



UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

ESCOLA POLITÉCNICA

DEPARTAMENTO DE ELETRÔNICA E DE

COMPUTAÇÃO

Introdução ao Aprendizado de Máquina

Trabalho 2

Preditor de preços de imóveis

Juan Coutinho Lima

3 de junho de 2021

Sumário

Lista de Figuras	1
Lista de Tabelas	1
1 Introdução	2
2 Estudo da base de dados	2
2.1 Variáveis categóricas	2
2.1.1 Descarte das variáveis categóricas	2
2.2 <i>One-hot encoding</i>	3
2.3 Binarização	3
2.4 Escolha dos atributos	4
3 Modelos Analisados	5
3.1 <i>LinearRegression - PolynomialFeatures</i>	5
3.2 <i>Ridge</i>	6
3.3 <i>LogisticRegression</i>	6
4 Conclusões	7

Lista de Figuras

1	Exemplo de One-hot encoding	3
2	One-hot encoding RPA	3
3	Exemplo de binarização	4
4	Análise do coeficiente de correlação	4

Lista de Tabelas

1	Variáveis categóricas	2
2	Regressão polinomial	6
3	Resultados do regressor Ridge	6
4	Resultados LogisticRegression	6

1 Introdução

O objetivo deste trabalho é construir um preditor que seja capaz de, a partir de uma base histórica de venda de imóveis, definir o valor de um imóvel, com base em atributos fornecidos obtidos através da base de dados.

2 Estudo da base de dados

A nossa base de dados consiste em um arquivo CSV com 20 atributos, sendo o parâmetro “preco” a nossa variável alvo.

2.1 Variáveis categóricas

Para que seja possível avaliar a adequação dos nossos atributos, devemos primeiro atestar que todos eles se tratam de variáveis numéricas. O primeiro passo a ser feito é identificar quais são as variáveis categóricas. Isso pode ser realizado ao listar todas as colunas do tipo “object”.

Variável	Tipo	Número categorias
bairro	não ordinal	66
tipo	não ordinal	4
tipo_vendedor	binário	2
diferenciais	não ordinal	83

Tabela 1: Variáveis categóricas

2.1.1 Descarte das variáveis categóricas

O objetivo desta seção é identificar quais foram as variáveis descartadas e o motivo desta decisão.

Apenas o atributo diferenciais foi excluído. Isso se deu pois esta coluna de atributos possui um número grande de categorias, o que ampliaria consideravelmente a dimensão da nossa regressão. Além disso, algumas das categorias apresentadas nessa coluna são redundantes, pois já estão especificadas numa coluna diferente.

2.2 One-hot encoding

Agora que as variáveis relevantes foram separadas, nos resta atribuir um tipo numérico para os seus campos. Essa atribuição de valor sera feita através da técnica de *one-hot encoding*. Tomemos o atributo “tipo” como exemplo. Para esta coluna temos 4 categorias possíveis: “Casa”, “Apartamento”, “Loft”, e “Quitinete”. Com o uso da técnica de *one-hot encoding* teremos o seguinte resultado:

tipo_Apartamento	tipo_Casa	tipo_Loft	tipo_Quitinete
0	0	0	1
0	0	0	1
0	1	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0

Figura 1: Exemplo de One-hot encoding

A mesma técnica foi aplicada ao atributo “bairro”, com a diferença de que antes foi necessário agrupar os estados em regiões político-administrativas (RPA). Esta consulta foi realizada através do site da prefeitura de recife, disponível em: <http://www2.recife.pe.gov.br/servico/perfil-dos-bairros>. Uma pequena amostra desta operação pode ser vista na figura abaixo.

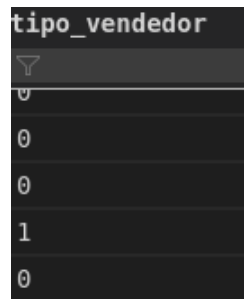
RPA_RPA1	RPA_RPA2	RPA_RPA3	RPA_RPA4	RPA_RPA5	RPA_RPA6
1	0	0	0	0	0
1	0	0	0	0	0
0	0	0	0	0	1
0	0	1	0	0	0
0	1	0	0	0	0
0	0	0	0	0	1
0	1	0	0	0	0
1	0	0	0	0	0

Figura 2: One-hot encoding RPA

2.3 Binarização

A ultima parte do tratamento das variáveis categóricas consiste em transformar as variáveis com apenas duas categorias em campos com 1 e 0. Um exemplo da necessi-

dade dessa aplicação é o atributo “tipo_vendedor” que possui os valores “Pessoa Física” e “Imobiliária”. Com a aplicação da técnica obtivemos o seguinte resultado:



tipo_vendedor
0
0
0
1
0

Figura 3: Exemplo de binarização

2.4 Escolha dos atributos

Para dar sequência ao estudo foi necessário definir quais atributos seriam considerados pelo nosso modelo de regressão. O critério de escolha desses atributos foi realizado primeiramente através do coeficiente de correlação de pearson e depois através de testes para verificar se a seleção contribuiu com um resultado favorável. A comparação entre cada coluna e o preço ficou da seguinte forma:



Correlacao Preço x Coluna:	
preco	= 1.0000
tipo_vendedor	= 0.1240
suites	= 0.0421
vagas	= 0.0290
area_util	= 0.0276
quartos	= 0.0209
RPA_RPA3	= 0.0194
churrasqueira	= 0.0159
RPA_RPA2	= -0.0142
playground	= -0.0122
tipo_Casa	= 0.0090
tipo_Apartamento	= -0.0085
vista_mar	= 0.0077
s_festas	= 0.0075
piscina	= 0.0068
estacionamento	= -0.0064
RPA_RPA6	= -0.0059
RPA_RPA5	= 0.0055
RPA_RPA1	= -0.0049
quadra	= -0.0044
RPA_RPA4	= -0.0028
s_jogos	= -0.0025
s_ginastica	= -0.0023
tipo_Quitinete	= -0.0015
tipo_Loft	= -0.0014
sauna	= 0.0012
area_extra	= 0.0011

Figura 4: Análise do coeficiente de correlação

A partir desta lista e de experimentação a seguinte lista de atributos foi alcançada:

Atributos considerados:

- tipo_Casa
- tipo_vendedor
- quartos
- suites
- vagas
- area_util
- estacionamento
- playground
- s_ginastica
- sauna
- vista_mar
- RPA_RPA3
- RPA_RPA5

3 Modelos Analisados

Como critério de comparação dos modelos foi utilizada a métrica RMSPE, que não é afetada pela magnitude da série.

$$RMSPE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{y}_i - y_i}{y_i} \right)^2}$$

3.1 *LinearRegression - PolynomialFeatures*

Para que fosse feita uma primeira tentativa, foi decidido que usaríamos o modelo de regressão polinomial. Para isso, passamos o valor dos parâmetros deste classificador através de um laço de iteração, com o objetivo de descobrir os parâmetros ótimos para este modelo. através deste método o melhor RMSPE foi de 6,16 com um polinômio de grau 1.

K	RMSPE
1	6.1621
2	99529991153.2926
3	8961071513.3371

Tabela 2: Regressão polinomial

3.2 *Ridge*

Para implementar o uso deste regressor foi utilizada a estrutura já existente para a obtenção dos parâmetros de um polinômio de ordem K. Assim, de maneira iterativa obtivemos o seguinte resultado:

K	RMSPE
1	4.8119
2	7.8617
3	12.5862
4	9.7843

Tabela 3: Resultados do regressor Ridge

3.3 **LogisticRegression**

O último regressor investigado foi o de regressão logística, onde o algoritmo de otimização que apresentou o melhor resultado foi o liblinear. Variando o parâmetro “C” de 10×10^{-5} a 10×10^3 obtivemos o seguinte resultado:

K	RMSPE
-5	0.4412
-4	0.4306
-3	0.4491
-2	0.4463
-1	0.3990
0	0.3590
1	52.8567
2	61.6272
3	41.6847

Tabela 4: Resultados LogisticRegression

4 Conclusões

Pelos resultados alcançados podemos afirmar que o melhor regressor encontrado foi o que utiliza o modelo de regressão logística, o que nos proporcionou um RMSPE de 0,3590 no melhor caso.