Case 2 – Regressão de Preços de Hospedagem

Descrição: Este é o nosso segundo projeto prático do AI Bootcamp. Desta vez, vamos explorar conceitos de Machine Learning, com foco em um problema de regressão. O objetivo é que vocês possam aprofundar o entendimento sobre temas como: viés-variância, regularização, engenharia de features, tratamento de variáveis categóricas, avaliação de modelos de regressão, divisão entre dados de treino e teste, entre outros.

Problema: Seu time faz parte de uma start-up de grande sucesso no setor de aluguel de casas para curtas durações. A plataforma vem experimentando um crescimento significativo, com o site e o aplicativo atraindo cada vez mais usuários.

Atualmente, o mercado principal da empresa é a Europa, onde a oferta de acomodações continua aumentando. Com esse crescimento acelerado, a demanda por decisões baseadas em dados tem se intensificado, tanto para análises preditivas quanto prescritivas.

Neste momento de expansão, foi criada uma equipe, a de **Data Product Management**. O primeiro objetivo dessa equipe é entregar um produto de dados: um modelo de regressão capaz de prever o valor esperado do aluguel com base em informações sobre a acomodação (como número máximo de pessoas, quantidade de quartos, distância ao centro da cidade, latitude, longitude, dia da semana, entre outras features).

A base de dados disponível conta com informações de cerca de 10 cidades: **Amsterdã, Atenas, Barcelona, Berlim, Budapeste, Lisboa, Londres, Paris, Roma e Viena**. Sua equipe **deve escolher três dessas cidades** e realizar a análise para elas. Os arquivos estão organizados por cidade e por dias da semana (dia útil e fim de semana).

Objetivo: O principal objetivo é explorar os dados, identificar relações entre as variáveis e construir **um único** modelo de regressão capaz de prever os preços das acomodações. A avaliação do modelo será feita utilizando a métrica **MAE (Mean Absolute Error).**

Além disso, é necessário apresentar visualmente as características mais relevantes para o modelo de regressão, destacando quais informações impactam mais o preço das hospedagens.

Dicas:

- **Tratamento de features categóricas**: Durante sua análise das diferentes cidades, considere como incorporá-las como features. Explore técnicas de tratamento de variáveis categóricas, como o One-Hot Encoding.
- **Divisão entre treino e teste**: Separe seus dados em conjuntos de treino e teste. Avalie a performance do seu modelo apenas no conjunto de dados de teste, ou seja, nos dados em que o modelo não foi treinado.
- **Criação de novas features**: Pense em maneiras de criar features combinando duas ou mais variáveis do seu dataset existente. Isso pode ajudar a melhorar a performance do modelo.
- **Conceitos de Underfit e Overfit**: Explore e estude os conceitos de Underfitting e Overfitting. Procure entender como esses conceitos se correlacionam com o trade-off entre viés e variância.
- **Técnicas de regularização**: Explore técnicas de regularização para evitar o overfitting, como a regularização L1 (Lasso) e L2 (Ridge).
- **Normalização de features**: Dependendo do modelo utilizado, não se esqueça de normalizar as features, especialmente para modelos sensíveis à escala.
- **Avaliação de modelos**: Avalie diferentes modelos de regressão, como Regressão Linear, Lasso, Ridge, ElasticNet, Random Forests, XGBoost, LightGBM, CatBoost, SVRs, entre outros! Não é necessário testar todos, mas explore ao menos 2 a 3 modelos e compare os resultados das suas predições.