MODELO DE REGRERSSÃO LINEAR

Autora: JULIANA CRUZ

22/02/2024

Florianópolis, SC

Sumário

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*
- 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

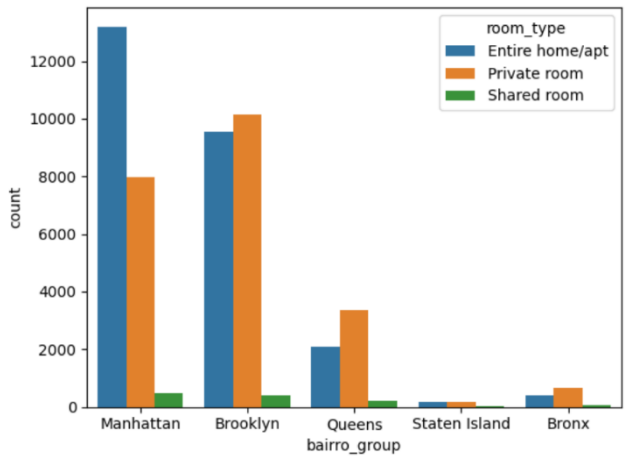
- Entire home/apt 0.519675
- Private room 0.456600
- Shared room 0.023725

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

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

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



Valores percentuais

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

Figura 1: Elaborada pela autora

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.

	price			
	min	max	sum	mean
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. host_name distribuição dos registros

São 47904 registros com nomes anúncios (nome) diferentes no atributo nome, já no atributo usuário que hospedou (host_name) são 11452 registros diferentes, neste caso a média de anúncios por host_name é de aproximadamente 4.

Os cinco principais 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
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
```

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 pata as cinco primeiras com frequência igual ou maior que 20: Luxury, Townhouse, Private, Loft e Park.

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.

a) O atributo “room_type” possui três categorias, neste sentido, foi alterado para atributo discreto resultando nas novas colunas: 'room_type_Entire_home_apt', 'room_type_Privat_room', 'room_type_Shared_room'.
b) 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 algoritimos 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.

a) Preenchimento dos reviews_por_mes com valor aleatório entre 0.01 e 1;
b) Substituição dos zeros de disponibilidade_365 por 365 que é valor da moda;
c) Exclusão dos valores com preço superior a 500 e 0;
d) Exclusão do mínimo de noites maior que 15;
e) Exclusão do numero_de_reviews maior que 100 e zero;

Os registros passaram de 48894 para 31318 com as exclusões.

3. CORRELAÇÃO

A baixo a correlação entre os atributos quantitativos

- price 1.000000
- room_type_Entire_home_apt 0.572529
- bairro_Manhattan 0.297992
- calculado_host_listings_count 0.118009
- minimo_noites 0.071660
- latitude 0.053642
- reviews_por_mes -0.013372
- numero_de_reviews -0.029052
- disponibilidade_365 -0.036054
- bairro_Staten_Island -0.043885
- bairro_Bronx -0.090260
- room_type_Shared_room -0.123614
- bairro_Queens -0.154543
- bairro_Brooklyn -0.156894
- longitude -0.282587
- room_type_Privat_room -0.536201

Nos dados de correlação acima em relação ao “price”, pode-se destacar o tipo de hospedagem “Entire_home_apt” que possui correlação de 57,25%, o grupo do bairro Manhattan com correlação de 29,79%, ambas positiva. Disponibilidade de dias possui correlação de 3,6% positiva e mínimo de noites 7,16% de correlação positiva, ambas consideradas correlações baixas. Na correlação negativa pode-se destacar o tipo de acomodação “Privat_room” com -53,62% e o grupo de bairros “Queens” e ”Brooklyn” 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:

LinearRegression MODELO 01

- Intercept: -16193.255977498979
- Predição amostra de 3: [38.58261609 125.19828191 159.96021222]
- Erro absoluto médio (MAE): 44.39
- Erro quadrado médio (MSE): 4277.74
- R2: 0.417041

Random Forest Regressor MODELO 02

- Predição amostra de 3: [32.9 156.66666667 112.46666667]
- Erro absoluto médio (MAE): 40.63
- Erro quadrado médio (MSE): 3694.49
- R2: 0.496525

O melhor R2 foi do segundo modelo com o valor de 49,65% de explicação da variável alvo em relação aos atributos.

TESTE MODELO 02


```
{'id': 2595,
 'nome': 'Skylit 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}
```


- Valor real: 225
- Valor predito: 257.4
- Erro Absoluto = 36.25, ficando abaixo do erro do modelo que é de 40.63.