Regressão

Lucas Gessner da Silva
Centro Tecnológico de Joinville (CTJ)
Universidade Federal de Santa Catarina
Joinville, Brasil

Daniel Augusto S. Porto
Centro Tecnológico de Joinville (CTJ)
Universidade Federal de Santa Catarina
Joinville, Brasil

Lucas Dantas Igarashi
Centro Tecnológico de Joinville (CTJ)
Universidade Federal de Santa Catarina
Joinville, Brasil

Resumo— Este estudo aplica e avalia diferentes métodos de regressão em uma base de dados para prever o preço de revenda de veículos usados. A fase de pré-processamento envolveu a análise de atributos, correção de sintaxe, preenchimento dinâmico de lacunas para variáveis contínuas e conversão de variáveis categóricas. Foram utilizados três modelos de regressão: Regressão Linear, Random Forest Regressor e Gradient Boosting Regressor. Os modelos foram avaliados utilizando validação cruzada com cinco pastas, considerando métricas como o Erro Médio Absoluto (MAE) e o R^2. Os resultados fornecem insights sobre o desempenho e a precisão de cada modelo. **(**resumo**)**

Palavras-chave—regressão, pré-processamento, veículos usados, Regressão Linear, Random Forest, Gradient Boosting, validação cruzada, Erro Médio Absoluto, R^2 (palavras-chaves**)**

I. Pré-Processamento dos Dados

A. Analisando os Atributos

Durante a etapa de análise exploratória, identificamos que alguns atributos poderiam ser removidos da base de dados sem prejudicar a qualidade das previsões. A exclusão desses atributos foi baseada na falta de relevância, redundância ou pela baixa contribuição na capacidade preditiva do modelo. Abaixo, listamos os atributos deletados e os motivos para sua exclusão.

TABELA I. Atributos Deletados

Atributo	**Motivo**
ID	Identificador único, irrelevante para a previsão do preço.
Débitos	Informações financeiras específicas, introduzem ruído.
Fabricante, Modelo	Redundantes, capturados por classificação e combustível.
Categoria, Couro, Volume_motor, Tipo_cambio, Tração, Portas, Rodas, Airbags, Numero_proprietarios, Adesivos_personalizados, Radio_AM_FM, Codigo_concessionaria	Detalhes específicos ou com variância mínima, não contribuem significativamente.
Cor	Subjetiva, impacto variável.
Data_ultima_lavagem, Historico_troca_oleo	Sem relação direta com o valor de revenda.

Os atributos que foram mantidos na base de dados foram selecionados devida sua relevância e influência direta no preço de revenda dos veículos. São eles:

TABELA II. Atributos Permanecidos

Atributo	**Motivo**
Ano, Km	Relevante na depreciação do valor do veículo.
Combustível	Impacta diretamente o seu valor de mercado.
Cilindros	Influencia a performance e o valor do veículo.
Classificacao_veiculo	Inclui informações sobre o tipo e segmento do veículo.
Faixa_preço	Intervalos que captura variações do preço dentro de segmentos específicos.
Preco	Alvo

B. Arrumando as Sintaxes

Entre os atributos restantes foi preciso arrumar a sintaxe na coluna “Combustível” e “Km”. Na característica combustível foi padronizado as diferentes formas que estavam escritos “Gasolina” e “Diesel”. Já na coluna Km, os dados foram mudados de string para float.

C. Linhas com Lacunas

Para manter a consistência dos dados, as lacunas de variáveis contínuas foram preenchidas com a média da coluna. Além disso, as lacunas na variável “Faixa_Preco” foram preenchidas dinamicamente com base nos valores de Preco. Os limites para cada faixa de preço (Econômico, Médio, Luxo, Muito Luxo) foram calculados com base nos valores máximos de “Preco” dentro de cada faixa existente. Isso garante que os valores preenchidos para “Faixa_Preco” sejam consistentes com os dados existentes. Porém, as linhas com lacunas nos demais atributos discretos foram eliminados para não serem preenchidas de forma enviesadas.

D. Discreta para Contínua

Para os atributos discretos ordinários, como a faixa de preço, foram atributos valores inteiros crescentes dada sua ordem. Entretanto, o mesmo não pode ser feito para atributos nominais, não havendo uma hierarquia intrínseca. Nesses casos, levando em conta que o número de categorias em cada coluna nominal era baixo, foi usado o método One Hot Encoding, transformando cada categoria em um atributo diferente.

E. Padronizando e Removendo os Outliers

Para melhorar o desempenho dos modelos e a retirada dos outliers, todos os atributos foram padronizados. Após esse processo a identificação de dados anômalos pelo método One Class SVM foi mais eficaz.

II. Modelos e Avaliações

Foram utilizados três modelos de regressão: Regressão Linear, Random Forest e Gradient Boosting. Com validação cruzada em 5 pastas, avaliamos a Média do Erro Absoluto (MAE) e o R^2 a fim de medir o erro médio e a precisão das previsões. Dessa forma, os resultados obtidos são:

A. *Regressão Linear*

TABELA III. Regressão Linear - Peformace

Métrica	Valor
R2 Médio	0.7682 (±0.0094)
MAE Médio	6146.2630 (±115.0881)

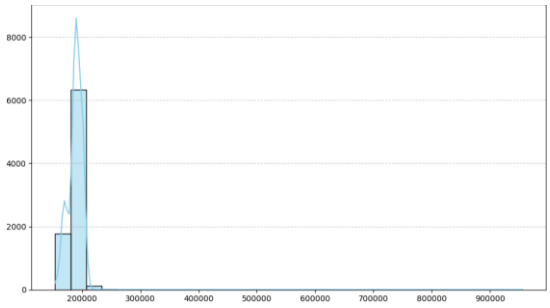

Fig 1. Histograma – Erro (Preço Real X Preço Previsto)

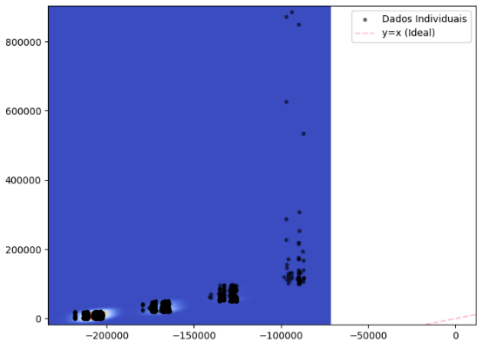

Fig 2. Heatmap – Preço Real X Preço Previsto

A análise de regressão linear para a base de dados multidimensional apresentou resultados que indicam sua limitação. Embora o R^2 médio e o MAE médio sejam relativamente altos, esses resultados não são confiáveis devido ao overfitting. O modelo se ajusta excessivamente aos dados de treinamento, capturando o ruído em vez de identificar padrões reais. Os gráficos evidenciam essas limitações: o histograma do erro (Fig. 1) mostra uma distribuição com erros elevados, e o mapa de calor (Fig. 2) revela discrepâncias entre os preços reais e previstos. Portanto, a regressão linear não é adequada para a complexidade dos dados não-lineares desta base, sendo incapaz de generalizar para novos dados.

B. *Random Forest Regressor*

TABELA IV. *Random Forest* - Peformace

Métrica	Valor
R2 Médio	0.8108 (±0.0135)
MAE Médio	4751.2926 (±139.8813)

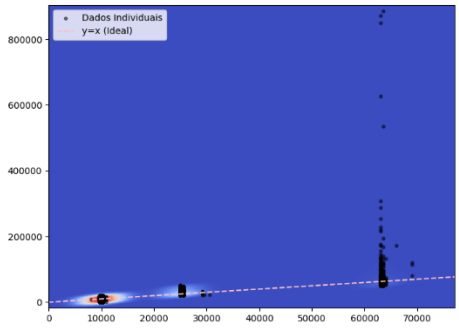

Fig 3. Heatmap – Preço Real X Preço Previsto

O Random Forest demonstrou uma performance sólida na previsão de preços de revenda de veículos. O mapa de calor (Fig. 3) mostra que para valores entre 100.000 e 800.000, o erro aumenta significativamente devido à evolução brusca do preço e à diminuição da quantidade de dados nessa faixa, dificultando a modelagem precisa dos preços mais altos. Apesar disso, como a grande maioria dos dados está entre 0 e 70.000, o modelo ainda apresenta uma alta concentração de pontos ao longo da linha y=x, indicando previsões precisas.

C. *Gradient Boosting Regressor*

TABELA V. *Gradient Boosting* - Peformace

Métrica	Valor
R2 Médio	0.8184 (±0.0092)
MAE Médio	5286.2781 (±127.9089)

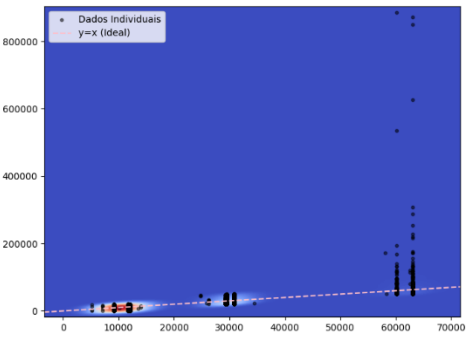

Fig 4. Heatmap – Preço Real X Preço Previsto

Por fim, o Gradient Boosting apresentou resultados ligeiramente melhores que o Random Forest devido à sua capacidade de reduzir erros de previsão de maneira mais eficiente ao combinar múltiplos modelos fracos de maneira sequencial, corrigindo os erros do modelo anterior. Assim, melhorando a precisão, especialmente em dados complexos e não lineares.

References

[1] *Medium: Read and write stories.*

[2] scikit-learn: machine learning in Python — scikit-learn 1.5.2 documentation

[3] pandas documentation — pandas 2.2.3 documentation

[4] Matplotlib — Visualization with Python

[5] seaborn: statistical data visualization — seaborn 0.13.2 documentation