UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DE DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

LEONARDO HENRIQUE OLIVEIRA PENA

IMPUTAÇÃO DE VALORES FALTANTES EM SERIE HISTÓRICA NO CENÁRIO AUTOMOTIVO

DOIS VIZINHOS
2022

LEONARDO HENRIQUE OLIVEIRA PENA

IMPUTAÇÃO DE VALORES FALTANTES EM SERIE HISTÓRICA NO CENÁRIO AUTOMOTIVO

Trabalho apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Francisco Carlos Monteiro Souza

DOIS VIZINHOS
2022

LEONARDO HENRIQUE OLIVEIRA PENA

IMPUTAÇÃO DE VALORES FALTANTES EM SERIE HISTÓRICA NO CENÁRIO AUTOMOTIVO

Trabalho apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 22/junho/2022

Francisco Carlos Monteiro Souza
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

Rodolfo Adamshuk Silva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

Rafael Alves Paes de Oliveira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

DOIS VIZINHOS
2022

AGRADECIMENTOS

Primeiramente gostaria ao meu orientador Francisco Carlos Monteiro Souza, pelo apoio e dedicação fundamental na construção desse trabalho.

Aos meus familiares e amigos, pelo carinho, incentivo e compreensão em todas as jornadas que tracei até aqui.

Aos professores pela imensa contribuição e dedicação na transferência do conhecimento.

Agradecer a Universidade Tecnológica Federal do Paraná pela oportunidade de participar deste curso, sempre proporcionado experiências de alto nível no meu desenvolvimento pessoal e profissional.

RESUMO

PENA, Leonardo Henrique Oliveira. Imputação de Valores Faltantes em serie histórica no cenário automotivo. 2022. 25 f. – Curso de Especialização em Ciência de Dados, Universidade Tecnológica Federal do Paraná. Dois Vizinhos, 2022.

Modelos de machine learning ajudam a criar soluções há anos, nas mais diversas áreas do conhecimento. Seja na criação de uma predição de venda, ou na classificação de uma imagem. Em series temporais especificamente, conseguimos predizer o valor de um produto nos próximos meses, e com isso antecipar qualquer ação para ser mais acertiva. Porém, um dos problemas que mais frequentes encontrados é a falta de dados, isto é, quando temos uma base de dados com pouca informação ou com buracos seu histórico. Devido a isso, alguns modelos não funcionam bem e perdem valor pela falta de dados. Esse trabalho traz a proposta de solucionar o problema de dados faltantes em uma base de dados com informações de veículos. Por meio de algorímos de machine learning é feito o preenchimento dos dados ausentes, com base nos poucos dados preenchidos.
Palavras-chave: Series Temporais, Classificação, Machine Learning, Algorítmos.

ABSTRACT

PENA, Leonardo Henrique Oliveira. Missing Value Imputation in historical series in the automotive scenario. 2022. 25 f. – Curso de Especialização em Ciência de Dados, Universidade Tecnológica Federal do Paraná. Dois Vizinhos, 2022.

Machine learning models have been helping to create solutions for years, in many different areas of knowledge. Whether in the creation of a sales prediction, or the classification of an image. In time series specifically, we can predict the value of a product in the coming months, and thus anticipate any action to be more accurate. However, one of the most frequent problems encountered is the lack of data, that is, when we have a database with little information or with holes in its history. Due to this, some models do not work well and lose value due to lack of data. This work brings the proposal to solve the problem of missing data in a database with vehicle information. Through machine learning algorithms the missing data is filled in, based on the few filled in data.

Keywords: Time Series, Classification, Machine Learning, Algorithms.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de máquina
ML	Machine Learning
IA	Inteligência Artificial
C2B	Consumer to Business
B2B	Business to Business
TCC	Trabalho de Conclusão de Curso
B2C	Business to Consumer
KNN	K-Nearest Neighbors

SUMÁRIO

1 INTRODUÇÃO

O trabalho com dados faltantes é um grande obstáculo na concepção de modelos preditivos, uma vez que a maioria dos métodos estatísticos se baseiam em dados completos sem falta de valores. Considerando que os dados raramente podem ser perfeitamente recolhidos e muitos problemas, tais como falhas no processo, e mesmo erros humanos podem ocorrer, aprender a imputar dados é um passo importante na análise de dados (HANG; FONG; CHEN, 2011). Nos casos em que é necessário juntar dados sequenciais de diferentes fontes, o conjunto de dados obtidos torna-se, frequentemente, com uma quantidade considerável de dados em falta (KIM; CHI, 2018).

No caso especial das séries temporais univariadas, as técnicas populares não podem ser utilizadas, uma vez que se baseiam em correlações inter-variáveis, para estimar os valores em falta. Assim, as características das séries temporais devem ser levadas em consideração, para desenvolver uma estratégia adequada para quando se trata de dados em falta.

Muitos dos modelos de machine learning experam dados completos para funcionar, quando não há esses dados, ou seja, quando há a ausência em alguma parte, é deletado toda a observação onde possui valor faltante. O que pode causar um problema grande quando temos porções de falta muito altas em relação a base de dados, como por exemplo viés nos dados ou pouca informação.

Este trabalho tem como objetivo propor uma abordagem, com diversas técnicas para o preenchimento de séries temporais em uma base de dados que possui informações de veículos e o preço de venda nos canais C2B (Cliente para empresa), B2B (Empresa para empresa) e B2C (Empresa para cliente).

Assim, cada carro possui 3 preços, com uma lacuna de valores faltantes maiores nos canais B2B e B2C, portanto tem-se como objetivo a imputação de valores faltantes nos canais B2B e B2C com a utilização de *machine learning* e técnicas de imputação.

Este trabalho está estruturado da seguinte forma: O Capítulo 2 apresenta a fundamentação teórica. Então, os métodos de coleta de dados e todas as etapas da elaboração da solução do problema de pesquisa são detalhados no Capítulo 4. Os resultados experimentais são discutidos no Capítulo 5. No Capítulo 6 são apresentadas as conclusões e as futuras direções de pesquisa.

2 ASPECTOS CONCEITUAIS

Neste capítulo são apresentados os aspectos conceituais para a construção desta pesquisa, com base em uma revisão bibliográfica na produção acadêmica sobre o tema. Quando se lida com dados em falta nas séries temporais, é importante identificar o mecanismo de falta dos dados para escolher o método mais apropriado de imputação (KIM; CHI, 2018).

Uma série temporal é uma sequência de realizações (observações) de uma variável ao longo do tempo. Dito de outra forma, é uma sequência de pontos (dados numéricos) em ordem sucessiva, geralmente ocorrendo em intervalos uniformes. Portanto, uma série temporal é uma sequência de números coletados em intervalos regulares durante um período de tempo.

Uma série temporal pode possuir tendência, o que significa que apresenta uma longa diminuição ou aumento com o tempo, linear ou não, e pode mesmo mudar de crescente para decrescente.

Neste capítulo também será falado sobre a teoria dos dados faltantes. Quando temos ausência de valores em nossa base, há diversos casos e motivos para essa ocorrência. O estudo de dados faltantes é um problema e é pesquisado a séculos. Porém foi nos anos 1980 em que dois livros foram publicados com fundamentos e que serão as bases para a teoria (LITTLE; RUBIN, 1987) e (RUBIN, 1987).

2.1 Teoria dos dados faltantes

Um padrão de dados em falta descreve onde se encontram os valores em falta no conjunto de dados. No caso de series temporais uni-variadas, os dados faltantes estão em uma única variável.

Entender os motivos da ausência de dados é necessário para saber lidar com eles corretamente. Dado isso, abaixo é apresentado as diferentes abordagens em que é possível termos dados ausentes: MCAR, MAR e MNAR.

2.1.1 Missing Completely at Random (MCAR)

No caso de *Missing Completely at Random*, a probabilidade dos dados estarem em falta é a mesma para todos os dados. Esta probabilidade não depende nem dos dados observados nem dos não observados. Isto significa que não há lógica por detrás da perda. Os dados são perdidos por um processo aleatório, por exemplo, como resultado de uma falha do sensor.

Então supondo que a variável Y possua valores ausentes. Dizemos que esses valores são do tipo MCAR se a probabilidade de haver dados ausentes em Y não é relacionada com os valores de Y por si ou por quaisquer variáveis nos dados.

Alguns exemplos são de um questionário ter sido perdido no passado, ou de uma amostra de exame de sangue tenha sido prejudicada no laboratório.

2.1.2 Missing at Random (MAR)

No caso de *Missing at Random (MAR)* a probabilidade de perda dos dados não é a mesma para todos os dados. Esta probabilidade depende dos dados observados, mas não dos não observados. Isso significa que a ausência dos dados pode ser prevista pelos dados observados.

Por exemplo, considerando um conjunto de dados sobre rendimento e nível de educação, as pessoas com baixo nível de educação tendem a não informar os seus rendimentos. Isto implica que a ausência dos dados sobre os rendimentos pode ser prevista pela informação sobre o nível de educação.

2.1.3 Missing Not at Random (MNAR)

Os dados em falta estão relacionados com os dados não observados. Por outras palavras, a falta de dados está relacionada com fatores não tidos em conta. Por exemplo, num conjunto de dados sobre rendimentos, tanto os rendimentos mais baixos como os mais elevados não são divulgados pelos inquiridos. Não é possível definir a qual das extremidades pertencem os dados em falta, nem prever se os dados serão ou não informados.

No caso de series temporais uni-variadas, imputação MCAR e MAR são similares e desenham inferências no comportamento das variáveis sob o tempo (MORITZ et al., 2015).

2.2 Técnicas para lidar com dados ausentes

A ausência de dados traz resultados imprecisos as análises, por reduzir a representatividade da amostra e assim distorcer inferencias sobre a população.

Existe três diferentes abordagens para o tratamento de dados faltantes, eles são: Imputação, Omissão e Analise.

Essas tecnicas, respectivamente são para, respectivamente, preencher valores vazios, discartar os valores vazios da base e aplicar análises que não são afetadas pelos valores ausentes.

Neste trabalho, o foco será nas técnicas de imputação de valores ausentes.

2.3 Métodos de Imputação de dados

A imputação de dados faltantes é um campo vasto da literatura, onde muitas pesquisas estão em andamento. As técnicas populares são a Imputação Múltipla (LITTLE; RUBIN, 1987), Expectativa-Maximização (DEMIRHAN; RENWICK, 2018), Vizinho mais próximo, e os métodos *Hot Deck*. No entanto, os métodos populares de imputação dependem de correlações inter-variáveis para imputar valores em falta. No caso de séries temporais uni-variadas, variáveis adicionais não podem ser empregadas diretamente. Em vez disso, os algoritmos precisam de explorar séries temporais a fim de estimar os valores dos dados em falta. (MORITZ et al., 2015).

2.3.1 Filtro de Kalman

O Filtro de Kalman é baseado nos modelos de espaço de estados, e seguem dois estágios (FUNG, 2006).

$$x_t = F_t x_{t-1} + \epsilon_t \tag{1}$$

$$y_t = H_t x_t + \omega_n \tag{2}$$

Onde

- x_t é o vetor de estado do sistema no momento t;
- y_t é o vetor de medição correspondente ao tempo t;
- F_t é um parâmetro de transição de estado;
- ϵ_t é o termo de ruído do estado;
- H_t é um parâmetro de medição;
- ω_n é o termo de erro de medição.

Na primeira etapa, o estado do sistema e variância associada é estimada em relação a primeira expressão 1. Então, no segundo estágio, o estado estimado é atualizado com base em ambas as expressões 1 e 2 (DEMIRHAN; RENWICK, 2018).

2.3.2 `Imputação Multivariada por cadeias encadeadas`

MICE (Multiple imputation by Chained Equation) é uma técnica particular de imputação múltipla (BUUREN; GROOTHUIS-OUDSHOORN, 2011a). *MICE* funciona sob o pressuposto de que, dadas as variáveis utilizadas no procedimento de imputação, os dados em falta estão em falta ao acaso (MAR), o que significa que a probabilidade de um valor estar em falta depende apenas dos valores observados e não dos valores não observados (SCHAFER; GRAHAM, 2002).

A imputação por *MICE* segue 6 passos:

- Passo 1: Uma simples imputação, como a imputação da média, é efetuada para cada valor em falta no conjunto de dados. Estas imputações médias podem ser pensadas como *"place holder"*
- Passo 2: As imputações médias *"place holder"* para uma variável são definidas novamente para o valor em falta.
- Passo 3: Os valores observados da variável no Passo 2 são regredidos nas outras variáveis do modelo de imputação, que podem ou não consistir em todas as variáveis do conjunto de dados. Por outras palavras, a variável dependente num modelo de regressão e todas as outras variáveis são variáveis independentes no modelo de regressão. Estes modelos de regressão funcionam sob as mesmas hipóteses que se faria ao executar modelos de regressão linear, logística, ou *poison* fora do contexto de imputação de dados em falta.

- Passo 4: Os valores em falta para a variável são então substituídos por previsões (imputações) do modelo de regressão. Quando a variável é subsequentemente utilizado como variável independente nos modelos de regressão para outras variáveis, tanto os valores observados como estes valores imputados serão utilizados.
- Passo 5: Os passos 2-4 são então repetidos para cada variável que tenha dados em falta. O ciclo através de cada uma das variáveis constitui uma iteração ou *"ciclo"*. No final de um ciclo, todos os valores em falta foram substituídos por previsões a partir de regressões que refletem as relações observadas nos dados.
- Passo 6: Os passos 2-4 são repetidos durante vários ciclos, com as imputações a serem atualizadas em cada ciclo.

3 TRABALHOS RELACIONADOS

Na literatura há diversos trabalhos quanto a imputação de valores faltantes. Hoje algumas técnicas são conhecidas e utilizadas por muitos profissionais para o tratamento de dados faltantes. Há as técnicas que possuem métodos de imputação simples, como: imputação pela média ou pela mediana de uma variável, imputação pelo método do vizinho mais próximo, imputação por *hot deck*, imputação por regressão linear e máxima verossimilhança. Essas técnicas, por serem chamadas de simples, são aplicadas apenas uma vez aos dados faltantes, isso faz com que a incerteza associada ao procedimento não seja agregada as estimativas geradas pelo banco completo, o que traz uma grande limitação aos métodos (ENDERS, 2010).

Técnicas para imputação múltipla surgiram na década de 70, em que cada valor faltante é substituído por um conjunto de valores plausíveis. Com isso, várias abordagens surgiram para imputação de dados multivariados, como a especificação totalmente condicional (FCS), também conhecida como imputação múltipla por equações encadeadas (MICE) (BUUREN; GROOTHUIS-OUDSHOORN, 2011a).

Alguns trabalhos foram feitos para o estudo de valores faltantes, e para que sejam feitas abordagens para tratá-los, como ao utilizar de imputação de valores. Em 2010 foi utilizado de técnicas de estatística e *machine learning* para o auxilio na imputação de valores faltantes em dados de câncer de mama (JEREZ et al., 2010). Neste artigo foi utilizado das técnicas mencionadas acima, como: Média, hot-deck e também para imputação multipla como: *multi-layer perceptron (MLP)*, *self-organisation maps (SOM)* e *k-nearest neightbour (KNN)*. O que gerou resultados satisfatórios como conclusão. Nesse trabalho haviam 45.61% de valores faltantes na base de câncer de mama e houve uma melhor obtenção de *missing* com o método AUC.

Em 2022 foi feito um estudo com diversos artigos para entender como a academia está tratando dados faltantes em estudos clínicos. Foi concluído que apesar do problema de dados faltantes ser muito comum em qualquer pesquisa medida. A maioria dos modelos preditivos usando *machine learning* não reportam informação suficiente sobre a presença desses dados faltantes. Estratégias nas quais os dados do pacientes são simplesmente omitidos são, infelizmente, as mais comuns e mais utilizadas. Apesar disso poder gerar um viés e perca no poder analítico da predição (NIJMAN et al., 2022).

Em 2007 foram feitos comparações de métodos para a identificação de valores faltantes em problemas de classificação (GARCÍA-LAENCINA; SANCHO-GÓMEZ; FIGUEIRAS-VIDAL, 2009), para assim sumarizar e comparar métodos conhecidos para o tratamento de dados faltantes. Foi concluído que os dados ausentes podem vir de diversas formas, e que o tratamento deles é fundamental para classificação de padrões, uma vez que um tratamento inapropriado dos dados faltantes pode causar grandes erros e classificações falsas nos resultados. No trabalho foi listado as diferentes abordagens para tratamento de dados faltantes, sendo a mais fácil a *listwise*

e *pairwise deletion*, ou seja, omitindo os casos que há dados faltantes, em que nesse caso a perca de informação é a pior desvantagem. Em segundo, estariam os métodos de imputação, que seria preencher os dados ausentes com outros estimados. Nesse último pode-se dividir em duas abordagens: Abordagens estatísticas (imputação pela média ou múltipla imputação), ou por técnicas de *machine learning*, como redes neurais.

Baseado nesses estudos, como aprendizado, entende-se que o tratamento de dados faltantes é algo necessário para o presente estudo. Uma vez que pode gerar viés e problemas com nossa modelagem, como perca de informação. Com isso, esse estudo trata de métodos de imputação que sejam eficientes para series temporais, como foi feito para dados climáticos em (AFRIFA-YAMOAH et al., 2020) e utilizado abordagens apropriadas, como a suavização de kalman.

4 Metodologia da Pesquisa

A Tabela 1 possui informações de carros (marca, modelo, versão, região_uf e ano do carro) e 3 canais de venda com preço para cada carro (C2B, B2B e B2C), além do mês e ano do preço praticado. Porém, nos canais B2B e B2C há uma quantidade considerável de valores faltantes. Uma vez que tem-se 100% preenchido o canal C2B, os outros dois possuem menos de 10% preenchido.

O problema abordado é o preenchimento dos dois campos, com a solução de utilizar de métodos de imputação (filtro de kalman e *MICE*), para a obtenção de estimativas para os valores não observados.

A base de entrada se dá como, na forma (dados fictícios):

Tabela 1 – Base de Entrada

marca	modelo	versao	regiaouf	ano	valor_c2b	valor_b2b	valor_b2c
bmw	x3	2.0	SP	2003	100000	NA	NA
bmw	x3	2.5	MG	2003	120000	130000	NA
bmw	x1	2.0	SP	2005	150000	NA	NA
volkswagen	fox	1.6	SP	2003	55000	NA	60000
volkswagen	gol	1.6	SP	2003	60000	NA	NA
volkswagen	polo	1.0	DF	2014	65000	70000	75000

Fonte: Autoria própria

4.1 Linguagens de programação e pacote base

Para resolução do problema, foi feito todo o *script* na linguagem R, com auxilio do pacote *ImputeTS* (MORITZ; BARTZ-BEIELSTEIN, 2017) e do pacote *MICE* (BUUREN; GROOTHUIS-OUDSHOORN, 2011b), também da linguagem R. O pacote *ImputeTS*, de 2021, traz diversas formas de imputação de dados ausentes para series temporais univariadas, como médias móveis, interpolação e suavização de kalman (método utilizado no trabalho).

O código para imputação com kalman é feito como:

```
for(i in unique(df_final$vehicle_id)){
    filtro <- na_kalman(df_final %>% filter(vehicle_id == i))
    df_temp <- bind_rows(df_temp, filtro)
  }
```

Em que é rodado a função na_kalman, para preenchimento da base (denominada df) filtrada por carro (denominado vehicle_id), e assim concatenado em uma base chamada df_temp.

4.2 Coleta e Tratamento de Dados

Os dados são coletados diretamente do portal de vendas e comercio de carros da empresa Auto Avaliar. São valores transacionais de carros, com ano e mês de cada carro e valor de venda em cada canal.

Com os dados crus e transacionais, é feito um tratamento dos dados para que se possa ter a média de valor de cada carro mês a mês, assim gerando a tabela final com os 3 canais (C2B, B2B e B2C) com o objetivo de preencher os valores faltantes e alcançar o objetivo do trabalho. Assim, a base de dados foco deste estudo para o preenchimento de valores faltantes possui as seguintes colunas:

- marca: marca do veículo;
- modelo: modelo do veículo;
- versão: versão do veículo;
- ano_modelo: ano modelo do veículo;
- REGIAO_UF: localização do veículo, contemplando registros de estados, regiões e nível nacional (BR);
- media_c2b: valor de referência de avaliação (C2B) do veículo;
- media_b2b: valor de referência de repasse (B2B) do veículo;
- media_b2c: valor de referência de venda final (B2C) do veículo;
- ano_mês: data de referência para aquele registro. marca, modelo, versão, ano_modelo, id, região_uf, media_c2b, media_b2b, media_b2c, data.

E por fim, como saída após o fluxo de preenchimento, é criado as variáveis abaixo (explicado no tópico a seguir)

- nível_preenchimentob2b: relacionado a qual tipo de preenchimento foi utilizado para ter o valor B2B (A, B, C, D ou E);
- nível_preenchimentob2c: relacionado a qual tipo de preenchimento foi utilizado para ter o valor B2C (A, B, C, D ou E);
- faixa_carro: campo utilizado para classificar o veículo de acordo com o seu preço para realizar a imputação via nível D.

4.3 Preenchimento com Modelos de Machine Learning

Para realizar o processo de preenchimento de valores nos canais B2B e B2C, foram determinadas cinco classificações de níveis de preenchimento: A, B, C, D e E. O nível A determina que o específico registro da Tabela (identificado pela marca, modelo, versão, ano-modelo, região e data) já apresenta valores preenchidos para os 3 canais (C2B, B2B e B2C). Sendo assim, para registros com nível A, não há necessidade de passar por nenhuma modelagem tendo em vista que já possuem os valores preenchidos. Em relação aos registros de nível B, é realizada uma modelagem via método Filtro de Kalman para o preenchimento do canal que não apresenta registro para aquele mês (seja B2B, B2C ou ambos). Sobre essa modelagem, ela

se baseia em modelos de séries temporais e foi utilizada a biblioteca *imputeTS* no R.

Para os casos em que não for possível realizar o preenchimento via Kalman, é necessário estabelecer o nível C de preenchimento. Para esse nível, a modelagem estatística é realizada via Imputação Multivariada por Equações Encadeadas pelo pacote "mice"no R. Já para situações em que não seja viável a modelagem via Kalman ou *MICE* (níveis B e C), é estabelecido o nível de preenchimento D. Para esse caso, os valores para os canais B2B e B2C são estimados utilizando a taxa média de crescimento entre os canais para a marca do veículo em questão e sua respectiva faixa de preço, ou seja, é utilizada a taxa histórica obtida do crescimento do valor médio do C2B para o B2B e a taxa histórica do do crescimento do valor médio do B2B para o B2C. Em último caso, em que não seja viável e/ou não passe pelas validações a estimação dos valores B2B e B2C pelos níveis de preenchimento anteriores, é estabelecido o nível E. Nessa parte, a estimação dos preços é realizada de forma semelhante ao nível D, sendo que o único detalhe é a utilização de uma outra marca parecida para estimar, tendo em vista que não há amostra suficiente para criar a taxa de variação daquela mesma marca.

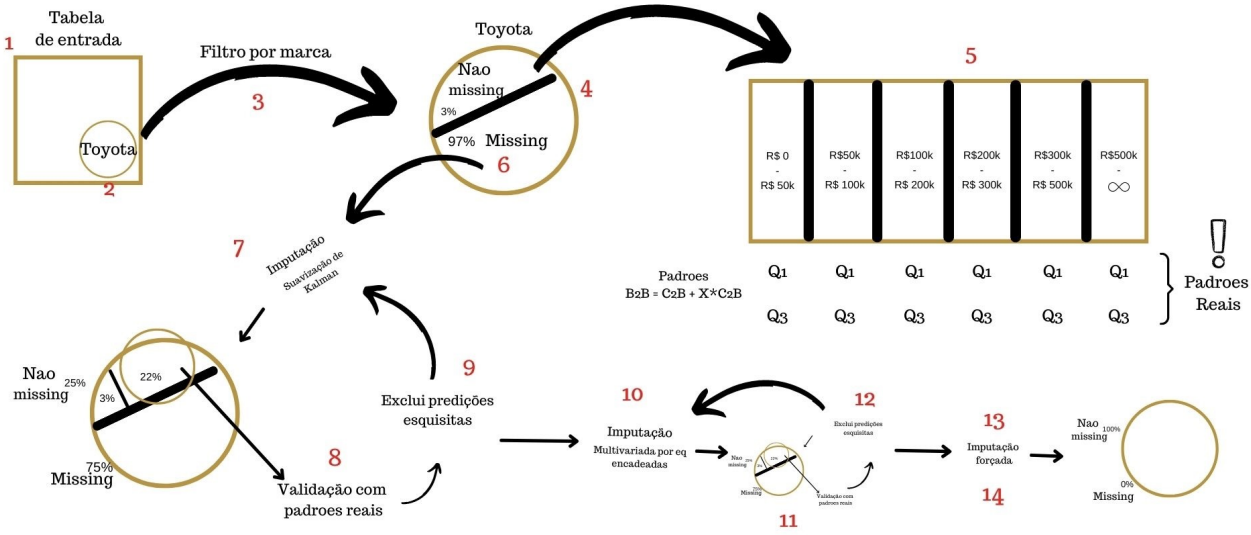
Tendo exposto todas as modelagens e formas de estimação, é importante ressaltar que para todas as etapas são utilizados critérios de validação de negócio, tais como: verificação de NA´s e a garantia do critério lógico de Valor C2B $<$ Valor B2B $<$ Valor B2C. Isto é, um canal é sempre maior que o outro na sequência (B2C $>$ B2B $>$ C2B) e a validação se dá para verificar se o preenchimento de valor faltante foi feito adequadamente de tal forma que siga essa lógica.

4.3.1 Fluxo de Preenchimento

O preenchimento de valores faltantes segue os seguintes passos, ilustrados pela Figura 1:

1. Abre as tabelas de entrada para serem preenchidas;
2. Escolhe-se uma marca m para o estudo;
3. Filtra-se a base por essa marca;
4. Divide-se a nova base entre duas sub-bases. Uma com todos os canais (c2b, b2b, b2c) preenchidos e outra com um ou mais *missing values* em algum dos canais;
5. Com a base de canais preenchida, é estudado o comportamento de crescimento entre os canais, i.e, quantos porcento normalmente sobe no valor de carros entre o canal C2B e B2B e também entre o B2B para o B2C. Esse estudo é feito individualmente para cada faixa de carro e é utilizado os quantis de resultado (quartil 1 e quartil 3);
6. Com base de valores *missing* selecionada no item 4, começamos o estudo de preenchimento;
7. Na primeira etapa é feito o preenchimento por uma Suavização de Kalman, com ajuda do *imputeTS* do R;
8. Os valores preenchidos são analisados, baseado nos conhecimentos adquiridos no item 5 e são validados para ver se faz sentido, com algumas validações do tipo: Canal C2B não

Figura 1 – Fluxo de Preenchimento.



Fonte: Autoria própria

pode ser maior que B2B, canal B2B não pode ser maior que B2C, e além disso, valores preenchidos por canais devem seguir o estudo do item 5 baseado nos quartis;

9. Os valores que não passaram pelas regras acima são excluídos e preenchidos em uma nova etapa para o item 7 e 8.
10. São realizadas i iterações no item 8, preenchendo e excluindo conforme necessário, até que não se consegue mais ter ganho com a Suavização de Kalman. Com isso é feito o preenchimento por imputação baseado no pacote "mice"do R;
11. É feito novamente as validações dos preenchimentos do *MICE*, pelo fluxo do item 8;
12. Novamente os valores que não passam no crivo são excluídos e iteramos o *MICE* j vezes;
13. Após isso, os valores restantes são preenchidos com imputação "forçada";
14. Por fim, valores que não passaram ainda em todos os anteriores, são feitos por outra imputação forçada, agora baseada em outra marca.

5 RESULTADOS

Perante o exposto, foram obtidos resultados muito bons e satisfatórios para os valores B2B e B2C via modelagem estatística ou método de imputação por taxa de variação entre os canais, sendo importante ressaltar que os resultados finais passaram por todos os critérios de validação mencionados anteriormente, garantindo a confiabilidade da informação para que se reflitam valores aderentes aos praticados no mercado.

Foi feito utilização de algoritmos de imputação e também com regras de negócio, para por fim termos uma tabela completa e preenchida 100% em todos os canais.

Para validar a quantidade de dados que foram preenchidos por cada um dos processos de modelagem, foi criada uma classificação, sendo:

- A: Dados não-ausentes, isto é, observados.
- B: Dados ausentes e imputados pelo método de kalman.
- C: Dados ausentes e imputados pelo método MICE para cada mês.
- D: Dados ausentes e imputados por variação na taxa da marca do carro.

Na Tabela 2 é possível ver a quantidade de dados que foram imputados por cada um dos níveis para o canal referência B2B. Sendo uma base de 9.718.559 dados.

Tabela 2 – Preenchimentos no canal B2B

A	B	C	D
456.961	1.929.320	1.068.338	6.263.940

Fonte: Autoria própria

E, de forma análoga, para o canal B2C, os resultados foram, conforme Tabela 3:

Tabela 3 – Preenchimentos no canal B2C

A	B	C	D
604.349	2.009.591	932.805	6.171.814

Fonte: Autoria própria

Diante disso, nota-se que foi possível preencher uma grande parcela dos dados, chegando a cerca de 3 milhões de dados com modelagem via as técnicas apresentadas (preenchimento B e C). Ou seja, anteriormente tínhamos 95% de dados faltantes no canal B2B (Soma de preenchimento B, C e D, dividido pelo número total de linhas), e tínhamos 93% de dados ausentes no canal B2C, o que foram todos preenchidos.

Porém, vale ressaltar que 6 milhões foram preenchidos apenas pela taxa de variação do mercado, diante de carros semelhantes.

Vale notar que, os resultados de preenchimentos estão dentro da faixa limitante já estipulados para que o valor do carro não seja anômalo ao mercado. Isto é, ao imputar o valor ausente ao carro X_i, é analisado se, a marca e modelo desse carro X_i (com valores de mercado e não ausentes), faz sentido o valor imputado.

6 CONCLUSÃO

Neste trabalho, buscou-se compreender o fenômeno da ausência de dados, estudando toda a teoria de ausência de dados. Assim, também estudou-se as diversas abordagens para lidar com os dados faltantes, chegando a conclusão de usar métodos de imputação para preencher os valores faltantes.

Com base nisto, foi feito algoritmos de imputação para conseguir levar as variáveis dos canais B2B e B2C ao seu preenchimento completo. Dado que é uma base de series temporais, foi feito um preenchimento baseado em suavização de kalman, e também foi utilizado, com a base filtrada no mês, de Imputação Múltipla por Equações encadeadas. Os resultados foram satisfatórios, visto que por fim conseguiu-se preencher toda a base de dados, em que tinha-se 95% de dados ausentes. Somado-se a isso, todos os preenchimentos foram validados com os valores praticados no mercado.

Apesar de toda a base estar preenchida, para trabalhos futuros, é visto como importante a diminuição de valores imputados via variação de mercado, pois uma vez que a confiabilidade é menor referente aos valores por técnicas consolidadas. Outras abordagens também seriam bem-vindas, como o estudo e a utilização de outros algoritmos.

Referências

AFRIFA-YAMOAH, E. et al. Missing data imputation of high-resolution temporal climate time series data. **Meteorological Applications**, Wiley, v. 27, n. 1, jan. 2020. Disponível em: <https://doi.org/10.1002/met.1873>. Citado na página 16.

BUUREN, S. van; GROOTHUIS-OUDSHOORN, K. bmice/b: Multivariate imputation by chained equations inir/i. **Journal of Statistical Software**, Foundation for Open Access Statistic, v. 45, n. 3, 2011. Disponível em: <https://doi.org/10.18637/jss.v045.i03>. Citado 2 vezes nas páginas 13 e 15.

BUUREN, S. van; GROOTHUIS-OUDSHOORN, K. bmice/b: Multivariate imputation by chained equations inir/i. **Journal of Statistical Software**, Foundation for Open Access Statistic, v. 45, n. 3, 2011. Disponível em: <https://doi.org/10.18637/jss.v045.i03>. Citado na página 17.

DEMIRHAN, H.; RENWICK, Z. Missing value imputation for short to mid-term horizontal solar irradiance data. **Applied Energy**, Elsevier BV, v. 225, p. 998–1012, set. 2018. Disponível em: <https://doi.org/10.1016/j.apenergy.2018.05.054>. Citado 2 vezes nas páginas 12 e 13.

ENDERS, C. K. **Applied missing data analysis.** [S.l.]: The Guilford Press, 2010. Citado na página 15.

FUNG, L. S. **Methods for the estimation of missing values in time series**. Cowan University, 2006. (Edith Cowan University). ISBN 9780471802549. Disponível em: <https://ro.ecu.edu.au/theses/63>. Citado na página 13.

GARCÍA-LAENCINA, P. J.; SANCHO-GÓMEZ, J.-L.; FIGUEIRAS-VIDAL, A. R. Pattern classification with missing data: a review. **Neural Computing and Applications**, Springer Science and Business Media LLC, v. 19, n. 2, p. 263–282, set. 2009. Disponível em: <https://doi.org/10.1007/s00521-009-0295-6>. Citado na página 15.

HANG, Y.; FONG, S.; CHEN, W. Aerial root classifiers for predicting missing values in data stream decision tree classification. In: . [S.l.: s.n.], 2011. Citado na página 10.

JEREZ, J. M. et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. **Artificial Intelligence in Medicine**, Elsevier BV, v. 50, n. 2, p. 105–115, out. 2010. Disponível em: <https://doi.org/10.1016/j.artmed.2010.05.002>. Citado na página 15.

KIM, Y. J.; CHI, M. Temporal belief memory: Imputing missing data during rnn training. In: **Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18**. International Joint Conferences on Artificial Intelligence Organization, 2018. p. 2326–2332. Disponível em: <https://doi.org/10.24963/ijcai.2018/322>. Citado 2 vezes nas páginas 10 e 11.

LITTLE, R.; RUBIN, D. **Statistical Analysis With Missing Data**. Wiley, 1987. (Wiley Series in Probability and Statistics). ISBN 9780471802549. Disponível em: <https://books.google.com.br/books?id=w40QAQAAIAAJ>. Citado 2 vezes nas páginas 11 e 12.

MORITZ, S.; BARTZ-BEIELSTEIN, T. imputets: time series missing value imputation in r. **R J.**, v. 9, n. 1, p. 207, 2017. Citado na página 17.

MORITZ, S. et al. **Comparison of different Methods for Univariate Time Series Imputation in R**. arXiv, 2015. Disponível em: <https://arxiv.org/abs/1510.03924>. Citado na página 12.

NIJMAN, S. et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. **Journal of Clinical Epidemiology**, v. 142, p. 218–229, 2022. ISSN 0895-4356. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0895435621003759>. Citado na página 15.

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. [S.l.]: Wiley, 1987. 258 p. Citado na página 11.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: Our view of the state of the art. **Psychological Methods**, American Psychological Association (APA), v. 7, n. 2, p. 147–177, 2002. Disponível em: <https://doi.org/10.1037/1082-989x.7.2.147>. Citado na página 13.