**Relatório de Análise estatística e EDA**

Autor: Rafael Negrão de Souza

Data: 06/02/2022

Base de dados: Teste indicium predições

Ferramentas utilizadas: (numpy, pandas, sklearn, redes neurais, matplotlib, seaborn)

**1. Introdução**

O desafio é desenvolver um modelo de predição a partir de um dataset e avaliar o modelo, além de fazer uma análise exploratória nos dados fornecidos para um melhor compreensão.

**2. Descrição dos dado**

Foram observados dezesseis dados presentes no conjunto, id, nome, host\_id, host\_name, bairro\_group, bairro, latitude, longitude, type\_room, price, minimo\_noites, numero\_de\_reviews, ultima\_review, review\_por\_mes, calculado\_host\_listings\_count e disponibilidade\_365.

O id representa o identificador único do registro, nome representa o nome do lugar, host\_id o identificador do proprietário, host\_name o nome do proprietário, bairro\_group a região que agrupa bairros, bairros os bairros da cidade de Nova York, latitude e longitude são as coordenadas geográficas no imóvel, type\_room o tipo de imóvel, price o preço, valor, minimo\_noites é o mínimo de noites para aluguel, numero\_de\_reviews a quantidade de reviews que o imóvel teve, ultima\_review o registro da última review, review\_por\_mes média de reviews por mês, calculado\_host\_listings\_count contagem de imóveis por um anfitrião e disponibilidade\_365 a disponibilidade do imóvel ao ano.

**2. Analise exploratória**

Foi observado valores bem discrepantes, outliers, quando se observa os preços pelos bairros, então, observei por cada região e mesmo assim ainda continuava a ter diversos valores fora do padrão, devido a imóveis de luxo.

Manhattan seguia o topo das regiões onde possuíam mais imóveis para aluguel e logo em seguida o Brooklyn. A hipótese é devido a serem os mais populosos.

Staten Island possui o primeiro e segundo dos três imóveis com a media de preços mais cara de NovaYork, Fort Wadsworth e Woodrow, com Manhattan ocupando o terceiro lugar com Tribeca. A hipótese de Staten Island estar entre os mais caros e o que menos aluga imóvel, pode ser porque é menos populoso que os demais, que fornecem imóveis mais compactos, indo fazendo diferente e fornecendo imóveis mais espaçosos.

E entre os três coma media de preços mais baratos o primeiro também é de Staten Island, Bull’s Head e o segundo e terceiro estão no Bronx, Hunt’s Point e Tremont.

Quando dividimos as regiões, ou os bairros, com a media dos dez mais caros e dez mais baratos, percebemos, no histograma, que a diferença de preços da media dos bairros entre os mais caros é bem maior que a diferença da media dos preços entre os bairros mais baratos.

Preços de alugueis de imóveis do tipo *“Entire home/apt”* pertencente ao campo *“type\_room”* são os mais caros que os demais tipos em todos as regiões e o mais barato em todas é *“shared room”.* A hipótese que o quarto compartilhado tem a média mais barata se da devido a privacidade e o espaço que é dividido com outras pessoas que não são conhecidas ou parentes, consequentemente diminuindo o conforto.

A região do Bronx e Staten Island possuem mais imóveis com disponibilidade próxima de um ano.

**2.1 Tipo de dados**

|  |  |
| --- | --- |
| **Coluna** | **Tipo** |
| id | int64 |
| nome | object |
| host\_id | int64 |
| host\_name | object |
| bairro\_group | object |
| bairro | object |
| latitude | float64 |
| longitude | float64 |
| type\_room | object |
| price | int64 |
| minimo\_noites | int64 |
| numero\_de\_reviews | int64 |
| ultima\_reviews | object |
| review\_por\_mes | float64 |
| calculado\_host\_listings\_count | int64 |
| disponibilidade\_365 | int64 |

**3 Processamento de dados**

Foram observados 7 dados do tipo int64, 6 dados do tipo object e 3 dados do tipo float64 e os nomes das colunas estavam misturando inglês com português, então foi padronizada na língua inglesa.

Dentre esses dados, destacasse *“ultima\_review”* e *“review\_por\_mes”* com 10052 valores nulos, isso ocorre porque no campo *“numero\_de\_review”* há imóveis que não possuem nenhuma review e com isso, *“numero\_de\_reviews”* e *“review\_por\_mes”* apresentavam esses valores faltantes, assim, os campos de *“review\_por\_mes”* nulos receberam o valor zero e *“ultima\_review”* recebeu um texto No Review.

Os campos *“nome”* e “*host\_name”* também apresentavam valores nulos, 16 e 21 nulos respectivamente, porém é uma coluna pouco relevante para a criação do modelo, então foram retirados, assim como o campo *“id”* e *“ultima\_review”*.

Foi também observado que o campo *“price“* alguns registros possuíam o valor zero, não se sabe o por que, se devido alguma propaganda, promoção, divulgação, porém, como somente eram onze registros, eles foram retirados da base para o treinamento do modelo para não atrapalhar no treinamento do modelo.

Além disso, três dados categóricos, *“bairro\_group”*, *“bairro”* e *“type\_room”* que precisaram ser transformados e representados numericamente.

**3.1 Observações**

Foi observado que os campos que utilizam o tipo int64, não apresentam grande valores para necessitar de tal estrutura. A mudança em código não foi alterada devido a ser um dataset com poucos registros e tempo, mas caso necessário diminuir o consumo de memória e processamento mais eficiente, poderia ser mudado pata int8 ou int16.

**4. Perguntas**

a)

b) A correlação entre o número mínimo de noites e o preço é 0.04, indicando que não há uma correlação linear forte e entre disponibilidade e preço é 0.08 indicando também que não há uma correlação linear forte.

c) Sim, imóveis com os preços mais altos tem tendem a ter palavras como, luxury, suite.

**5. Modelo de Aprendizado de Máquina - Explicação**

O tipo de problema se trata de regressão, onde precisamos prever preços de imóveis a partir de uma entrada, com base em dados já fornecidos para aproximar um valor. Para as variáveis categóricas, realizei a transformação de encode, para serem representadas numericamente. Utilizei as variáveis latitude, longitude, minimo\_noites, numero\_de\_reviews, reviews\_por\_mes, calculado\_host\_listings\_count, disponibilidade\_365 e os encode de bairro\_group, room\_type, bairro e o price como classe. No treinamento também fiz uma escala dessa varáveis para que nenhum valor maior que o outro tivesse um peso maior no treinamento.

Realizei treinamento com três modelos diferentes, redes neurais, knn e gxboost que é baseado em árvores de decisões, as métricas de escolha foram o mean square error (MSE) - que calcula a media ao quadrado do valor previsto pelo real e mean absolute error (MAE) – que calcula a diferença absoluta entre o valor previsto é o real. Dentre essas métricas foi escolhido o modelo de rede neural pois obteve um melhor desempenho no treinamento diante das métricas.

**6. Entrada de características**

Para a entrada do apartamento com as entradas sugeridas no documento de apoio, o modelo sugeriu um preço de 105.94.