A logo with blue lines

AI-generated content may be incorrect.



**INtrodução à metodologia do data science**

Data Science aplicado

José Eduardo Pécora Junior, PhD.

Apostila didática

**APRESENTAÇÃO DA DISCIPLINA**

Sejam bem-vindos à apostila de Data Science do MBA do projeto de extensão LIGA DATA SCIENCE da Universidade Federal do Paraná (UFPR)!

A Inteligência Artificial veio para mudar o modo como fazemos as coisas. Isso se aplica no processo produtivo, no processo de compras, na gestão de recursos, na gestão financeira e em diversas áreas da administração. Esta apostila é um bom exemplo de como a inteligência artificial mudou o modo de se fazer as coisas. Para redigi-la, eu usei um algoritmo que transforma a minha voz no texto, ou seja, um algoritmo de inteligência artificial que faz a transcrição. Após, eu usei um software de revisão de texto que também é baseado em inteligência artificial.

Ao ler a apostila, vocês notarão que usei um estilo de redação de como se eu estivesse conversando com vocês. Isso foi intencional, pois este tema é inovador e pode ficar difícil rapidamente. Então uma linguagem mais pessoal facilita o aprendizado. Outro ponto importante desta apostila é que eu uso os termos ciência de dados e o anglicismo ‘Data Science’ como sinônimos, isso também foi intencional. Visto que a maioria esmagadora da literatura está em inglês, facilita para o aluno que deseja aprofundar seus conhecimentos.

A apostila foi concebida para uma pessoa leiga em ciência de dados ou linguagens de programação. Com este intuito, todos os códigos, em Python, são parte integrante dessa apostila e estão à disposição dos alunos. Eu sugiro que os alunos leiam a apostila, façam o download dos arquivos quando forem apresentados.

O objetivo desta disciplina é fazer uma introdução ao processo e à metodologia da ciência de dados, mas também ilustrar como a tomada de decisão pode ser mais eficaz.

Unidade 1 – Processo da Ciência de Dados.

Unidade 2 – Modelos para previsão de preço de vendas.

Até breve,

Prof. Eduardo Pécora

|  |  |
| --- | --- |
| Resultado de imagem para lattes | http://lattes.cnpq.br/1700029990089437 |
| Resultado de imagem para linkedin | https://www.linkedin.com/in/eduardopecora/ |
| Resultado de imagem para facebook | https://www.facebook.com/eduardo.pecora.1 |
| Ícone  Descrição gerada automaticamente | https://www.instagram.com/edupek/ |

Sumário

[Unidade 1- Processo da Ciência de Dados 4](#_Toc190290488)

[Ciência de Dados 4](#_Toc190290489)

[Metodologia de Ciência de Dados 5](#_Toc190290490)

[Unidade 2 – Modelos para previsão de preços de venda 8](#_Toc190290491)

[Antes de começar 8](#_Toc190290492)

[Nosso primeiro modelo 8](#_Toc190290493)

[Entendimento do negócio. 8](#_Toc190290494)

[Abordagem analítica. 8](#_Toc190290495)

[Requerimento de dados. 8](#_Toc190290496)

[Coleta de dados. 8](#_Toc190290497)

[Entendimento dos dados. 9](#_Toc190290498)

[Preparação dos dados. 11](#_Toc190290499)

[Modelagem 11](#_Toc190290500)

[Avaliação 11](#_Toc190290501)

[Modelagem Melhorada 13](#_Toc190290502)

[Implementação e Feedback 15](#_Toc190290503)

[Referências 16](#_Toc190290504)

[Apêndice 1 – Habilitando o Google COLAB 17](#_Toc190290505)

# Unidade 1- Processo da Ciência de Dados

## Ciência de Dados

O Analytics é o campo da ciência aplicado à coleta, processamento, análise e interpretação de dados, com o objetivo de extrair insights significativos e embasar a tomada de decisões. Se olharmos por esta lente, podemos dizer que o Analytics combina estatística, matemática, ciência da computação, a administração, a engenharia e até o design.

Tenho plena noção de que isso amplia um pouco da definição formal de ciências de dados, que é basicamente o uso de técnicas de estatística, matemática e computação para encontrar padrões nos dados. Por isso, usei a expressão Analytics ao invés de Ciência de Dados.

O analytics (Figura 1) combina o rigor da matemática e da estatística com o poder de cálculo e a organização da computação, a praticidade e a resolução de problemas da engenharia, a tomada de decisão da administração, a apresentação de gráficos e storytelling do design. Todas essas habilidades são necessárias em um cientista de dados, mas é claro que pouquíssimas pessoas no mundo têm todas elas. O mais comum é encontrar um profissional especialista em uma delas, mas como um conhecimento básico nas demais.

A blue hexagon with white text

AI-generated content may be incorrect.

Figura - Dimensões do Analytics.

Para os que buscam uma definição mais rigorosa de Ciência de Dados, podemos citar as definições da IBM ou da Oracle.

“A **Ciência de Dados** combina matemática e estatística, programação especializada, análises avançadas, inteligência artificial (IA) e aprendizado de máquina com conhecimento específico da área foco, para descobrir insights nos dados de uma organização. Esses insights podem ser utilizados para orientar a tomada de decisões e o planejamento estratégico.” IBM.[[1]](#footnote-1)

“A **Ciência de Dados** abrange a preparação dos dados para análise, incluindo a limpeza, agregação e manipulação dos dados para a realização de análises avançadas. Os cientistas de dados podem então revisar os resultados para identificar padrões através das aplicações analíticas permitindo que líderes empresariais obtenham insights bem fundamentados.” Oracle.[[2]](#footnote-2)

## Metodologia de Ciência de Dados

A diagram of a process

AI-generated content may be incorrect.

Figura - Metodologia da Ciência de Dados (Rollings, 2015)

Antes de entrarmos no detalhamento de cada uma das fases da metodologia da ciência de dados, é importante destacar que ela possui um formato circular (Figura 2). Diferente das metodologias tradicionais, que seguem uma sequência linear de passos, a metodologia de ciência de dados envolve várias idas e vindas, representadas no diagrama metodológico pelas setas em sentidos opostos. Isso significa que, muitas vezes, será necessário reavaliar decisões anteriores e, consequentemente, revisar os passos subsequentes. Sempre enfatizo que essa metodologia é baseada em experimentação e testes contínuos. Esse conceito ficará mais claro quando realizarmos nosso primeiro projeto aplicado.

Qualquer projeto de ciência de dados se inicia com uma clara definição e entendimento do negócio. Este momento é crucial para todo o desenvolvimento subsequente. Um claro entendimento do negócio normalmente vem do patrocinador do projeto. É essencial definir qual problema será tratado, quais são os objetivos do projeto e os requisitos da solução, tudo isso a partir de uma visão da organização.

A fase descrita acima vai determinar todo o alicerce da solução de ciência de dados. Ela será a base de toda a validação, avaliação e revisão. Todas as próximas fases serão sempre vistas com a ótica desse entendimento do negócio.

O erro mais comum no desenvolvimento de soluções em ciência de dados é partir direto para a modelagem. Existe um deslumbre quanto à utilização de ciência de dados e de data analytics no mundo corporativo. Não podemos nos entregar a este deslumbre; toda solução de ciência de dados deve seguir rigorosamente a metodologia. Por isso, a metodologia começa com um bom entendimento do problema e do negócio que está sendo abordado.

O próximo passo da metodologia da ciência de dados é a abordagem analítica. Uma vez que o problema estiver claramente identificado, os cientistas de dados podem definir a abordagem analítica para esse problema. Nesse estágio, expressamos o problema em um contexto de modelagem de técnicas estatísticas e de data analytics de forma geral, ajudando a organização a identificar quais seriam as técnicas mais adequadas para resolver esse problema.

O próximo passo da metodologia, são os requisitos de dados. Aqui, o cientista de dados define quais dados e em quais formatos serão utilizados, ou melhor ainda, quais dados são desejáveis. Em qualquer problema prático de ciência de dados, existe uma diferença entre os dados que seriam ideais para o nosso modelo e os dados que estão disponíveis na empresa.

Isso nos leva ao próximo passo, que é a coleta de dados. Aqui, são verificados todos os recursos de dados disponíveis dentro da empresa, sejam eles dados estruturados ou não estruturados. Esses dados podem ser provenientes de diversas fontes, como o próprio ERP (Enterprise Resourse Planning) da empresa, um WMS (Warehouse Management System), ou até planilhas e sistemas legados. Também pode ser identificada a necessidade de investimentos adicionais para manter e estruturar o banco de dados da empresa, ou até uma mudança no processo de coleta desses dados. Quero enfatizar que, mesmo com o melhor modelo de inteligência artificial do mundo, se utilizarmos dados ruins, faltantes, não tratados e cheios de erros, o resultado será uma conclusão ruim e cheia de erros. Pois o modelo busca padrões matemáticos nos dados. Em inglês, temos o princípio do GIGO (Garbage In, Garbage Out), entra lixo, sai lixo; que se aplica de forma óbvia em qualquer projeto de ciência de dados. Portanto, uma empresa que mantém seus dados bem estruturados e bem mantidos tem uma visão de futuro, pois é inegável que os dados são um ativo hoje em dia.

Na fase de entendimento dos dados, os cientistas de dados utilizam estatísticas descritivas e técnicas de visualização gráfica para compreender o conteúdo dos dados empresariais. Além disso, avaliam a qualidade desses dados, identificando possíveis lacunas e falta de padronização. É importante destacar que, durante esse processo, pode ser necessário revisar os processos de requerimento e coleta de dados, reforçando a natureza iterativa da metodologia.

Chegamos então à fase de preparação de dados, a qual sem dúvida, é a mais demorada do processo de implementação de uma solução de data analytics. Esta fase deve ser executada com uma visão na próxima etapa, a modelagem em si. O cientista de dados prepara e organiza todos os dados em um formato que o modelo possa ler. Portanto, um bom entendimento tanto de dados quanto de modelagem é necessário. As etapas características dessa fase incluem a limpeza de dados, a combinação de dados de diversas fontes e a transformação de dados em variáveis mais úteis. Se os dados estiverem em formato de texto, como logs de uma central de atendimento ou opiniões de clientes, eles precisam ser tratados por meio de uma análise de texto para serem transformados em variáveis estruturadas, enriquecendo o resultado da modelagem.

Finalmente, chegamos à tão esperada etapa da modelagem. Aqui o foco é o desenvolvimento de modelos preditivos ou descritivos, conforme definido na etapa de abordagem analítica. Normalmente, reserva-se uma parte dos dados cujo resultado é conhecido para validação e avaliação do modelo. Este processo de modelagem é raramente feito de uma única vez; sendo muito mais comum algo iterativo, com os cientistas de dados experimentando diferentes algoritmos e configurações de parâmetros, avaliando o modelo a cada passo para obter um resultado mais adequado às necessidades da organização.

Como mencionado na etapa de modelagem, a avaliação do modelo deve ser vista em comum acordo com a modelagem, em um processo interativo. Aqui, os cientistas de dados, juntamente com a empresa ou o financiador do projeto, definem os melhores indicadores para avaliar o modelo. Esses modelos são avaliados e sua eficácia testada, sempre mantendo o foco nos objetivos definidos no primeiro passo, o entendimento dos negócios.

Finalmente, chegamos à etapa de implementação e feedback. Quando o modelo é julgado satisfatório pela organização, ele deve ser implementado em um ambiente de produção para que os analistas da organização possam utilizá-lo, melhorando assim a tomada de decisões. A implementação pode variar desde a geração de um relatório com recomendações até a integração do modelo em um fluxo de trabalho ou ERP. Esta etapa pode requerer uma grande interação com o pessoal de TI da empresa.

O feedback é a última etapa da modelagem de ciência de dados, onde os cientistas de dados obtêm um retorno sobre o desempenho do modelo e seu impacto no ambiente organizacional. É importante enfatizar que a implementação de um projeto de ciência de dados é contínua, semelhante a um ciclo PDCA (Plan, Do, Control, Act). Imagine uma organização que implementou com sucesso um projeto de ciência de dados, como uma previsão de demanda. Este novo modelo de previsão, baseado em inteligência artificial, melhorará o planejamento de compras, produção, distribuição e o planejamento estratégico da organização. Com o novo modelo em operação, as decisões serão mais acertadas, o contexto da organização e os resultados melhorarão. Portanto, uma nova análise ou desenvolvimento pode ser necessário, fechando assim o ciclo PDCA.

|  |
| --- |
| **SAIBA MAIS**  [**https://www.ibm.com/downloads/documents/pt-pt/10a99803d3afd122**](https://www.ibm.com/downloads/documents/pt-pt/10a99803d3afd122) |

# Unidade 2 – Modelos para previsão de preços de venda

## Antes de começar

Como o foco deste curso é na análise e tomada de decisão, e não a programação em si, eu colocarei todas os resultados dos modelos em arquivos separados para o aluno analisar e redigir o relatório. Caso o aluno queira executar os modelos, sugiro a plataforma chamada Google Colab. Neste link temos um PDF[[3]](#footnote-3) explicando passo a passo como habilitar o Google Colab no seu computador. Caso a empresa em que você trabalha não permita o uso do Google Drive ainda existe a possibilidade de instalar o Python no seu computador, mas isso deve ser feito por alguém da TI da sua empresa.

## Nosso primeiro modelo

### Entendimento do negócio.

Imagine que você queira vender um carro, mas você precisa saber qual o valor que você poderá vender esse carro. Aparentemente, é uma decisão simples, mas se pedirmos um valor muito alto, não o vendemos, por outro lado, se pedirmos um valor muito baixo, deixamos de ganhar dinheiro.

Então podemos definir o nosso problema, como determinar o preço de venda de um veículo.

### Abordagem analítica.

A pergunta que se deve fazer neste momento seria: é possível resolver esse problema utilizando uma abordagem analítica?

A resposta curta é um simples e categórico SIM. Mas esse, sim, depende de alguns fatores, como, por exemplo, termos acesso a um banco de dados de venda de veículos. No entanto, se tivermos acesso a esse banco de dados, podemos utilizar algum método de regressão para calcular o preço de venda do veículo.

### Requerimento de dados.

Quais seriam os dados dos sonhos para que possamos rodar esse modelo?

Alguns exemplos são: fabricante, modelo, ano, estado geral do veículo, cor, potência, etc. Essa lista pode ser aumentada com inúmeras outras características. No entanto, quanto mais características colocamos nesse momento da nossa modelagem, mais difícil será a coleta.

### Coleta de dados.

|  |
| --- |
| **Para esta parte você precisará do arquivo**    **LDS\_Metodologia\_RegressaoLinear\_01.ipynb (Python)**  **A qr code on a white background  AI-generated content may be incorrect.**  **https://github.com/EduPekUfpr/PythonProject/tree/main/Metodologia** |

Os dados que usaremos vêm de uma plataforma, chamada Kaggle, onde temos vários conjuntos de dados (Data Sets) disponíveis. Especificamente para este problema, usaremos uma versão do ‘Car\_Price\_Prediction\_data’com os dados já tratados e organizados e em formato CSV que está disponível no material de aula neste link:

<https://github.com/EduPekUfpr/PythonProject/tree/main/Dados>

Na Figura 3 temos um exemplo de código em Python que coletará os dados que usaremos.

A computer screen with text and green text

AI-generated content may be incorrect.

Figura - Coleta de Dados no Python

Para coletar os dados, usaremos o método `read\_csv` da biblioteca Pandas do Python. O Pandas é uma biblioteca de leitura e manipulação de dados em Python. Como dito na introdução, esta apostila não tem a aspiração de ensinar Python, mas para quem quer aprender um pouco mais, sugiro os vídeos da Liga Data Science da UFPR, neste link:

<https://www.youtube.com/watch?v=xhdVh2Hyv3I>

### Entendimento dos dados.

Para o entendimento dos dados, usaremos a biblioteca PandasProfiling[[4]](#footnote-4)[[5]](#footnote-5), que é uma ferramenta de código aberto em Python que permite aos cientistas de dados gerar relatórios descritivos e abrangentes sobre seus conjuntos de dados de forma rápida e fácil. Ela fornece uma análise estendida de um DataFrame, incluindo estatísticas do conjunto de dados, distribuição de valores, valores ausentes, uso de memória, entre outras informações úteis para explorar e analisar dados de maneira eficiente. A Pandas Profiling é amplamente utilizada na Análise Exploratória de Dados (EDA), pois automatiza muitas das tarefas que normalmente exigiriam várias linhas de código, gerando relatórios interativos em HTML que facilitam a compreensão e a visualização dos dados. Usaremos o arquivo `DS\_Metodologia\_RegressaoLinear\_01.ipynb` na pasta:

<https://github.com/EduPekUfpr/PythonProject/tree/main/Metodologia>

Na Figura 4 temos o perfil da variável city-mpg (consumo médio do carro em milhas por galão), do canto superior esquerdo nós conseguimos ver o nome da variável e o domínio dela, que neste caso são números reais, também vemos 11 marcador dizendo que ela tem uma alta correlação com alguma outra variável. A tabela abaixo mostra algumas informações, por exemplo, a quantidade e porcentagem de valores distintos, a quantidade e porcentagem de dados faltantes, a existência de valores muito altos, a média, o mínimo e o máximo. Também mostra a quantidade de valores nulos e valores negativos, além da quantidade de memória que ele utiliza para guardar essa informação. O gráfico à direita é um histograma onde podemos ver que existem três valores que são mais frequentes do que o resto e a concentração de valores está abaixo de 30, com poucas observações acima deste valor.

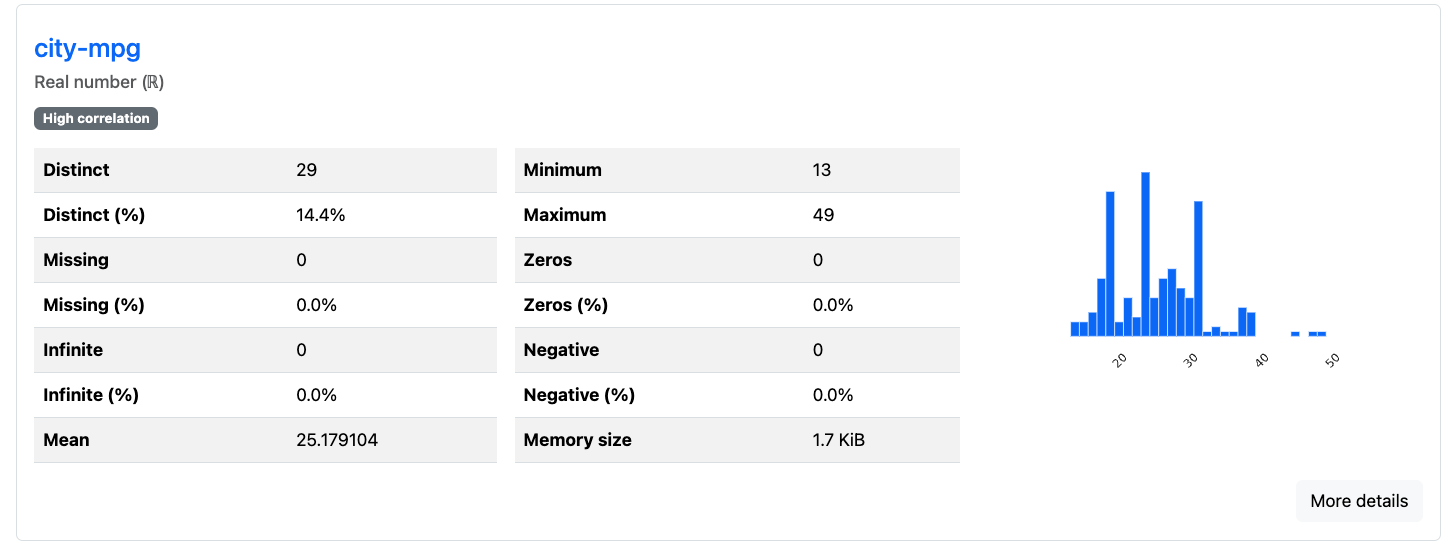


Figura - Perfil da variável city-mpg

Da mesma forma que temos o perfil desta variável city-mpg o Pandas Profiling mostra o perfil de todas as outras variáveis dos nossos dados.

Outra análise interessante é o diagrama de correlação, na Figura 5 podemos observar que tanto nas linhas quanto nas colunas temos os nomes de todas as variáveis começando com body-style e terminando com o symboling, as cores representam a correlação azul-escuro são uma correlação positiva, já o vermelho-escuro mostra uma correlação negativa. Lembra da análise anterior em que o city-mpg indicava uma alta correlação?

No diagrama da Figura 5, podemos ver que o city-mpg tem uma alta correlação negativa com as variáveis curb-weight, horsepower e price e uma correlação média positiva com highway-mpg. Confirmando e detalhando assim a informação anterior.

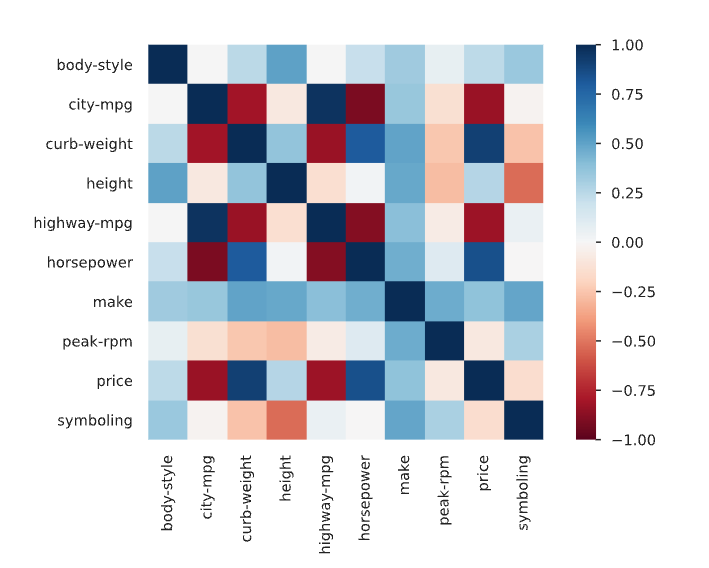


Figura - Diagrama de Correlação

### Preparação dos dados.

|  |
| --- |
| **Para esta parte você precisará do arquivo**  **LDS\_Metodologia\_RegressaoLinear\_02.ipynb**  **A qr code on a white background  AI-generated content may be incorrect.**  **https://github.com/EduPekUfpr/PythonProject/tree/main/Metodologia** |

Nesta fase, a primeira coisa que faremos será deletar as colunas que não contêm dados numéricos.

Quando trabalhamos com modelos de data Science, queremos garantir que eles funcionem bem, não apenas com os dados que usamos para treiná-los, mas também com novos dados que eles nunca viram antes. Para fazer isso, dividimos nossos dados em duas partes: treino e teste.

1. **Dados de Treino**: Usamos essa parte dos dados para ensinar o modelo. É como se estivéssemos treinando um atleta, dando a ele várias práticas para que ele aprenda e melhore suas habilidades.
2. **Dados de Teste**: Depois que o modelo foi treinado, precisamos verificar se ele realmente aprendeu bem. Para isso, usamos os dados de teste, que são como uma prova final. Esses dados não foram mostrados ao modelo durante o treinamento, então eles ajudam a avaliar se o modelo pode generalizar bem e fazer previsões precisas em novos dados.

Separar os dados dessa maneira ajuda a evitar um problema chamado **overfitting**, que ocorre quando o modelo se ajusta muito bem aos dados de treino, mas não consegue performar bem com novos dados. Em resumo, essa separação nos dá uma medida mais realista de quão bem nosso modelo irá funcionar no mundo real.

### Modelagem

Para exemplificar, usaremos um modelo simples, na verdade, o mais simples que existe, a regressão linear.

Revisão de Regressão Linear Múltipla.

Sejam:

* : é o valor da variável dependente calculada pela regressão (resultado)
* : se refere ao intercepto da regressão linear, ou seja, o valor de Y quando todos X são 0.
* : são os coeficientes de regressão, que representam a influência de cada variável X independente em

### Avaliação

Para avaliação do modelo, inicialmente, usaremos os dois indicadores, a raiz do erro quadrático médio e o R2.

* **Raiz do Erro Quadrático Médio (RMSE)**: A RMSE é uma medida da diferença entre os valores previstos por um modelo e os valores observados. Ela é calculada como a raiz quadrada da média dos quadrados dos erros. Em termos simples, quanto menor o valor da RMSE, melhor o modelo está ajustado aos dados observados.
* **Coeficiente de Determinação (R²)**: O R², também conhecido como coeficiente de Pearson, mede a proporção da variabilidade dos dados explicada pelo modelo. Ele varia de 0 a 1, onde 1 indica que o modelo explica toda a variabilidade dos dados e 0 indica que o modelo não explica nenhuma variabilidade. Em outras palavras, quanto mais próximo de 1, melhor o modelo está ajustado aos dados.

Antes de interpretá-los vamos deixar claro que os valores que aparecerem no seu computador podem ser ligeiramente distintos, pois quando fizermos a separação em dados de treino e teste o algoritmo usa números aleatórios, e portando a separação pode ser distinta se executada em computadores diferentes.

No meu caso, os valores são:

RMSE = 5547

R² = 0.748

Neste momento, não tem como sabermos se esse RMSE é alto ou baixo, pois precisaríamos comparar com um segundo método. Faremos isso em breve.

Já R² explica 74,8% da variância, o que, em termos de previsão, é considerado bom, não uma previsão excelente, mas é razoável. Visto que o modelo é o mais simples que existe, digamos que temos margem para melhorar.

Na Figura 6 temos um gráfico de densidade, chamado KDE.

* Um gráfico KDE (Kernel Density Estimate) é uma técnica de visualização de dados que estima a densidade de uma variável contínua. Ele é usado para mostrar a distribuição de dados de uma maneira suave e contínua, ao contrário de um histograma que mostra a distribuição em intervalos discretos. O gráfico KDE é útil para identificar padrões e tendências nos dados, como picos e vales, e pode ser usado para comparar distribuições de diferentes conjuntos de dados.

Interpretando o KDE da Figura 6, temos duas linhas: a linha laranja é o valor observado, ou seja, o valor atual, já a linha azul é o valor previsto pelo modelo. O eixo-x contêm os preços dos carros em dólares, no eixo-y nós temos a densidade que é uma característica do KDE. Algo que chama atenção nesse gráfico são os valores negativos tanto do atual quanto do previsto, na verdade, não temos nenhuma observação negativa, o que nós temos é um comportamento do KDE que continua a curva. Portanto, na nossa análise, nós simplesmente ignoraremos a parte negativa.

Analisando a coerência entre as duas curvas, podemos notar que para carros baratos, digamos até 6000 USD, o erro é relativamente pequeno, para carros por volta dos 10000 USD temos um erro grande, isso fica evidente devido à separação das curvas. Já para carros perto dos 15000 USD, temos um erro pequeno novamente. Finalmente, carros acima desse valor, o modelo é completamente ineficaz. O que conseguimos observar no KDE é um detalhamento de quais intervalos de preço o meu modelo é mais eficaz. Levando em conta o processo de tomada de decisão e que nunca teremos um modelo perfeito, o tomador de decisão poderia usar o modelo do modo que está para estimar os carros baratos e os carros perto de 15000 USD.

A graph of a graph

AI-generated content may be incorrect.

Figura – KDE da Regressão Linear

### Modelagem Melhorada

|  |
| --- |
| **Para esta parte você precisará do arquivo**  **LDS\_Metodologia\_RegressaoLinear\_03.ipynb**  **A qr code on a white background  AI-generated content may be incorrect.**  **https://github.com/EduPekUfpr/PythonProject/tree/main/Metodologia** |

Visto que os resultados são razoáveis e o modelo é simples, a conclusão do cientista de dados é que se usarmos um modelo mais avançado podemos melhorar os resultados. E isso é aquele ciclo na metodologia que fica entre a modelagem e avaliação. Então vamos rodar esse modelo novamente usando outro algoritmo, o XGBoost (Extreme Gradient Boosting)[[6]](#footnote-6)

O XGBoost (Extreme Gradient Boosting) é um algoritmo de aprendizado de máquina baseado em árvores de decisão, amplamente utilizado para tarefas de classificação e regressão. Ele é conhecido por sua eficiência, flexibilidade e desempenho superior em competições de ciência de dados. O XGBoost utiliza uma técnica chamada boosting, onde múltiplas árvores de decisão são treinadas sequencialmente, e cada nova árvore tenta corrigir os erros das árvores anteriores. Isso resulta em um modelo robusto e preciso.

Os resultados comparativos dos dois modelos estão na Tabela 1, onde temos três colunas, o indicador, a regressão linear e o XGBoost. Podemos ver que o RMSE do XGBoost é melhor, na verdade, ele é aproximadamente (75%) do RMSE da regressão linear. Quanto ao R2, também o XGBoost apresenta um resultado melhor, explicando mais de 90% da variância dos dados.

|  |  |  |
| --- | --- | --- |
|  | Regressão Linear | XGBoost |
| RMSE | 5547,19 | 3413,11 |
| R2 | 0,748 | 0,905 |

Tabela - Comparação de Algoritmos

Na figura 7 temos o gráfico do KDE comparando os dois algoritmos. Claramente, o XGBoost (linha verde) é melhor que a regressão linear (linha laranja) em todo o intervalo. Uma análise mais profunda revela que XGBoost tem um ótimo desempenho, até carros perto dos 21000 USD, acima disso ele tem um erro mais alto.

A graph of different colored lines

AI-generated content may be incorrect.

Figura - KDE da regressão e XGBoost

### Implementação e Feedback

Finalizadas as fases de modelagem e avaliação, temos a conclusão de que o algoritmo XGBoost é melhor, portanto, é aconselhável à empresa passar à fase de implementação deste algoritmo, utilizando-o para fazer a previsão do preço de venda dos carros. Esse novo processo de previsão do preço de venda dos carros precisa ser usado por algum tempo até que se tenha um feedback para o cientista de dados, então poderemos melhorar ainda mais a análise**.**

|  |
| --- |
| **SAIBA MAIS**  [**https://www.youtube.com/@LigaDataScience/**](https://www.youtube.com/@LigaDataScience/)  [**https://www.linkedin.com/company/liga-data-science-ufpr**](https://www.linkedin.com/company/liga-data-science-ufpr) |

# Referências

Rollings, J. B. (2015). *Metodologia de Base para Ciência de Dados*. https://www.ibm.com/downloads/documents/pt-pt/10a99803d3afd122

# Apêndice 1 – Habilitando o Google COLAB

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A white card with blue text

AI-generated content may be incorrect.

1. https://www.ibm.com/think/topics/data-science [↑](#footnote-ref-1)
2. https://www.oracle.com/what-is-data-science/ [↑](#footnote-ref-2)
3. Apêndice desta apostila [↑](#footnote-ref-3)
4. https://pypi.org/project/pandas-profiling/ [↑](#footnote-ref-4)
5. https://www.influxdata.com/blog/pandas-profiling-tutorial/ [↑](#footnote-ref-5)
6. https://medium.com/data-hackers/forecast-gradient-boost-5f91104566e2 [↑](#footnote-ref-6)