NYC Airbnb Open Data: Exploração dos dados e Previsão do preço dos aluguéis

1st Caio Martim Barros
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
caiobarros@alu.ufc.br

2nd Lucas Vitoriano
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
lucasvitoriano25@gmail.com

3rd Mario Cesar F. D. Filho
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
mariocesarfreire@alu.ufc.br

***Abstract*—Airbnb é uma empresa operante no mercado online de hospdagens, de forma a permitir cadastro e busca de casas e apartamentos por meio da listagem e validação dos dados, proporcionando uma maior segurança e gerando um crescimento na sua busca dentro desse setor devido a inovação. Por conta de sua ascensão, é possível recolher dados sem muitas complicações e assim fazer estudos amplos, realizando uma análise exploratória desses dados a fim de entender mais sobre suas aplicações e criar materiais para uma abordagem mais específica utilizando esses dados disponíveis. Nesse sentido, nosso propósito para este artigo é, a partir da utilização de regressões lineares, fazer previsões utilizando as informações do *dataset* para estabelecer o valor do preço dos aluguéis dos imóveis e assim comparar com o valor real com intuito de entender se os valores cobrados são compreensíveis. Nesse âmbito, foram-se feitas previsões sobre os valores dos aluguéis com uma margem de erro média de 1.26 dólares sobre o valor real. Ademais, é realizada uma análise exploratória dos dados, efetuando gráficos e tabelas para uma abordagem aprimorada da situação do dataset.**

***Index Terms*—Airbnb, dataset, análise, previsão, regressão linear.**

I. Introdução

No contexto de grande volume de dados, a Airbnb é uma empresa em que se utiliza bastante como objeto de estudo e análise por conta da alta disponibilidade dos dados provenientes, tornando-o, assim, uma grande percussora no âmbito da Inteligência Artificial e do Processamento de Dados. Com a utilização da linguagem Python e do *dataset* "New York City Airbnb Open Data" [1] foi possível realizar uma previsão dos preços de aluguéis com o uso da tecnica de regressão linear. Além disso realizamos operações para identificação de padrões e comportamento dos dados colunares por meio da utilização de gráficos, como *scatter* e *biplot*. Dessa forma, é possível fazer análises como: quais regiões de Nova Iorque estão mais suscetíveis a usar esse tipo de tecnologia e quais são os locais mais caros de aluguéis.

Com os dados disponivéis, foi necessário realizar uma conversão entre dados categóricos e dados numéricos em três colunas do *dataset* [1], que foram *"neighbourhood_group"*, *"neighbourhood"* e *"room_type"*. Para esse propósito, foi-se utilizado duas técnicas, o cat.codes, que é uma opção da biblioteca pandas e o LabelEncoder, da biblioteca sklearn.

No que se diz a respeito das regressões lineares utilizadas para a criação das previsões dos aluguéis dos imóveis de Nova Iorque, pode-se destacar a utilização da regressão linear simples, regressão linear penalizada *(PLR)*, além das PLS e PCR. Dentre essas destacadas, foram feitas operações de comparação para identificar qual regressão se encaixa melhor no contexto e também foi utilizado as previsões do preço do aluguel por meio do RMSE e R^2. Nesse raciocínio, foi empregada o *k-fold cross validation* nas regressões com intuito de melhoria do modelo.

II. Análise dos dados

Nesta seção vamos abordar os dados que estamos trabalhando, explicar as variáveis que vamos considerar nas nossas regressões, técnicas que utilizamos para averiguar os dados e as transformações que neles fazeremos.

A. Conhecimento necessário

Nesse trabalho focamos em analisar os dados principalmente com foco no sua média, skewness e desvio padrão.

O desvio padrão indica se os valores de um preditor estão próximos ou muito espaçados, um alto valor de desvio padrão pode tornar um modelo impraticável, pois ele tentará abranger todos os valores, podendo gerar um overfitting. Para calcular o desvio padrão utilizamos a fórmula 1:

$$s = \sqrt{\frac{\sum_{i}^{n}(x_i - \mu_x^2)}{n-1}} \tag{1}$$

Outro valor importante é o Skewness, skewness é uma forma de analisar se o preditor está centralizado, se moda e a media estão na mesma posição, caso não esteja, métodos como PLS são afetados negativamente [1] Para calcular a skewness utilizamos a fórmula 2

$$\frac{\sum(x_i - \overline{x})^3}{(n-1)s^(3)} \tag{2}$$

Para descobrir qual a melhor regressão comparamos 5 tipos de modelos, regressão linear, regressão de ridge, regressão de lasso, PCR e PLS. A seguir faremos uma breve explicação sobre cada modelo.

Regressão linear traça uma reta que tenha ter o menor erro quadrático, a expressão da regressão possui fórmula 3:

$$y_i = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \tag{3}$$

Onde cada b_n é o beta que define a importância daquele preditor na definção da reta.
A regressão linear simples possui como medida de sucesso o menor erro quadrático, já a regressão de ridge adiciona um outro elemento a fórmula de erro quadrático, chegando a forma 4:

$$SSE_{L_2} = \sum_{n=1}^{n}(y_i - \overline{y_i})^2 + \lambda\sum_{j=1}^{P}\beta_j^2 \tag{4}$$

Na regressão de ridge temos que o lambda é uma forma de penalizar a bias em pro de obter uma menor variância no final, assim pode-se reduzir alguns preditores a valores próximos de 0, mas nunca iguais a ele.

Já no modelo de lasso temos uma regressão similar a de ridge, penalizada, mas com uma alteração na fórmula do SSE.

$$SSE_{L_1} = \sum_{n=1}^{n}(y_i - \overline{y_i})^2 + \lambda\sum_{j=1}^{P}|\beta_j| \tag{5}$$

Com essa alteração temos um modelo muito parecido com o de ridge, mas com a diferença que alguns preditores podem ser zerados.

Outro modelo que utilizamos foi o PCR, que diferente dos modelos anteriores ele utiliza primeiro uma PCA nos dados, para reduzir suas dimensões. Com o novo dataset após a redução é então realizado uma regressão linear para encontrar o coeficientes da reta.

O último modelo de regressão que utilizaremos é o PLS , partial least squares, que é um método supervisionado, a variável de saída é considerada no pré processo dos dados[2].

PLS é um método de regressão que primeiramente utiliza uma redução de dimensão,assim como o PCR, ele recebe os dados originais, reduz a dimensão dos preditores, transformando-os em novas colunas X1',X2',...,X'n, e em seguida realiza a regressão usando como método os mínimos quadrados unido com uma penalidade, para achar a regressão com menor erro.

B. Matriz Airbnb

No nosso dataset recebemos uma matriz 48895 linhas x 16 colunas, com as seguintes colunas:

- Id
- name
- host_id
- host_name
- neighbourhood_group
- neighbourhood
- latitude
- longitude
- room_type
- price
- minimum_nights
- number_of_reviews
- last_review
- reviews_per_month
- calculated_host_listings_count
- availability_365

Como queremos predizer o valor do aluguel para um cliente interessado em passar um período em Nova York podemos retirar colunas que não são interessantes para o cliente, assim, retirando essas colunas não oportunas podemos trabalhar em cima dos dados que nos são interessante.

C. Análise dos dados

Para os métodos que usaremos nesse trabalho é necessário que trabalhemos com dados numéricos, assim precisamos transformar os dados categóricos.

Para nossa matriz temos somente 3 colunas categóricas, como a coluna com a maior variedade de dados , neighbouhood, tem apenas 220 valores diferentes, podemos utilizar o método *labelencoder()*, assim garantimos que todas as colunas que estamos trabalhando são numéricas.

D. Padronização dos dados

Fazendo primeiramente uma análise no preditor price, que será o principal valor do nosso trabalho, vemos que ele possui um skewness muito positivo, oque pode levar a divergências no modelo de regressão linear que usaremos.

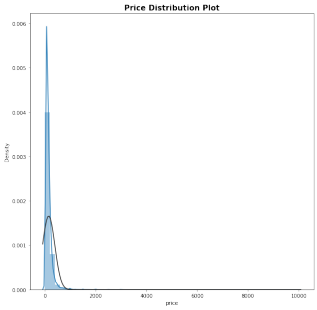

Fig. 1: Preditor Price com um alto skewness

Como vemos em 1, é necessário centralizar os dados, diminuir o skewness, a técnica de transformar a unidade de normal para logarítma foi aplicada, tornando os dados mais centralizados.

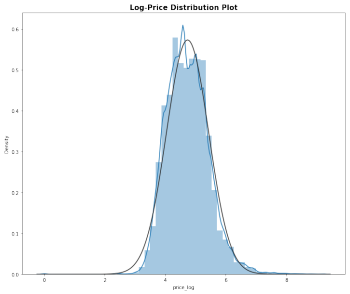

Fig. 2: Preditor Price na base log e

Para as outras variáveis calculamos o desvio padrão skewnesse e média,para que possamos fazer uma análise mais detalhada do dataframe que estamos trabalhando, oque gerou a tabela I.

TABLE I: Média desvio e skewness antes da normalização

variável	média	desvio padrão	skewness
latitude	40.728949	0.054530	0.237167
longitude	-73.952170	0.046157	1.284210
número_reviews	23.274466	44.550582	3.690635
reviews_por_mês	1.373221	1.497775	3.511906
grupo_bairro	1.675345	0.735816	0.373464
bairro	107.122732	68.743096	0.256005
tipo_quarto	0.504060	0.545379	0.422704
Residência por host	7.143982	32.952519	7.933174
Disponibilidade	112.781327	131.622289	0.763408
price_log	4.736885	0.695344	0.553105
noites_mínimas	7.029962	20.510550	21.827275

Vemos que em certos preditores o desvio padrão e a média estão com valores muito altos, valores esses que precisam ser normalizados para ter média igual a 0 e desvio padrão igual a 1, assim quando fizermos operações como a análise de componente principal teremos maior precisão nos resultados. Para realizarmos essa normalização, utilizamos a função *standardscaler*, com esses dados trabalhados, obtivemos os seguintes novos valores:

TABLE II: Média desvio e skewness depois da normalização

variável	média	desvio padrão	skewness
latitude	3.949141e-14	1.000010e+00	0.237167
longitude	662286e-13	1.000010e+00	1.284210
número_reviews	-2.315687e-14	1.000010e+00	3.690635
reviews_por_mês	-3.349251e-15	1.000010e+00	3.511906
grupo_bairro	-8.991399e-15	1.000010e+00	0.373464
bairro	5.305490e-16	1.000010e+00	0.256005
tipo_quarto	7.635292e-15	1.000010e+00	0.422704
Residência por host	3.545441e-14	1.000010e+00	7.933174
Disponibilidade	1.387977e-14	1.000010e+00	0.763408
price_log	2.582160e-15	1.000010e+00	0.553105
noites_mínimas	-7.665145e-16	1.000010e+00	21.827275

Como vemos em II, agora nossos preditores estão normalizados. Entretanto, notamos que o skewness de longitude, numero reviews, reviews por mês, residência por host e noites mínimas são positivos, representando que a moda é menor q a média, oque é um valor aceitável, pois, com exceção de longitude, que é uma característica da região de nova york , as variáveis estão sujeitas a ações humanas, oque pode gerar outliers, deslocando a média para a direita, mas, como esses outliers são dados importantes para considerar na nossa análise, optamos por não tirarmos.

E. Análise bi-variada

Como nós estamos tratando com mais de 10 variáveis, seria muito custoso abordamos todas as relações entre as variáveis aqui, portanto, optamos por mostrar ao leitor apenas as 4 variáveis que possuem mais influência no nosso output, o preço. Para obtermos essa ordem de influência pegaremos os maiores valores obtidos em módulo na tabela de correlação.

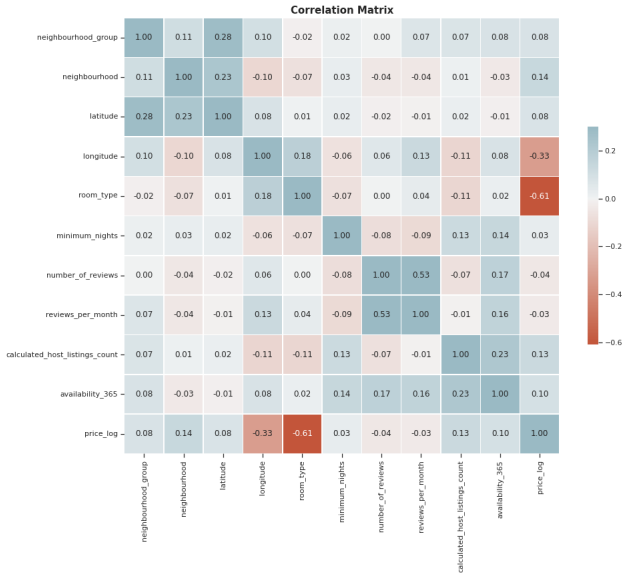

Fig. 3: Matriz de correlação entre todos os preditores

Observando a tabela 3, notamos que existem duas relações moderadas ,tipo_quarto e longitude, o resto das relações são fracas, sendo vizinhança e residências por host as duas maiores desse grupo. Assim, essas quatro são os que tem as maiores correlações em módulo com price_log, para mostrarmos mais detalhes essas relações traremos o scatterplot 4.

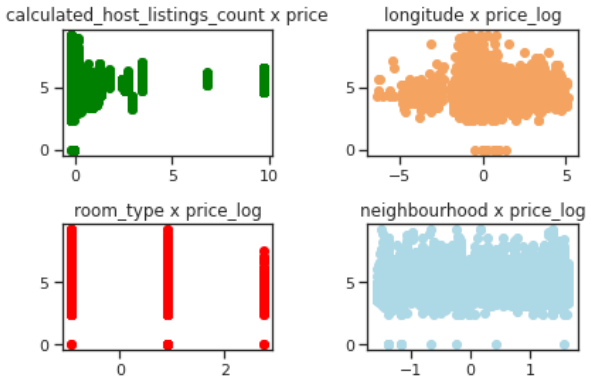

Fig. 4: ScatterPlot's preço x preditores

Dentre os plots que obtivemos conseguimos retirar informações importantes sobre cada um dos 4 preditores mais correlacionados com o preço.

Os valores de longitude intermediários alcançam uma média maior comparada aos valores de longitude das extremidades, oque é explicado pela própria geografia de Nova York, onde a região central possui diversos pontos turísticos como o central park e a times square .

Para tipo de quarto (room_type) e neighbourhood podemos realizar uma análise mais detalhada se mudarmos o plot para um do tipo de barra.

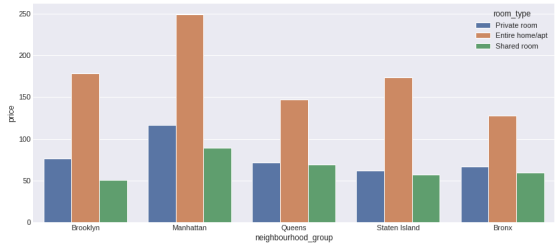

Fig. 5: Preço por quarto e bairro

Com esse novo gráfico 5, conseguimos visualizar que os alugueis do tipo entire room/apt são aqueles com maior média e aqueles do tipo shared room os com menor média. Além disso também conseguimos visualizar que os alugueis locados em Manhattam são os mais caros de todos os 5.

Com os outros preditores não conseguimos ter realizar uma boa análise somente com o plot preditor x price, oque pode resultar em um péssimo fit da regressão, perdendo precisão.

F. Análise da PCA

Como vimos na última sessão, os dados que estamos tratando possuem diversas variáveis, assim, torna-se útil realizarmos uma PCA, para extrair as informações mais importantes da tabela [5], reduzir a dimensão, usar menos variáveis e manter uma variedade de dados suficente para podermos gerar uma boa regressão.

Realizando o principal component analysis com a restrição de dimensão n = 2 chegamos a dois autos valores. Para checarmos se 2 componentes principais é o suficiente para termos uma boa variância calculamos tabém a variância associada a cada componente, chegando a tabela III:

TABLE III: Autovalores e variância associada

autovalores	variância
1.74048852	0.16601841
1.55728554	0.14854339

Como vemos, apenas dois componentes não é o suficiente para representar bem os dados, assim, vamos ilustrar melhor a variância associada a cada componente com o uso do scree plot 6.

Fig. 6: Variância por número de componentes

Exemplificado pelo gráfico 6, para alcançar uma variância de aproximadamente 60% seria necessário ao menos 5 componentes, mostrando assim que nossos dados não possui uma boa separação.

Outra forma de vermos que nossos dados não são bem separados é com o uso do scatter plot 7.

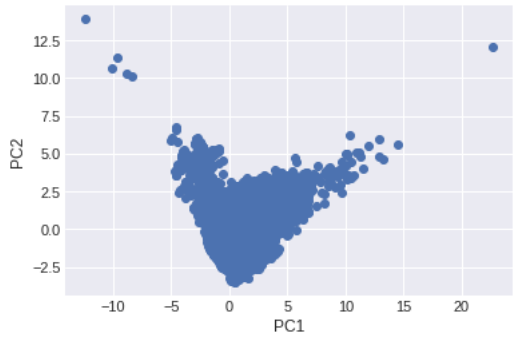

Fig. 7: ScatterPlot

Como os dados estão muito próximos vemos que apenas duas dimensões não é o suficente para bem representa-los pois temos muita sobreprosição de dados entre PC1 e PC2, principalmente na regiao x $\approx$ 0 e y $\approx$ -1.

III. Resultados

Nesta seção será discutidos gráficos sobre o dataset e valores encontrados dos aluguéis por meio da utilização de regressões lineares. Para isso, nesse caso, foi-se utilizada as 10 colunas de preditores mais a coluna do *log_price*, como já mencionado em seções anteriores. Nesse dataset, então, não foi necessário fazer uma limpeza dos dados pois as colunas utilizadas não possuíam valores nulos, mas técnicas como a transformação dos dados categóricos para númericos utilizando a cat.codes do Pandas foi utilizada, como também a aplicação do método de normalização, para padronização dos dados, utilizando o *StandardScaler()*.

Por fim, tem-se um dataframe normalizado, com 11 colunas no total e 48895 linhas de dados. Para a divisão do treino e teste, foi-se utilizada a própria biblioteca do sklearn, *train_test_split()*, onde os dados foram divididos entre 70% para treino e 30% para teste.

A. Regressão Linear Simples

TABLE IV: Resultados obtidos na regressão linear

	Regressão Linear Simples	**Regressão Linear 5-fold**
RMSE	0.536052	0.538368
R^2	0.501930	0.497618

Pela tabela IV acima, percebe-se uma comparação entre a regressão linear simples e a regressão linear utilizando o cross validation com 5 fold. Dessa tabela, inclui-se que o RMSE para a regressão 5-fold teve um valor um pouco maior e o R^2 um pouco menor, valores estes tão pequenos que poderia considerar irrelevante para o processo de diferença entre as duas regressões lineares.

Para a validação da corretude da previsão, foi-se criada uma tabela em que compara os valores reais dos preços dos aluguéis

TABLE V: Margem de erro entre os valores encontrados na regressão linear 5-fold e os valores reais dos aluguéis

	AV (log)	PV (log)	Error (Dolar)
0	6.111467	5.084200	2.793422
1	5.298317	4.440909	2.357045
2	4.709530	4.474065	1.265498
3	4.510860	5.191789	0.506146
...	...	...	...
9776	4.753590	5.117566	0.694908
9777	4.025352	3.802666	1.249427
9778	4.510860	4.512676	0.998185

e os valores encontrados na previsão desses preços utilizando a regressão linear 5-fold. É importante ressaltar, então, a pouca diferença entre as colunas destacadas, onde o valor do erro médio entre a diferença do valor real e valor encontrado é de apenas 1.2598.

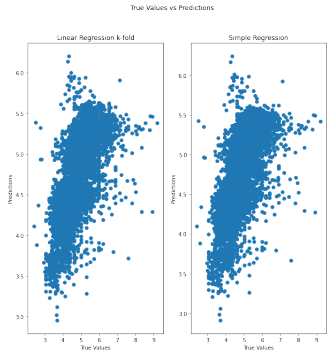

Fig. 8: Regressão Simples vs. Regressão 5-fold

B. Regressão Linear Penalizada

Nesse tópico, foi comparada três modelos de regressão linear penalizada: Ridge, Lasso e ElasticNet, a efeito de identificar qual encaixaria melhor na nossa previsão de aluguéis.

Nesse sentido, primeiramente foi-se necessário encontrar o valor ótimo do parâmetro de penalização (λ) da regressão linear desses três modelos e, para isso, criou-se um algoritmo autoral para a identificação desse parâmetro. Esse algoritmo realiza um loop no teste dessas regressões onde ele vai buscando o melhor parâmetro para cada parâmetro colocado nesse teste, onde inicia-se o valor do parâmetro em 0 e vai acrescentando esse valor em 0.01 até chegar no valor 1, fazendo comparações com o RMSE e atualizando o valor do parâmetro quando esse for o melhor até o momento. Da mesma forma, existe um algoritmo próprio da biblioteca sklearn que busca esses melhores valores e, em comparação ao nosso, mostra um pouco de diferença de valor.

No final, o algoritmo realiza todos os processos e retorna os valores ótimos (λ) desses parâmetros, em que são: ridge = 0.01, lasso = 0.001 e elasticnet = 0.01. Nesse sentido, é o fator diferenciador entre os reguladores e, sua não utilização implica em valores iguais de RMSE e R^2.

Para constatar qual dos três modelos é realmente o melhor para o propósito do artigo, foi realizada uma comparação entre cada modelo para a previsão dos aluguéis utilizando o cross validation 5-fold e o melhor λ para cada situação.

1) Ridge erro médio (em dólar): 1.2620

TABLE VI: Margem de erro entre os valores encontrados na regressão Ridge e os valores reais dos aluguéis

	AV (log)	PV (log)	Error (Dolar)
0	6.111467	5.077259	2.812880
1	5.298317	4.443183	2.351690
2	4.709530	4.476372	1.262581
3	4.510860	5.187131	0.508510
...	...	...	...
9776	4.753590	5.112807	0.698223
9777	4.025352	3.811455	1.238494
9778	4.510860	4.514263	0.996603

2) Lasso erro médio (em dólar): 1.2653

TABLE VII: Margem de erro entre os valores encontrados na regressão Lasso e os valores reais dos aluguéis

	AV (log)	PV (log)	Error (Dolar)
0	6.111467	5.080281	2.804392
1	5.298317	4.436944	2.366409
2	4.709530	4.472182	1.267883
3	4.510860	5.190518	0.506790
...	...	...	...
9776	4.753590	5.114348	0.697148
9777	4.025352	3.801700	1.250636
9778	4.510860	4.510538	1.000322

3) ElasticNet erro médio (em dólar): 1.2888

TABLE VIII: Margem de erro entre os valores encontrados na regressão ElasticNet e os valores reais dos aluguéis

	AV (log)	PV (log)	Error (Dolar)
0	6.111467	5.061609	2.857248
1	5.298317	4.422529	2.400768
2	4.709530	4.466036	1.275699
3	4.510860	5.183244	0.510490
...	...	...	...
9776	4.753590	5.099358	0.707676
9777	4.025352	3.802469	1.249674
9778	4.510860	4.503026	1.007864

Como mostrado nas tabelas acima, os valores são muito próximos e percebe-se uma pequena diferença entre os três modelos. Para poder destacar um modelo exclusivo, a tabela IX compõe os valores de RMSE e R^2 a qual será usada para identificar qual é o melhor modelo.

TABLE IX: Resultados obtidos de cada modelo prescrito

	Ridge	Lasso	ElasticNet
RMSE	0.538921	0.538550	0.540712
R^2	0.496584	0.497277	0.493233

Dessa forma, percebe-se, pelos valores da tabela IX, que o modelo Ridge é o que melhor se encaixa no nosso propósito de previsão dos preços de aluguéis dentre os três mencionados, visto que possui um erro médio menor, 1.2620, e uma composição de RMSE e R^2 menor em comparação aos outros dois modelos.

No ponto de vista estatístico, os três modelos contém fórmulas parecidas, diferenciando apenas por conta de um termo, onde para o Lasso é a soma dos valores absolutos dos coeficientes 5, para o Ridge é a soma dos quadrados dos coeficientes 4 e para o ElasticNet é a soma dos outros dois juntos [4].

Em conclusão, a figura 9 abaixo demonstra a distribuição em um modelo scatter das previsões dos valores em função de cada modelo e também em função da regressão linear 5-fold, onde consegue-se visualizar poucas variações entre os pontos em cada modelo, porém existe uma diferença mínima onde faz com que o modelo de regressão linear simples 5-fold ser o melhor no caso, pois possui um erro médio 1.2598.

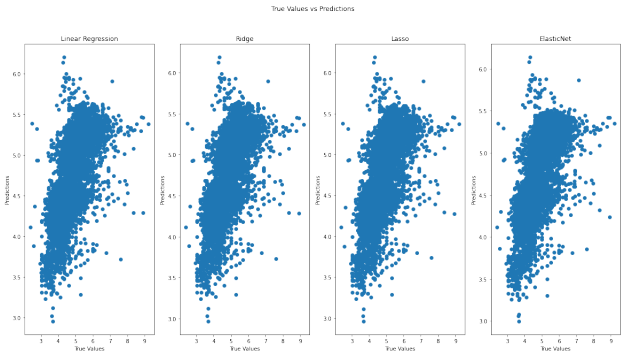

Fig. 9: Comparação entre as regressões

C. PLS vs. PCR

Para escolher qual regressão é melhor nesse caso, entre PCR (Regressão por Componente Principal) e PLS (Regressão parcial de mínimos quadrados), será feita uma comparação com os valores de RMSE e R^2 encontrados no dataset utilizando uma configuração padrão para ambos, que é com o número de componentes igual a 1. Dessa forma, o que apresentar os melhores valores de ambos parâmetros, será escolhido como o modelo de revisão e, a partir dele, será feita uma análise para identificar qual é o número ótimo de componentes.

TABLE X: PCR vs. PLS

	PCR	**PLS**
RMSE	0.649	0.530
R^2	0.307	0.262

Desse modo, percebe-se que PLS obteve um valor significativamente melhor do que o PCR e, por conta disso, será o modelo utilizado nesse tópico. É valido ressaltar, também, que a regressão parcial de mínimos quadrados (PLS) é uma técnica que reduz os preditores a um conjunto menor de componentes não correlacionados e efetua regressão de mínimos quadrados para esses componentes no lugar dos dados originais, sua maior diferença, desse modo, é a redução da dimensão dos dados, transformando o dataset em colunas menores, mas mantendo uma relação de importância com o resultado. Em comparação ao PCR, ele faz a redução da dimensão com a utilização do PCA em conjunto com o cross validation, porém não se tem a segurança da dessas novas variáveis estarem relacionadas com o resultado previsto [4].

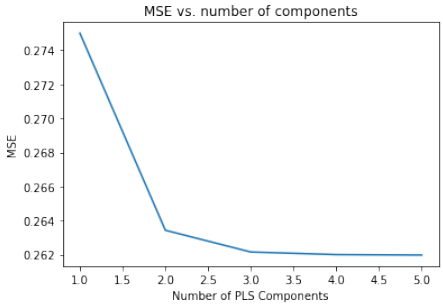

Fig. 10: Comparação entre as regressões

Pela figura 10, pode-se garantir que o número ótimo de componentes deve ser 3, pois na curva mostrada na figura, percebe-se que os melhores valores para mse (baixo valor) é quando o número de componentes é igual a 3.

Portanto, utilizando esse número de componentes igual a 3 no algoritmo, encontrou-se os valores de RMSE = 0.517 e R^2 = 0.262

IV. Conclusão

Em suma, pode-se destacar que existe sim diferença entre o treinamento e teste dos dados, a qual este último é usado apenas como efeito de comparação com o que foi usado durante o treinamento, em que no treino são utilizadas técnicas e métodos para chegar a uma previsão próxima ao que existe no teste. Nesse raciocínio, o treinamento dos dados é para a criação do nosso modelo previsor, no caso de aluguéis do Airbnb, e o teste é para qualificar a perfomance do modelo construído.

Pelos resultados das regressões mostradas, percebe-se que a maioria delas obtiveram valores semelhantes, porém, a efeito de destaque, é importante ressaltar a regressão por mínimos quadrados (PLS), a qual obteve os melhores RMSE e R^2 em comparação ao restantes das regressões realizadas e, por conta disso, ela é o modelo referência para esse tipo de estudo realizado. Á efeito de enfatizar esse melhor valor, o erro médio, que é a margem entre o valor real e o valor encontrado pela regressão PLS é de 1.2015 e, se formos comparar com o melhor anteriormente, que foi de 1.2598 pela regressão linear simples 5-fold, mostra que o PLS é realmente o melhor modelo para o nosso caso.

References

[1] Dgomonov, "New York City Airbnb Open Data".

[2] Max Kuhn, Kjell Johnson, "Applied Predictive Modeling".

[3] SWAMY, Vijay. "Lasso Versus Ridge Versus Elastic Net". Disponível em: https://medium.com/@vijay.swamy1/lasso-versus-ridge-versus-elastic-net-1d57cfc64b58

[4] kassambara. "Principal Component and Partial Least Squares Regression Essentials". Disponível em: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/152-principal-component-and-partial-least-squares-regression-essentials/

[5] Herve Abdi and Lynne J. Williams. "Principal component analysis".