Uso de modelos de regressão não-linear na precificação de venda de veículos usados

Douglas Farias Cordeiro, Renata Moreira Limiro, Anelise Souza Rocha, Núbia Rosa da Silva

Resumo A precificação de produtos é um tema de considerável complexidade, que leva em conta uma série de fatos, os quais vão desde questões relacionadas à própria manufatura em si, quanto fatores tributários e logísticos. No contexto de produtos usados, existe ainda impactos que remetem a questões regionais, de manutenção, assim como interesse enquanto objeto de apreciação ou coleção. No mercado de veículos usados, uma das principais referências de precificação é a tabela disponibilizada pela Fundação Instituto de Pesquisas Econômicas (FIPE). Entretanto, tal tabela não leva em conta algumas especificidades de veículos, principalmente em termos de regionalidade. Neste sentido, se torna interessante avaliar o uso de soluções alternativas, como é o caso da aplicação de métodos computacionais inteligentes. Neste artigo, é proposto a avaliação de modelos baseados em regressão linear e não-linear, considerando um conjunto de dados de anúncios extraídos em um portal de vendas online. Os resultados obtidos mostram que métodos de regressão não-linear apresentam ganhos consideráveis em termos de acurácia, se destacando uma interessante solução para o problema.

1 Introdução

A precificação de produtos é um tema de considerável complexidade, que leva em conta uma série de aspectos, os quais permeiam desde questões relacionadas à demanda e oferta, até fatores ligados à qualidade, marca e regionalidade. Neste contexto, a determinação de preços de vendas de veículos usados se destaca como um problema de grande interesse, uma vez que este tipo de mercado sofre uma série de impactos, onde se destaca a variação entre a desvalorização em face do tempo de uso, para uma determinada faixa de tempo, e a posterior valorização, quando um veículo alcança uma idade específica de fabricação.

Douglas Farias Cordeiro
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: cordeiro@ufg.br

Renata Moreira Limiro
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: renatamlimiro@gmail.com

Anelise Souza Rocha
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: anelisesrocha@gmail.com

Núbia Rosa da Silva
Instituto de Biotecnologia - Universidade Federal de Catalão, Catalão, Goiás, Brasil.
e-mail: nubia@ufg.br

Anais do XV Encontro Anual de Ciência da Computação (EnAComp 2020). ISSN: 2178-6992.
Catalão, Goiás, Brasil. 25 a 27 de Novembro de 2020.

Neste sentido, o preço e seus aspectos de volatilidade sempre foram alvo dos olhares e questionamentos da ciência. Vários são os conceitos que procuram definir, simular ou ainda prever preços ideais considerando diferentes contextos e cenários para sua aplicação. Segundo Smith (2017), o fator determinante do preço seria o custo relacionado à mão de obra dedicada em produzir aquele bem específico. Ricardo (2018), no entanto, trouxe uma perspectiva mais ampliada para o conceito de produção, propondo que o custo de máquinas e equipamentos utilizados para a produção do bem também sejam considerados no cálculo de preço.

Segundo Marshall (2018), a definição do preço de um bem não está condicionada somente à visão da indústria, dos produtores ou ofertantes do bem, mas também à utilidade que este tem ao comprador. Nesse sentido, o conceito de elasticidade de preço explora a variação do preço em função da maximização dos lucros do ofertante do bem em relação a busca por máxima utilidade por parte do demandante. Desta maneira, o comprador estaria sempre buscando o menor preço enquanto o fornecedor visa maiores margens de lucro, chegando assim, a um ponto onde o preço do bem, desconsiderando fatores externos, tais como produtos concorrentes, atende as expectativas que tangem a demanda e a oferta, alcançando um equilíbrio parcial de mercado.

Outro aspecto que pode ser considerado na definição do preço de um bem, sendo ele produto, serviço ou experiência, é o hedonismo, ou seja, o quão agradável, satisfatório, ou prazeroso, é ao comprador adquirir aquele determinado item. Segundo Besanko, Dranove e Shanley (2012) os preços hedônicos podem ser considerados como preços sensíveis a variações decorrentes da conciliação de conjuntos de atributos. Como, por exemplo, ao se analisar o preço de um imóvel é possível observar o tamanho do terreno, quantidade de banheiros, a localização e outros atributos, os quais, existindo em diferente níveis, podem impactar o interesse do comprador e consequentemente o valor de mercado daquele bem. Neste contexto, Fávero, Belfiore e Lima (2008) sugere que, para uma melhor análise e entendimento de preços hedônicos e das características ou atributos de um bem que causam mais impacto em seu preço, são necessárias aplicações de métodos de regressão para que, assim, possam ser observadas as variações de preço em função de determinadas características.

No contexto de mercado de automóveis, no Brasil, a Fundação Instituto de Pesquisas Econômicas (FIPE) realiza continuamente o acompanhamento e divulgação dos preços médios dos veículos anunciados por vendedores através da tabela FIPE. Apesar de ser um referencial interessante de precificação de veículos usados, a tabela não leva em conta especificidades dos produtos, tais como as supracitadas questões de regionalidade. Neste sentido, a aplicação de modelos computacionais inteligentes para a precificação de veículos usados se torna consideravelmente relevante. Considerando os aspectos referentes aos atributos deste tipo de problema, onde o principal objetivo é se obter o preço de um produto, modelos de inferência baseados em regressão se tornam uma interessante alternativa, uma vez que possibilitam, a partir de um conjunto de valores independentes, se obter o valor alvo.

O presente artigo apresenta um estudo relacionado à aplicação e avaliação de modelos de regressão não-linear, especificamente polinomiais, para solução do problema da precificação de veículos usados. Para tanto, é constituída uma base de dados obtida a partir da extração de anúncios de veículos usados publicados em portais de classificados digitais. Os resultados alcançados demonstram a efetividade do uso de modelos de regressão não-linear em face de modelos lineares, com um ganho considerável em termos de acurácia.

2 Revisão bibliográfica

O mercado de carros usados no Brasil tem uma forte representação oriunda das altas tarifas e impostos na aquisição de veículos novos. Entretanto, a informalidade desta área de atuação, trouxe consigo a necessidade de profissionalizar a revenda destes, sendo possível verificar as necessidades dos usuários e a crescente do mercado. A FENAUTO (Federação Nacional das Associações dos Revendedores de Veículos Automotores) é a responsável por gerenciar e medir o mercado de seus associados.

Os números de 2019 apurados pela FENAUTO contam com cerca de 48 mil revendas em todo o Brasil, girando 380 bilhões de reais por ano e gerando 620 mil empregos diretos e indiretos. No primeiro semestre de 2020, um estudo realizado mostrou que para cada carro novo vendido, quatro usados são negociados, e que destes, 39% são com fabricação de 4 a 8 anos, e 15% de seminovos (menos de quatro anos de uso) (RODRIGUES, 2020). Por meio destes dados é possível verificar a potência mercadológica de usados e a necessidade de se especializar ainda mais seus processos de venda e o hedonismo de seus preços.

A precificação dos usados, na teoria baseia-se no uso da Tabela FIPE, entretanto, a utilização deste meio se distancia veemente dos valores praticados no mercado, pois o cálculo é feito pela média nacional como descrito pela própria fundação. Tal descrição deixa em aberto a possibilidade de negociação por parte do mercado, entretanto essa prática dificulta o processo de precificação e padronização.

Com a utilização de modelos estatísticos e computacionais pertinentes à análise de dados, é possível obter valores de forma preditiva a partir de dados existentes, ou seja, o modelo elucidará o caminho correto a seguir dos preços através das características correlacionadas. De acordo com Carmo (2014), a aplicação de técnicas de precificação hedônicas admite a valoração implícita nos atributos do bem. Em um estudo sobre a precificação de casas, Boye, Mireky-Gyimah e Okpoti (2017) utilizou os atributos da construção, aplicando regressão linear múltipla, e obtiveram um coeficiente de regressão (R^2) de 99%, validando assim o uso da regressão linear. Entretanto estudos relacionados por Rosa, Oliveira e Pinto (2019), também no ramo habitacional, finalizou a pesquisa com um coeficiente (R^2) de 42%, e na tentativa de melhora desse índice, com troca de previsores, e não avaliando o uso de outro tipo de modelo.

Regressando ao ramo de carros, Ozgur et al. (2016) buscou predizer o preço de veículos de uma marca específica, aplicando regressão linear e realizando testes do modelo também apenas alterando previsores, a fim de diminuir a quantidade de *outliers*, os quais se mantiveram em todos os testes, a pesquisa não cogitou a mudança de modelo, uma vez que o coeficiente R^2 estava satisfatório, variando de 76% a 81%. Segundo Chernick e Friis (2003), os *outliers* podem influenciar não só no coeficiente R^2, como no interceptor, pois a linha de melhor ajuste é feita pela soma do quadrado dos resíduos, ou seja, como um número maior de extremos a magnitude aumenta. Neste contexto, e corroborando com a proposta deste estudo, testar qual o modelo mais propício para os dados se faz de grande valia para eficácia dos resultados uma vez que a acurácia, as avaliações de desvio padrão é que fazem a confiabilidade do modelo, principalmente em um ambiente de preços hedônicos, como o mercado de veículos.

3 Metodologia

A necessidade de geração de informações úteis e que agreguem valor aos negócios é algo fundamental e de grande interesse. Neste sentido, a aplicação de soluções computacionais inteligentes, as quais possam lidar com grandes volumes de dados, se torna essencial. O processo denominado de Descoberta do Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases - KDD*), proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996) consiste em um conjunto ordenado de passos compostos por: seleção, pré-processamento, transformação, mineração e interpretação de dados. Uma das principais vantagens do uso deste processo na geração de informação se refere ao fato de ser interativo e iterativo, permitindo alterações e avaliações ao longo da realização de todas as etapas.

No âmbito dos experimentos realizados no presente trabalho, o KDD será utilizado como metodologia. Diante disso, considerando o objetivo principal do artigo, o qual se refere a um estudo sobre a aplicação de modelos de regressão não-linear na precificação de veículos usados, serão utilizados dados extraídos diretamente de anúncios veiculados em portais de classificados na Internet. Para tanto, foi desenvolvida uma aplicação baseada em Web Scraping (LAWSON, 2015), utilizando a linguagem de programação Python.

Os atributos considerados para a realização dos experimentos foram: modelo, ano de fabricação, ano do modelo, tipo de câmbio e tipo de combustível. Os dados foram persistidos em uma base estruturada em formato CSV[1], diretamente após sua extração. Uma das vantagens do uso deste formato é a sua flexibilidade e adaptabilidade em termos da construção de soluções usando a linguagem de programação Python, a qual foi utilizada durante todas as fases do estudo.

Os dados obtidos foram submetidos à etapa de pré-processamento para verificação de possíveis erros e *outliers*. Para tanto, foram feitas análises verificando-se a distribuição dos valores em função de seus quartis, sendo utilizada a regra interquartil $1.5 \cdot FIQ$, a qual define que *outliers* superiores possuem valor maior que $Q_3 + 1.5 \cdot FIQ$, e *outliers* inferiores valor menor que $Q_1 - 1.5 \cdot FIQ$, sendo Q_1 o primeiro quartil, Q_3 o terceiro quartil, e $FIQ = Q_3 - Q_1$, denominado como faixa interquartil. Além disso, originalmente as instâncias de ano de fabricação e ano de modelo são obtidas de forma concatenada, isto é, em um mesmo atributo, sendo necessário realizar a separação e consecutiva associação aos atributos devidos.

Uma vez que os valores foram persistidos em uma base de dados estruturada CSV, adequada para os propósitos de processamento e análise dos dados, não houve necessidade de transformação ou integração, sendo possível a aplicação dos modelos de mineração de dados selecionados para o estudo, a saber, regressão linear e não-linear. Conforme descrito por Weisberg (2013), a análise de regressão se refere ao estudo do relacionamento entre uma variável alvo, denominada variável dependente, com relação a outras variáveis, denominadas de variáveis independentes, através de um modelo matemático linear. Para o caso em que exista apenas uma variável independente, o modelo é denominado de regressão linear simples, definido pela Equação 1:

$$\gamma = \alpha x + \beta + \varepsilon, \tag{1}$$

onde α é dado por:

[1] CSV é o acrônimo em inglês para *comma separated values* (valores separados por vírgulas).

$$\alpha = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \tag{2}$$

e β por:

$$\beta = \overline{y} - \alpha\overline{x}. \tag{3}$$

Para problemas em que exista mais de uma variável independente, o modelo é denominado de regressão linear múltipla, sendo, neste caso, definido pela Equação 4:

$$\gamma = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \beta + \varepsilon, \tag{4}$$

onde n é o número de variáveis independentes.

Uma forma de avaliar a efetividade dos resultados obtidos pela aplicação dos métodos de regressão, considerando uma amostra com m elementos, nos quais se conhece o valor real da variável independente y, é através do uso do coeficiente R^2, o qual é definido como (Equação 5):

$$R^2 = 1 - \frac{\sum_i (\gamma_i - \overline{\gamma}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \tag{5}$$

onde o melhor valor possível é 1.0

É interessante observar que a predição de valores através de métodos de regressão linear pode não ser adequada para situações onde a distribuição de dados se distancia de uma distribuição linear, ou seja, em determinadas regiões os dados se afastam consideravelmente do que seria esperado em uma distribuição aqui denominada de linear. Uma possível solução é o uso de modelos de regressão não-linear. No âmbito deste estudo são considerados modelos não-lineares polinomiais, os quais se referem a funções que apresentem a forma (Equação 6):

$$\gamma = \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + ... + \alpha_n x^n + \beta, \tag{6}$$

sendo que os coeficientes podem ser obtidos através do seguinte sistema de equações (Equação 7):

$$\begin{bmatrix} \sum_i x_i^{2m} & ... & \sum_i x_i^{m+2} & \sum_i x_i^{m+1} & \sum_i x_i^m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_i x_i^{m+2} & ... & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^{m+1} & ... & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^m & ... & \sum_i x_i^2 & \sum_i x_i & \sum_i n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \sum_i y_i x_i^m \\ \vdots \\ \sum_i y_i x_i^2 \\ \sum_i y_i x_i \\ \sum_i y_i \end{bmatrix} \tag{7}$$

4 Resultados e discussões

Através da aplicação da solução de extração de dados desenvolvida, foram obtidos 229 registros de veículos anunciados através de um portal de classificados na Internet. Para os propósitos de análise foram extraídos dados referentes à uma marca e modelo específico de veículo, considerando suas derivações e variações em termos de geração e motor. Além disso, levando-se em conta a premissa de que os aspectos de regionalidade são um influenciador no preço médio de venda, os registros obtidos são apenas de veículos anunciados para o estado de Goiás.

Através do cálculo dos quartis para os atributos numéricos da base de dados, a saber: valor de venda, anos de fabricação e de modelo, foram verificadas possíveis existências de *outliers*. A

Figura 1 apresenta os resultados obtidos para os valores de venda, nos quais, assim como os demais atributos, não foram detectados *outliers*. A partir disso, foi calculada a correlação entre os atributos numéricos, apresentada na Figura 2, onde, tomando-se como referência a variável alvo, o valor do veículo, é possível observar que a maior correlação se refere ao ano de fabricação. É interessante destacar que a correlação com o ano de modelo é consideravelmente próxima ao ano de fabricação, o que é justificado pelo fato de este valor ser sempre igual ou maior em uma unidade. Por outro lado, o atributo referente à categoria de modelo (*modelo_c*) possui baixo valor de correlação, não sendo adequado para propósitos de inferência de valores através de modelos de regressão.

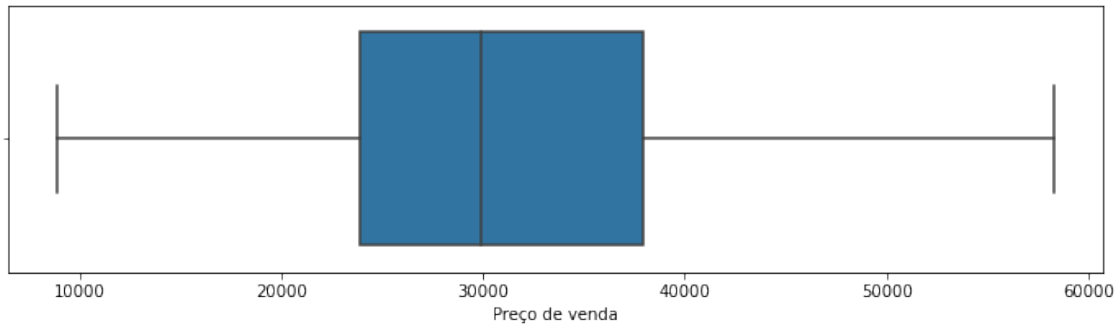

Figura 1: Boxplot para verificação de *outliers*.

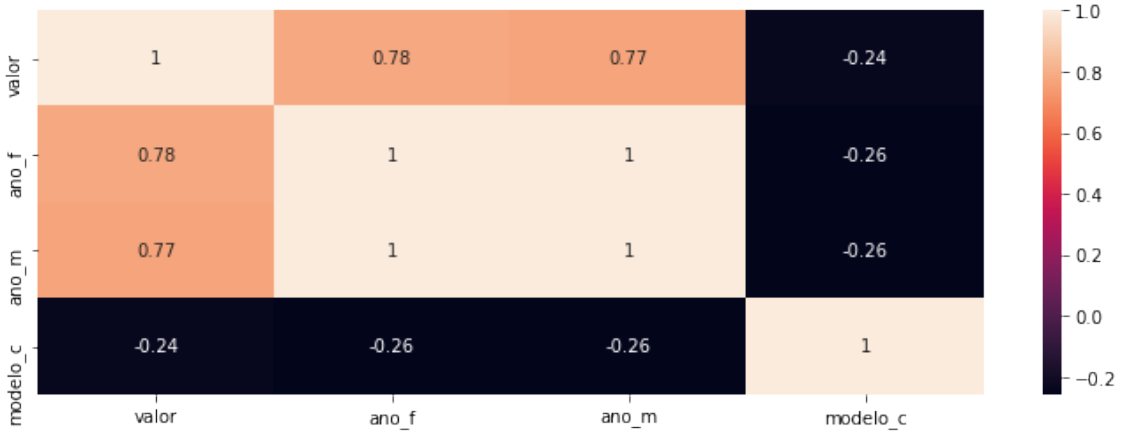

Figura 2: Correlação dos atributos candidatos.

Com base nas análises obtidas através do cálculo das correlações, o melhor atributo a ser utilizado no modelo como variável independente é o ano de fabricação. Diante disso, a Figura 3 apresenta a distribuição dos valores de venda em relação ao ano de fabricação. É interessante observar neste gráfico que, partindo-se do maior ano para os menores, existe uma desvalorização contínua até um determinado ano, sendo que os anos menores apresentam valores de venda maiores, uma vez que tais veículos passam a ser objetos de coleção ou considerados como raridades.

Uso de modelos de regressão não-linear na precificação de venda de veículos usados

Douglas Farias Cordeiro, Renata Moreira Limiro, Anelise Souza Rocha, Núbia Rosa da Silva

Resumo A precificação de produtos é um tema de considerável complexidade, que leva em conta uma série de fatos, os quais vão desde questões relacionadas à própria manufatura em si, quanto fatores tributários e logísticos. No contexto de produtos usados, existe ainda impactos que remetem a questões regionais, de manutenção, assim como interesse enquanto objeto de apreciação ou coleção. No mercado de veículos usados, uma das principais referências de precificação é a tabela disponibilizada pela Fundação Instituto de Pesquisas Econômicas (FIPE). Entretanto, tal tabela não leva em conta algumas especificidades de veículos, principalmente em termos de regionalidade. Neste sentido, se torna interessante avaliar o uso de soluções alternativas, como é o caso da aplicação de métodos computacionais inteligentes. Neste artigo, é proposto a avaliação de modelos baseados em regressão linear e não-linear, considerando um conjunto de dados de anúncios extraídos em um portal de vendas online. Os resultados obtidos mostram que métodos de regressão não-linear apresentam ganhos consideráveis em termos de acurácia, se destacando uma interessante solução para o problema.

1 Introdução

A precificação de produtos é um tema de considerável complexidade, que leva em conta uma série de aspectos, os quais permeiam desde questões relacionadas à demanda e oferta, até fatores ligados à qualidade, marca e regionalidade. Neste contexto, a determinação de preços de vendas de veículos usados se destaca como um problema de grande interesse, uma vez que este tipo de mercado sofre uma série de impactos, onde se destaca a variação entre a desvalorização em face do tempo de uso, para uma determinada faixa de tempo, e a posterior valorização, quando um veículo alcança uma idade específica de fabricação.

Douglas Farias Cordeiro
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: cordeiro@ufg.br

Renata Moreira Limiro
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: renatamlimiro@gmail.com

Anelise Souza Rocha
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: anelisesrocha@gmail.com

Núbia Rosa da Silva
Instituto de Biotecnologia - Universidade Federal de Catalão, Catalão, Goiás, Brasil.
e-mail: nubia@ufg.br

Anais do XV Encontro Anual de Ciência da Computação (EnAComp 2020). ISSN: 2178-6992.
Catalão, Goiás, Brasil. 25 a 27 de Novembro de 2020.

Neste sentido, o preço e seus aspectos de volatilidade sempre foram alvo dos olhares e questionamentos da ciência. Vários são os conceitos que procuram definir, simular ou ainda prever preços ideais considerando diferentes contextos e cenários para sua aplicação. Segundo Smith (2017), o fator determinante do preço seria o custo relacionado à mão de obra dedicada em produzir aquele bem específico. Ricardo (2018), no entanto, trouxe uma perspectiva mais ampliada para o conceito de produção, propondo que o custo de máquinas e equipamentos utilizados para a produção do bem também sejam considerados no cálculo de preço.

Segundo Marshall (2018), a definição do preço de um bem não está condicionada somente à visão da indústria, dos produtores ou ofertantes do bem, mas também à utilidade que este tem ao comprador. Nesse sentido, o conceito de elasticidade de preço explora a variação do preço em função da maximização dos lucros do ofertante do bem em relação a busca por máxima utilidade por parte do demandante. Desta maneira, o comprador estaria sempre buscando o menor preço enquanto o fornecedor visa maiores margens de lucro, chegando assim, a um ponto onde o preço do bem, desconsiderando fatores externos, tais como produtos concorrentes, atende as expectativas que tangem a demanda e a oferta, alcançando um equilíbrio parcial de mercado.

Outro aspecto que pode ser considerado na definição do preço de um bem, sendo ele produto, serviço ou experiência, é o hedonismo, ou seja, o quão agradável, satisfatório, ou prazeroso, é ao comprador adquirir aquele determinado item. Segundo Besanko, Dranove e Shanley (2012) os preços hedônicos podem ser considerados como preços sensíveis a variações decorrentes da conciliação de conjuntos de atributos. Como, por exemplo, ao se analisar o preço de um imóvel é possível observar o tamanho do terreno, quantidade de banheiros, a localização e outros atributos, os quais, existindo em diferente níveis, podem impactar o interesse do comprador e consequentemente o valor de mercado daquele bem. Neste contexto, Fávero, Belfiore e Lima (2008) sugere que, para uma melhor análise e entendimento de preços hedônicos e das características ou atributos de um bem que causam mais impacto em seu preço, são necessárias aplicações de métodos de regressão para que, assim, possam ser observadas as variações de preço em função de determinadas características.

No contexto de mercado de automóveis, no Brasil, a Fundação Instituto de Pesquisas Econômicas (FIPE) realiza continuamente o acompanhamento e divulgação dos preços médios dos veículos anunciados por vendedores através da tabela FIPE. Apesar de ser um referencial interessante de precificação de veículos usados, a tabela não leva em conta especificidades dos produtos, tais como as supracitadas questões de regionalidade. Neste sentido, a aplicação de modelos computacionais inteligentes para a precificação de veículos usados se torna consideravelmente relevante. Considerando os aspectos referentes aos atributos deste tipo de problema, onde o principal objetivo é se obter o preço de um produto, modelos de inferência baseados em regressão se tornam uma interessante alternativa, uma vez que possibilitam, a partir de um conjunto de valores independentes, se obter o valor alvo.

O presente artigo apresenta um estudo relacionado à aplicação e avaliação de modelos de regressão não-linear, especificamente polinomiais, para solução do problema da precificação de veículos usados. Para tanto, é constituída uma base de dados obtida a partir da extração de anúncios de veículos usados publicados em portais de classificados digitais. Os resultados alcançados demonstram a efetividade do uso de modelos de regressão não-linear em face de modelos lineares, com um ganho considerável em termos de acurácia.

2 Revisão bibliográfica

O mercado de carros usados no Brasil tem uma forte representação oriunda das altas tarifas e impostos na aquisição de veículos novos. Entretanto, a informalidade desta área de atuação, trouxe consigo a necessidade de profissionalizar a revenda destes, sendo possível verificar as necessidades dos usuários e a crescente do mercado. A FENAUTO (Federação Nacional das Associações dos Revendedores de Veículos Automotores) é a responsável por gerenciar e medir o mercado de seus associados.

Os números de 2019 apurados pela FENAUTO contam com cerca de 48 mil revendas em todo o Brasil, girando 380 bilhões de reais por ano e gerando 620 mil empregos diretos e indiretos. No primeiro semestre de 2020, um estudo realizado mostrou que para cada carro novo vendido, quatro usados são negociados, e que destes, 39% são com fabricação de 4 a 8 anos, e 15% de seminovos (menos de quatro anos de uso) (RODRIGUES, 2020). Por meio destes dados é possível verificar a potência mercadológica de usados e a necessidade de se especializar ainda mais seus processos de venda e o hedonismo de seus preços.

A precificação dos usados, na teoria baseia-se no uso da Tabela FIPE, entretanto, a utilização deste meio se distancia veemente dos valores praticados no mercado, pois o cálculo é feito pela média nacional como descrito pela própria fundação. Tal descrição deixa em aberto a possibilidade de negociação por parte do mercado, entretanto essa prática dificulta o processo de precificação e padronização.

Com a utilização de modelos estatísticos e computacionais pertinentes à análise de dados, é possível obter valores de forma preditiva a partir de dados existentes, ou seja, o modelo elucidará o caminho correto a seguir dos preços através das características correlacionadas. De acordo com Carmo (2014), a aplicação de técnicas de precificação hedônicas admite a valoração implícita nos atributos do bem. Em um estudo sobre a precificação de casas, Boye, Mireky-Gyimah e Okpoti (2017) utilizou os atributos da construção, aplicando regressão linear múltipla, e obtiveram um coeficiente de regressão (R^2) de 99%, validando assim o uso da regressão linear. Entretanto estudos relacionados por Rosa, Oliveira e Pinto (2019), também no ramo habitacional, finalizou a pesquisa com um coeficiente (R^2) de 42%, e na tentativa de melhora desse índice, com troca de previsores, e não avaliando o uso de outro tipo de modelo.

Regressando ao ramo de carros, Ozgur et al. (2016) buscou predizer o preço de veículos de uma marca específica, aplicando regressão linear e realizando testes do modelo também apenas alterando previsores, a fim de diminuir a quantidade de *outliers*, os quais se mantiveram em todos os testes, a pesquisa não cogitou a mudança de modelo, uma vez que o coeficiente R^2 estava satisfatório, variando de 76% a 81%. Segundo Chernick e Friis (2003), os *outliers* podem influenciar não só no coeficiente R^2, como no interceptor, pois a linha de melhor ajuste é feita pela soma do quadrado dos resíduos, ou seja, como um número maior de extremos a magnitude aumenta. Neste contexto, e corroborando com a proposta deste estudo, testar qual o modelo mais propício para os dados se faz de grande valia para eficácia dos resultados uma vez que a acurácia, as avaliações de desvio padrão é que fazem a confiabilidade do modelo, principalmente em um ambiente de preços hedônicos, como o mercado de veículos.

3 Metodologia

A necessidade de geração de informações úteis e que agreguem valor aos negócios é algo fundamental e de grande interesse. Neste sentido, a aplicação de soluções computacionais inteligentes, as quais possam lidar com grandes volumes de dados, se torna essencial. O processo denominado de Descoberta do Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases - KDD*), proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996) consiste em um conjunto ordenado de passos compostos por: seleção, pré-processamento, transformação, mineração e interpretação de dados. Uma das principais vantagens do uso deste processo na geração de informação se refere ao fato de ser interativo e iterativo, permitindo alterações e avaliações ao longo da realização de todas as etapas.

No âmbito dos experimentos realizados no presente trabalho, o KDD será utilizado como metodologia. Diante disso, considerando o objetivo principal do artigo, o qual se refere a um estudo sobre a aplicação de modelos de regressão não-linear na precificação de veículos usados, serão utilizados dados extraídos diretamente de anúncios veiculados em portais de classificados na Internet. Para tanto, foi desenvolvida uma aplicação baseada em Web Scraping (LAWSON, 2015), utilizando a linguagem de programação Python.

Os atributos considerados para a realização dos experimentos foram: modelo, ano de fabricação, ano do modelo, tipo de câmbio e tipo de combustível. Os dados foram persistidos em uma base estruturada em formato CSV[1], diretamente após sua extração. Uma das vantagens do uso deste formato é a sua flexibilidade e adaptabilidade em termos da construção de soluções usando a linguagem de programação Python, a qual foi utilizada durante todas as fases do estudo.

Os dados obtidos foram submetidos à etapa de pré-processamento para verificação de possíveis erros e *outliers*. Para tanto, foram feitas análises verificando-se a distribuição dos valores em função de seus quartis, sendo utilizada a regra interquartil $1.5 \cdot FIQ$, a qual define que *outliers* superiores possuem valor maior que $Q_3 + 1.5 \cdot FIQ$, e *outliers* inferiores valor menor que $Q_1 - 1.5 \cdot FIQ$, sendo Q_1 o primeiro quartil, Q_3 o terceiro quartil, e $FIQ = Q_3 - Q_1$, denominado como faixa interquartil. Além disso, originalmente as instâncias de ano de fabricação e ano de modelo são obtidas de forma concatenada, isto é, em um mesmo atributo, sendo necessário realizar a separação e consecutiva associação aos atributos devidos.

Uma vez que os valores foram persistidos em uma base de dados estruturada CSV, adequada para os propósitos de processamento e análise dos dados, não houve necessidade de transformação ou integração, sendo possível a aplicação dos modelos de mineração de dados selecionados para o estudo, a saber, regressão linear e não-linear. Conforme descrito por Weisberg (2013), a análise de regressão se refere ao estudo do relacionamento entre uma variável alvo, denominada variável dependente, com relação a outras variáveis, denominadas de variáveis independentes, através de um modelo matemático linear. Para o caso em que exista apenas uma variável independente, o modelo é denominado de regressão linear simples, definido pela Equação 1:

$$\gamma = \alpha x + \beta + \varepsilon, \tag{1}$$

onde α é dado por:

[1] CSV é o acrônimo em inglês para *comma separated values* (valores separados por vírgulas).

$$\alpha = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, \tag{2}$$

e β por:

$$\beta = \overline{y} - \alpha\overline{x}. \tag{3}$$

Para problemas em que exista mais de uma variável independente, o modelo é denominado de regressão linear múltipla, sendo, neste caso, definido pela Equação 4:

$$\gamma = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \beta + \varepsilon, \tag{4}$$

onde n é o número de variáveis independentes.

Uma forma de avaliar a efetividade dos resultados obtidos pela aplicação dos métodos de regressão, considerando uma amostra com m elementos, nos quais se conhece o valor real da variável independente y, é através do uso do coeficiente R^2, o qual é definido como (Equação 5):

$$R^2 = 1 - \frac{\sum_i (\gamma_i - \overline{\gamma}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \tag{5}$$

onde o melhor valor possível é 1.0

É interessante observar que a predição de valores através de métodos de regressão linear pode não ser adequada para situações onde a distribuição de dados se distancia de uma distribuição linear, ou seja, em determinadas regiões os dados se afastam consideravelmente do que seria esperado em uma distribuição aqui denominada de linear. Uma possível solução é o uso de modelos de regressão não-linear. No âmbito deste estudo são considerados modelos não-lineares polinomiais, os quais se referem a funções que apresentem a forma (Equação 6):

$$\gamma = \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + ... + \alpha_n x^n + \beta, \tag{6}$$

sendo que os coeficientes podem ser obtidos através do seguinte sistema de equações (Equação 7):

$$\begin{bmatrix} \sum_i x_i^{2m} & ... & \sum_i x_i^{m+2} & \sum_i x_i^{m+1} & \sum_i x_i^m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_i x_i^{m+2} & ... & \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^{m+1} & ... & \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^m & ... & \sum_i x_i^2 & \sum_i x_i & \sum_i n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \sum_i y_i x_i^m \\ \vdots \\ \sum_i y_i x_i^2 \\ \sum_i y_i x_i \\ \sum_i y_i \end{bmatrix} \tag{7}$$

4 Resultados e discussões

Através da aplicação da solução de extração de dados desenvolvida, foram obtidos 229 registros de veículos anunciados através de um portal de classificados na Internet. Para os propósitos de análise foram extraídos dados referentes à uma marca e modelo específico de veículo, considerando suas derivações e variações em termos de geração e motor. Além disso, levando-se em conta a premissa de que os aspectos de regionalidade são um influenciador no preço médio de venda, os registros obtidos são apenas de veículos anunciados para o estado de Goiás.

Através do cálculo dos quartis para os atributos numéricos da base de dados, a saber: valor de venda, anos de fabricação e de modelo, foram verificadas possíveis existências de *outliers*. A

Figura 1 apresenta os resultados obtidos para os valores de venda, nos quais, assim como os demais atributos, não foram detectados *outliers*. A partir disso, foi calculada a correlação entre os atributos numéricos, apresentada na Figura 2, onde, tomando-se como referência a variável alvo, o valor do veículo, é possível observar que a maior correlação se refere ao ano de fabricação. É interessante destacar que a correlação com o ano de modelo é consideravelmente próxima ao ano de fabricação, o que é justificado pelo fato de este valor ser sempre igual ou maior em uma unidade. Por outro lado, o atributo referente à categoria de modelo (*modelo_c*) possui baixo valor de correlação, não sendo adequado para propósitos de inferência de valores através de modelos de regressão.

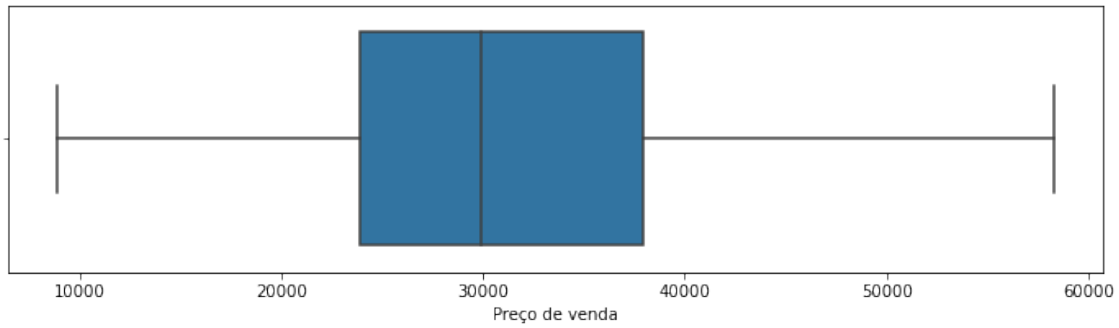

Figura 1: Boxplot para verificação de *outliers*.

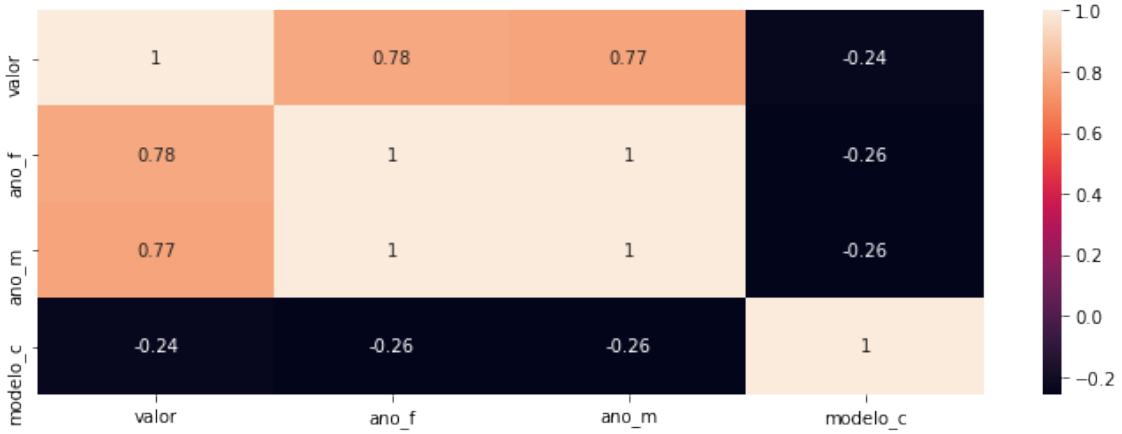

Figura 2: Correlação dos atributos candidatos.

Com base nas análises obtidas através do cálculo das correlações, o melhor atributo a ser utilizado no modelo como variável independente é o ano de fabricação. Diante disso, a Figura 3 apresenta a distribuição dos valores de venda em relação ao ano de fabricação. É interessante observar neste gráfico que, partindo-se do maior ano para os menores, existe uma desvalorização contínua até um determinado ano, sendo que os anos menores apresentam valores de venda maiores, uma vez que tais veículos passam a ser objetos de coleção ou considerados como raridades.

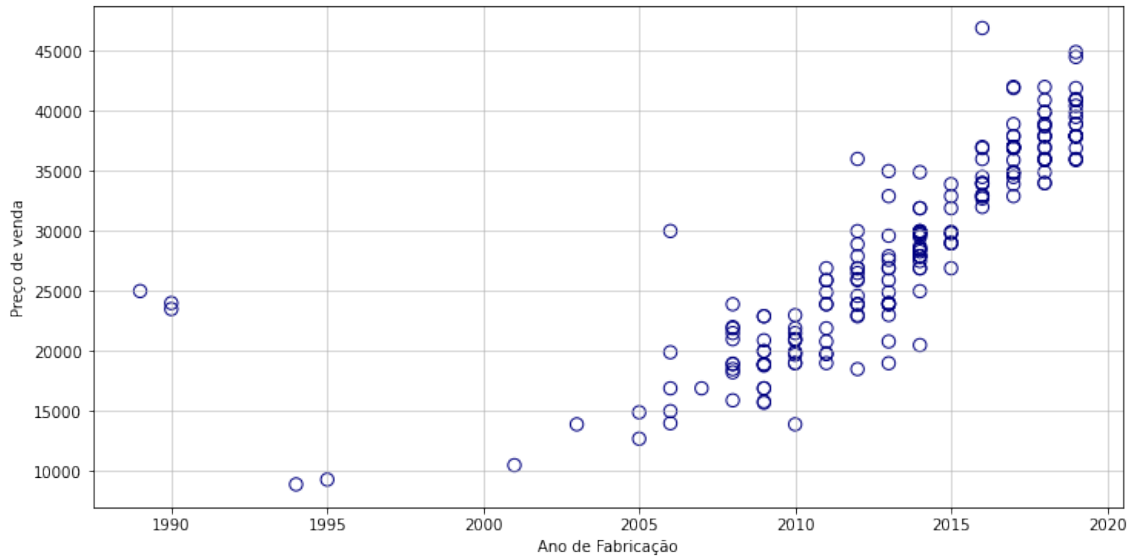

Figura 3: Distribuição dos valores de venda em relação ao ano de fabricação.

Considerando os atributos valor de venda e ano de fabricação, foram gerados modelos de regressão linear simples e de regressão não-linear polinomial, onde foram avaliadas soluções com diferentes graus. A avaliação foi realizada utilizando o coeficiente R^2 (Equação 5). A Tabela 1 apresenta os resultados obtidos, onde nota-se uma melhora considerável na aplicação de modelos de regressão não-linear para problemas deste tipo, onde o valor de R^2 subiu de 0.61 para 0.85, com um modelo de grau três.

Grau	R^2
1	0.6156685419977836
2	0.8481529153203329
3	0.8547499965132850
4	0.8547377101105589
5	0.8547253109852020

Tabela 1: R^2 em função do grau do modelo de regressão.

A Figura 4 apresenta os resultados obtidos com os modelos gerados para regressão linear simples e regressão não linear polinomial de grau dois e três. É possível observar que os modelos de regressão não linear apresentam um potencial mais adequado para a distribuição dos dados, minimizando os erros de predição, inclusive para situações onde se tem uma variação nas tendências de desvalorização ou valorização dos automóveis.

Além dos testes realizados sobre o conjunto de dados em face de sua distribuição original, foram calculadas as médias dos valores de venda levando-se em conta o ano de fabricação dos automóveis. Essa abordagem foi verificada devido ao fato de que a principal referência de valores venais de automóveis, a tabela FIPE, se baseia no cálculo da média. A partir disso, os modelos de regressão foram aplicados, apresentando uma melhoria nos resultados para o uso dos modelos não lineares, como pode ser observado na Tabela 2. A Figura 5 apresenta os resultados obtidos considerados os valores médios de venda para regressão linear simples e regressão não linear polinomial de grau dois e três.

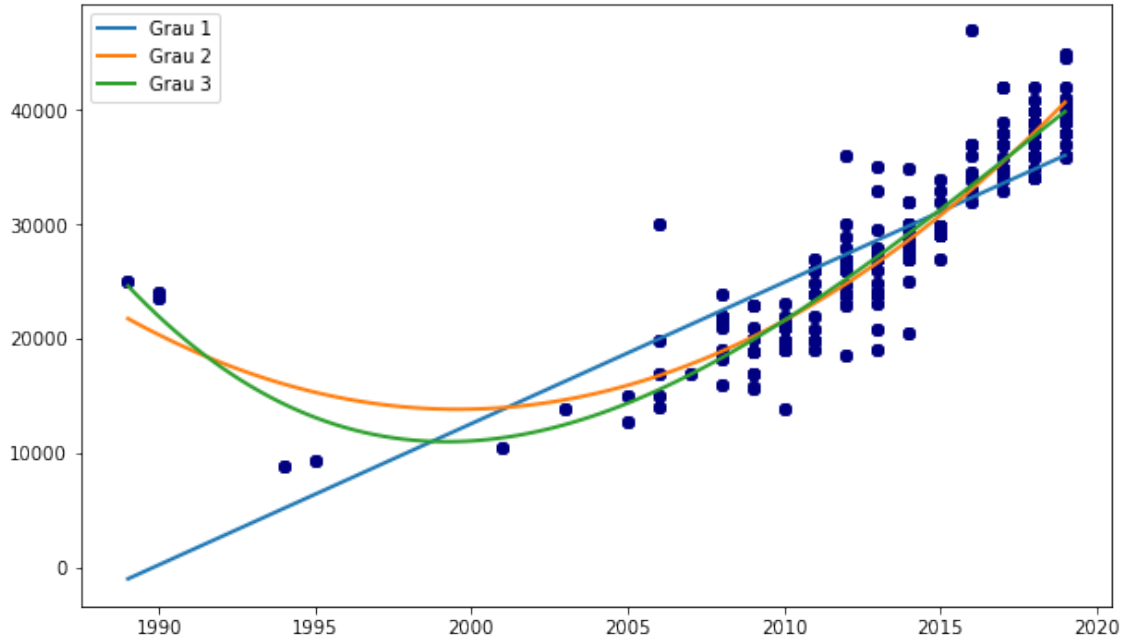

Figura 4: Curvas de predição obtidas pelos modelos de regressão.

Grau	R^2
1	0.46195858128076006
2	0.9288910051147128
3	0.9685642555641376
4	0.9684335352597673
5	0.9683013022516503

Tabela 2: R^2 em função do grau do modelo de regressão para valores médios de venda.

5 Conclusões

O comércio de veículos usados é um setor de grande relevância para a economia no Brasil, uma vez que a cada veículo novo vendido, há a comercialização de quatro veículos usados. Como este é um mercado que sofre com diversos fatores peculiares, como a região de venda, estado de conservação, entre outros, exige-se que modelos matemáticos sejam aplicados a fim de obter resultados padronizados e acurados.

Neste trabalho, os dados de comercialização de 229 veículos usados foram obtidos de classificados na Internet para verificar a aplicabilidade de modelos de regressão linear e não-linear para a classificação desses veículos. Os atributos modelo, valor de venda, anos de fabricação e de modelo foram extraídos, no entanto, o ano de fabricação resultou na maior de taxa de correlação dos atributos sendo utilizado como parâmetro da precificação. Os resultados mostraram que os modelos de regressão não-linear com grau três são mais robustos, apresentando um aumento de 0.24 no valor de R^2 quando comparado ao modelo linear.

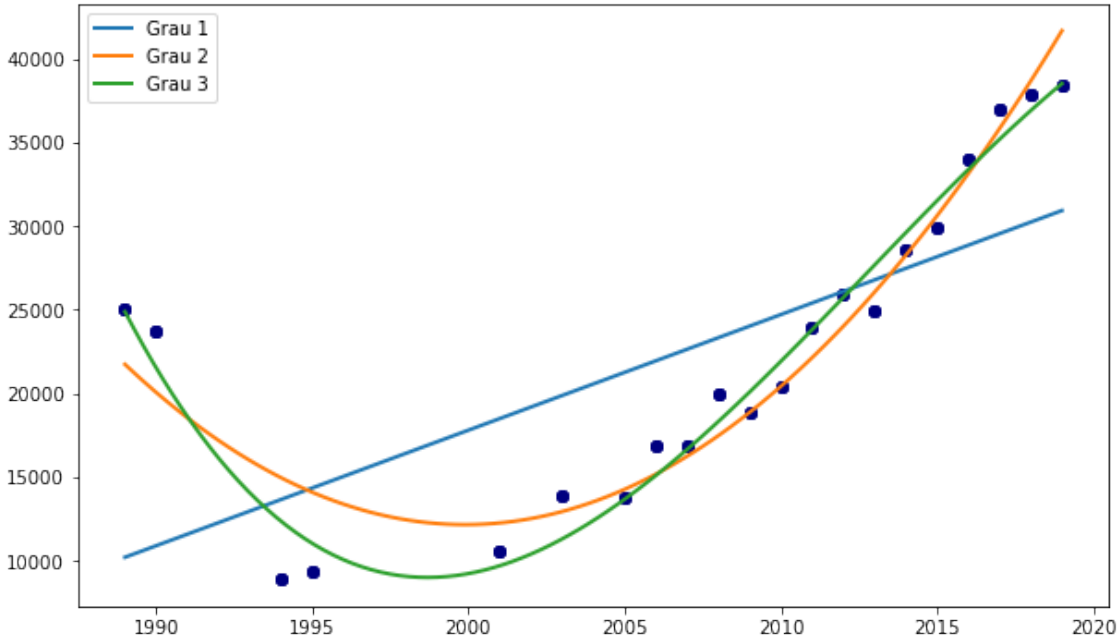

Figura 5: Curvas de predição obtidas pelos modelos de regressão para valores médios de venda.

Referências

BESANKO, David; DRANOVE, David; SHANLEY, Mark. **Economics of Strategy**. Hoboken, NJ, USA: John Wiley & Sons, 2012.

BOYE, P.; MIREKY-GYIMAH, D.; OKPOTI, C. A. Multiple Linear Regression Model for Estimating the Price of a Housing Unit. **Ghana Mining Journal**, v. 17, n. 2, p. 66–77, 2017.

CARMO, C. Precificação imobiliária baseada em modelagem hedônica e externalidades: um estudo aplicado a terrenos urbanos. **ReFAE – Revista da Faculdade de Administração e Economia**, v. 5, n. 2, p. 2–23, 2014.

CHERNICK, M. R.; FRIIS, R. H. **Introductory biostatistics for the health sciences: modern applications including bootstrap**. New Jersey: John Wiley & Sons, 2003.

FÁVERO, Luiz Paulo Lopes; BELFIORE, Patrícia Prado; LIMA, Gerlando A. S. Franco de. Modelos de precificação hedônica de imóveis residenciais na região metropolitana de São Paulo: uma abordagem sob as perspectivas da demanda e da oferta. **Estudos Econômicos (São Paulo)**, v. 38, n. 1, p. 73–96, 2008.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

LAWSON, R. **Web Scraping with Python**. Birmingham, UK: Packt Publishing, 2015.

MARSHALL, A. **Principles of Economics**. [S.l.]: Franklin Classics Trade Press, 2018.

OZGUR, C. et al. Multiple Linear Regression Applications Automobile Pricing. **International Journal of Mathematics and Statistics Invention (IJMSI)**, v. 4, n. 6, p. 1–10, 2016.

RICARDO, David. **Princípios de Economia Política e Tributação**. São Paulo: Lebooks Editora, 2018.

RODRIGUES, H. **Para cada carro novo vendido no Brasil, quatro usados são negociados**. [S.l.: s.n.], 2020. Revista Quatro Rodas. Disponível em: https://quatrorodas.abril.com.br/noticias/para-cada-carro-novo-vendido-no-brasil-quatro-usados-sao-negociados/. Acesso em 16 de Setembro de 2020.

ROSA, V. S.; OLIVEIRA, P. B.; PINTO, R. L. M. Modelos de precificação para locação e venda de imóveis residenciais na cidade de João Monlevade-MG via regressão linear multivariada. **Gestão da Produção, Operações e Sistemas**, v. 14, n. 3, p. 151–167, 2019.

SMITH, Adam. **A Riqueza das Nações**. Rio de Janeiro: Nova Fronteira, 2017.

WEISBERG, S. **Applied Linear Regression**. Hoboken, NJ, USA: John Wiley & Sons, 2013.