências de Dados, localizado nas competições do site Kaggle e intitulado House Prices: Advanced Regression Techniques.*

1. Introdução

1.1. Motivação e contextualização

Com o objetivo de complementar a formação dos alunos de graduação em Computação interessados no uso de técnicas de Ciência de Dados, em particular, Aprendizado de Máquina e Mineração de Dados, em problemas reais, utilizando para isso as técnicas e ferramentas vistas em disciplinas relacionadas[Jupiterweb 1999]. Dessa maneira, com o projeto em questão, vê-se a possibilidade de aplicar tais conceitos e desenvolver novas habilidades.

1.2. Descrição do problema

Peça a um comprador de casas para descrever a casa dos seus sonhos e provavelmente ele não começará pela altura até o teto do porão, ou pela proximidade à ferrovia de leste-oeste. No entanto, o conjunto de dados dessa competição mostra que há muito mais influências nos preços das negociações do que quantidade de quartos ou cercas brancas[Kaggle 2010]. Desse modo, busca-se resolver o problema de predição dos preços de casas, por meio de técnicas de regressão, com base nos atributos fornecidos no conjunto de dados.

1.3. Objetivos

O objetivo geral deste projeto é analisar o conjunto de dados obtido do site Kaggle, de modo a aplicar conceitos de ciência de dados (CCD) para extrair informações e conhecimentos deste a fim de obter dados concisos e que atendam aos requisitos para que se possa aplicar técnicas de aprendizado de máquina (AM) a fim de alcançar uma solução plausível para o problema.

Os objetivos específicos deste projeto são:

1. Analisar o conjunto de dados referente ao problema em questão.
2. Fornecer aos alunos o conhecimento sobre CCD e aplicação da mesma para extração de conhecimento acerca dos dados.
3. Fornecer aos alunos o conhecimento sobre diferentes técnicas de AM que utilizam a regressão para classificação.
4. Solidificar o conhecimento da Linguagem de programação Python, que será utilizada nos experimentos realizados pelos alunos.
5. Conhecer e utilizar procedimentos comumente empregados para planejamento de experimentos e análise de resultados em AM.
6. Realizar comparações entre diferentes algoritmos e analisar os resultados segundo a metodologia proposta pela comunidade de aprendizado de máquina para classificação de dados.

2. Conjunto de dados

O conjunto de dados consiste principalmente em dois arquivos, um arquivo representando instâncias de treino e o outro com instâncias de teste (train.csv e test.csv), além de um arquivo descrevendo o significado dos atributos contidos no mesmo (data_description.txt).

Tanto para os arquivos de treino quanto para os de teste, existem 79 atributos referentes às características das casas que estão sendo vendidas, como o tamanho do terreno, bairro em que se localiza, tipo do terreno em que está construída (regular, ligeiramente irregular, moderadamente irregular ou irregular), ano de construção, entre outras características que vão desde as que influenciam muito no preço da casa até as que quase não influenciam no mesmo. Além disso, no arquivo de treinamento são fornecidos os valores das casas correspondentes às tuplas fornecidas.

Ademais, os conjuntos de treino e teste são compostos de 1460 exemplos.

3. Pré-processamento dos dados

Inicialmente, para iniciar o pré-processamento dos dados, realizou-se um estudo a respeito das variáveis que compunham o problema estudado. No mesmo, existem 79 atributos referentes às condições das casas, sendo estes desde os mais relevantes, usados para predizer o preço final, até os menos importantes, sendo descritos por valores discretos e categóricos.

A partir destas informações foram propostos modos de tratamento dos dados utilizando modelos estatísticos para calcular correlação, além da utilização de métricas para a discretização de variáveis categóricas. Assim, podemos dividir essa etapa de pré-processamento em duas sub-etapas, compostas de uma tentativa inicial e uma consequente tentativa mais detalhada.

3.1. Primeira fase do pré-processamento

Após o entendimento do significado de cada atributo do conjunto de dados, buscou-se discretizar as variáveis categóricas que pareciam, a priori, exercer maior influência sobre o preço final da casa, dessa maneira, variáveis como LotShape, que poderia ser regular, ligeiramente irregular, moderadamente irregular ou irregular, foram discretizadas em valores iguais a 3, 2, 1 e 0, respectivamente.

Além disso, também foram discretizadas as variáveis *MSZoning, Street, Alley, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, RoofStyle, RoofMatl, Exterior1st, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, Electrical, KitchenQual, Functional, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PavedDrive, PoolQC e Fence.*

Por fim, foram retiradas as variáveis *Exterior2nd, MiscFeature, MiscVal, SaleType e SaleCondition.* Também foram preenchidos com 0 os valores faltantes e os representados por *NaN* ou *None.*

3.2. Segunda fase do pré-processamento

Ao terminar a primeira fase, avaliou-se o conjunto de dados obtidos por meio do pipeline utilizado para realização de regressão, com técnicas como Regressão Linear, Lasso e Ridge, que são técnicas utilizadas para problemas de regressão. Com isso, ao comparar-se o erro obtido, utilizando-se validação cruzada e a raiz quadrada do erro médio para o cálculo, com o erro dos usuários do leaderboard, verificou-se que o mesmo estava alto e, assim, decidiu-se realizar um aprofundamento na realização do pré-processamento. Para isso, utilizou-se técnicas estatísticas como a correlação para decidir os elementos mais relevantes a serem mantidos no conjunto de dados, bem como formas de discretização melhores, métodos estes que serão explicados abaixo.

Sabendo-se que praticamente metade dos atributos são compostos por variáveis numéricas e a outra metade por variáveis do tipo *Object*, retirou-se, em primeiro momento, as do tipo *Object* para realizar um estudo com base na correlação das variáveis numéricas. A partir disso, foi gerado um *heatmap*[Minitab], observado na figura 1, para se visualizar a relação entre as variáveis que apresentavam uma correlação maior do que 1/2.

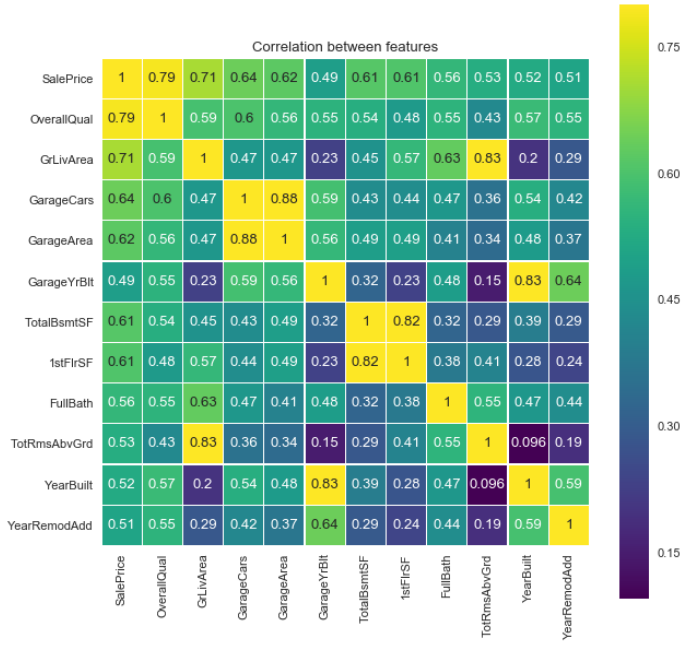

Figure 1. Heatmap gerado com base na correlação entre as variáveis numéricas

Assim, pode-se observar, pelo gráfico apresentado, que os atributos *(GarageCars e GarageArea), (TotRmsAbvGrd e GrLivArea), (TotalBsmtSF e 1stFlrSF), (YearBuilt* e *GarageYrBlt*) possuem uma correlação alta, isso quer dizer que essas variáveis podem ser redundantes, ou seja, podem exercer a mesma influência para o resultado final, desse modo, removeu-se uma delas de suas respectivas tuplas a fim de manter a consistência.

Dessa maneira, considerou-se apenas as variáveis *GarageArea, GrLivArea, TotalBsmtSF, YearBuilt* e assim, com a retirada das variáveis, as *features* numéricas que serão utilizadas para a regressão são as seguintes: *OverallQual, GrLivArea, GarageArea, TotalBsmtSF, FullBath, YearBuilt* e *YearRemodAdd*

Feito o processamento dos dados numéricos, foi realizada a análise das variáveis categóricas e, para isso, foi utilizado um procedimento estatístico chamado teste de hipótese, em que aceita ou rejeita-se uma hipótese. Para o problema em questão, foi utilizado a distribuição ANOVA, em que calculou-se o p-valor dos atributos categóricos em relação ao preço de venda e, caso esse valor fosse menor que 0.05, este atributo fora rejeitado.

Dessa maneira, após a utilização da técnica descrita, selecionou-se os seguintes atributos categóricos: *Utilities, LandSlope* e *Street*

4. Regressão

Após toda a etapa de pré-processamento analisada, partiu-se para etapa de regressão, em que diferentes algoritmos utilizados comumente para regressão foram aplicados para avaliar o desempenho de cada algoritmo, realizando diversos testes com variações dos parâmetros.

Assim, realizou-se testes inicialmente com algoritmos como *Logistic Regression, Linear Regression* e *Random Forest*. Após estes testes iniciais, viu-se que estes modelos eram muito básicos e não conseguiam lidar muito bem com o problema em questão. Também verificou-se que as técnicas lineares obtinham melhores resultados, podendo-se afirmar com um certo grau de certeza que o problema é linearmente separável, desse modo, investiu-se os esforços em outras técnicas de regressão linear[sklearn]. Selecionou-se assim, os modelos *LassoCV, RidgeCV, ElasticNetCV* e por fim, também utilizou-se a técnica *XGBoost* para regressão(*XGBRegressor*)

Vale destacar que os gráficos que serão apresentados abaixo, referentes aos modelos utilizados, foram gerados com base no conjunto de treinamento, pois, como o conjunto de teste não fornece os resultados, não há como construí-los.

4.1. LassoCV

O modelo Lasso consiste em um método de regressão linear utilizado para modelos complexos, com vários coeficientes regressores e este, por sua vez, atua utilizando um mecanismo de penalização dos coeficientes com alto grau de correlação de acordo com seu valor absoluto[Lasso].

Para o projeto em questão, utilizou-se o regressor *LassoCV*, que consiste no modelo *Lasso* juntamente com uma validação cruzada que permite a variação de seus parâmetros a fim de encontrar a melhor configuração dos mesmos, dados os vetores com diferentes valores dos parâmetros. Assim, variou-se em dois momentos distintos

o parâmetro alpha e gerou-se gráficos, tanto para o primeiro conjunto de treino quanto para o segundo, os quais podem ser observados nas figuras 2 e 3.

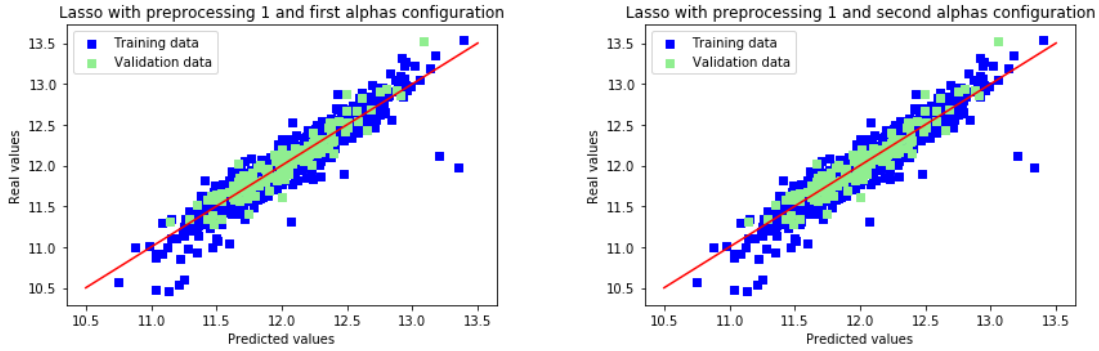

Figure 2. Gráfico dos resultados obtidos utilizando o método Lasso com o primeiro pré-processamento realizado.

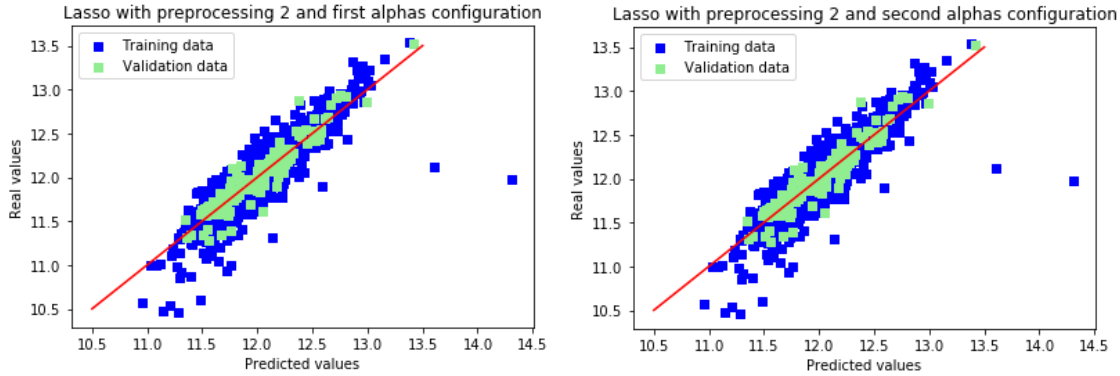

Figure 3. Gráfico dos resultados obtidos utilizando o método Lasso com o segundo pré-processamento realizado.

4.2. RidgeCV

O modelo Ridge consiste em um método de regularização do modelo que tem como principal objetivo suavizar atributos que sejam relacionados uns aos outros e que aumentam o ruído no modelo, a isso dá-se o nome de multicolinearidade[Ridge].

Para o projeto em questão, utilizou-se o regressor *RidgeCV*, que consiste no modelo *Ridge* juntamente com uma validação cruzada que permite a variação de seus parâmetros de modo a encontrar a melhor configuração dos mesmos dados vetores com diferentes valores dos parâmetros. Desse modo, variou-se em dois momentos distintos o parâmetro alpha e gerou-se gráficos tanto para o primeiro conjunto de treino quanto para o segundo, que podem ser observados nas figuras 4 e 5

4.3. ElasticNetCV

O método Elastic Net foi escolhido por ser composto de uma combinação linear entre as restrições dos métodos Lasso e Ridge.

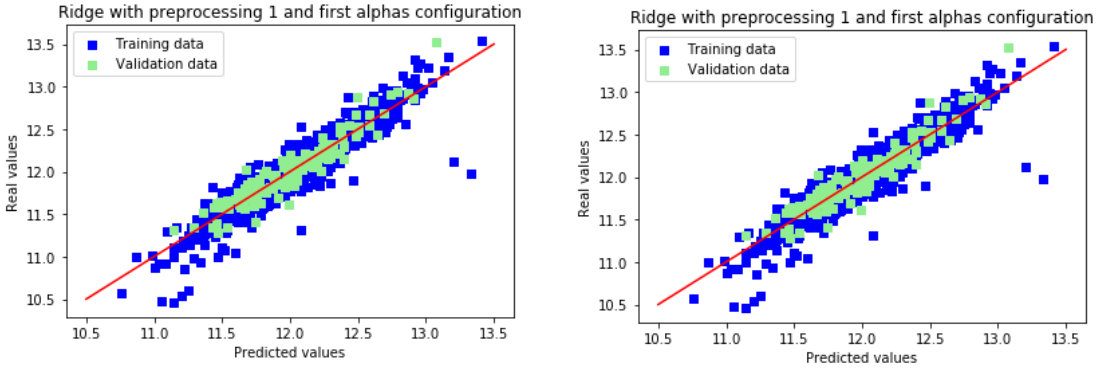

Figure 4. Gráfico dos resultados obtidos utilizando o método Ridge com o primeiro pré-processamento realizado.

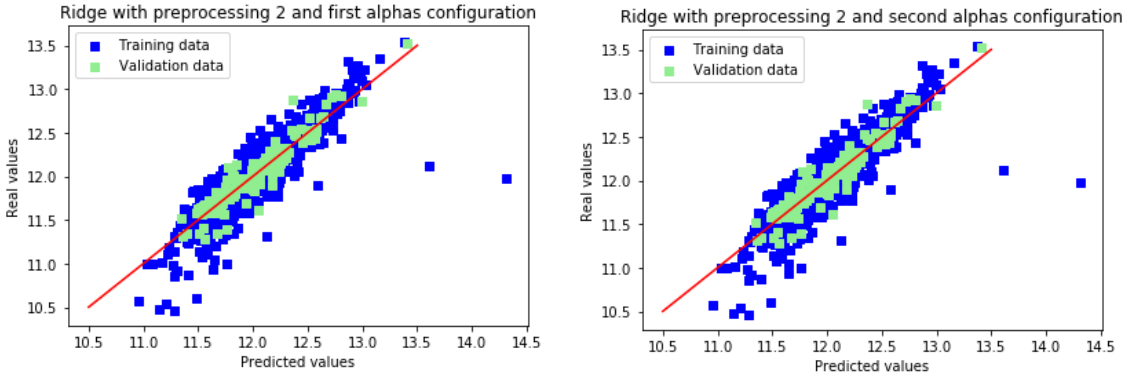

Figure 5. Gráfico dos resultados obtidos utilizando o método Ridge com o segundo pré-processamento realizado.

Dessa maneira,, utilizando-se da *ElasticNetCV* que permite a variação dos parâmetros da *ElasticNet*, variou-se as configurações alpha e l1_ratio a fim de obter melhores modelos e assim, podem ser observados nos gráficos 6 e 7 os resultados obtidos para o conjunto de pré-processamento 1 e 2, respectivamente.

4.4. XGBoost

Por fim, a técnica XGBRegressor, da biblioteca xgboost, foi aplicada para verificar sua eficácia, pois após os testes com as técnicas mencionadas anteriormente serem feitas e seus resultados computados, verificou-se, no site Kaggle que haviam alguns tutoriais de como realizar os experimentos e que método XGBRegressor era comumente utilizado pelos competidores. Assim, era interessante aplicá-lo para análises, então o mesmo foi inserido para realizar novas predições que seriam enviadas ao site.

5. Kaggle

Por fim, após as etapas de pré-processamento e regressão, os resultados gerados foram submetidos ao website de competições Kaggle, na competição intitulada de Houses Prices[Kaggle 2010] e os resultados obtidos podem ser observados na tabela 1

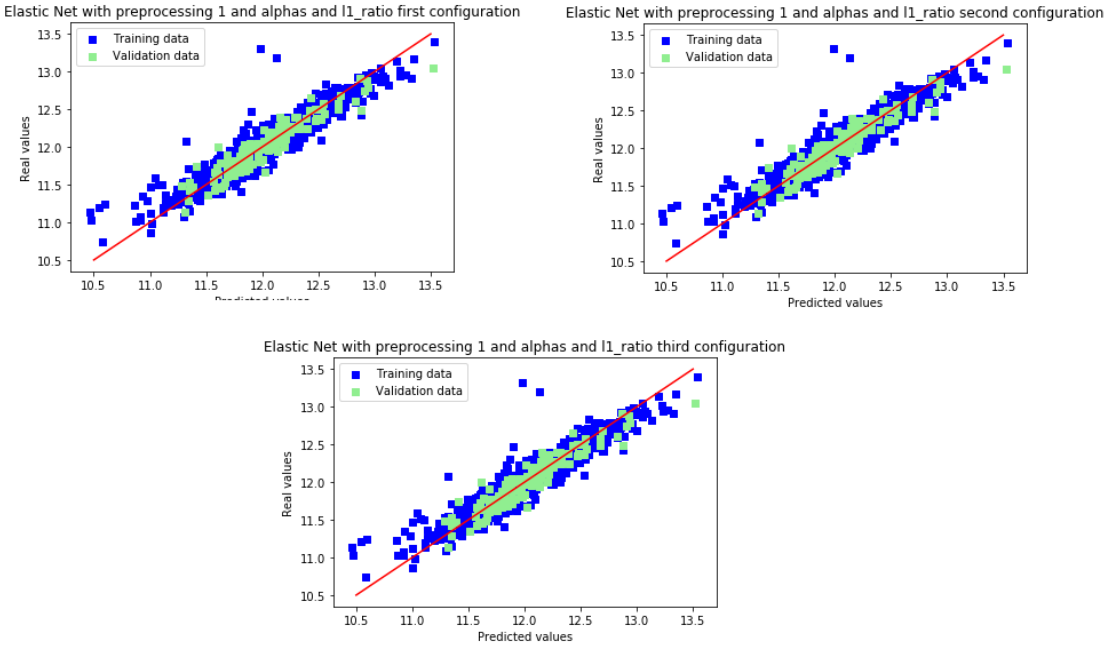

Figure 6. Gráfico dos resultados obtidos utilizando o método ElasticNet com o primeiro pré-processamento realizado.

Table 1. Tabela referente aos resultados obtidos nas submissões ao site Kaggle[Kaggle 2010].

	Preprocessing 1	*Preprocessing 2*
Lasso - Alphas 1	0.13397	0.16179
Lasso - Alphas 2	0.13348	0.16183
Ridge - Alphas 1	0.13367	0.16160
Ridge - Alphas 2	0.13383	0.16161
ElasticNet - Params 1	0.13449	0.16189
ElasticNet - Params 2	0.13416	0.16185
ElasticNet - Params 3	0.13427	0.16184
XGBRegressor	0.13344	0.15665

6. Conclusão

Após a análise dos resultados obtidos, concluimos que, com um pré-processamento mais aprimorado os resultados mostraram-se pior, isso pode ser atribuído ao método de avaliação das variáveis categóricas e/ou numéricas. Dessa maneira, sugere-se, para um estudo de maior prazo, métodos mais efetivos, como, por exemplo, o uso de diferentes métodos estatísticos como skewness, que é utilizado nos tutoriais da competição Houses Prices, no site Kaggle.

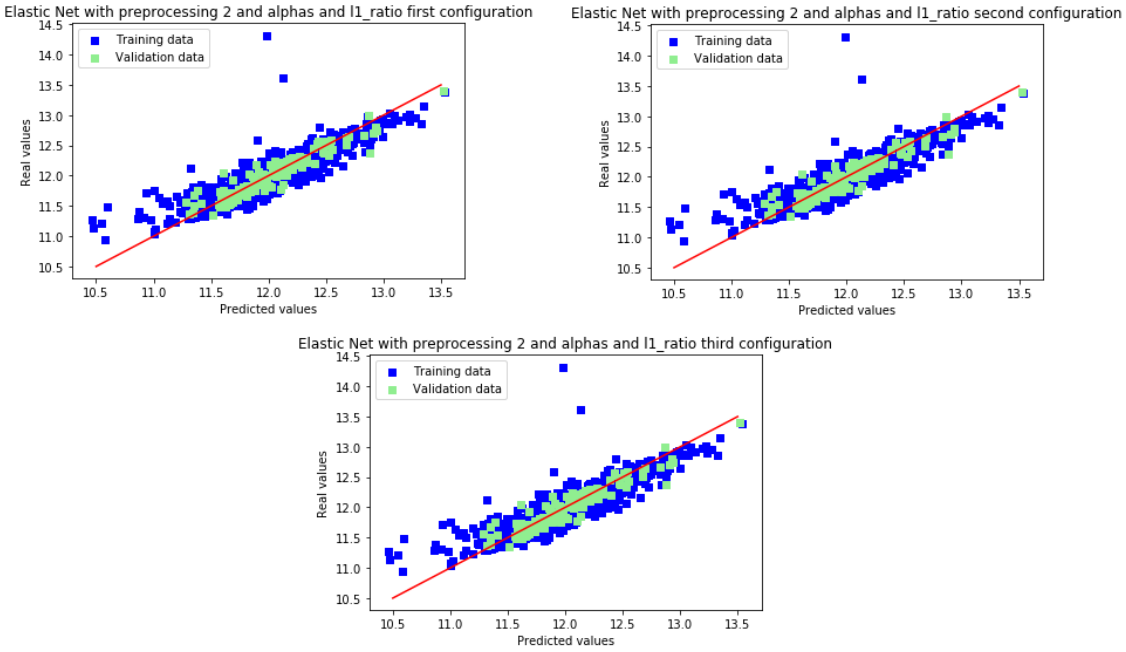

Figure 7. Gráfico dos resultados obtidos utilizando o método ElasticNet com o segundo pré-processamento realizado.

References

Jupiterweb (1999). Available at https://uspdigital.usp.br/jupiterweb/obterDisciplina?sgldis=SCC0277&codcur=55041&codhab=0.

Kaggle (2010). Available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

Lasso. Available at https://mineracaodedados.wordpress.com/2015/06/20/qual-a-diferenca-entre-lasso-e-ridge-regression/.

Minitab. Available at http://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-ana

Ridge. Available at https://mineracaodedados.wordpress.com/2015/06/20/qual-a-diferenca-entre-lasso-e-ridge-regression/.

sklearn. Available at http://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model.