Projeto Desafio Cientista de Dados – Indicium

1. Introdução

Neste projeto, desenvolvi um modelo preditivo de preços para uma plataforma de aluguéis temporários em Nova York. Meu objetivo foi responder às seguintes questões de negócio:

- Onde é mais vantajoso investir na compra de um apartamento para alugar?
- O número mínimo de noites e a disponibilidade ao longo do ano influenciam o preço?
- Existe algum padrão no texto dos anúncios que indique um valor mais alto?
- Como posso prever o preço e qual modelo se aproxima melhor dos dados?

2. Metodologia

2.1. Coleta e Limpeza dos Dados

- Carreguei os dados a partir do arquivo listings_ny.csv.
- Verifiquei a presença de valores ausentes e duplicatas, utilizando a biblioteca Missingno para uma visualização rápida.
- Realizei o tratamento dos dados: removi duplicatas, preenchi os valores nulos em "nome" e "host_name" com "Sem Nome" e descartei a coluna "ultima_review" devido à alta incidência de valores ausentes.

2.2. Análise Exploratória (EDA)

- Analisei a distribuição dos preços por meio de histogramas e boxplots para identificar outliers.
- Calculei a matriz de correlação entre variáveis numéricas usando o coeficiente de Spearman.
- Criei a feature "tem_luxo" a partir dos nomes dos anúncios (buscando termos como "luxury", "lux", "elegant", "premium", "penthouse", "view", "modern" e "spacious") e comparei os preços médios entre anúncios com e sem essa característica.
  Resposta à pergunta 3: O teste t mostrou uma diferença estatisticamente significativa entre os preços, o que indica que anúncios com termos indicativos de luxo tendem a ter preços mais altos.
- Desenvolvi gráficos de barras para visualizar a distribuição dos anúncios por "bairro_group" e os 15 bairros com maior preço médio.
  Resposta à pergunta 1: Esses gráficos sugerem que, embora bairros com preços elevados (como em Manhattan) indiquem alta demanda, eles também podem ter concorrência intensa. Assim, investir em bairros com preços moderados pode ser uma estratégia mais vantajosa, dependendo do equilíbrio entre custo e retorno.

2.3. Modelagem Preditiva

- Desenvolvi pipelines para quatro modelos de regressão: Regressão Linear, Random Forest, Gradient Boosting e uma rede neural (MLPRegressor).
  Resposta à pergunta 4: Considerei o problema como uma regressão, aplicando transformações (OneHotEncoding para variáveis categóricas e StandardScaler para

variáveis numéricas) e realizando seleção de features. Incluí a rede neural para testar a captura de relações complexas, embora ela exija cuidados para evitar overfitting. Cada modelo tem suas vantagens: a Regressão Linear é mais simples e interpretável, enquanto os modelos de ensemble (RF e GB) e a rede neural demonstram robustez para relações não-lineares.

- Avaliei os modelos com uma divisão treino/teste, utilizando métricas como RMSE, MAE e R^2.
- Realizei validação cruzada para comparar a estabilidade e a performance dos modelos.
- Executei tuning de hiperparâmetros com RandomizedSearchCV para otimizar Random Forest, Gradient Boosting e a rede neural, comparando os resultados antes e depois da otimização.

2.4. Interpretação dos Resultados

- Usei SHAP para interpretar o modelo Random Forest otimizado, o que me permitiu entender a contribuição de cada feature na predição dos preços.

2.5. Previsão de Caso Específico

- Apliquei o modelo otimizado para prever o preço de um apartamento com características específicas (detalhadas no desafio). A previsão obtida foi apresentada no relatório, e o modelo final foi salvo em formato .pkl para uso futuro.

3. Resultados e Discussão

- **Análise Exploratória (EDA):**
  Meus gráficos e testes estatísticos indicaram que anúncios com termos de luxo têm, em média, preços mais altos. Além disso, os gráficos de preço médio por "bairro_group" e por "bairro" sugerem que bairros com preços elevados podem ser indicativos de alta demanda, mas também de alta concorrência.
- **Modelagem:**
  Ao comparar os modelos, observei que cada abordagem possui pontos fortes e limitações. A Regressão Linear apresentou resultados surpreendentemente bons em termos de RMSE, mas modelos como Random Forest e Gradient Boosting também se destacaram por sua robustez. A rede neural mostrou potencial para capturar relações complexas, embora o tuning de hiperparâmetros não tenha reduzido significativamente o RMSE, o que pode indicar que, com o tamanho atual do dataset, aumentar a complexidade não traz ganhos expressivos.
- **Previsão de Preço:**
  Com base no modelo que obtive melhor desempenho (no meu caso, a Regressão Linear apresentou os melhores resultados), a sugestão de preço para o apartamento de id 2595 foi de aproximadamente USD 316,62.
- **Interpretação dos Resultados:**
  O gráfico SHAP me permitiu identificar quais variáveis têm maior influência na predição dos preços. Essa análise é fundamental para validar se as features selecionadas estão de fato contribuindo para o modelo, e para guiar decisões de negócio.

4. Conclusões

- Realizei uma análise completa do dataset, identificando padrões importantes e respondendo às questões de negócio propostas.
- A interpretação dos resultados com SHAP me ajudou a compreender a importância relativa das variáveis na predição dos preços.
- Recomendo que, para decisões de investimento, sejam considerados não apenas os preços médios, mas também fatores como a taxa de ocupação e características regionais.
- Sugiro a integração de outros datasets, como informações sobre criminalidade, disponibilidade de serviços e áreas de lazer, para refinar ainda mais a estratégia de precificação.

5. Dependências

As dependências estão listadas no arquivo requirements.txt.

6. Considerações Finais

Este projeto demonstra meu domínio em análise exploratória, modelagem preditiva, tuning de modelos e interpretação dos resultados. Através de uma abordagem integrada, pude responder às questões de negócio propostas e sugerir direções para melhorar a estratégia de precificação. Meu trabalho baseou-se na experiência adquirida como estagiário em Ciência de Dados, e acredito que, com a integração de dados externos e análises geoespaciais, o modelo poderá se tornar ainda mais robusto.