Aprendizado de Máquina na Precificação de Carros Usados: desenvolvimento de uma base de dados para modelos de regressão

Machine Learning in Used Car Pricing: Development of a Database for Regression Models

Pedro P. O. Moura[1], **Raphael Barbosa Holmes**[1], **Sheyla Natália de Medeiros**[1]

[1] Análise e Desenvolvimento de Sistemas – Instituto Federal de Pernambuco
Paulista – PE – Brasil

ppom@discente.ifpe.edu.br, rhb@discente.ifpe.edu.br,

sheyla.medeiros@paulista.ifpe.edu.br

Resumo. *A compra e venda de carros usados é uma atividade vital para a economia brasileira e para o acesso à mobilidade urbana, especialmente em um cenário de aumento dos preços de veículos novos. No entanto, determinar o preço de um carro usado pode ser desafiador. Este trabalho explora o uso de técnicas de aprendizado de máquina para aprimorar a acurácia na precificação de automóveis usados. Desenvolveu-se um conjunto de dados abrangente, extraído da plataforma OLX Brasil por meio de raspagem de dados utilizando Scrapy. A seguir, realizou-se uma análise exploratória dos dados, além de limpeza e preparação para o treinamento de diversos modelos de aprendizado supervisionado voltados para regressão. Os resultados mostraram-se promissores, especialmente com o modelo de Floresta Aleatória, que alcançou um coeficiente de determinação de 0,9434 e um erro médio absoluto de 4855,27. Estes resultados indicam que modelos de regressão podem ser eficazes na previsão de preços de veículos com base em suas características, e sugerem a necessidade de investigações adicionais para aprimorar ainda mais as técnicas de precificação de carros usados.*
Palavras-chave: Aprendizado de Máquina, Aprendizado Supervisionado, Raspagem de Dados, Análise de Dados, Regressão

Abstract. *The buying and selling of used cars is a vital activity for the Brazilian economy and for access to urban mobility, especially in a scenario of rising new vehicle prices. However, determining the fair price of a used car can be challenging. This final year project explores the use of Machine Learning techniques to improve the accuracy of used car pricing. A comprehensive dataset was developed, extracted from the OLX Brazil platform through Web Scraping using Scrapy. Subsequently, exploratory data analysis was carried out, as well as cleaning and preparation for training various supervised learning models focused on regression. The results were promising, especially with the Random Forest model, which achieved a determination coefficient of 0.9434 and a mean absolute error of 4855.27. These results indicate that regression models can be effective in predicting vehicle prices based on their characteristics, and suggest the need for further investigations to further improve used car pricing techniques.*
Keywords: Machine Learning, Supervised Learning, Web Scraping, Data Analysis, Regression

1. Introdução

A compra de veículos usados oferece uma alternativa mais acessível para quem busca mobilidade, sendo essencial para a autonomia em diversas regiões do Brasil. Em 2024, as vendas de carros usados cresceram 14% em relação ao ano anterior, totalizando 766.558 unidades em janeiro, enquanto os carros novos somaram apenas 118.507 no mesmo período, quase sete vezes menos (DREHMER, 2024). Essa disparidade reflete a alta demanda por veículos mais acessíveis no mercado automotivo brasileiro.

Nesse contexto, a tabela FIPE desempenha um papel crucial. Criada para fornecer uma referência confiável na determinação do valor de mercado de veículos usados, a tabela FIPE é atualizada mensalmente com base em dados coletados de transações reais. Esta ferramenta não só influencia as negociações comerciais entre compradores e vendedores, mas também é utilizada para calcular o valor do prêmio de seguros automotivos e o valor do Imposto sobre Propriedade de Veículos Automotores (IPVA) (QUATRO RODAS, 2023).

No entanto, apesar de ser muito útil, a tabela FIPE tem algumas limitações. Por exemplo, ela oferece uma estimativa geral do valor de mercado com base em uma média de preços, sem levar em consideração características individuais de um veículo. Ou seja, quilometragem, estado de conservação geral do veículo e equipamentos instalados após a compra não são levados em conta na hora de definir a tabela FIPE de um carro.

Dessa forma, a tecnologia pode ajudar a superar essas limitações, destacando-se o uso de raspagem de dados e aprendizado de máquina. A raspagem de dados é uma técnica que permite a coleta automatizada de informações de páginas web, mantendo a estrutura original dos dados (REITZ, 2016). Essa coleta inicial pode ser a base para que um ator independente, após realizar um processo cuidadoso de limpeza, tratamento e organização dos dados, realize análises em larga escala, como a comparação de preços no mercado de carros usados, sem depender de canais oficiais. Com os dados prontos, modelos de aprendizado de máquina podem ser treinados. Aprendizado de máquina é uma disciplina que se dedica a desenvolver sistemas computacionais capazes de melhorar automaticamente com a experiência, além de investigar as leis fundamentais estatísticas, computacionais e teóricas que regem todos os sistemas de aprendizado, abrangendo computadores, humanos e organizações (MITCHELL; JORDAN, 2015). No contexto deste trabalho, esses modelos aprendem a associar características de veículos a preços, permitindo a previsão precisa do valor ideal de um carro baseado em seus atributos .

Assim, este trabalho tem como objetivo principal desenvolver uma plataforma de coleta e análise de dados, utilizando raspagem de dados para extrair informações do mercado de carros usados e modelos de regressão supervisionada de aprendizado de máquina para prever o preço desses veículos com maior precisão. Como objetivos específicos, enumera-se: desenvolver um conjunto de dados abrangente utilizando raspagem de dados, realizar análise exploratória e limpeza dos dados coletados e treinar e avaliar diferentes modelos de aprendizado de máquina supervisionado. A proposta busca não apenas melhorar a precisão da precificação ao considerar características específicas dos veículos, mas também fornecer uma ferramenta acessível que possa auxiliar vendedores e compradores em negociações mais justas.

2. Referencial Teórico

2.1. Aprendizado de Máquina

O conceito de aprendizado de máquina é definido como um processo em que um programa de computador aprende a realizar tarefas a partir de experiências, utilizando uma métrica de desempenho

como referência (MITCHELL, 1997). O aprendizado ocorre quando há uma melhora na performance da tarefa com o acúmulo de experiência. Em seu trabalho, Mitchell exemplifica várias aplicações possíveis para algoritmos de aprendizado de máquina, como o reconhecimento de palavras faladas, a condução de veículos autônomos, a classificação de estruturas astronômicas e a prática de jogos de tabuleiro, como damas ou gamão. No contexto de um jogo de damas, podemos identificar várias variáveis relevantes. A tarefa principal é jogar damas, enquanto a métrica de performance é a porcentagem de jogos ganhos. Além disso, a experiência de treinamento pode ser obtida por meio de jogos simulados contra si mesmo.

Complementando essa definição, afirma-se que uma máquina aprende sempre que modifica sua estrutura, programa ou dados com base em *inputs* ou informações externas, de modo que sua performance futura esperada melhore (NILSSON, 2015). Um exemplo claro disso é uma máquina de reconhecimento de voz que aprimora seu desempenho após processar diversas amostras de fala de uma pessoa específica, caracterizando o processo de aprendizado.

O autor também enfatiza a importância do aprendizado de máquina, mesmo quando é possível desenvolver sistemas que atinjam um desempenho satisfatório desde o início. Uma das razões é que muitas atividades não podem ser descritas de forma precisa sem a utilização de exemplos. Algumas tarefas são tão complexas que exigem que as máquinas aprendam a associar entradas às saídas desejadas a partir de exemplos concretos. Um exemplo clássico é o reconhecimento de imagens, onde as máquinas aprendem a identificar objetos em fotos por meio de grandes conjuntos de imagens previamente rotuladas.

Além disso, grandes volumes de dados podem conter informações valiosas, frequentemente ocultas sob a forma de padrões e correlações. Técnicas de aprendizado de máquina são empregadas para extrair essas informações e identificar relações significativas. Essa capacidade permite que as máquinas não apenas aprendam com a experiência, mas também se adaptem a ambientes complexos, superando as limitações de projetos humanos iniciais. Ademais, como esses ambientes podem evoluir ao longo do tempo, a habilidade de adaptação reduz a necessidade de constantes reconfigurações manuais.

2.1.1. Aprendizado Supervisionado

Dentro do campo do aprendizado de máquina há três abordagens principais, segundo Shagan Sah (2020): supervisionado, não supervisionado e por reforço. Este trabalho focará exclusivamente no aprendizado supervisionado.

O aprendizado supervisionado é definido como um processo em que uma máquina é alimentada com um conjunto de exemplos rotulados, denominado dados de treinamento, e utiliza essas informações para realizar previsões em pontos não rotulados (MOHRI et al, 2012).

Dentro do aprendizado supervisionado, tem-se dados na forma de variáveis de *input* e uma coluna de "alvo". O objetivo é que o algoritmo aprenda a prever ou estimar esse valor de alvo durante o treinamento, usando os outros dados do conjunto de dados como referência. É uma abordagem onde o algoritmo aprende uma função que mapeia as variáveis de entrada para o alvo desejado com base nos exemplos de treinamento fornecidos. Ainda dentro do aprendizado supervisionado, há duas principais categorias: Classificação e Regressão. Nos algoritmos de classificação, são categorizadas as variáveis de *input* em um conjunto de categorias pré-estabelecidas: por exemplo, treinar um algoritmo para diferenciar homens e mulheres. Então, com base nas características físicas de um grupo

de indivíduos, treinaremos o programa para identificar o sexo desses indivíduos. Já utilizando algoritmos de regressão, buscamos prever valores numéricos como preços de ações ou propriedades, scores de cartão de crédito, temperatura e vários outros. Já que o presente trabalho busca prever preços de carros usados, sendo esse um dado numérico, técnicas de regressão são as mais adequadas para essa tarefa.

2.1.2. Regressor de Árvore de Decisão

As árvores de decisão são definidas como um método para aproximar funções de destino de valores discretos, no qual a função aprendida é representada por uma árvore de decisão (MITCHELL, 1997). As árvores de decisão são estruturas de aprendizado de máquina compostas por nós interligados por arestas, sendo cada nó representativo de uma decisão a ser tomada com base nos dados, e as arestas indicando possíveis caminhos a serem seguidos.

Tradicionalmente, árvores de decisão eram utilizadas para problemas de classificação, como indica o próprio Mitchell ao dizer que as funções destino têm valores discretos, ou seja, categóricos. No entanto, na atualidade, regressores de Árvores de Decisão são amplamente utilizados. Para adaptar seu uso à regressão, a métrica de desvio padrão assume um papel crucial (LOH, 2014). Ela mede a dispersão dos dados, indicando a variabilidade dos valores em um conjunto. Ao minimizar o desvio padrão em cada nó, o algoritmo cria grupos de dados mais homogêneos, facilitando a previsão de valores futuros.

2.1.3. Regressor de Florestas Aleatórias e Árvores Extremamente Aleatórias

Florestas Aleatórias são definidas como uma combinação de preditores de árvore de forma que cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta (BREIMAN, 2001). Sendo uma combinação de preditores, se trata então de uma técnica de *ensemble* (conjunto): trabalham no mesmo problema de aprendizado diversos modelos diferentes de Árvore de Decisão.

Utilizar Árvores de Decisão individuais pode ser problemático, já que a literatura mostra que esse algoritmo é sensível a mudanças nos dados (RAJ, 2020) e tipicamente apresentam alta variância e tendem ao *overfit*. A introdução de aleatoriedade no processo de construção das florestas permite que as árvores corrijam os erros umas das outras, mitigando esses problemas e, frequentemente, resultando em um modelo com melhor desempenho.

As Árvores Extremamente Aleatórias, como o nome sugere, aumenta a aleatoriedade ao utilizar o conjunto de treino inteiro em todas as árvores, ao invés de treiná-las com uma amostra, e também por definir de forma aleatória a ramificação de nós em cada árvore (GEURTS et al, 2006). Sendo assim, a aleatoriedade vem da forma como essas árvores são ramificadas. O intuito desse modelo é apresentar uma variância ainda menor, tendo o benefício adicional de um tempo de execução menor, já que as Árvores Aleatórias tentam ramificar seus nós de forma óptima, enquanto o método do outro modelo é menos computacionalmente intensivo (AZNAR, 2020).

2.1.4. Regressor de Bagging

Preditor *Bagging* (Bootstrap Aggregating) é mais um método de *ensemble*. É definido como um método para gerar múltiplas versões de um preditor e usá-la para conseguir um preditor agregado. O preditor *Bagging* consiste em gerar múltiplas versões de um mesmo modelo de aprendizado de máquina e combinar suas previsões para obter um resultado final mais preciso e robusto (BREIMAN, 1996).

Primeiro, o conjunto de dados original é dividido em diversas amostras aleatórias com reposição, chamadas de "substituições". Cada substituição possui o mesmo tamanho do conjunto original, mas com alguns pontos repetidos e outros ausentes. Um modelo de aprendizado de máquina, como uma árvore de decisão, é treinado em cada substituição. Como cada substituição possui uma composição diferente de dados, os modelos treinados terão perspectivas ligeiramente distintas sobre os dados.

As previsões individuais de cada modelo são combinadas, geralmente por meio da média aritmética, para gerar uma única previsão final. Essa agregação ajuda a reduzir o erro geral e a aumentar a confiabilidade do modelo.

2.2. Raspagem de dados web

Raspagem de dados web é definido como uma técnica para extrair dados da *World Wide Web* e salvá-la num sistema de arquivos ou banco de dados para análise futura (ZHAO, 2017). Ele cita como um processo que pode ser realizado manualmente, mas nos dias atuais as técnicas de raspagem web são praticamente sinônimas com o uso de *bots* para coleta de dados automatizada: devido a escala do volume de dados que as indústrias necessitam capturar e analisar, realizar essa atividade de forma manual é no melhor dos casos uma tarefa infrutífera, e no pior uma tarefa impossível.

A raspagem de dados tem aplicações em pesquisas de mercado, ao coletar dados sobre preços, produtos e tendências de mercado, no monitoramento de mídias sociais, ao acompanhar menções à uma marca e, é claro, na pesquisa acadêmica. É importante utilizar técnicas de raspagem de dados de forma responsável. O uso de *bots* para realizar requisições a um domínio pode impactar negativamente a experiência de outros usuários e sobrecarregar a infraestrutura do site.

2.3. Métricas de Avaliação

As métricas de avaliação desempenham um papel crucial na análise da qualidade dos modelos que utilizamos para resolver problemas específicos. Elas servem como ferramentas para verificar a eficácia dos modelos em relação aos dados disponíveis. Além disso, essas métricas nos permitem realizar comparações detalhadas entre diferentes modelos, facilitando a identificação do modelo que melhor se ajusta às características e necessidades dos nossos dados.

2.3.1. R^2

O coeficiente de determinação, também conhecido como R^2, é uma medida que expressa a proporção da variância da variável dependente que pode ser explicada pelas variáveis independentes (CHICCO et al, 2021). O valor de R^2 varia de 0 a 1, onde valores mais próximos de 1 indicam um melhor ajuste do modelo aos dados. O R^2 pode interpretado também como uma "Porcentagem de Variação Explicada", multiplicando o valor de R^2 por 100 para expressá-lo como uma porcentagem. Por exemplo, se o

coeficiente de determinação da regressão linear for 0,80, consideramos que esse modelo explica 80% da variação nos dados.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\overline{Y} - Y_i)^2} \tag{1}$$

A fórmula 1 define o R² como 1 menos a razão entre dois somatórios. No numerador, está a soma dos quadrados dos erros, que é a diferença entre os valores previstos pelo modelo (***Yi***) e os valores observados (***Xi***). No denominador, temos a soma total dos quadrados, que mede a variabilidade total dos dados em relação à média. Quanto maior for a soma dos quadrados dos erros, pior será o ajuste do modelo, pois as previsões estarão mais distantes dos valores reais. Já a soma total dos quadrados reflete a variabilidade natural dos dados em torno da média.

2.3.2. *Mean Squared Error*

O Erro Médio Quadrático (MSE) é definido como a média das diferenças ao quadrado entre o *output* real e o *output* previsto num modelo de regressão cujos valores alvos são contínuos (CHICCO et al, 2021).

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \tag{2}$$

A fórmula 2 calcula o somatório dos quadrados das diferenças entre os valores observados na amostra (***Xi***) e os valores previstos pelo modelo (***Yi***). O termo ***m*** representa o número total de observações. Após obter o somatório, dividimos o resultado pelo número de observações, o que nos dá o valor médio dos erros quadráticos (MSE). Quanto menor o MSE, melhor o ajuste do modelo, indicando erros menores nas previsões. Um MSE elevado, por outro lado, significa que as previsões estão, em média, mais distantes dos valores observados.

2.3.3. *Root Mean Squared Error*

A Raiz do Erro Médio Quadrático (RMSE) é a raiz quadrada do valor do MSE, descrito anteriormente (CHICCO et al, 2021). A intenção aqui é demonstrar o valor do MSE na mesma unidade de medida da variável de alvo, no nosso caso, reais.

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \tag{3}$$

2.3.4. Mean Absolute Error

Pode-se definir o Erro Médio Absoluto (MAE) como uma métrica amplamente utilizada em problemas de regressão para avaliar a diferença média absoluta entre os valores previstos e os valores reais (AHMED, 2023). O MAE é intuitivo e de fácil interpretação, pois expressa o erro na mesma unidade da variável alvo, que neste trabalho é o real, facilitando a compreensão dos resultados financeiros. Além disso, o MAE é menos sensível a *outliers*, tornando-o uma métrica robusta para avaliar a precisão das previsões.

Embora o RMSE compartilhe a característica de expressar o erro na mesma unidade da variável alvo, o MAE complementa a análise ao fornecer uma visão linear das discrepâncias, sem dar peso adicional a grandes desvios. Incluir o MAE, portanto, oferece uma perspectiva mais abrangente da performance do modelo.

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i| \tag{4}$$

Matematicamente, como o próprio nome sugere, o MAE é calculado como a média do valor absoluto das diferenças entre os valores previstos e os valores reais para cada observação no conjunto de dados.

3. Trabalhos Relacionados

Foi utilizado o Google Scholar como base para busca de referências. Inicialmente, as *keywords* de pesquisa utilizadas foram "Machine Learning Price Prediction", porém, descobrimos que os temas dominantes eram relacionados à previsão de preços de ações na bolsa de valores e de criptomoedas. O intuito dessa pesquisa era encontrar trabalhos que utilizassem aprendizado de máquina para prever o preço de venda de mercadorias baseado em suas características, enquanto trabalhos focados em criptomoedas e ações são baseados na análise de séries temporais para realizar sua previsão de preços. Portanto, a *string "machine learning price prediction -bitcoin -stock"* foi utilizada para eliminar esses trabalhos da busca. O recorte de tempo utilizado para procurar esses trabalhos foi de 2022 até 2024.

O estudo da previsão de preços é uma das aplicações mais comuns de técnicas de aprendizado de máquina, portanto, há diversos trabalhos que abordam o tema. Por exemplo, Silva (2023) realizou uma análise para prever o preço de imóveis em Recife, utilizando dados da OLX e o algoritmo de Floresta Aleatória. Embora tenha alcançado uma média de Erro Médio Absoluto Percentual de apenas 19,71%, o estudo oferece contribuições relevantes por meio de uma análise exploratória de dados. Esse artigo é relevante porque utiliza a mesma fonte de dados que o presente trabalho, além de algoritmos de regressão semelhantes.

Outro exemplo é o trabalho de Dias (2023), que buscou prever os preços de terrenos em Brasília em sua tese de mestrado. Utilizando o *HistGradientBoostingRegressor*, obteve um R^2 de 0,922 e um RMSE de 0,39, demonstrando um bom ajuste do modelo aos dados.

No estudo conduzido por Bukvic et al. (2022), os pesquisadores alcançaram métricas notáveis de 1373,2 para Erro Médio Absoluto e 0.95 para R^2 utilizando um algoritmo de Regressão Linear. Esses resultados são significativos porque indicam um forte relacionamento linear entre as variáveis dependentes e independentes no conjunto de dados analisados. Esse tipo de relação é relevante para o presente trabalho, pois sugere que a aplicação de modelos lineares pode ser eficaz na previsão de

preços de carros usados. Ao entender essa relação, é possível escolher algoritmos adequados que maximizem a precisão na precificação de veículos, um dos objetivos centrais desta pesquisa.

No contexto da previsão de preços de carros usados, que é o foco central deste trabalho, é importante destacar a dissertação de mestrado de Magalhães (2023). Ao empregar a técnica *XGBoost*, Magalhães alcançou um R^2 de 0,96432 e um MAE de 0,12892, demonstrando a alta eficácia de seu modelo. A relevância dessa dissertação reside na similaridade metodológica com a do presente trabalho, especialmente na coleta de dados a partir de uma plataforma de compra e venda de veículos e na construção de um conjunto de treinamento para modelos de aprendizado supervisionado. No entanto, a principal diferença está na abordagem para obtenção dos dados: enquanto utilizamos técnicas de raspagem de dados, Magalhães obteve os dados diretamente do website, mediante solicitação formal.

A revisão da literatura demonstra uma base consolidada para a aplicação de algoritmos de aprendizado de máquina supervisionado na previsão de preços de objetos cujos valores são influenciados por suas características individuais, como terrenos, apartamentos e carros. Esse conjunto de estudos oferece suporte teórico robusto para a implementação de metodologias semelhantes no presente trabalho, que visa prever o valor de carros usados com base em características que variam significativamente entre os veículos. A extração de dados de plataformas de venda online se destaca como uma estratégia amplamente utilizada, sendo a OLX Brasil um possível candidato para aplicação no contexto brasileiro. Além disso, os trabalhos analisados fazem uso de algoritmos de aprendizado supervisionado, principalmente algoritmos de regressão, para realizar essas predições.

4. Materiais e Métodos

Para alcançar o objetivo geral desta pesquisa, optou-se por compilar um conjunto de dados a partir dos principais portais de compra e venda de carros usados no Brasil. Foram consideradas quatro fontes de dados distintas: Facebook Marketplace, OLX, Webmotors e iCarros.

Inicialmente, o Facebook Marketplace foi descartado como fonte viável de dados devido ao alto número de anúncios fraudulentos ou imprecisos, que poderiam introduzir ruídos no conjunto de dados. Com base em experiência empírica, foi observado que alguns anúncios incluíam no preço do veículo o valor da entrada de um possível financiamento ou apresentavam valores muito abaixo do real, presumivelmente para atrair compradores.

Esses anúncios, considerados *outliers* em uma análise exploratória de dados, poderiam comprometer o treinamento dos modelos de previsão. Embora problemas semelhantes também possam ocorrer na OLX, a base de dados da OLX já foi utilizada em trabalhos relacionados, conforme demonstrado na sessão anterior. Além disso, os dados são mais facilmente coletados na plataforma da OLX, o que motivou ainda mais seu uso.

A segunda plataforma analisada foi a OLX Brasil, que possui aproximadamente 800.000 carros anunciados em todo o país. A extração de dados desta plataforma não apresentou desafios significativos, como bloqueios ou sistemas de detecção de acessos automatizados. Tanto a interface de busca quanto os anúncios em si mostraram-se adequados para a realização da raspagem de dados.

Após desenvolvermos as soluções de extração de dados para a OLX, avaliou-se também a Webmotors e a iCarros, que possuem menos anúncios, mas adotam critérios mais rigorosos na publicação, o que poderia melhorar a qualidade do nosso conjunto de dados. Contudo, houve dificuldades de acesso devido às medidas rigorosas dessas plataformas contra acessos automatizados. O acesso à Webmotors foi temporariamente bloqueado, enquanto o acesso à iCarros foi permanentemente negado.

Diante dessas limitações, a OLX tornou-se a única fonte de dados utilizada. Apesar da intenção de incluir outros portais para produzir um conjunto de dados mais abrangente, os aproximadamente 800.000 anúncios disponíveis na OLX foram suficientes para treinar os modelos de previsão.

4.1. Conjunto de Dados

O conjunto de dados que produzimos ao aplicar raspagem de dados na OLX é composto por diversas características valiosas que permitem uma análise abrangente do mercado de carros usados. Cada carro é descrito por um conjunto de atributos que fornecem uma visão holística de suas características e funcionalidades.

É armazenado o **título do anúncio**, a **descrição** feita pelo anunciante, a **unidade federativa** em que o anúncio foi feito e a **URL** do anúncio. Embora essas informações não sejam utilizadas diretamente no treinamento de modelos de regressão, elas são armazenadas para viabilizar análises futuras, caso sejam necessárias.

O **preço** anunciado do carro é um dos atributos mais importantes para a decisão de compra. Permite identificar a faixa de preço dos carros, analisar a depreciação ao longo do tempo e comparar preços entre diferentes modelos e marcas. É a variável alvo a ser modelada, a qual o modelo aprenderá a prever com base nas demais características do carro, e se trata de um dado numérico contínuo.

Captura-se também a identificação específica do **modelo** do carro, como ”Gol 1.0 2023”. É uma variável categórica, e devido ao altíssimo número de modelos presentes no conjunto de dados, não foi feito um pré-tratamento para transformá-los em dados numéricos, já que isso poderia dificultar analises futuras. Um pesquisador utilizando o presente conjunto de dados acharia mais conveniente. É um indicador importante do valor do carro, pois diferentes modelos dentro da mesma marca geralmente possuem preços distintos. Outra característica categórica que não passou por pré-tratamento pelo mesmo motivo do modelo é a **marca** do carro. É um indicador da reputação e do valor geral da marca, influenciando significativamente o preço do carro.

Para as características categóricas **tipo de carro**, **cilindrada**, **combustível**, **GNV**, **câmbio**, **cor**, **portas** e **direção**, foi realizado um pré-processamento dos dados, no qual cada valor categórico foi convertido para um mapeamento numérico, conforme apresentado na Tabela 1. Os dados originais estavam em formato textual; por exemplo, o tipo de carro podia assumir valores como ”Hatch”ou ”Picape”. Durante o armazenamento, esses valores foram representados numericamente: ”Hatch”foi codificado como o número 5, enquanto ”Picape”foi associado ao número 7.

O significado dessas variáveis é bastante claro. O tipo de carro refere-se à categoria do automóvel, conforme mencionado anteriormente. A cilindrada relaciona-se à capacidade volumétrica do motor, sendo um indicativo de desempenho e custo operacional. Veículos com maior cilindrada, em geral, apresentam valores de mercado mais elevados devido à maior potência. GNV indica se o carro possui kit de gás natural veicular instalado. Um carro com GNV instalado normalmente é indicativo de que é um carro que foi utilizado para Táxi ou Uber, indicando que foi um automóvel utilizado de forma exaustiva e talvez tenha desgastado um número elevado de peças. Por isso, é esperado que um carro com o kit GNV instalado desvalorize. Essa relação pode ser evidenciada na figura 7. O câmbio representa o tipo de transmissão do veículo, como manual, automático ou automatizado, enquanto a direção indica o sistema de direção utilizado, como mecânica, hidráulica ou elétrica. Por fim, as variáveis cor e portas são autoexplicativas, representando, respectivamente, a tonalidade externa do automóvel e a quantidade de portas disponíveis.

O conjunto de dados conta também com os opcionais, incluindo **airbags**, **alarme**, **ar-condicionado**, **trava elétrica**, **vidro elétrico**, **som**, **sensor de ré**, **câmera de ré** e **blindagem**, todos representados como dados categóricos, indicados por 1 para verdadeiro e 0 para falso. Se verdadeiro, indica que o carro possui esse opcional, se falso, não o possui. Itens como alarme, ar-condicionado, travas elétricas e vidros elétricos são frequentemente encontrados em veículos mais novos, mas talvez possam representar um diferencial em modelos populares mais antigos, especialmente o vidro elétrico. Já sistemas como som, sensores de ré, câmeras de ré e blindagem agregam valor de acordo com o segmento e a idade do veículo.

A decisão de armazenar os dados categóricos como números mapeados às suas respectivas categorias, em vez de mantê-los com os nomes originais, foi motivada por benefícios específicos para os algoritmos de regressão que serão utilizados nesse trabalho. Algoritmos de regressão exigem que as entradas sejam numéricas, e ao converter categorias em números, permitimos que o modelo de regressão interprete essas variáveis de forma compatível com suas operações matemáticas.

Este conjunto de dados rico e detalhado oferece uma base sólida para a análise e modelagem preditiva do valor de mercado de carros usados. As variáveis selecionadas, tanto numéricas quanto categóricas, fornecem uma ampla gama de informações que permitem capturar as nuances do mercado automobilístico, desde aspectos técnicos e mecânicos até preferências estéticas e de conforto.

Tabela 1. Mapeamento numérico das variáveis categóricas, com os índices representando os valores no conjunto de dados. Por exemplo, "Tipo de Carro: Antigo"foi mapeado como 1, e "Cor: Laranja"como 9.

	Tipo de Carro	Cilindrada	Combustível	Câmbio	Cor	Portas	Direção
1	Antigo	1.0	Gasolina	Manual	Preto	2 Portas	Hidráulica
2	Buggy	1.2	Álcool	Automático	Branco	4 Portas	Elétrica
3	Caminhão Leve	1.3	Flex	Semi-Automático	Prata		Mecânica
4	Conversível	1.4	Diesel	Semi-automático	Vermelho		Assistida
5	Hatch	1.5	Híbrido		Cinza		
6	Passeio	1.6	Elétrico		Azul		
7	Pick-up	1.7	Gás Natural		Amarelo		
8	Sedã	1.8			Verde		
9	SUV	1.9			Laranja		
10	Van/Utilitário	2.0-2.9			Outra		
11		3.0-3.9					
12		4.0 ou mais					

4.2. Técnicas de Raspagem de Dados

Aqui, será descrito o processo de coleta de dados da plataforma OLX utilizando a biblioteca Scrapy em Python. O objetivo principal é extrair informações relevantes de anúncios de carros e armazená-las em um formato estruturado para análise e modelagem subsequentes.

A coleta de dados é realizada por meio de duas *spiders* do Scrapy: *OlxCars* e *CarPageOLX*. A função da *OlxCars* é percorrer as páginas de busca da OLX e extrair links de anúncios de carros. Devido à limitação da paginação da OLX em 99 páginas, foi necessário implementar uma estratégia para capturar todos os anúncios. Os estados brasileiros foram divididos em três grupos com base na densidade de anúncios (baixa, média e alta). No caso de estados com baixa densidade, a *spider* coleta

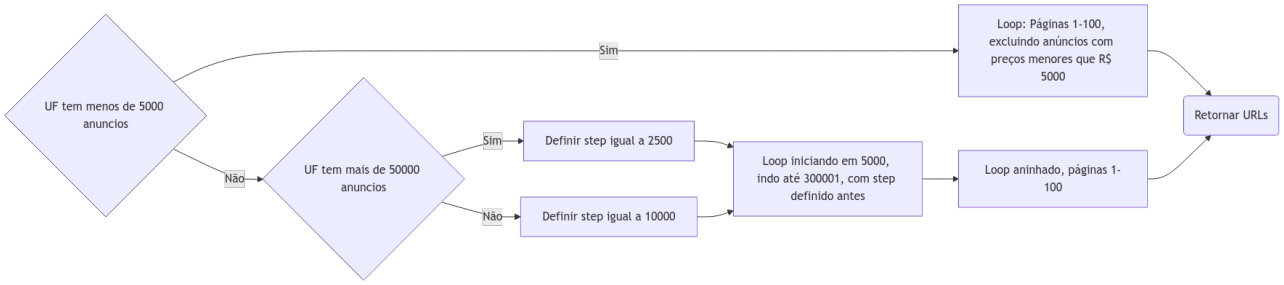

Figura 1. Representação visual do funcionamento da spider OlxCars. Fonte: elaborada pelos autores

todos os anúncios com preço mínimo de R$ 5.000, pois o número total de anúncios é menor que o limite da paginação da OLX. Evitou-se anúncios com preços menores que esse para não coletar ruído. Para estados de densidade média, a *spider* inicia um *loop* com passo de 10.000, definindo um preço mínimo de R$ 5.000 e o critério adicional do passo para limitar o número de anúncios por requisição. Para estados com alta densidade, a *spider* segue a mesma lógica dos estados com densidade média, porém com um passo de 2.500 para lidar com o maior volume de anúncios. Dessa forma, foi possível capturar todos os anúncios de carros com valor acima de R$ 5.000 presentes na plataforma. A figura 1 demonstra o funcionamento dessa spider de forma mais visual.

A *spider CarPageOLX* é encarregada de extrair e processar os dados de cada anúncio de carro. Toda a informação relevante está encapsulada em um objeto *JSON* presente no código fonte *HTML* de cada página. A *spider* então processa esse objeto *JSON*, realizando tratamentos como conversão de colunas categóricas em formatos numéricos. Isso não apenas reduz o espaço ocupado pelo arquivo *CSV* no disco, mas também economiza tempo durante a etapa de análise exploratória de dados. Por fim, o dicionário resultante é agregado ao já citado arquivo *CSV* para possibilitar a análise e o treinamento dos modelos.

O código utilizado para essa extração de dados está disponível no Github e pode ser acessado através do seguinte link: https://encurtador.com.br/0GkWj. Utilizando os *scripts* disponíveis, qualquer um pode reproduzir o mesmo processo de coleta que foi realizado no presente trabalho.

5. Resultados e Discussões

5.1. Análise Exploratória de Dados

Para melhorar o ajuste dos modelos utilizados, foram realizadas operações de tratamento no nosso conjunto de dados. Primeiramente, decidimos filtrar palavras do título e descrição que indiquem que não se trata de uma venda comum de um carro, o que poderia introduzir um ruído nos dados. Para isso, utilizamos a lista de palavras:

'repasse', 'sucata', 'retirada', 'peças', 'quebrado','sinistrado', 'leilão', 'recuperado', 'avariado', 'desmanche', 'problema mecânico', 'defeito', 'sem documentação', 'financiado com dívida', 'recuperável', 'emergência', 'motor fundido', 'lataria danificada', 'sem motor', 'queima de estoque', 'quitado com alienação'.

Essas palavras indicam que o veículo em questão será vendido significativamente abaixo de seu valor de mercado. Isso se deve ao fator que veículos presentes com tais palavras-chave possuem uma alta probabilidade de incorporarem dados, principalmente em relação ao preço, discrepantes com a média de tal veículo. Com base nisso, foram excluídas as colunas "Url", "Título"e "Descrição", pois

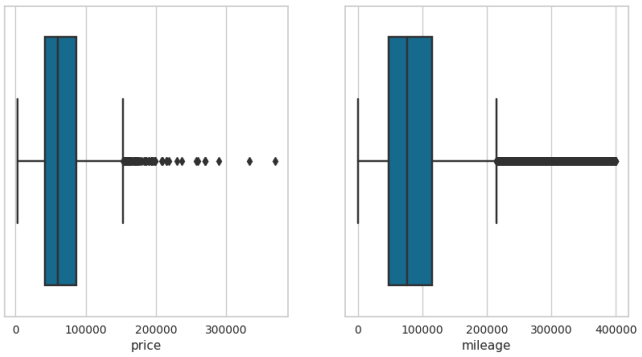

Figura 2. Boxplot das variáveis preço e quilometragem antes da função IQR. Fonte: elaborada pelos autores

não são essenciais para a preparação do conjunto de dados para a etapa de aprendizado de máquina.

Em seguida, foram removidos veículos com ano de fabricação igual ou inferior a 2023 e quilometragem de até 1000 km. Dado que veículos com esses traços normalmente apresentam dados consonantes com a venda de um carro novo, ou por agências particulares, que não se enquadrariam com os critérios da construção de nossa base de dados. Ademais, carros com quilometragem muito baixa ou praticamente nula podem ser considerados atípicos e, portanto, influenciar negativamente o ajuste do modelo, justificando sua exclusão.

Além disso, foram excluídos veículos cujo preço fosse igual ou inferior a R$ 3.000 ou superior a R$ 400.000, como parte de um tratamento preliminar de *outliers*. Da mesma forma, veículos com quilometragem superior a 400000 km também foram removidos. Isso foi feito sendo considerado a média de preços-alvo do mercado de usados algo como em média R$ 80.000, e a média de quilometragem anual algo como 13500 km/ano, é considerável a disparidade e ruído que esses artigos trariam à base, e portanto, a retirada dos mesmos foi feita como parte de um pré tratamento dos dados extraídos.

Após esses tratamentos, modelos e marcas com pouca representatividade dentro do conjunto de dados foram eliminados. Apenas as 20 maiores marcas e os 500 modelos mais frequentes foram mantidos no conjunto de dados. As figuras 5 e 6 apresentam a distribuição das marcas e modelos mais representativos. As colunas categóricas relacionadas a opcionais foram consolidadas em uma única coluna numérica, representando a soma do número de opcionais que cada veículo possui, facilitando o processamento pelo modelo de aprendizado de máquina.

Também foi realizada uma análise mais aprofundada de *outliers* no conjunto de dados. A figura 2 destaca a alta incidência de veículos com quilometragem superior a 200000 km, bem como alguns veículos com preço acima de R$ 150.000. Para filtrar esses *outliers*, foi desenvolvida uma função baseada no Intervalo Interquartil (IQR).

5.1.1. Função Baseada em IQR

Foi criada uma função para calcular o Intervalo Interquartil (IQR) que recebe como entrada um *dataframe* e o nome de uma coluna. A função calcula o primeiro (Q1) e o terceiro quartil (Q3) da coluna especificada, e o IQR é determinado pela diferença entre Q3 e Q1. A partir disso, os limites inferior e superior são definidos como:

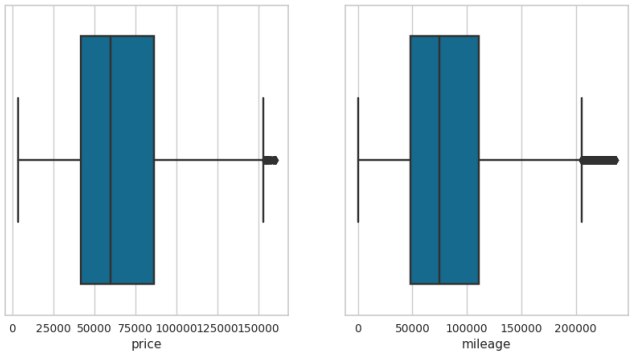

Figura 3. Boxplot das variáveis preço e quilometragem após função IQR. Fonte: elaborada pelos autores

$$LimiteInferior = Q1 - 1,5 \cdot IQR \tag{5}$$

$$LimiteSuperior = Q3 + 1,5 \cdot IQR \tag{6}$$

Estes limites representam o intervalo esperado dentro do qual a maioria dos pontos de dados reside. A função retorna um novo *dataframe* contendo apenas linhas onde o valor na coluna especificada está dentro dos limites calculados. O código percorre os valores únicos da coluna ”Modelo”e, para cada modelo, cria um subconjunto do *dataframe* contendo apenas as linhas correspondentes a esse modelo. Em seguida, aplica a função IQR nas colunas ”Preço”e ”Quilometragem”. Os resultados filtrados pela função IQR são concatenados em um novo *dataframe*. A figura 3 mostra o *dataframe* após aplicação da função.

5.1.2. Correlação entre as variáveis

Na Figura 4, é apresentado um mapa de calor que ilustra a correlação entre as diversas variáveis do conjunto de dados. Valores próximos a 1 indicam uma correlação positiva forte, enquanto valores próximos a -1 apontam para uma correlação negativa forte. Correlações próximas a 0 sugerem pouca ou nenhuma relação entre as variáveis. Uma correlação positiva indica que a presença ou o aumento dos valores dessa variável está associado a um aumento no preço, e vice-versa.

Observou-se que as variáveis com maior influência sobre o preço de um veículo são, em ordem de importância: o tipo de transmissão, o ano de fabricação e a cilindrada do motor. Por outro lado, apenas duas variáveis apresentaram uma correlação negativa significativa: a quilometragem e a presença de um kit de Gás Natural Veicular (GNV). Conforme mencionado anteriormente, a instalação de um kit GNV geralmente indica que o veículo foi utilizado para trabalho, o que pode ter submetido o carro a um desgaste mais intenso do que sua quilometragem poderia sugerir.

Destaca-se que variáveis como cor, número de portas e tipo de direção não apresentaram correlação significativa com o preço do veículo. As demais variáveis exibiram correlações positivas, embora de forma modesta.

As únicas duas variáveis categóricas remanescentes após o processo de filtragem e tratamento dos dados foram ”Modelo”e ”Marca”. No entanto, devido à grande diversidade de modelos de carros

presentes no conjunto de dados, que abrange mais de 800 modelos distintos, optou-se por excluir a coluna referente ao "Modelo". A aplicação de uma técnica como o *one-hot encoding* nesse contexto resultaria em um *dataframe* extremamente volumoso e complexo, com mais de 800 variáveis distintas, o que poderia comprometer a eficiência e a eficácia do modelo de aprendizado de máquina, além de aumentar significativamente a demanda por recursos computacionais.

Por outro lado, após o tratamento anterior, foram identificadas 20 marcas de veículos diferentes dentro do conjunto de dados, um número que se mostrou adequado para a aplicação do *one-hot encoding*. Essa técnica foi utilizada para transformar as categorias de marcas em variáveis numéricas binárias, o que permite ao modelo de aprendizado de máquina interpretar e processar essas informações de forma mais eficaz.

Além disso, foi realizado o *label encoding* nessas novas variáveis categóricas, convertendo-as em valores numéricos que mantêm a representatividade das categorias originais. Esse passo é fundamental para garantir que o modelo consiga lidar com os dados categóricos de maneira eficiente, transformando-os em uma forma que o algoritmo consiga compreender e utilizar durante o processo de aprendizado.

Subsequentemente, foi aplicada uma função de otimização que tenta converter as colunas numéricas do *dataframe* para o menor tipo de dado numérico possível, garantindo que esses valores sejam representados de maneira precisa, sem perda de informação. Essa etapa é especialmente relevante dado o tamanho do conjunto de dados, uma vez que a otimização do uso de memória é crucial para o desempenho durante a fase de treinamento do modelo. Ao minimizar o espaço ocupado pelos dados, possibilita-se um processamento mais rápido e eficiente, o que é essencial para a construção de modelos de aprendizado de máquina robustos e escaláveis.

5.2. Avaliação e Métricas

Foram avaliados nove modelos de regressão: Regressão Linear, Regressor de Árvore de Decisão, Floresta Aleatória, Regressor *Bagging*, Regressor *Extra Trees*, Regressão *Adaboost*, *Ridge Regressor*, Gradiente Descendente Estocástico e SVM. A seleção desses modelos fundamentou-se na diversidade das abordagens que cada um oferece para o problema de regressão, possibilitando uma análise comparativa abrangente das diferentes técnicas de aprendizado de máquina. Foi utilizada uma proporção de 20% do conjunto de dados para os testes. Não foi utilizada uma semente para reprodução desses resultados. Foram utilizados os modelos disponíveis na biblioteca Scikit Learn, na versão 1.2.2 e hiperparâmetros padrão foram utilizados para todos os modelos. As métricas obtidas estão apresentadas na Tabela 2.

Conforme esperado, a Regressão Linear apresentou desempenho limitado. Esse modelo depende de uma suposição rígida: a existência de uma relação linear entre as variáveis independentes e a variável alvo. Como também utilizamos o one-hot enconding para features categóricas como modelo, que ocasiona, exceto em modelos de regressão em árvore, uma certa desapropriação da relação entre variáveis já que o valor indicado na nova coluna modelo não é ordenal. Outro fator que indica ter contribuido para isso é a assimetria de dados como ano e quilometragem enviesada à esquerda, indicando uma distribuição majoritária em carros mais novos, o que pode afetar modelos mais antigos e impactar substancialmente a acurácia. Em cenários reais, especialmente em conjuntos de dados complexos e com um grande número de variáveis, essa suposição raramente se verifica. Como resultado, o modelo linear não conseguiu capturar adequadamente as nuances e interações presentes nos dados.

Em contraste, a Floresta Aleatória destacou-se como o modelo de melhor desempenho. Tendo em vista que se adaptou melhor a grande complexidade de dimensões da tabela, assim como des-

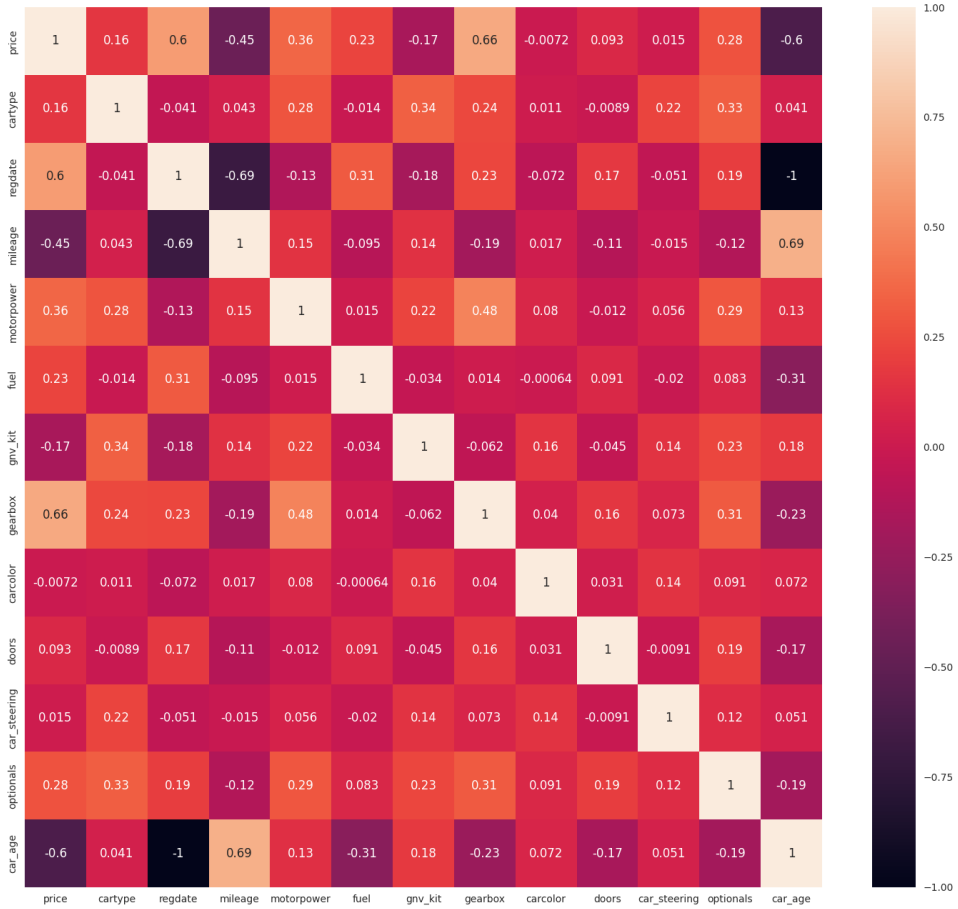

Figura 4. Mapa de calor de correlação entre as variáveis usadas no treinamento dos modelos. Fonte: elaborada pelos autores

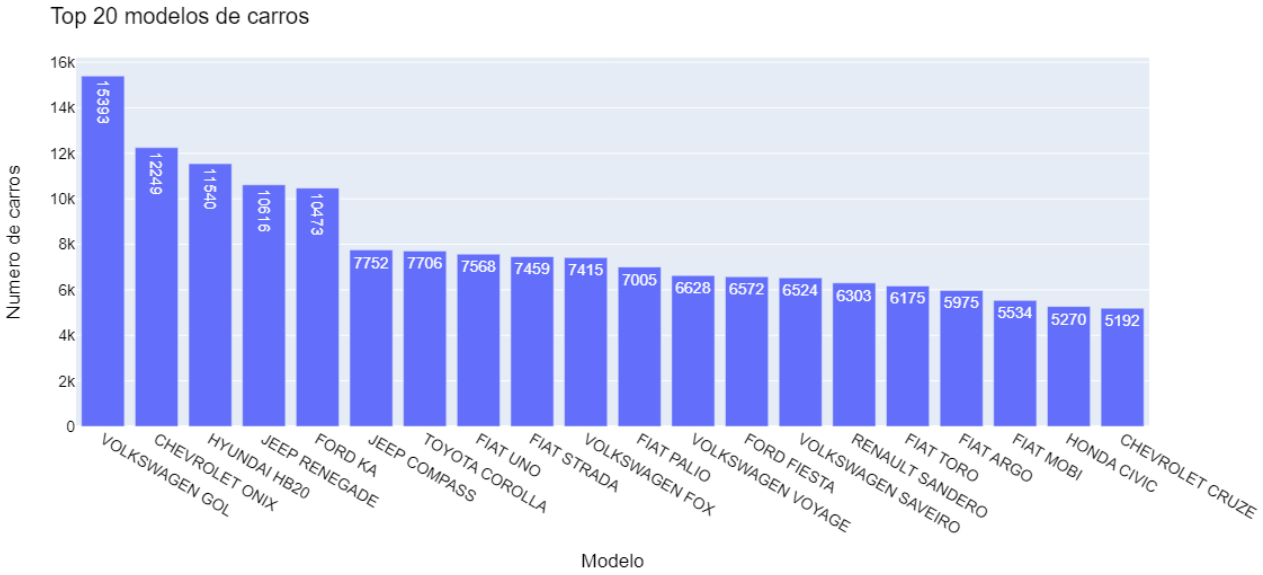

Figura 5. Gráfico de barras mostrando quantitativo dos 10 modelos com maior número de anúncios. Fonte: elaborada pelos autores

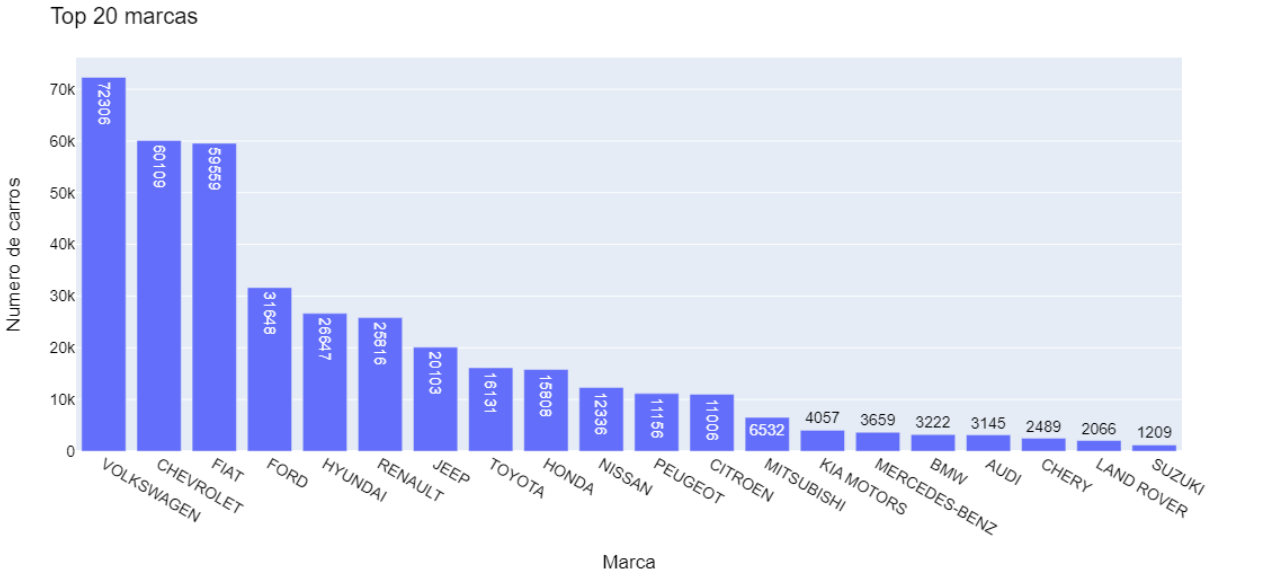

Figura 6. Gráfico de barras mostrando quantitativo das 10 marcas com maior número de anúncios. Fonte: elaborada pelos autores

Tabela 2. Comparação das métricas dos modelos utilizados

Modelo	R2	MSE	MAE	RMSE
Linear Regression	0,7553	231778907,11	11185,16	15223,96
Decision Tree Regression	0,9045	9041447,30	6133,26	9508,86
Random Forest	0,9434	54074407,04	4855,27	7320,21
Bagging Regressor	0,9393	57541349,31	5034,07	7585,60
Extra Trees Regressor	0,9382	58578644,20	5019,38	7653,67
Adaboost Regression	0,7552	231839225,73	11183,09	15226,27
Ridge Regression	0,7553	231772513,38	11185,18	15224,08
Stochastic Gradient Descent	0,7545	232502277,21	11210,94	15248,03
Support Vector Machines	0,1445	810346207,91	21976,06	28466,58

considerou o fator ordenal da tabela modelo pode alcançar um certo nível de acurácia. A eficácia desse modelo decorre da combinação de múltiplas árvores de decisão, o que lhe permite capturar de maneira mais precisa as interações complexas e a variabilidade intrínseca dos dados. A robustez e a capacidade de generalização da Floresta Aleatória tornam-na particularmente eficaz no contexto deste estudo. O desempenho desse modelo foi seguido de perto pelo Regressor *Bagging* e pelo Regressor *Extra Trees*, enquanto o Regressor de Árvore de Decisão apresentou uma performance ligeiramente inferior.

Os demais modelos apresentaram desempenho insatisfatório, com destaque negativo para o SVM, que obteve um R² de 0,14 e um MAE de 21.976,06, configurando-se como o pior entre os testados. Esse resultado é atribuído a uma série de fatores também aplicáveis a outros modelos de regressão. Contudo, no caso específico do SVM, destacam-se os seguintes dilemas: o kernel da matriz exige uma alocação de memória considerável devido à escala quadrática com a qual cresce em relação à cardinalidade de uma característica; além disso, assim como as regressões lineares, o modelo depende de uma certa ordenação das características, o que é inviável para colunas como a variável 'modelo'. Conforme mencionado anteriormente, em relação às limitações decorrentes do enviesamento dos dados, os hiperplanos também amplificam esse viés de forma quadrática, desbalanceando a proporção entre vetores de suporte positivos e negativos, o que compromete a predição (BATUWITA; PALADE, 2013)

6. Conclusões

Este trabalho demonstrou o potencial significativo da extração e análise de dados provenientes de plataformas online dedicadas à venda de automóveis usados. O processo de coleta permitiu a obtenção de métricas tangíveis e reais do mercado, utilizando técnicas de análise exploratória de dados para filtrar, quantificar e reclassificar os anúncios de acordo com diversos critérios.

A utilização de técnicas de aprendizado de máquina na precificação de carros usados foi particularmente relevante. A aplicação de métodos de regressão, com ênfase no modelo de Floresta Aleatória, destacou-se pela sua capacidade de fornecer previsões precisas. Com um coeficiente de determinação (R²) de 0,9434 e um erro absoluto médio (MAE) de 4855,27, o modelo demonstrou resultados promissores. Esses resultados evidenciam a eficácia do modelo em capturar as complexidades e variabilidades inerentes ao mercado de veículos usados, o que representa um avanço significativo na abordagem da precificação desse segmento.

Além disso, o estudo sublinha a importância de integrar técnicas de aprendizado de máquina

com dados extraídos de fontes diversas para aprimorar a precisão das estimativas de preços e para compreender melhor as dinâmicas do mercado. O trabalho realizado oferece um ponto de partida para o desenvolvimento de modelos mais sofisticados e para a realização de análises mais aprofundadas no futuro. O embasamento fornecido por este estudo abre caminho para novos trabalhos que busquem explorar mais detalhadamente as nuances do mercado de carros usados e suas variações regionais e temporais.

7. Trabalhos Futuros

Baseando-se nos resultados obtidos, foi possível desenvolver um conjunto de dados e um método de produção que podem ser utilizados tanto para o treinamento de modelos em fases posteriores quanto para análises preliminares, utilizando dados coletados por técnicas de raspagem de dados. Esse conjunto abrange uma ampla variedade de variáveis categóricas relacionadas à venda de automóveis usados em plataformas online, bem como uma descrição detalhada dos veículos, cujas características e valores são informações de difícil acesso em contextos acadêmicos ou pessoais.

No entanto, este trabalho apresenta algumas limitações que podem comprometer sua validade, particularmente na etapa de treinamento dos modelos de regressão. Uma estratégia experimental mais robusta poderia ter contribuído para aumentar a confiabilidade dos resultados. Além disso, o uso de hiperparâmetros padrão poderia ser substituído por técnicas de otimização, de modo a aprimorar o desempenho dos modelos. Recomenda-se que estudos futuros dentro do mesmo escopo abordem essas limitações para obter resultados mais sólidos e precisos.

Diante desse cenário, abre-se uma gama de possibilidades para trabalhos futuros. Uma delas é a utilização do modelo desenvolvido para gerar estimativas de faixas de preço de veículos usados, focadas em categorias específicas de automóveis. Embora as métricas dos modelos tenham sido promissoras, observou-se que as estimativas diretas de preços individuais não atingiram o desempenho esperado. Por isso, sugere-se trabalhar com faixas de preço, estratégia que poderia ser mais viável. Além disso, a variável de descrição dos anúncios pode ser explorada para identificar padrões linguísticos, destacando as palavras mais frequentes em contextos de compra e venda online.

Outra linha de investigação seria o uso de dados geográficos para mapear perfis de consumo em diferentes estados do Brasil, analisando preferências de marcas e tipos de veículos, com foco nos hábitos regionais de compra. Análises comparativas entre veículos novos e seus equivalentes usados também são sugeridas, levando em conta mudanças em características como opcionais, câmbio e motorização ao longo dos anos. Há também a possibilidade de realizar uma análise aprofundada sobre o impacto de carros importados e nacionais no mercado de usados, explorando diferenças de preço, demanda e preferências regionais, o que pode oferecer *insights* sobre as dinâmicas do mercado e revelar oportunidades para segmentos específicos. Incluir carros elétricos nessa análise, conforme os dados se tornem disponíveis nas plataformas de venda, pode trazer informações importantes para o futuro do mercado automotivo.

Além dessas possibilidades, futuros estudos poderiam se concentrar na integração de algoritmos mais complexos, como *deep learning*, para aprimorar a precisão das previsões e identificar padrões mais sutis nas características dos veículos. A inclusão de dados adicionais, como histórico de manutenção ou sinistros, também poderia enriquecer o conjunto de variáveis e gerar estimativas ainda mais robustas. Por fim, a criação de ferramentas interativas ou *APIs* para acesso aos modelos preditivos poderia democratizar o uso dessas tecnologias no mercado automotivo, oferecendo uma aplicação prática dos resultados desta pesquisa.

Referências

AHMED, M. Waqar. Understanding Mean Absolute Error (MAE) in Regression: A Practical Guide. Medium. 2023. Disponível em: https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df. Acesso em: 03/06/2024.

AZNAR, Pablo. What is the difference between Extra Trees and Random Forest?. Junho de 2020. Disponível em: https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/. Acesso em: 07/06/2024.

BATUWITA, R.; PALADE, V. Class imbalance learning methods for support vector machines. In: HE, H.; MA, Y. (Eds.). Imbalanced learning. [S.l.]: Wiley, 2013.

BREIMAN, Leo. Bagging Predictors. Machine Learning, v. 24, p. 123-140, 1996.

BREIMAN, Leo. Random Forests. Machine Learning, v. 45, p. 5-32, 2001.

Bukvic et al. Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. Sustainability, v. 14, 2022.

CHICCO, Davide et al. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 2021.

DIAS, Willamy Mamede Da Silva. MACHINE LEARNING E A PREVISÃO DE PREÇOS DE TERRENOS EM BRASÍLIA. Tese (Mestrado Profissional em Economia) - Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa - IDP. Brasília. 2023.

DREHMER, Vitória. Venda de carros usados cresce 14% e é 7 vezes maior que a de novos. AutoEsporte. 2024. Disponível em: https://autoesporte.globo.com/carros/usados-e-seminovos/noticia/2024/02/venda-de-carros-usados-cresce-14percent-e-e-7-vezes-maior-que-a-de-novos.ghtml. Acesso em: 13 de março de 2024.

GEURTS et al. Extremely randomized trees. Machine Learning, v. 63, p. 3-42, 2006.

IBM. What is Machine Learning?. 2023. Disponível em: https://web.archive.org/web/20231227153910/https://www.ibm.com/topics/machine-learning. Acesso em: 25 de setembro de 2024.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. Science, v. 349, n. 6245, p. 255–260, 2015.

LOH, Wei-Yin. Fifty Years of Classification and Regression Trees. International Statistical Review, v. 82, n. 3, p. 329-348, 2014.

MAGALHÃES, Tomás Silva de. Sistema de Previsão de Preço de Carros Usados através de Machine Learning. Tese (Mestrado em Engenharia de Inteligência Artificial) - Departamento de

Engenharia e Informática, Instituto Superior de Engenharia do Porto.

MITCHELL, Tom M. Machine Learning. McGraw-Hill Science/Engineering/Math. 1997.

NILSSON, N. J. Introduction to Machine Learning. Draft of Incomplete Notes, 2015. Disponível em: https://web.archive.org/web/20190816182600/http://ai.stanford.edu/people/nilsson/mlbook.html. Acesso em: 12 de novembro de 2023.

OLX Brasil. Disponível em: https://www.olx.com.br/. Acesso em: 12 de novembro de 2023.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

RAJ, Ashwin. A Quick and Dirty Guide to Random Forest Regression. Towards Data Science. Junho de 2020. Disponível em: https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8. Acesso em: 06/05/2024.

REITZ, Kenneth. HTML Scraping. The Hitchhiker's Guide to Python. 2016. Disponível em: https://docs.python-guide.org/scenarios/scrape/#web-scraping. Acesso em: 25 de setembro de 2024.

SALÁRIO MÍNIMO - Tabela de Preço. Disponível em: https://www.guiatrabalhista.com.br/guia/salario_minimo.htm. Acesso em: 10 de agosto de 2024.

Sci-Kit Learn. Ensembles: Gradient boosting, random forests, bagging, voting, Stacking. Disponível em: https://scikit-learn.org/stable/modules/ensemble.html. Acesso em: 28 de novembro de 2024.

Scrapy documentation. Disponível em: https://docs.scrapy.org/en/latest/. Acesso em: 28 de novembro de 2024.

SILVA, Thiago César de Miranda. Uso de Machine Learning para Previsão de Valores de Apartamentos no Município do Recife. Tese (Bacharelado em Sistemas de Informação) - Universidade Federal Rural de Pernambuco - UFPE. Recife. 2023.

Tabela FIPE: Entenda o que é e como funciona. Quatro Rodas. Disponível em: https://quatrorodas.abril.com.br/tabela-fipe. Acesso em: 12 de novembro de 2023.

Tabela Fipe Kwid. Disponível em: https://tabelacarros.com/anos_modelos/carros/renault/kwid. Acesso em: 10 de agosto de 2024.

ZHAO, Bo. Web Scraping. Encyclopedia of Big Data, p. 1-3, 2017.