

MODELO DE REGRERSSÃO LINEAR

Versão 2

Autora: JULIANA CRUZ

23/02/2024

Florianópolis, SC

Sumário

1. ANÁLISE EXPLORATÓRIA DE DADOS – AED	3
1.1. Informações Gerais.....	3
1.2. Ajuste dos dados antes da análise.....	3
1.3. Tipos de Atributos	3
1.4. Identificação dos dados faltantes.....	3
1.5. Análise dos registros nominais/categóricos.....	4
1.5.1. room_type distribuição dos registros por categoria	4
1.5.2. bairro_group e bairro distribuição dos registros por categoria	4
1.5.3. distribuição dos registros por usuário.....	5
1.5.4. Avaliação da frequência de palavras do nome com “price” >= 1000	6
2. PRÉ-PROCESSAMENTO DOS DADOS	6
2.1.1. Transformação das variáveis categorias em booleanas	6
2.1.2. Tratamento dos dados faltantes e outlier	6
3. CORRELAÇÃO.....	7
4. REGRESSÃO.....	8
4.1.1. MODELO 01 LinearRegression	8
4.1.2. MODELO 02 Random Forest Regressor.....	8

1. ANÁLISE EXPLORATÓRIA DE DADOS – AED

1.1. Informações Gerais

Os dados analisados referem-se a aluguel de estabelecimentos por diária, sendo possível alugar a casa/apartamento inteiro, quarto privativo ou quarto compartilhado. O total de atributos são 16 e total de registros são 48.894, a variável alvo para a realizar a predição de precificação será “price”.

Abaixo a lista dos atributos e suas respectivas descrições:

- **id** – Atua como uma chave exclusiva para cada anúncio nos dados do aplicativo
- **nome** - Representa o nome do anúncio
- **host_id** - Representa o id do usuário que hospedou o anúncio
- **host_name** – Contém o nome do usuário que hospedou o anúncio
- **bairro_group** - Contém o nome do bairro onde o anúncio está localizado
- **bairro** - Contém o nome da área onde o anúncio está localizado
- **latitude** - Contém a latitude do local
- **longitude** - Contém a longitude do local
- **room_type** – Contém o tipo de espaço de cada anúncio
- **price** - Contém o preço por noite em dólares listado pelo anfitrião
- **minimo_noites** - Contém o número mínimo de noites que o usuário deve reservar
- **numero_de_reviews** - Contém o número de comentários dados a cada listagem
- **ultima_review** - Contém a data da última revisão dada à listagem
- **reviews_por_mes** - Contém o número de avaliações fornecidas por mês
- **calculado_host_listings_count** - Contém a quantidade de listagem por host
- **disponibilidade_365** - Contém o número de dias em que o anúncio está disponível para reserva

1.2. Ajuste dos dados antes da análise

- substituição de “ ; ” por "vazio" de 121 erros na coluna “nome”
- retirada de quebra de linha de 185 erros na coluna “nome”
- correção dos nomes de colunas retirando espaços e caracteres especiais

1.3. Tipos de Atributos

Qualitativo		Quantitativo	
Categórico/Nominal	Ordinal	Discreto	Contínuo
nome, host_name, bairro_group, bairro, room_type		ID, host_id, minimo_noites, numero_de_reviews, calculado_host_listings_count, disponibilidade_365	latitude, longitude, price, reviews_por_mes,

1.4. Identificação dos dados faltantes

- ultima_review e reviews_por_mes contém 10.052 registros *Null*, isso pode levar a uma hipótese de serem de anúncios novos ou que ainda não foram avaliados.
- nome contém 16 registros *Null*
- host_name contém 21 registros *Null*

Na etapa de pré-processamento será realizado o tratamento dos dados faltantes. Os registros não possuem dados duplicados a partir do atributo “id”.

1.5. Análise dos registros nominais/categóricos

A seguir a distribuição dos registros dos atributos relevantes:

1.5.1. *room_type* distribuição dos registros por categoria

- Entire home/apt 0.519675
- Private room 0.456600
- Shared room 0.023725

O principal tipo de acomodação é Entire home/apt com 51,96% de registros.

1.5.2. *bairro_group* e *bairro* distribuição dos registros por categoria

Grupo de bairro

- Manhattan 0.443020
- Brooklyn 0.411155
- Queens 0.115883
- Bronx 0.022314
- Staten Island 0.007629

Os bairros de Manhattan e Brooklyn representam 85,41% do total de anúncios.

Bairro

- Williamsburg (Brooklyn) 0.080173
- Bedford-Stuyvesant (Brooklyn) 0.075960
- Harlem (Manhattan) 0.054362
- Bushwick (Brooklyn) 0.050415
- Upper West Side (Manhattan) 0.040312

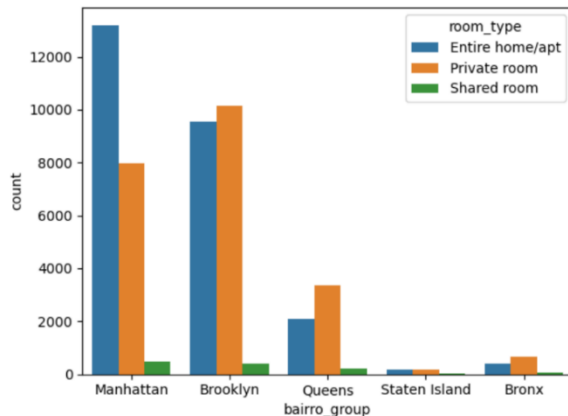
Os cinco principais bairros representam 30,12% do total de anúncios.

Valores percentuais

A figura ao lado demonstra a distribuição dos registros entre os grupos de bairros comparado ao tipo de acomodação, os grupos que possuem a maior concentração dos aluguéis são Manhattan e Brooklyn, no primeiro a acomodação “Entire home/apt” é 60,93% do total, e no segundo a acomodação “Private room” representa 50,4% do total.

room_type	Entire home/apt	Private room	Shared room
bairro_group			
Bronx	0.347388	0.597617	0.054995
Brooklyn	0.475501	0.503955	0.020544
Manhattan	0.609344	0.368496	0.022160
Queens	0.369926	0.595129	0.034945
Staten Island	0.471850	0.504021	0.024129

Valores inteiros



Ao lado a distribuição dos anúncios por tipo de acomodação e grupo de bairros.

Valores percentuais

Ao analisar os registros quanto a localização dos aluguéis, o grupo de bairro Manhatan tem média de aluguel de 196,87 e o Brooklyn a média de aluguel é 124,38, seriam as melhores opções para adquirir imóvel para investimento em aluguel.

	min	max	sum	price
bairro_group				
Bronx	0	2500	95459	87.496792
Brooklyn	0	10000	2500451	124.381983
Manhattan	0	10000	4264527	196.875814
Queens	10	10000	563867	99.517649
Staten Island	13	5000	42825	114.812332

1.5.3. distribuição dos registros por usuário

São 48.894 registros de nomes anúncios (nome), já no atributo usuário que hospedou (host_id) são 37.457 registros diferentes, neste caso a média de anúncios por usuário é de aproximadamente 1,3 anúncios.

Os cinco principais ids de usuários representam 1,79% do total dos registros:

host_id

1. 219517861 0.006688
2. 107434423 0.004745
3. 30283594 0.002475
4. 137358866 0.002107
5. 12243051 0.001963

Os cinco principais nomes de usuários representam 3,52% do total dos registros:

1. Michael 0.008532
2. David 0.008246
3. Sonder (NYC) 0.006691
4. John 0.005995
5. Alex 0.005709

1.5.4. Avaliação da frequência de palavras do nome com “price” >= 1000

Luxury	28	Seleção de nomes de anúncio com valores de “price” maior ou igual a 1000. O conjunto de dados foi transformado em uma série com a frequência das palavras que mais aparecem nos 298 registros filtrados. Ao lado a lista das 20 principais palavras, destaque para as cinco primeiras com frequência igual ou maior que 20: Luxury, Townhouse, Private, Loft e Park.
Townhouse	27	
Private	25	
Loft	25	
Park	20	
NYC	19	
Bedroom	17	
Village	17	
Manhattan	16	
East	16	
apartment	15	
West	15	
Central	14	
Brooklyn	13	
Airbnb	13	
Beautiful	13	
bedroom	13	
New	12	
Hidden	12	
Penthouse	12	

2. PRÉ-PROCESSAMENTO DOS DADOS

2.1.1. Transformação das variáveis categorias em booleanas

Para algoritmos de regressão é necessário que os atributos sejam de valores quantitativos, desde modo foi transformado as colunas room_type e bairro_group.

- O atributo “room_type” possui três categorias, neste sentido, foi alterado para atributo discreto resultando nas novas colunas: 'room_type_Entire_home_aprt', 'room_type_Private_room', 'room_type_Shared_room'.
- O atributo “bairro_group” possui cinco categorias, neste sentido, foi alterado para atributo discreto resultando nas novas colunas: 'bairro_Bronx', 'bairro_Brooklyn', 'bairro_Manhattan', 'bairro_Queens', 'bairro_Staten_Island'.

2.1.2. Tratamento dos dados faltantes e outlier

Os algoritmos de regressão precisam receber dados com distribuição parecida com a normal e sem dados ausentes, foram realizadas o preenchimento de valores zero ou vazio, além de seleção de valores de algumas colunas para excluir outliers.

- Preenchimento dos reviews_por_mes com valor aleatório entre 0.01 e 1;
- Exclusão dos valores com preço superior a 300 e 0;
- Exclusão do mínimo de noites maior que 15;
- Exclusão do numero_de_reviews maior que 100 e zero;

Os registros passaram de 48.894 para 29.955 com as exclusões.

Abaixo a descrição dos dados ajustados.

	minimo_noites	numero_de_reviews	disponibilidade_365	calculado_host_listings_count	price	reviews_por_mes
count	29955.000000	29955.000000	29955.000000	29955.000000	29955.000000	29955.000000
mean	2.712001	19.109664	96.599432	3.457052	116.861259	1.269117
std	1.811918	23.141492	123.889292	22.450594	64.978452	1.538027
min	1.000000	1.000000	0.000000	1.000000	10.000000	0.010000
25%	1.000000	3.000000	0.000000	1.000000	65.000000	0.180000
50%	2.000000	9.000000	26.000000	1.000000	100.000000	0.690000
75%	3.000000	26.000000	177.000000	2.000000	150.000000	1.845000
max	10.000000	100.000000	365.000000	327.000000	300.000000	27.950000

3. CORRELAÇÃO

Abaixo a correlação entre os atributos quantitativos

- price 1.000000
- room_type_Entire_home_apartment 0.630215
- bairro_Manhattan 0.304572
- calculado_host_listings_count 0.107163
- minimo_noites 0.075023
- latitude 0.051673
- disponibilidade_365 0.016464
- numero_de_reviews -0.016683
- reviews_por_mes -0.033504
- bairro_Staten_Island -0.043409
- bairro_Bronx -0.101719
- room_type_Shared_room -0.143101
- bairro_Brooklyn -0.152650
- bairro_Queens -0.159712
- longitude -0.295421
- room_type_Private_room -0.586301

Nos dados de correlação acima em relação ao “price”, pode-se destacar o tipo de hospedagem “Entire_home_apartment” que possui correlação de 63,02%, o grupo do bairro Manhattan com correlação de 30,45%, ambas com correlação positiva. A disponibilidade de dias possui correlação de 1,64% positiva e o mínimo de noites 7,5% de correlação positiva, ambas consideradas correlações baixas.

Na correlação negativa pode-se destacar o tipo de acomodação “Private_room” com -58,63% e o grupo de bairros “Queens” e “Brooklyn” ambos acima de -15% ambas.

4. REGRESSÃO

A solução para o problema de previsão de valores contínuos pode ser resolvida com modelos de regressão. A classificação é indicada quando se quer um resultado categórico ou discreto e que envolva um certo grupo a ser classificado. Para este problema se utilizou o modelo de regressão linear e floresta aleatória, abaixo as métricas de cada modelo:

4.1.1. MODELO 01 *LinearRegression*

- Intercept: -17008.05354214767
- Real amostra de 3: [60 175 150]
- Predição amostra de 3: [60.6008392 153.60845886 171.11710202]
- Erro absoluto médio (MAE): 34.05
- Erro quadrado médio (MSE): 2087.95
- R2: 0.499535

4.1.2. MODELO 02 *Random Forest Regressor*

- Real amostra de 3: [60 175 150]
- Predição amostra de 3: [73. 169.73333333 155.2]
- Erro absoluto médio (MAE): 31.16
- Erro quadrado médio (MSE): 1867.27
- R2: 0.552430

O melhor R2 foi do segundo modelo com o valor de 55,24% de explicação da variável alvo em relação aos atributos.

TESTE MODELO 02

```
{ 'id': 2595,  
  'nome': 'Skyliit Midtown Castle',  
  'host_id': 2845,  
  'host_name': 'Jennifer',  
  'bairro_group': 'Manhattan',  
  'bairro': 'Midtown',  
  'latitude': 40.75362,  
  'longitude': -73.98377,  
  'room_type': 'Entire home/apt',  
  'price': 225,  
  'minimo_noites': 1,  
  'numero_de_reviews': 45,  
  'ultima_review': '2019-05-21',  
  'reviews_por_mes': 0.38,  
  'calculado_host_listings_count': 2,  
  'disponibilidade_365': 355 }
```

- Erro Absoluto = 14.90, ficando abaixo do erro do modelo que é de 31.16.

- Valor real: 225
- Valor predito: 210.10