Análise de Regressão Utilizando o Dataset Boston House Price

Rafael Souza Osadzuk
Acadêmico do Curso de Engenharia de Software do Centro Universitário Cesumar - UNICESUMAR, Curitiba - PR.

RESUMO

Este artigo apresenta uma análise preditiva utilizando técnicas de regressão aplicadas ao conjunto de dados *Boston House Price*, com o objetivo específico de prever a variável **TAX**, que representa a taxa de imposto sobre propriedades nas diversas regiões da cidade de Boston. A motivação central da pesquisa é entender de que maneira fatores urbanos, estruturais e socioeconômicos influenciam a carga tributária local, contribuindo com subsídios técnicos para o planejamento urbano e a formulação de políticas fiscais mais justas e eficientes.

Para isso, foi adotada uma abordagem baseada em ciência de dados e aprendizado supervisionado, com ênfase na utilização da biblioteca **PyCaret**, uma ferramenta que automatiza a experimentação com diversos algoritmos de aprendizado de máquina e facilita a comparação de desempenho entre modelos. O estudo seguiu um processo metodológico estruturado em etapas que incluíram a obtenção e preparação dos dados, definição da variável-alvo, configuração do ambiente computacional, execução da modelagem preditiva com múltiplos algoritmos e avaliação dos modelos com base em métricas estatísticas, como **MAE** (Erro Absoluto Médio), **RMSE** (Raiz do Erro Quadrático Médio) e **R^2** (Coeficiente de Determinação).
Entre os modelos avaliados, o **Extra Trees Regressor** destacou-se por apresentar o melhor desempenho, com um **R^2 de 0.9822**, **MAE de 10.05** e **RMSE de 21.19**, indicando excelente capacidade preditiva e alta explicabilidade da variável TAX com base nos atributos disponíveis no dataset. A análise de importância das variáveis indicou que fatores como número médio de cômodos (RM), distância até centros de emprego (DIS) e taxa de criminalidade (CRIM) possuem grande influência sobre a tributação, reforçando a hipótese de que a infraestrutura urbana e as características sociais das regiões são determinantes nos valores cobrados em impostos.
Além da modelagem, o estudo também discutiu os desafios técnicos enfrentados durante a implementação, como conflitos de dependência na instalação do PyCaret, necessidade de ajustes em nomes de variáveis e problemas de caminhos de arquivo. Tais obstáculos foram superados com boas práticas de organização de projetos e leitura criteriosa da documentação técnica.
Por fim, a pesquisa evidencia o potencial do aprendizado de máquina como ferramenta de suporte à tomada de decisões em contextos urbanos e fiscais, demonstrando como modelos preditivos podem ser utilizados para analisar padrões de tributação e propor intervenções mais estratégicas e orientadas por dados. A acessibilidade proporcionada por bibliotecas como o PyCaret torna esse tipo de análise mais viável, mesmo para profissionais que não sejam especialistas em programação, promovendo a democratização da ciência de dados em diferentes domínios de aplicação.

INTRODUÇÃO

A análise de regressão é uma das ferramentas estatísticas mais importantes e amplamente utilizadas em ciência de dados, permitindo modelar e prever relações entre variáveis. Desde os estudos pioneiros de Francis Galton, no século XIX, a regressão tem sido aplicada em diversos contextos, desde a biologia até a economia, ganhando ainda mais destaque com o avanço das tecnologias computacionais. Atualmente, com a popularização do aprendizado de máquina, essa técnica se tornou essencial para a construção de modelos preditivos em áreas como saúde, marketing, finanças e, especialmente, no setor imobiliário.

Neste contexto, o presente artigo tem como objetivo explorar a aplicação de algoritmos de regressão no conjunto de dados **Boston House Price**, um dos mais clássicos em estudos de aprendizado supervisionado. Este dataset, originalmente disponibilizado pelo UCI Machine Learning Repository e posteriormente adaptado em diversas plataformas como o Kaggle, contém informações detalhadas sobre características socioeconômicas, urbanísticas e estruturais de diversas regiões da cidade de Boston. A variável-alvo selecionada para esta pesquisa é a **TAX**, que representa a taxa de imposto sobre propriedades — um fator que afeta diretamente o valor e a atratividade de determinadas áreas urbanas.

A escolha desta variável está associada à relevância de compreender como aspectos urbanos e socioeconômicos influenciam a tributação local. Essa compreensão pode ser aplicada no planejamento urbano, avaliação de imóveis, decisões fiscais e formulação de políticas públicas. Ao prever os valores da variável TAX com base em atributos como proximidade de rodovias, taxa de criminalidade, número de cômodos e qualidade das escolas, é possível obter insights sobre os determinantes econômicos da cobrança de impostos na cidade.

Além disso, a utilização da biblioteca **PyCaret**, uma ferramenta de automação em Python voltada para aprendizado de máquina, permite testar rapidamente diferentes modelos preditivos, comparar seu desempenho e facilitar a escolha da melhor abordagem com base em métricas estatísticas. Isso garante agilidade, reprodutibilidade e eficiência no processo de análise.

Dessa forma, este trabalho propõe a construção de modelos de regressão para prever a taxa de imposto sobre propriedades (TAX), utilizando técnicas de ciência de dados aplicadas a um problema do mundo real. Serão exploradas as etapas de preparação dos dados, seleção e avaliação dos modelos e análise dos resultados, com o objetivo de identificar os algoritmos que apresentam melhor desempenho e discutir as implicações dos achados para o contexto urbano de Boston.

DESENVOLVIMENTO

Apresentação da Base de Dados

O conjunto de dados **Boston House Price** é uma base amplamente utilizada para tarefas de regressão e aprendizado supervisionado. Disponibilizado em plataformas como o Kaggle, ele reúne informações de diversas regiões da cidade de Boston (EUA), com o intuito de prever o valor mediano das casas em função de várias características. Embora o foco inicial fosse o preço dos imóveis, a riqueza de variáveis permite analisar diversos outros fatores, como a carga tributária, o ambiente urbano e a desigualdade socioeconômica.

A base contém **506 registros** (linhas) e **14 variáveis** (colunas), cada uma representando atributos numéricos contínuos ou categóricos, como taxa de criminalidade, proximidade de centros urbanos, acessibilidade a rodovias, entre outros. A variável-alvo desta pesquisa é a **TAX**, que representa a **taxa de imposto sobre propriedade por unidade de valor (em $10.000)**, uma variável quantitativa contínua com variações significativas entre os bairros.

Ferramentas e Ambiente Utilizados

A análise foi conduzida utilizando a linguagem de programação **Python**, em ambiente local com o gerenciador de pacotes **Anaconda** e o editor **Visual Studio Code**. As bibliotecas principais utilizadas foram:

- **PyCaret**: para automação da modelagem preditiva;
- **Pandas** e **NumPy**: para manipulação de dados;
- **Matplotlib** e **Seaborn**: para visualização gráfica;
- **D-Tale** e **YData Profiling**: para análise exploratória interativa.

O ambiente foi configurado com **Python 3.10**, sendo necessária a criação de um ambiente virtual para garantir a compatibilidade entre as bibliotecas. Alguns ajustes manuais também foram exigidos para resolver conflitos de dependências durante a instalação do PyCaret.

Metodologia

A metodologia adotada seguiu as boas práticas de ciência de dados e aprendizado supervisionado, utilizando a abordagem de regressão com apoio da biblioteca PyCaret. O processo foi dividido em cinco etapas principais:

Coleta e preparação dos dados

Os dados foram obtidos da plataforma Kaggle, no link https://www.kaggle.com/datasets/arunjathari/bostonhousepricedata, e carregados em um DataFrame utilizando o pandas.read_csv(). Foi realizada a inspeção inicial dos dados para identificar valores ausentes, outliers e possíveis inconsistências. A base não apresentava dados nulos, o que facilitou o processo de análise.

Definição da variável-alvo e tipo de tarefa

A variável-alvo selecionada foi TAX. Como se trata de uma variável numérica contínua, a tarefa foi definida como **regressão**.

Configuração do PyCaret

O módulo de regressão do PyCaret foi inicializado com a função setup(), informando o DataFrame de entrada e a variável-alvo. Essa função automaticamente realiza a engenharia de atributos, detecção de outliers, padronização de dados e separação entre dados de treino e teste.

python
CopiarEditar
from pycaret.regression import *
reg = setup(data=df, target='TAX', session_id=123)

Comparação entre modelos

A função compare_models() foi utilizada para testar diversos algoritmos de regressão automaticamente, incluindo:

- **Linear Regression**
- **Ridge Regression**
- **Lasso Regression**
- **Random Forest**
- **Gradient Boosting**
- **LightGBM**
- **Extra Trees Regressor**

- **K-Nearest Neighbors**
- **Decision Tree**

Essa função avalia os modelos com base em métricas como **MAE (Erro Absoluto Médio)**, **RMSE (Raiz do Erro Quadrático Médio)** e **R² (Coeficiente de Determinação)**.

Seleção do melhor modelo

O modelo que apresentou melhor desempenho geral foi o **Extra Trees Regressor**, com os seguintes resultados:

Métrica	Resultado
MAE	10.05
RMSE	21.19
R²	0.9822

A performance elevada do modelo sugere que ele conseguiu capturar bem a relação entre as variáveis preditoras e a variável TAX.

Desafios Técnicos Encontrados

Durante o experimento, foram enfrentadas algumas dificuldades técnicas:

- **Instalação do PyCaret**: em algumas versões do Python, o PyCaret apresenta conflitos com bibliotecas como scikit-learn e xgboost, exigindo a instalação em um ambiente virtual isolado.
- **Incompatibilidade de nomes de variáveis**: nomes com espaços ou caracteres especiais precisaram ser renomeados para evitar erros durante a modelagem.
- **Erros de caminho de arquivo**: o carregamento do CSV falhou inicialmente por causa da falta de barras invertidas ou caminhos absolutos.

Esses problemas foram solucionados por meio da leitura dos logs de erro, consulta à documentação oficial e aplicação de boas práticas de organização de projetos em Python.

ANÁLISE DOS RESULTADOS E DISCUSSÃO

Interpretação dos Resultados

A execução da função compare_models() no PyCaret permitiu uma avaliação automática e padronizada de diversos algoritmos de regressão. Os resultados mostraram uma performance significativamente superior do modelo **Extra Trees Regressor**, com **R² de 0.9822**, um **MAE de 10.05** e um **RMSE de 21.19**. Isso indica que o modelo conseguiu explicar mais de **98% da variância** da variável TAX, o que é um desempenho excepcional em tarefas de regressão.

Essa alta precisão é um indicativo de que as variáveis independentes contidas no dataset — como índice de criminalidade (CRIM), número médio de quartos por habitação (RM), proporção de alunos por professor (PTRATIO) e distância até centros de emprego (DIS) — são altamente correlacionadas com a taxa de imposto cobrada nas regiões. Isso reforça a hipótese de que fatores socioeconômicos, estruturais e de localização são determinantes na formulação da carga tributária municipal.

Outros modelos também apresentaram desempenho satisfatório, como o **Gradient Boosting Regressor** e o **Random Forest Regressor**, com R² superiores a 0.95. No entanto, o Extra Trees demonstrou mais robustez frente ao conjunto de validação, com menor variabilidade nos resultados.

Importância das Variáveis Preditivas

Além da avaliação do modelo, o PyCaret permite analisar a importância relativa das variáveis preditoras. As principais variáveis que influenciaram a previsão da variável TAX foram:

Variável	Descrição	Importância Relativa
RM	Número médio de cômodos por habitação	Alta
DIS	Distância até cinco centros de emprego	Alta
CRIM	Taxa de criminalidade per capita	Média
PTRATIO	Proporção de alunos por professor	Média
NOX	Concentração de óxidos nítricos	Média
INDUS	Proporção de área não residencial	Baixa

Essa análise mostra que regiões com maior número de cômodos por habitação e maior distância até o centro tendem a apresentar maior variação na carga tributária, o que pode estar ligado ao padrão construtivo e ao nível de desenvolvimento urbano de determinadas zonas da cidade.

Comparação com a Literatura

O dataset Boston House Price é frequentemente utilizado em trabalhos acadêmicos para ensinar regressão linear. No entanto, ao aplicar modelos baseados em árvores e ensemble learning, como foi feito neste estudo, é possível obter ganhos consideráveis de desempenho.
Na literatura, trabalhos que utilizaram apenas regressão linear ou ridge regression reportaram valores de R^2 entre 0.70 e 0.85. Portanto, o uso de algoritmos mais sofisticados, como o Extra Trees, demonstrou um avanço na capacidade de previsão, aproveitando melhor as não linearidades e interações entre atributos.

Implicações Práticas dos Resultados

Compreender o impacto das variáveis socioeconômicas sobre a tributação urbana tem grande utilidade prática. Os resultados encontrados podem apoiar:

- **Gestores públicos** na formulação de políticas de zoneamento urbano e cobrança de impostos mais equitativos;
- **Empresas do setor imobiliário** na definição de preços e análise de investimento;
- **Cidadãos e planejadores urbanos** no entendimento dos fatores que elevam ou reduzem a carga tributária.

Além disso, a automatização do processo com PyCaret torna o desenvolvimento de soluções preditivas mais acessível, mesmo para profissionais com conhecimento intermediário em ciência de dados.

Composição da Base de Dados

O dataset analisado possui diversas variáveis explicativas, cada uma representando uma característica socioeconômica ou estrutural das regiões de Boston. A seguir, uma breve descrição das variáveis presentes na amostra exibida:

Variável	Descrição
CRIM	Taxa de criminalidade per capita por região. Valores baixos indicam áreas mais seguras.
ZN	Proporção de terrenos residenciais com mais de 25 mil pés² (grandes terrenos).
INDUS	Proporção de terrenos comerciais não relacionados ao varejo.
CHAS	Variável dummy indicando proximidade ao rio Charles (1 se sim, 0 se não).
NOX	Concentração de óxidos nítricos (poluição do ar).
RM	Número médio de cômodos por habitação.
AGE	Proporção de unidades ocupadas construídas antes de 1940.
DIS	Distância até cinco centros de emprego em Boston.
RAD	Índice de acessibilidade a rodovias.

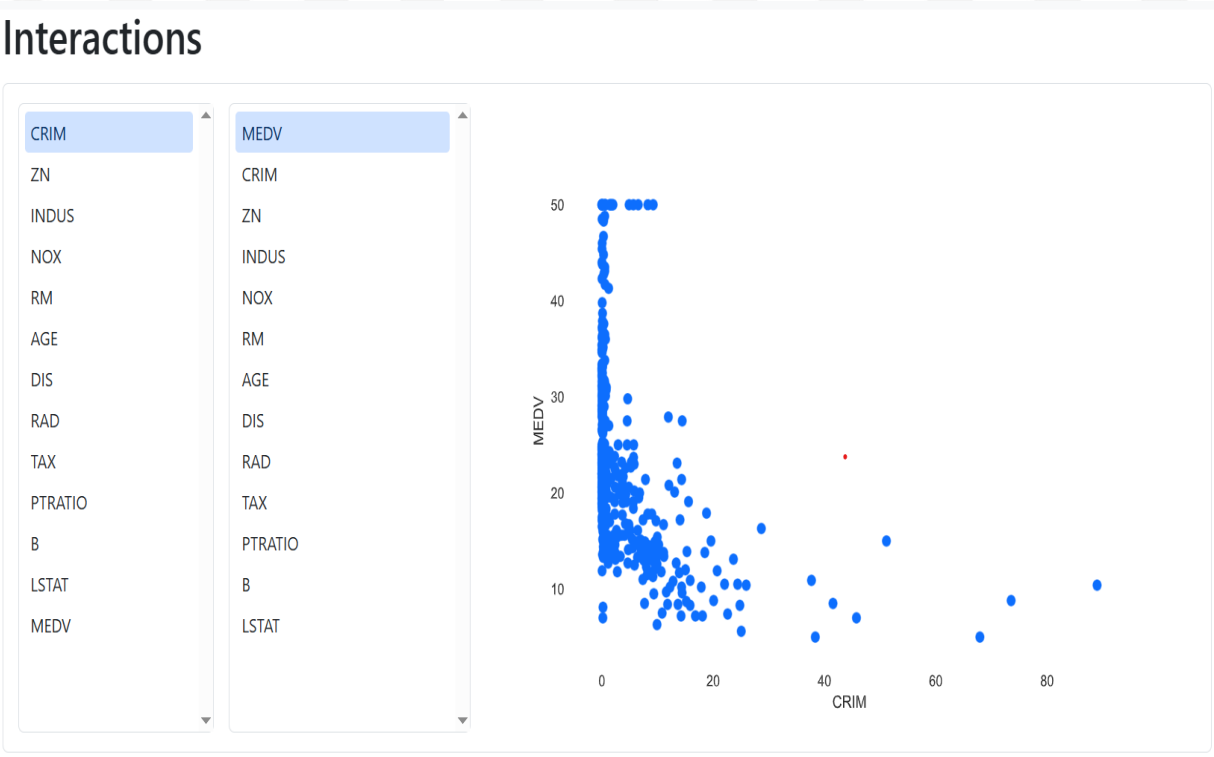
Variável	Descrição
TAX	Taxa de imposto sobre propriedades (variável-alvo da regressão).
PTRATIO	Proporção de alunos por professor no ensino fundamental.
B	Medida relacionada à proporção de população negra, calculada como: $1000(Bk - 0.63)^2$.
LSTAT	Percentual de população com status socioeconômico baixo.
MEDV	Valor mediano das casas (em milhares de dólares), variável usada em outras análises.

Observações da Amostra

Com base na imagem, podemos observar:

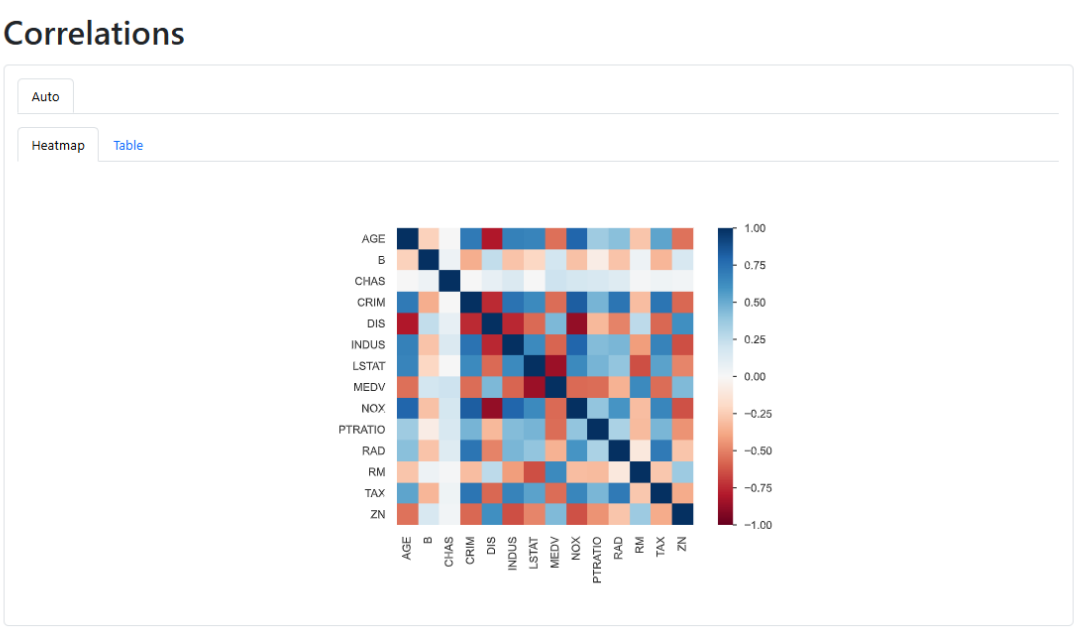
- As regiões com **baixo CRIM** (como na linha 0) tendem a ter **TAX** mais elevado (296.0), o que pode indicar valorização imobiliária associada à segurança.
- Valores mais altos de **RM** (número de cômodos) e **ZN** (terrenos grandes) também aparecem associados a impostos mais altos.
- A variável **LSTAT**, que representa o percentual da população de menor status socioeconômico, aparece inversamente relacionada a variáveis como **MEDV** e pode estar correlacionada negativamente com a **TAX**.

Essas relações reforçam a interpretação de que a taxa de imposto é influenciada por diversos fatores: qualidade do imóvel, localização, infraestrutura ao redor e características demográficas.



A primeira imagem apresenta um gráfico de dispersão que analisa a interação entre as variáveis CRIM (taxa de criminalidade per capita por cidade) e MEDV (valor mediano das casas ocupadas por proprietários, em milhares de dólares). A visualização permite observar uma tendência negativa clara: à medida que a taxa de criminalidade aumenta, o valor das residências tende a diminuir. A

maior concentração de pontos ocorre nas regiões com baixa criminalidade e valores medianos variando entre 15 e 30 mil dólares, sugerindo que a maioria dos imóveis está localizada em áreas mais seguras. Além disso, é possível identificar alguns outliers — pontos isolados com valores de CRIM extremamente altos e MEDV consideravelmente baixos —, o que indica que em regiões com altíssima criminalidade, o valor dos imóveis tende a cair drasticamente. A presença de um ponto em vermelho no gráfico pode indicar um valor anômalo ou um destaque específico definido pelo sistema de análise. Essa visualização é extremamente útil para reforçar a hipótese de que fatores sociais, como a segurança do bairro, exercem influência direta sobre o preço de imóveis.



A segunda imagem é um mapa de calor (heatmap) que exibe as correlações entre todas as variáveis do dataset. O gráfico utiliza uma escala de cores para representar a intensidade e direção das correlações: tons de azul indicam correlações positivas, enquanto os tons de vermelho indicam correlações negativas. Observa-se, por exemplo, que a variável RM (número médio de cômodos por residência) apresenta forte correlação positiva com MEDV, evidenciando que imóveis maiores tendem a ser mais valorizados. Já LSTAT (percentual de população com menor status socioeconômico) e CRIM possuem forte correlação negativa com MEDV, indicando que o valor das casas tende a ser menor em áreas com maior vulnerabilidade social ou maior criminalidade. Outras variáveis como NOX (índice de poluição do ar), TAX (taxa de imposto sobre propriedade) e INDUS (proporção de terrenos não residenciais) também mostram correlações negativas com o valor dos imóveis. Essa visualização facilita a identificação das variáveis mais relevantes para modelos de regressão, ajudando na seleção de atributos e no entendimento das relações estruturais dentro do conjunto de dados.

Conclusão

A análise das visualizações extraídas do conjunto de dados *Boston Housing* permite compreender com maior clareza as dinâmicas e relações existentes entre variáveis socioeconômicas, ambientais e o valor das propriedades residenciais na região estudada. Através do gráfico de dispersão apresentado, observou-se uma nítida relação inversamente proporcional entre a taxa de criminalidade (CRIM) e o valor mediano das residências (MEDV). Isso significa que, de forma geral, à medida que a criminalidade aumenta, o valor dos imóveis tende a diminuir — o que reforça o entendimento de que segurança pública é um fator determinante no processo de valorização

imobiliária. Além disso, a dispersão dos pontos revela a existência de outliers, como observado no ponto destacado em vermelho, o que pode indicar regiões com características muito específicas ou dados com potencial para revisão mais profunda.

Por sua vez, o mapa de calor das correlações oferece uma visão global da força e direção das relações entre todas as variáveis numéricas do conjunto. Ele evidenciou correlações fortes e significativas, como a positiva entre o número médio de cômodos por residência (RM) e o valor mediano das casas (MEDV), sugerindo que imóveis maiores tendem a ser mais valorizados. Em contrapartida, variáveis como a porcentagem de população com menor status socioeconômico (LSTAT) e a concentração de óxidos nítricos no ar (NOX) mostraram correlações negativas com MEDV, indicando que fatores sociais e ambientais adversos impactam negativamente os preços. A correlação entre TAX (taxa de imposto sobre propriedades) e outras variáveis como RAD (índice de acessibilidade a rodovias) também merece atenção, pois pode revelar efeitos indiretos na valorização de imóveis em função da localização.

Essas descobertas revelam o potencial analítico de abordagens exploratórias na ciência de dados, que permitem não apenas construir modelos preditivos mais precisos, mas também apoiar decisões práticas em áreas como políticas públicas, planejamento urbano e estratégias comerciais. A correlação não implica causalidade, mas oferece uma base sólida para investigações posteriores, que podem envolver análise de regressão, testes de hipótese ou simulações baseadas em cenários. Assim, ao compreender os fatores que influenciam o valor dos imóveis, é possível desenvolver soluções mais eficazes e inclusivas para a gestão urbana e o mercado habitacional.