Relatório - EDA

Primeiramente, gostaria de ressaltar que o processo de verificação desse relatório não foi linear, ou seja, tive que voltar algumas vezes para verificar outras variáveis para a resolução do problema proposto.

Após algumas análises, percebi que na coluna 'reviews_por_mes', alguns dados estavam como faltantes e na verdade queriam indicar que na verdade significava que não existia uma quantidade de review por mês pelo fato de que a residência não recebeu alguma avaliação. Para que isso não prejudique a nossa listagem, resolvi considerar esse dado como 0.

Ao analisar o dataframe, fiz uma primeira listagem, apenas com as colunas ['nome', 'room_type' e 'price'] estava dando o total de 180 contagens de elementos duplicados dentro do nosso banco de dados, mas com a latitude e longitude diferentes, o que pode indicar que se tratam de localizações diferentes, entretanto também pode significar um ruído significativo dentro de nossa análise, pois alguém pode ainda assim cadastrar o mesmo tipo de aluguel para a mesma localização e errar durante a medição de latitude e longitude, nesse caso, para resolver esse impasse se era duplicado ou não, levei em consideração que todas essas informações, podem ser colocadas pelo host como um padrão de aluguel, por exemplo, digamos que o seu João é dono de vários apartamentos dentro de um mesmo prédio ou setor residencial, ele pode querer padronizar os preços e nomes para os mesmos tipos de quarto, afinal de contas isso não mudaria em nada o aluguel de seus pontos de residência, mas ele também pode criar vários anúncios dentro da aplicação para um mesmo tipo de apartamento, com o intuito de elevar seus lucros. Para verificar se realmente existe alguns inputs duplicados, tendo a finalidade de evitar um viés indesejado, repeti a operação adicionando apenas a

coluna ['disponibilidade_365'] que são itens que não podem ser controlados pelo anunciante, e portanto, determinam uma possível duplicidade em nosso banco de dados, e para evitar erros dentro da nossa análise, é melhor retirar eles. Além de que também retirei os valores que ‘price’ eram iguais a 0 por não haver a possibilidade de existir aluguéis com esse valor.

Agora, para reduzir a quantidade de dados que estamos trabalhando para algo que vá impactar diretamente em nossa análise preditiva, reduzi a quantidade de colunas do nosso dataframe, levando em conta alguns motivos principais. Por exemplo, alguns itens são completamente descartáveis para a nossa análise, como por exemplo, o ID do anúncio, isso porque, seguindo a lógica de negócio, pois o usuário não vai interagir diretamente com esse tipo de dado para entender se deseja alugar o local ou não, também, não desejamos saber a quantidade de listagens por host nessa análise. Os dados escolhidos foram a localização tanto nominal qunato numeral, o mínimo de noites que se deve ficar, pois o anunciante pode optar por valores diferentes, se o a quantidade de noites foir maior ou menor, a quantidade de reviews totais e mensais, pois seguindo a lógica atual do negócio, as pessoas buscam antes saber sobre o local para poder alugar, e se as reviews forem boas, com certeza, o host irá se sentir à vontade para colocar um preço maior, e também a sua disponibilidade total, seguindo a lógica da procura e oferta, se não tem muitos dias disponíveis para locação, talvez o local seja muito procurado, e seguindo isso, o produto tende a ficar mais caro. Além disso tudo, também deixei a variável 'price' que é o que desejamos prever com esse modelo.

Também foi verificado que existia a presença de outliers em nosso modelo para a variável ‘price’ e variável ‘mínimo_noites’, para lidar com os outliers da variável price, primeiro eu calculei o seu IQR, e depois verifiquei se tinha algum dado maior, aliado a isso, verifiquei algo que eu estava pensando, alguns dos aluguéis estavam com o preço não convertido para noite, e isso estava causando um viés em nossa amostragem, devido a isso, fiz uma lógica para a retirada desses outliers dividindo o valor do aluguel pelo mínimo de noites, e com os que restaram e não puderam ser consertados, eu apenas os retirei da nossa tabela, para que eles não façam o preço dos imóveis aumentar. Aliado a isso, também tratei a variável ‘mínimo_noites’

seguindo uma outra lógica, pesquisei sobre o negócio de aluguéis na USA e pude perceber que é muito diifícil que o tempo mínimo passe de um ano, sendo que o comum seja justamente entre 6 meses e 1 ano, então, eu retirei todos os dados que extrapolaram os 365 dias dentro de nossa tabela. Depois disso, eu pude constatar que o tipo de quarto era um dado relevante para a construção do modelo, e assim o transformei em tabelas dummies para poder adicionar a nossa verificação, tentei fazer o mesmo com os bairros, pois também se mostram relevantes, mas eles não tiveram uma influência grande em nosso modelo preditivo, então optei por não utilizar os bairros para a medição.

De acordo com o gráfico de distribuição de locais de aluguel por bairro, Manhattan, seguida por Brooklyn liderou a maior quantidade de locais disponíveis para alugar, e portanto, terão mais peso para esse modelo. Além disso, pude constatar que o número de reviews e a quantidade de reviews por mês tem uma quantidade mínima de influência no preço do imóvel, seguidas por noites mínimas para serem alugadas, e de certa forma, quem é o host também influencia na decisão do preço, isso faz sentido seguindo a lógica de negócio que certas empresas trabalham alugando casas com caracteristicas semelhantes e portanto os preços também serão, além de que a disponibilidade também afeta no preço final.

Após realizar algumas plotagens de gráfico em nosso notebook, pude perceber que algumas variáveis que eu considerava importantes para a análise, como ‘numero_de_reviews’ e ‘reviews_por_mes’ não eram importantes para o modelo, pois no gráfico de covariância, elas tinham um impacto não só nulo, como negativo.

Perguntas propostas

Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

Segundo o gráfico gerado de preços para localização dos imóveis (bairros), o local ideal para a compra seria em Manhattan, pois é o local em que a média dos aluguéis é a mais cara

O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Sim, pelo gráfico de calor que foi gerado dentro de nossa EDA, pôde-se perceber que eles são um dos mais influentes dentro da precificação de certo imóvel.

Existe algum padrão no texto do nome do local para lugares de mais alto valor?

Levando em consideração a listagem dos nomes que possuem o valor mais alto, e colocando o valor mais alto como 3.000,00 pude perceber que muitos deles se localizam em Manhattan e utilizam palavras chave para promover o imóvel, como “SuperBowl”, “Luxury” e “Studio”.

Explique como você faria a previsão do preço a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

Para realizar a previsão do preço, levei em conta as variáveis que faziam a alteração do preço em nossas verificações realizadas dentro do EDA, sendo a localização, disponibilidade do imóvel, mínimo de noites que uma pessoa deve ficar, tipo de imóvel, e o id do host, nesse último pode ser menos visível a relação com o preço, entretanto, geralmente os anunciantes possuem um padrão de precificação em seus produtos e por aqui, não deveria ser diferente, e também utilizei o modelo de regressão linear, pois queríamos saber o comportamento de uma variável em relação a várias outras dentro de nosso banco de dados, os seus contras é que ela é

bastante afetada por outliers, além disso também não é recomendada para uso no mundo real, por sua simplicidade, e também não consegue definir importância de recursos utilizados no cálculo, sempre levando em conta que todos tem a mesma importância, isso aliado a sua suposição de que todos os recursos tem a mesma variância linear. Os seus prós para esse tipo de questão são a sua eficiência e velocidade na computação, justamente por ser um modelo simples, e também ela lida muito bem com esses tipos de problema em que uma variável pode ser influenciada por outras. A medida de perfomance escolhida foi a MAE (mean absolute error ou erro absoluto médio), ele foi utilizado porque ele realiza o tratamento de outliers dentro da própria média de correção, se adequando para que a média não seja afetada e consigamos obter um resultado mais preciso de correção, além disso a sua taxa de erro ficou em 38,43 U$D.

Supondo um apartamento com as seguintes características:

```
{'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',
 '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}
```

Qual seria a sua sugestão de preço?

O valor de preço previsto para um imóvel nessas condições foi de 180.48 U$D/noite

O vídeo requisitado pode ser acessado por aqui.