D2Fipe: Uma base de dados para previsão de desvalorização do preço médio de veículos a partir da tabela FIPE

Daniel A. P. Carvalho[1], Talisson C. Damião[1], Rosalvo F. de O. Neto[1]

Abstract
A indústria automotiva brasileira gerou aproximadamente R$ 80 bilhões em tributos em 2021 e representou 18% do Produto Interno Bruto (PIB). O estudo de modelos de previsão de desvalorização de veículos pode auxiliar vários setores da indústria automotiva como, por exemplo, locadoras e concessionárias podem definir o melhor momento para a venda do veículo. No entanto, apesar de relevante existem poucas base de dados de domínio público organizada em séries temporárias com a desvalorização dos preços dos veículos. Esse tipo de base de dados é essencial para construção de soluções de aprendizagem de máquina. O objetivo deste projeto foi construir uma base de dados pública com essa finalidade. O compartilhamento de dados de pesquisa possui vários benefícios dentre eles podemos destacar: é útil para treinar novos pesquisadores, investigar novos métodos, ajudar a identificar erros, e desencorajar fraudes.

Keywords
FIPE — Web scraping — ETL — Apache Airflow — MongoDB – Dataset

[1] *Engenharia da Computação, Universidade Federal do Vale do São Francisco, Juazeiro-BA, Brasil*
*__Daniel A. P. Carvalho, Talisson C. Damião, Rosalvo F. de O. Neto__: danielalencar746@gmail.com, talisoncaedoso@gmail.com, rosalvo.oliveira@univasf.edu.br

Contents

1 Critério de Amostragem **1**
2 Análises do Dataset **2**
2.1 Modelos por marca e categoria 2
2.2 Modelos por ano-modelo 2
2.3 Valores válidos 2
2.4 Depreciação dos modelos 3
3 Materiais e métodos **3**
3.1 Apache Airflow 3
3.2 Selenium 4
3.3 MongoDB 4
4 Exemplo de utilização do Dataset **4**
References **4**

Motivação

O estudo de modelos de previsão de desvalorização de veículos pode auxiliar vários setores da indústria automotiva como, por exemplo, locadoras e concessionárias podem definir o melhor momento para a venda do veículo. A construção de modelos de previsão usa algoritmos de aprendizado de máquina supervisionado. O treinamento desses algoritmos exige a utilização de base de dados histórica. No entanto, até o presente momento exitem poucas bases de dados pública com o histórico de desvalorização dos principais veículos vendidos no mercado brasileiro. Desta forma, este projeto teve como objetivo preencher essa lacuna.

1. Critério de Amostragem

A tabela FIPE é uma referência muito utilizada no Brasil para consulta de preços médios de veículos. Ela fornece valores atualizados mensalmente para veículos novos e usados de todas as marcas e modelos comercializados. Os modelos e preços dos veículos foram extraídos da Tabela FIPE para compor a base de dados.

Os modelos selecionados foram retirados do site **AUTOO**, que é um portal de notícias e informações sobre o mercado automobilístico no Brasil. No total, selecionou-se os 10 modelos mais vendidos, no período de 2015 a 2022, nas categorias mais populares de veículos. Muitas variantes dos mesmos modelos estão disponíveis na tabela FIPE, assim, para cada modelo, foram selecionados até 4 variantes com as maiores motorizações e os tipos de transmissão mecânica e automática. Esse critério de busca garante informações precisas e atualizadas, refletindo as tendências do mercado brasileiro de veículos.

A base de dados gerada contém informações detalhadas

sobre diferentes modelos e anos de veículos, abrangendo até 36 preços coletados para cada um deles. Essa diversidade foi registrada em diferentes datas, permitindo uma análise abrangente ao longo do tempo. Além disso, o Dataset é extremamente versátil, possibilitando o uso de modelos simples para prever preços futuros e realizar pesquisas adicionais.

A disponibilidade de informações sobre a variação de preços ao longo do tempo proporciona uma compreensão mais profunda das tendências do mercado automotivo. Além disso, a capacidade de realizar estudos comparativos entre os modelos mais populares no mercado brasileiro é um recurso valioso. Vale ressaltar que a seleção cuidadosa dos modelos mais vendidos em cada categoria garante a representatividade dos dados coletados, tornando a base de dados altamente relevante e útil para pesquisadores em várias áreas. Na **Tabela 1** encontra-se o dicionário de dados do Dataset.

Table 1. Dicionário de dados da Base de dados

Coluna	Significado
Categoria	Categoria do veículo
Marca	Marca do veículo
Modelo	Modelo do veículo
Ano-modelo	Ano-modelo do veículo
Mes 1	Preço do veículo no mês 1
Mes 2	Preço do veículo no mês 2
Mes 3	Preço do veículo no mês 3
...	...
Mes 36	Preço do veículo no mês 36
Depreciação	Depreciação do veículo em um período de 3 anos

2. Análises do Dataset

Após uma seleção e filtragem, o conjunto de dados foi reduzido para 413 registros (linhas). A partir disso, foram conduzidas análises estatísticas para obter informações relevantes sobre os preços dos veículos ao longo do tempo. Essas análises têm alto valor para compreender o comportamento dos preços e podem ser aplicadas no desenvolvimento de modelos de previsão precisos e eficientes. Portanto, o estudo minucioso das estatísticas contidas no conjunto de dados é crucial para obter insights valiosos e contribuir para pesquisas futuras.

2.1 Modelos por marca e categoria

Nessa etapa, foi analisada a quantidade de marcas e categorias diferentes de veículos. Essa relação é essencial ao criar um modelo preditivo em uma base de dados de preços de veículos. Isso ocorre porque a quantidade de modelos de uma marca pode impactar o comportamento dos preços ao longo do tempo.

As **Figuras 1** e **2** representam a quantidade de modelos por categoria e por marca na base de dados, respectivamente. Pode-se identificar que a marca Chevrolet possui a maior quantidade de modelos para a marca, com cerca de 12.6% de todo o Dataset. Em segundo lugar, temos as marcas Volkswagen e Ford com 10.7% dos modelos para cada uma.

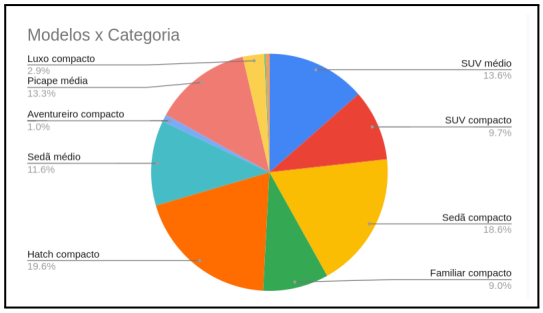

Figure 1. Quantidade de modelos por categoria.

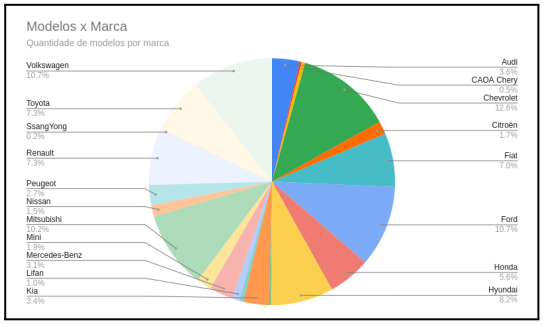

Figure 2. Relação da quantidade de modelos por marca.

2.2 Modelos por ano-modelo

A análise da quantidade de modelos de veículos por ano-modelo é útil para identificar padrões de comportamento nos preços ao longo do tempo. A **Figura 3** exibe a quantidade de modelos na base de dados gerada de 2015 a 2020 (os anos considerados). É possível perceber que a medida que o ano-modelo aumenta, menos registros de modelos são encontrados no Dataset.

2.3 Valores válidos

A análise da quantidade de valores válidos para cada mês de busca de cada ano-modelo ajuda a avaliar como os dados de preços estão distribuídos no Dataset. A **Figura 4** exibe um resumo das quantidades de valores válidos por ano e mês.

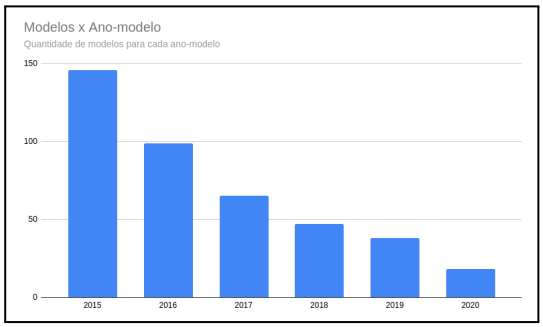

Figure 3. Relação da quantidade de modelos por ano-modelo.

Quantidade de valores válidos para cada mês e ano-modelo correspondente						
	Ano-modelo					
	2015	2016	2017	2018	2019	2020
Mês 1	137	88	46	40	35	14
Mês 2	140	89	52	41	35	16
Mês 3	142	90	56	44	35	17
Mês 4	143	94	57	44	36	17
Mês 5	146	95	57	45	36	17
Mês 6	146	96	59	45	36	17
Mês 7	146	96	60	45	36	17
Mês 8	145	96	61	45	36	17
Mês 9	145	97	61	45	36	17
Mês 10	145	97	61	45	36	18
Mês 11	145	97	61	45	36	18
Mês 12	145	97	61	45	36	18
Mês 13	145	97	61	45	36	18
Mês 14	145	97	61	45	36	18
Mês 15	144	97	61	45	36	18
Mês 16	144	97	61	46	36	18
Mês 17	145	97	61	46	37	18
Mês 18	143	97	62	46	37	18
Mês 19	143	97	62	46	37	18
Mês 20	143	98	63	47	37	18
Mês 21	143	98	63	47	37	18
Mês 22	143	98	63	47	37	18
Mês 23	143	98	63	47	37	18
Mês 24	143	98	63	47	38	18
Mês 25	143	98	64	47	38	18
Mês 26	143	98	64	47	38	18
Mês 27	143	98	64	47	37	18
Mês 28	143	98	64	47	37	18
Mês 29	143	99	64	47	37	18
Mês 30	143	99	64	47	37	18
Mês 31	143	99	64	47	37	18
Mês 32	143	99	63	47	37	18
Mês 33	143	99	63	47	37	18
Mês 34	143	99	63	47	37	18
Mês 35	143	98	63	47	37	18
Mês 36	143	98	63	47	37	18

Figure 4. Tabela de valores válidos do Dataset.

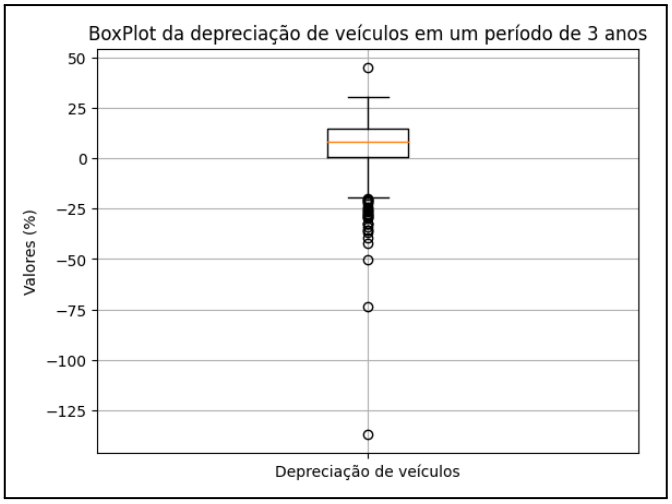

Figure 5. BoxPlot da Deprecição dos modelos.

2.4 Depreciação dos modelos

O Dataset possui uma coluna adicional que registra a depreciação dos veículos entre o valor do 1º mês e o 36º mês. A **Figura 5** exibe o boxplot com a variação da depreciação dos preços ao longo desse período. Como pode ser observado, existem alguns valores *outliers*.

3. Materiais e métodos

A construção da base de dados seguiu o processo de Extração, Transformação e Carga (ETL), também conhecido como Extract, Transform, Load [1] e [2]. Uma implementação de Web scraping foi necessário para a tarefa de extração dos dados do site da Tabela FIPE. O Web scraping é um processo automatizado de coleta de dados relevantes de sites, convertendo informações desestruturadas em estruturadas para análise posterior [3]. Essa técnica foi especialmente útil, pois eliminou a necessidade de coleta manual dos dados do site. O Web scraping foi implementado usando dois *frameworks*: Apache Airflow e o Selenium. Após o processo de Web scraping, os dados foram armazenados em um banco de dados NoSQL MongoDB e em arquivos CSV.

Os códigos do Web scrapping e MongoDB, juntamente com os arquivos gerados, estão disponíveis no **Github** para acesso público.

3.1 Apache Airflow

O Apache Airflow é uma plataforma open-source para criar, programar e monitorar fluxos de trabalho (workflows) complexos de processamento de dados. Ele permite que desenvolvedores e engenheiros de dados gerenciem e programem tarefas, que podem incluir transformação de dados, extração de dados de diferentes fontes, processamento em batch ou streaming, envio de notificações, entre outras ações.

O programa permite que o usuário crie fluxos de trabalho como DAGs *(Directed Acyclic Graphs)*, que são um conjunto de tarefas interligadas que são executadas em uma ordem

específica, permitindo que o usuário tenha maior controle e visibilidade sobre as tarefas executadas. Além disso, o Airflow possui uma interface web que permite a visualização dos workflows em tempo real, bem como o monitoramento do status das tarefas e o diagnóstico de problemas [4].

3.2 Selenium

Selenium é um *framework* de automação de testes para navegadores web. Ela permite que os usuários criem scripts em várias linguagens de programação para simular interações do usuário com um navegador web, como clicar em botões, preencher formulários e navegar por páginas.

O *framework* é comumente utilizado para automatizar testes de qualidade em sites e aplicativos web, mas também pode ser usado para coletar dados da web, como parte de projetos de Data Mining. Optou-se pelo Selenium, pois era preciso coletar informações de várias páginas da web e automatizar o processo economizaria tempo e esforço. Sua capacidade de interagir com o navegador em um nível mais profundo permite a coleta de dados inacessíveis por técnicas de mineração de dados básicas.

3.3 MongoDB

O MongoDB é um banco de dados NoSQL de código aberto que foi lançado em 2009. Ele é projetado para lidar com grandes volumes de dados [5] não estruturados e semiestruturados, que são comuns em aplicativos modernos da web e móveis [6].

4. Exemplo de utilização do Dataset

A base de dados gerada neste projeto contendo os preços de veículos mais vendidos nos últimos cinco anos pode ser usada para diversos objetivos. Modelos de aprendizagem de máquina podem ser aplicados para prever a degradação do valor de um veículo ao longo do tempo. Além de modelos preditivos, essa base de dados gerada pode ser usado para as seguintes finalidades:

- **Análise exploratória**: Realizar análises estatísticas descritivas para entender a distribuição dos preços, identificar tendências e padrões, e explorar relações entre as variáveis;
- **Análise de variação de preços**: Investigar a variação dos preços ao longo do tempo e identificar fatores que influenciam essas mudanças, como condições econômicas, sazonalidade e lançamento de novos modelos;
- **Estudos comparativos**: Comparar marcas, modelos ou categorias de veículos para identificar diferenças nos padrões de preços e comportamento de mercado;
- **Identificação de modelos populares**: Identificar os modelos de veículos mais populares com base nas vendas e preços ao longo do tempo, fornecendo insights sobre as preferências dos consumidores;
- **Análise de valor residual**: Estimar o valor residual dos veículos, ou seja, a diferença entre o preço de mercado e o valor depreciado, útil para proprietários, seguradoras e revendedores;
- **Projeção de demanda**: Utilizar os dados históricos de preços para projetar a demanda futura de modelos específicos de veículos, auxiliando na gestão de estoque e planejamento de produção.

References

[1] FERREIRA J., Miranda M., and Machado J. O processo etl em sistemas data warehouse. *INForum - II Simpósio de Informática*, pages 757–765, 2010.

[2] Shaker H. Ali EL-SAPPAGHA, Abdeltawab M. Ahmed HENDAWIB, and Ali Hamed El BASTAWISSYB. A proposed model for data warehouse etl processes. *Journal of King Saud University – Computer and Information Sciences*, pages 91–104, 2011.

[3] Igor Martins GALDINO, Erica de Lima GALLINDO, MOREIRA, and Mário W. L. Utilização de bots para obtenção automática de dados públicos usando as técnicas de web crawling e web scraping. *WORKSHOP DE COMPUTAÇÃO APLICADA EM GOVERNO ELETRÔNICO (WCGE)*, (8):172–179, 2020.

[4] DEIVE AUDIERES LEAL. Pentaho, airflow e python: AvaliaÇÃo de ferramentas para a criaÇÃo de pipeline de dados, 2023. Trabalho de Conclusão de Curso, INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO.

[5] Dharavath Ramesh, Ashay Sinha, and Suraj Singh. Data modelling for discrete time series data using cassandra and mongodb. In *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, pages 598–601, 2016.

[6] Allexandre Soares and Pablo Matos. Uma análise comparativa entre sistemas gerenciadores de bancos de dados nosql no contexto de internet das coisas. In *Anais do XXXII Simpósio Brasileiro de Bancos de Dados*, pages 306–311, 2017.