Fabio Dias Rezende Carvalho

email:fabior.carvalho@hotmail.com

Github: https://github.com/FabioDiasRC

Projeto LH_CD – EDA e modelo de machine learning

Para melhor informações considere a consulta do notebook completo dentro do link do meu github, github.com/FabioDiasRC. Nele está descrito passo a passo do projeto além do desenvolvimento e código de tudo que foi definido.

Objetivo: desenvolver um modelo de previsão de preços a partir do dataset oferecido, e avaliar tal modelo utilizando as métricas de avaliação que mais fazem sentido para o problema.

Análise exploratória dos dados:

Aqui podemos observar alguns pontos relevantes dentro do nosso conjunto de dados:

- **Price(preço)** - Aqui com o 'describe' do pandas podemos observar que existe um valor de zero como minimo, o que ja afeta nosso conjunto de dados, e um valor máximo muito acima em relação ao 75% superior, o que também eleva o valor resultante da média, também influenciando no desvio padrão. Farei uma análise visual dos boxplots em sequência para confirmar os outliers e tomar decisoes mais precisas;
- **disponibilidade_365** - Podemos observar que existem casas dentro do nosso dataset que não possui disponibilidade, isso é relevante de se observar para a pergunta sobre considerar um investimento em uma casa em NY para alugar, uma vez que considerando como um investimento o lucro máximo dessa casa vai se dar quando ela estiver ocupada em todos os dias do ano.
- **minimo_noites** - O dataset traz que a maior parte do nosso conjunto de dados possui alugueis de noites unicas, ou poucos dias em sequência, onde no 75% tem-se apenas 5 noite, e o máximo vai até 1250, ou seja, temos uma forte influencia do valor máximo e possiveis outiliers nessa coluna do nosso conjunto de dados.
- As colunas 'ultima_review' e 'reviews_por_mes' possuem mais de 20% de valores nulos dentro dessas colunas, para não ter perda das informações das linhas referentes decidi por remover ambas as colunas

As perguntas que norteiam o projeto de análise

- **Investimento Residencial:**

Considerando que uma pessoa busca investir em uma casa ela precisa focar em regiões que possuem menos dias disponiveis para aluguel, coluna denominada como 'disponibilidade_365', por que se tem menos dias disponiveis significa que ela esta ficando mais tempo alugada, o que representa maior retorno do investimento. a região do Brooklyn consegue ficar mais tempo ocupado que Manhattan, então a indicação seria de investimento em um imóvel no Brooklyn, mas um investimento em Manhattan

também é uma boa opção baseada nas regiões e também considerando que a região de Manhattan possui imóveis com aluguel mais elevado

- **Relação de preço com o número mínimo de noite e disponibilidade ao longo do ano:**

A relação com o preço tem-se que a coluna 'minimo_noites' tem uma influencia de 0.042805 no preço e a cada dia disponivel dentro do ano tem-se o fator de 0.081851, logo a disponibilidade durante o ano tem quase o dobro de influencia dentro do conjunto de dados em relação ao minimo de noites dentro do preço.

- **Padrão no nome do texto:**

Dentro dos imóveis de preço mais elevado encontramos que aproximadamente 25% deles levam o nome de 'apartament', 'apt' ou 'loft', ou seja, 1/4 levam esse tipo de nome dentro do nosso conjunto de dados

Previsão do preço

O conjunto de dados trazia uma grande quantidade de outliers em relaçãp ao preço dos imóvesi, entao era necessário buscar uma distribuição normal, que permite maiores possibilidades para nosso conjunto de dados (teste estatiscos e a definição do modelo de machine learning possuem melhor precisão em dados com distribuição normal). Para nosso conjunto de dados apliquei a transformação logaritmica. Nosso objetivo aqui é desenvolver um modelo de regressão linear, uma vez que queremos encontrar um valor numerico especifico baseado no nosso conjunto de dados de imóveis.

Após a transformação logaritimica eu defini as variavéis que iriam para o modelo de machine learning, apliquei a transformação ‘onehotenconding’ nas colunas ‘bairro_group’ e ‘room_type’ para melhor trabalhar essas colunas, no notebook eu cito os prós e contras dessas decisões. Para esse modelo eu defini como prós:

- Modelo mais simples: definido com uma equação, onde expressa a relação entre as variaveis escolhidas, além de levar em consideração todas as variáveis que colocamos dentro do nosso dataset e todas elas terem valores relativamente consideraveis e devem ser utilizadas dentro da nossa previsão;
- Eficiente: ele é eficiente e calculado mais rapidos que outros modelo;
- Consideravemente bem difundido e estudado: um tipo de modelo de ml bem difundido e bem utilizado em ciencia de dados;

Os pontos negativos foram:

- Outliers: apesar do tratamento dos outliers eles ainda podem prejudicar a estimativa, gerando forte impacto na reta de regressão;
- Dependencia de variaveis: ele pode assumir grande dependencia entre as variaveis;
- Variação: esse tipo de modelo define uma variação constante em torno da média;
- Quantidade de variaveis: temos uma quantidade consideravel de variaveis, onde pode afetar o modelo se fosse um conjunto de dados maior.

O modelo de machine learning utilizando o statsmodel encontrou aproximadamente 50% de acuracia no teste R^2, definindo o valor do imóvel de exemplo como $235.41 por noite. Na sequência eu salvei o modelo no formato '.pkl' como foi pedido no projeto