Case Data Science - Valuation Systems

Introdução

Na Loft, o squad de **Valuation Systems** é responsável por construir modelos de precificação de imóveis, ou AVMs (Automated Valuation Models). Para nós, um bom AVM é uma peça central no nosso negócio, pois permite que sejamos a única referência de preço do mercado, aproximando compradores, vendedores e corretores. Sem um AVM, não conseguimos expandir nossa área de atuação de forma escalável, controlar nosso risco financeiro e atender nossos clientes, fornecendo preços justos para seus imóveis.

Neste case você encontrará uma amostra dos desafios que enfrentamos na construção de AVMs no nosso dia a dia. Esperamos que seja desafiador, divertido, e útil como amostra do trabalho que você vai desempenhar caso se torne um Lofter.

Dados

Em anexo enviamos o arquivo `valuation_data.csv`, que contém informações obtidas em um site de anúncios de apartamentos (https://123i.uol.com.br), na cidade de São Paulo. A base contém algumas variáveis descrevendo prédios e apartamentos, assim como estimativas de valor.

Desafio e roteiro

Seu desafio será construir um AVM utilizando os dados fornecidos e outras bases públicas que desejar. Abaixo construímos um roteiro para ajudar no desenvolvimento do AVM. **Não é obrigatório responder todos os itens: a decisão sobre priorizar um item ou outro é sua, baseado no que aprendeu sobre a Loft até agora.** Segue o roteiro:

1. **Limpeza da base.** Explore a base e faça uma limpeza nos dados, se necessário. Que tratamentos você aplicou? Por quê?

2. **Enriquecimento da base e feature engineering.** Quais informações adicionais poderiam ser utilizadas além das fornecidas? Enriqueça a base com as informações que achar mais relevantes. Quais informações você adicionou?

3. **Construção do modelo.** Construa um modelo e monte um esquema de validação para testar o poder de generalização do seu modelo. Qual algoritmo foi utilizado? Como você validou o modelo? Qual é a métrica que você usou para avaliar o modelo? Qual foi o resultado?

4. **Interpretabilidade do modelo.** Suponha que os proprietários dos apartamentos com `id`'s 7818, 9315 e 18338 na base pediram uma avaliação a partir do modelo. Qual é o preço dado pelo seu modelo? Como você justificaria esses preços para os proprietários?

5. **Quantificação de risco.** Suponha que os proprietários dos apartamentos com `id`'s 7818, 9315 e 18338 na base estão pedindo, respectivamente, R$ 400.000, R$ 3.000.000 e R$

2.000.000 pelos seus apartamentos. Esses preços são razoáveis? Quão longe esses preços estão da realidade do mercado?

6. **Extrapolação do modelo.** Suponha que você só tenha dados de treino dentro da região geográfica dada pelo bounding box limitado em latitude pelo intervalo `(-23.5884, -23.5495)` e em longitude pelo intervalo `(-46.6817, -46.6379)`. Contudo, o seu modelo deve funcionar para toda a região de SP. Como você validaria o modelo? As variáveis que você coletou funcionariam nesse caso? Em quais regiões/apartamentos que estão fora do bounding box você espera que o modelo com dados limitados funcione bem?

7. **Eficiência de amostra.** Qual é o tamanho mínimo da base de treino para ter uma performance razoável (parecida com a performance da base completa)? Se você pudesse escolher somente 1000 linhas da base para treinar o modelo, quais linhas você escolheria?

Entregável

Você entregará seu case fazendo uma apresentação para a equipe de Data Science da Loft em uma data combinada previamente. Sua apresentação poderá ser feita usando a mídia que preferir (Jupyter Notebook, Slides, etc). Na apresentação, deve estar clara a lógica que você utilizou para resolver o case, mostrando as suposições feitas e as repercussões de suas escolhas para o negócio. Pedimos que envie o material com 24 horas de antecedência da apresentação para nos prepararmos.

Obrigado, e bom case!