

DEZEMBRO 2021



CARTA DE CONJUNTURA DA USCS

EDIÇÃO
20

Apresentação

Com alegria, lançamos, em 15/12/2021, a 20^a Carta de Conjuntura da USCS. Gradativamente, o Observatório de Políticas Públicas, Empreendedorismo, Inovação e Conjuntura da USCS (o Conjuscs) consolida e amplia um projeto iniciado em 2018, a partir de decisão e apoio da Reitoria da Universidade.

Nestes quatro anos (2018-2021), as vinte Cartas de Conjuntura publicadas reuniram 522 notas técnicas, que promoveram reflexões, debates e subsídios às políticas públicas e privadas em diferentes áreas do conhecimento.

No momento, o Observatório expande sua atuação. Além da publicação das Cartas de Conjuntura, o Conjuscs participa do projeto de implantação de um Hub de Inovação na USCS (o “Hub USCS Biosphere”), que intensificará a aproximação entre a universidade e o mercado, na busca de soluções inovadoras para problemas concretos da realidade.

A 20^a Carta de Conjuntura da USCS contém 244 páginas e 33 notas técnicas.

Participaram desta 20^a Carta 48 pesquisadores e pesquisadoras (entre permanentes e convidados) e 46 alunos e alunas de graduação da USCS e de outras instituições.

Reafirmando sua perspectiva multidisciplinar e plural, esta Carta organiza as notas técnicas em quatro blocos:

- a) Internacional;
- b) Economia, Gestão, Inovação, Negócios, Empreendedorismo e Legislação;
- c) Educação, Cultura, Políticas Urbanas, Meio ambiente e Sociedade;
- d) Saúde.

A Carta estará disponível em:

<https://seer.uscs.edu.br/index.php/conjuscs/index>

Todas as Cartas anteriores (da 1^a a 20^a) podem também ser acessadas em:

<https://www.uscs.edu.br/noticias/cartasconjuscs>

Por fim, ao se encerrar este ano de 2021, agradecemos a todos os pesquisadores e parceiros pelas colaborações voluntárias ao Observatório na forma de notas técnicas, artigos assinados em colunas e blogs de veículos de comunicação, geração de mídia espontânea na imprensa, *lives*, entre outras.

Desejamos que 2022 seja repleto de saúde, esperança, alegrias e muitos avanços individuais e coletivos.

Coordenação do Observatório Conjuscs



OBSERVATÓRIO DE POLÍTICAS PÚBLICAS, EMPREENDEDORISMO, INOVAÇÃO E CONJUNTURA DA USCS (CONJUSCS)

Sob a Direção da Pró-Reitoria de Graduação e da Pró-Reitoria de Pós-Graduação, o Observatório é formado por professores, alunos e parceiros convidados. O Observatório tem como objetivo elaborar e publicar, periodicamente, notas técnicas no campo das Políticas Públicas, Empreendedorismo, Inovação e Conjuntura.

Expediente –20ª Carta de Conjuntura (dezembro de 2021)

Reitor: Prof. Dr. Leandro Campi Prearo

Pró-Reitora de Pós-Graduação: Profª. Drª. Maria do Carmo Romeiro

Pró-Reitor de Graduação: Prof. Ms. Silton Marcell Romboli

Pró-Reitor Administrativo e Financeiro: Prof. Me. Orlando A. Bonfatti

Pró-Reitor de Inovação em Ensino: Prof. Dr. Nonato Assis de Miranda

Líder do Grupo de Pesquisa CNPQ do Observatório: Prof. Dr. Jefferson José da Conceição

Coordenação Geral do Observatório:

Prof. Dr. Jefferson José da Conceição

Equipe de Coordenação do Observatório:

Prof.Drª. Camila Faustinoni Cabello

Prof. Dr. Jefferson José da Conceição

Prof. Me. Francisco Rozsa Funcia

Prof. Esp. Ricardo Trefíglie

Equipe de Pesquisadores Permanentes do Observatório:

Prof. Drª Camila Faustinoni Cabello.

Prof. Dr. Eduardo de Camargo Oliva

Prof. Dr. Enio Moro Júnior

Prof. Dr. Jefferson José da Conceição

Prof. Dr. José Turíbio de Oliveira

Prof. Dr. Lúcio Flávio da Silva Freitas

Prof. Dr. Milton Carlos Farina

Prof. Dr. Roberto Vital Anav

Prof. Dr. Volney Aparecido de Gouveia

Equipe de Professores Técnicos do Grupo de Pesquisa do Observatório:

Prof. Me. Daniel Giatti de Sousa

Profª. Me. Alessandra Santos Rosa

Prof. Me. Daniel Vaz

Prof. Me. David Pimentel Barbosa de Siena

Prof. Me. Luiz Felipe Xavier

Profª. Me. Marta Angela Marcondes

Profª. Me. Rosana Marçon da C. Andrade

Prof. Me. Vinícius Oliveira Silva

Profª Me. Sandra Collado

Equipe de Estudantes do Grupo de Pesquisa do Observatório:

Doutorando Adhemar S. Mineiro (UFRRJ)

Doutorando Álvaro Francisco Fernandes Neto (USCS)

Doutorando André Ximenes de Melo (USCS)

Doutorando Francisco Rozsa Funcia (USCS)

Doutoranda Gisele Yamauchi (USJT)

Prof. Me. Gustavo Kaique Araújo Monea (USP)

Doutorando Ricardo Makoto Kawai (USCS)

Pesquisadores participantes desta edição - membros integrantes e convidados do Observatório Conjusc's:

Adhemar S. Mineiro
 Adriana Letícia dos Reis
 Adriana Paulino de Oliveira
 Alessandra Santos Rosa
 Amanda Marta Jardim Souza
 Antonio Aparecido de Carvalho
 Aristogiton Moura
 Bruno Luiz Castro da Conceição
 Camila Faustinoni Cabello
 Carlos João Schaffhausen Filho
 Celoy Sene Rodrigues Silva
 Cíntia Testa José
 Cláudia Rejane de Lima
 Claudio Pereira Noronha
 Clayton Vinicius Pegoraro de Araujo
 Denise Poiani Delboni
 Eliana Rigoni
 Enio Moro Junior
 Eric Klingenhoff Berno
 Erico Filev Maia
 Fabio Luis Falchi de Magalhães
 Francisco R. Funcia
 Henrique Farias dos Santos
 Hugo do Nascimento
 Ianní Muliterno
 Inez Galardinovic
 Jefferson José da Conceição
 Karen de Matos Zampieri Campos
 Laura Cristina Pereira Maia
 Lúcia Navegantes Bicalho
 Luiz Lopes Schmidt
 Marta Angela Marcondes
 Maxime Ndecky
 Patrícia Aparecida Montanheiro
 Paulo Henrique de Mello Santana
 Rafael Marques
 Rafael Rubim de Castro Souza
 Regina Albanese Pose
 Roberto Carvalho Junior
 Robson Palma Thomé dos Santos
 Rogério Lopes
 Sidnéia Sassi
 Stefanie Sussai
 Thiago Yokoyama Matsumoto
 Vânia Viana
 Vinicius Oliveira Silva
 Vívian Machado
 Volney Gouveia

Graduandos da USCS e de outras instituições participantes desta edição

Ana Luiza Soares dos Santos
 Ana Rhara Bergemann Souza Oliva Lima
 André Centoamore Antunes
 Anna Carolina de Araújo Cassão
 Beatriz Biagioli Bertanha Bozze
 Beatriz Bom Cirello
 Beatriz Pereira de Góes
 Bruna Maria Rodrigues Yochida
 Caique Fernando de Oliveira
 Carla Petravicius Bomfim
 Carolina Rezende Barbosa Ribeiro do Vale
 Dayra Zanetti da Silva
 Eduardo Tessarolo Filho

Nota Técnica

17. UMA CONVERSA ENTRE O R E O POWER BI: A VIDA COMO ELA É NAS EMPRESAS

Adriana Letícia dos Reis⁶⁵
Eliana Rigoni⁶⁶
Ianní Muliterno⁶⁷
Regina Albanese Pose⁶⁸

Resumo Executivo

Esta nota técnica pretende discutir algumas ideias e conceitos sobre o cotidiano do mundo corporativo e como o profissional de Ciência de Dados transforma as empresas. Apresenta também um pouco da história e da vida na comunidade das R-Ladies, e ainda, alguns fundamentos estatísticos dentro de aplicações de R e de Power BI. Conceitos de estatística básica que poderiam ser explorados no mundo corporativo com o objetivo de orientar os tomadores de decisão.

Palavras-chave: Cientista de Dados; Modelos de Machine Learning; Análise de Modelos.

“Um modelo, afinal de contas, nada mais é do que a representação abstrata de algum processo, seja um jogo de beisebol, a cadeia logística de uma petroleira, as ações de um governo estrangeiro, ou o público de um cinema. Esteja ele rodando dentro de um computador ou na nossa cabeça, o modelo pega o que sabemos e usa isso para prever respostas em situações variadas. Todos nós carregamos milhares de modelos em nossas cabeças. Eles nos dizem o que esperar, e guiam nossas decisões”

O’Neil, Cathy, 2020⁶⁹

Em 15 de outubro p.p., a revista exame⁷⁰ ressalta a importância desta “nova” área para a economia digital, reforçando, que dados são, cada vez mais recursos importantes para as empresas. Decisões pautadas em negócios, produtos, gestão de pessoas, cada vez mais, são tomadas no formato chamado “data driven”, e, de forma mais “complexa”.

⁶⁵ **Adriana Letícia dos Reis.** BI Consultant at Vivo - Bacharel em Sistema de Informação – Mackenzie - MBA Economia e Gestão Empresarial – FGV - <https://www.linkedin.com/in/reis-al/>

⁶⁶ **Eliana Rigoni.** Data Analyst at Ambev Tech - Bacharel em ADM – FACAMP - Analista de Dados – IGTI - MBA ESALQ USP – cursando - <https://www.linkedin.com/in/elianarigoni/>

⁶⁷ **Ianní Muliterno.** DS/Estatística at Unilever - Bacharel em Estatística – UFPE - Técnico em automação industrial SENAI - <https://www.linkedin.com/in/iannimuliterno/>

⁶⁸ **Regina Albanese Pose.** Professora USCS e gestora do Curso de Bacharelado em Estatística e Ciência de Dados USCS -Licenciada em Matemática – FSA -Psicopedagoga clínica e institucional/lato sensu – São Marcos Mestre em Ciências FMUSP - Especialista em Poluição atmosférica e saúde humana/lato sensu – FMUSP - Bacharel em Estatística – UNICAPITAL - <https://www.linkedin.com/in/regina-albanese-pose-2300b4110/>

⁶⁹ O’Neil, Cathy. Algoritmos de Destruição em Massa - Editora Rua do Sabão, Santo André – SP, 2020 - Edição eletrônica Kindle.

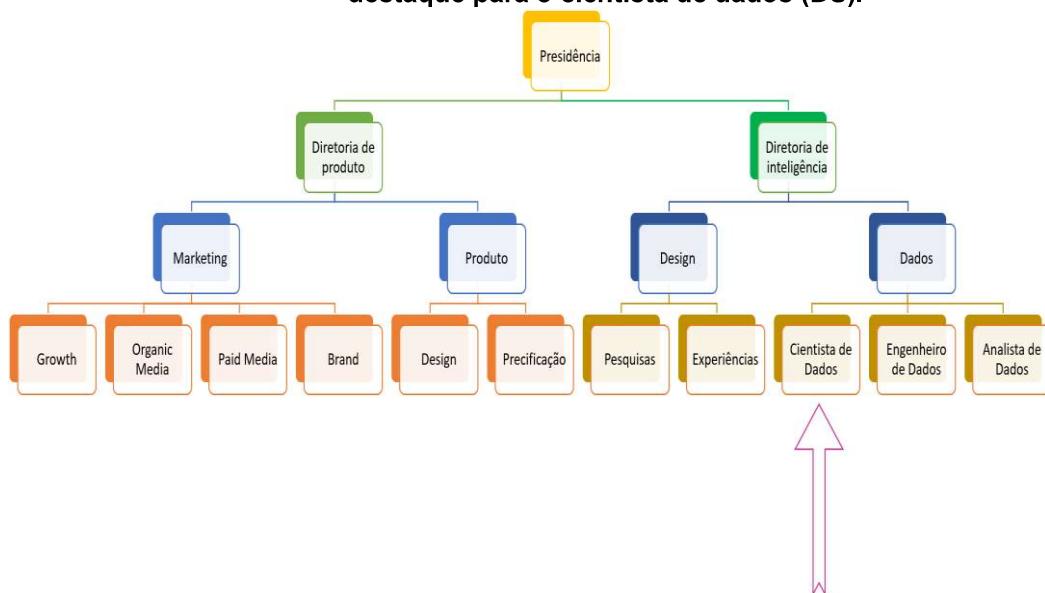
⁷⁰ <https://exame.com/carreira/como-se-tornar-cientista-dados-salario/>

O cientista de dados é um dos atores deste “novo” cenário, e, segundo relata a revista⁷¹, no relatório de 2020 do Fórum Econômico Mundial sobre profissões do futuro, além de especialistas em inteligência artificial e em Big Data. Tais profissionais são a promessa para a maior demanda de empregos, pelo menos, até 2025 a previsão é de 97 milhões de novos empregos no mundo todo. E quem são esses profissionais?

Um possível cenário pode ser compreendido, como, uma instituição que atue na área de negócios, e que, em geral, tem o engenheiro e o arquiteto de dados atuando na criação de sistemas que processam os dados da empresa, com uma intensa rotina de programação; o analista de BI (Business Intelligence), profissional com expertise em administração, produtos e negócios, que atua também com relatórios e visualização de dados; o engenheiro de ML (machine learning) que deve buscar padrões nos dados, criar modelos mais inteligentes e automatizados. Todos os atores trabalham de forma sincronizada, integrada e colaborativa⁷² (**Figura 1**).

Dentro deste cenário, pode-se compreender ciência de dados, como uma área multidisciplinar, responsável também, pela organização e análise de dados em uma instituição, seja ela qual for, de negócios, de saúde, de educação, ou outra qualquer. Ações como capturar, processar, transformar e analisar dados, estão no rol de atividades do cientista de dados. É de suma importância também, que os profissionais envolvidos, tenham habilidades para observar padrões nos diversos tipos de dados gerados e fornecidos, a fim de permitir a todos os envolvidos, que possam ter insights e então, que tomadas de decisões otimizadas possam ser realizadas⁷³.

Figura 1: Possível Cenário de atuação para profissionais da área de dados com destaque para o cientista de dados (DS).



Fonte: Autores

Uma possível história dentro da instituição (na área de business intelligence [BI]), poderia ser contada a partir de demandas de um presidente, que, em reunião com o *board* de investidores para determinar/estipular as metas de crescimento de cada diretoria, delibera algumas ações.

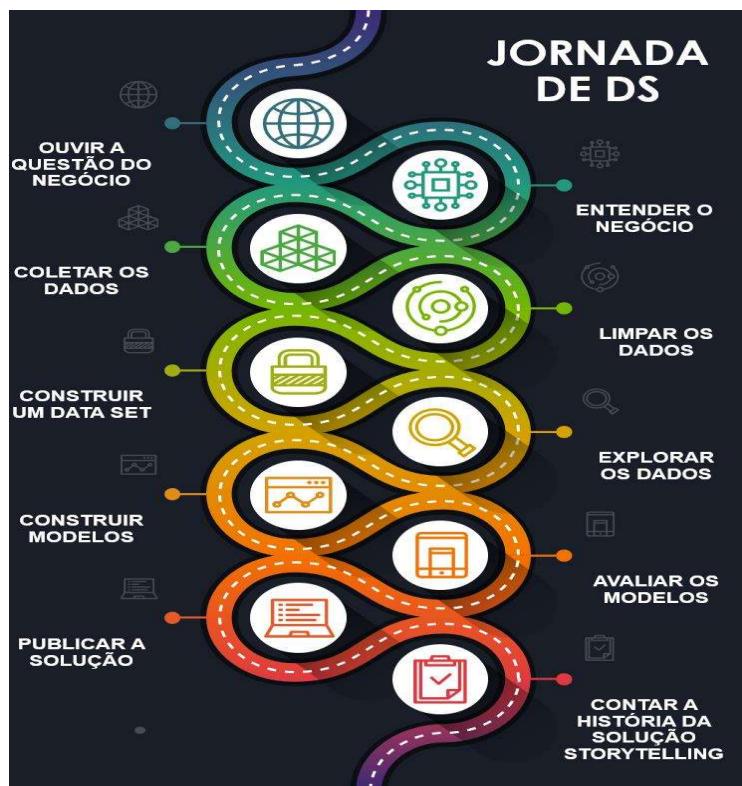
⁷¹ <https://exame.com/carreira/como-se-tornar-cientista-dados-salario/>

⁷² <https://exame.com/carreira/como-se-tornar-cientista-dados-salario/>

⁷³ <https://exame.com/carreira/como-se-tornar-cientista-dados-salario/>

As diretorias por sua vez, “quebram” as metas para os departamentos, que as fragmentam entre os times. E então, três situações, no limite, podem ocorrer, o departamento de dados pode ter a função de desenvolver esses projetos, ou ainda, pode existir um time de suporte pra outros times, que desempenhe as tarefas, ou ainda, pode existir um time específico de dados, que tem como serviço, desenvolver seu próprio produto, e que nem esteja muito vinculado a outros times. Contudo, de qualquer forma, a depender da deliberação da reunião com a presidência, o cientista de dados e seu time (ou o conjunto de cientistas de dados), devem desenvolver os projetos necessários, e, de forma integrada, colaborativa, e escalonada, as reuniões devem ocorrer na mesma hierarquia (ou não). E assim, fica a questão, qual pode ser uma possível e necessária jornada do cientista de dados? (**Figura 2**)

Figura 2: Uma jornada possível e necessária para um cientista de dados



Fonte: Autores

E, é neste espetáculo apresentado, que os atores que atuam no palco “data driven” devem olhar sempre para a plateia que os assistem, para que as cenas sejam sempre “modelos dinâmicos”, e que como concebidos e construídos por “humanos”, sempre carreguem possíveis erros e falhas, além de que, por sua própria natureza, apresentem “simplificações”. Ou seja, ainda não existem modelos capazes de incluir “TODA” a complexidade do mundo real ou as nuances da comunicação humana, inevitavelmente alguma informação sempre ficará de fora⁷⁴.

O modelo sempre é construído a partir de escolhas “humanas” sobre o que seja considerado importante para a instituição, o negócio, os consumidores, com o objetivo ainda, de “simplificar e facilitar” o entendimento e a compreensão desses consumidores e dos responsáveis pela demanda, a fim de que possam ser feitos insights e inferências para fatos e ações possíveis e necessários. Considerar sempre, quais podem ser os pontos cegos do modelo, dos resultados falso-positivos, lembrando, que, em determinadas situações, valores e desejos

⁷⁴ O’Neil, Cathy. Algoritmos de Destruição em Massa - Editora Rua do Sabão, Santo André – SP, 2020 - Edição eletrônica Kindle.

“PODEM” influenciar escolhas, ou seja, modelos refletem valores. E então, esse movimento de “avaliar a avaliação” do modelo sob os aspectos éticos, considerando, a escala, o tempo e a “plateia” que o modelo deve alcançar e em que velocidade e intencionalidade, é imprescindível.

No limite, modelos devem ser “transparentes”, interpretáveis, explicáveis e, continuamente atualizados, bem como os pressupostos utilizados, os insights das equipes, inferências realizadas, as conclusões e tomadas de decisão finalizadas. Tal prática permite que tanto os atores como a plateia possam compreender o processo todo e desta forma o objetivo do modelo pode ser compartilhado e escalável ao longo do tempo e espaço, lembrando que sempre deve ser atualizado dentro de uma política de *compliance* adotada à priori⁷⁵.

Quando o modelo “ganha escala” pode ser adotado em algum aplicativo e a metodologia toda utilizada será de alguma forma oferecida/imposta a talvez centenas de milhões de pessoas. Questões sobre “juízo de valor” devem sempre ser feitas ao longo do processo, quais sejam, o modelo funciona em que sentido/direção em relação aos consumidores do mesmo? Pode ser injusto? Pode danificar ou destruir vidas? O modelo pode perfilar uma pessoa pelas suas circunstâncias, e então pode criar um ambiente que justifique alguma premissa?⁷⁶

É importante não criar um ciclo destrutivo, pois, a cada volta completa do uso do mesmo, o processo do modelo pode, a cada novo ciclo, tornar-se mais e mais injusto⁷⁷.

Um pensamento livre desses autores, (do original “Yo soy yo y mi circunstancia, y si no la salvo a ella no me salvo yo”, de Ortega y Gasset, 1914/1966, p. 322, traduzido livremente pelos autores como, “Eu sou eu e minha circunstância, e se não salvo a ela não salvo a mim”), pode sugerir que, cada pessoa ao ser tocada por um aplicativo,— vivenciando a sua atual circunstância como um estado de ser neste momento e tempo, o que não necessariamente indica que essa pessoa SEJA como está, nessa circunstância —, pode ser modificada pelo aplicativo, de mesma forma, que a sua circunstância pode ser tocada, influenciada e modificada pela pessoa influenciada pelo aplicativo, para o bem/bom ou para o mal/mau⁷⁸.

Para o entendimento dos autores, quando Ortega Y Gasset, sugere que a circunstância deve ser salva, talvez ele esteja sugerindo que cada pessoa deva ser salva, por si mesmo, como destacado no livro *Meditaciones del Quijote*, em que o autor sugere que cada pessoa deve “buscar o sentido do que o rodeia (do original, “Es decir, buscar el sentido de lo que nos rodea” - tradução livre dos autores)⁷⁹.

Pautado nessas reflexões e estudos, os autores sugerem mais uma possível jornada para o cientista de dados (e todos os profissionais que, de alguma forma estejam no processo de desenvolvimento do modelo), que deve ser acrescentada à primeira, e a todas as demais, bem como, às regras de compliance, carregada de valores e intencionalidades (**Figura 3**).

⁷⁵ O’Neil, Cathy. Algoritmos de Destrução em Massa - Editora Rua do Sabão, Santo André – SP, 2020 - Edição eletrônica Kindle

⁷⁶ O’Neil, Cathy. Algoritmos de Destrução em Massa - Editora Rua do Sabão, Santo André – SP, 2020 - Edição eletrônica Kindle

⁷⁷ O’Neil, Cathy. Algoritmos de Destrução em Massa - Editora Rua do Sabão, Santo André – SP, 2020 - Edição eletrônica Kindle

⁷⁸ Ortega y Gasset, J. (1966). *Meditaciones del Quijote*. In *Obras completas de José Ortega y Gasset* (7a ed., Vol. 1, pp. 310-400). Madrid: Revista de Occidente. (Trabalho original publicado em 1914)

⁷⁹ Ortega y Gasset, J. (1966). *Meditaciones del Quijote*. In *Obras completas de José Ortega y Gasset* (7a ed., Vol. 1, pp. 310-400). Madrid: Revista de Occidente. (Trabalho original publicado em 1914)

Nesta jornada, o DS vive no mundo real agressivo e competitivo, e todo o tempo está envolvido no desenvolvimento e na compreensão de modelos que ele faz ou que ele deve entender e renovar, que, de forma geral, podem ser divididos em modelos não supervisionados e modelos supervisionados, ou, como são chamados no “jargão” da área, de “caixa preta” e “caixa branca”, respectivamente.

Figura 3: Uma jornada possível e necessária para o cientista de dados agregando valores.



Fonte: Autores

Os modelos de aprendizagem (ML, Machine Learning), muito utilizados hoje em dia, dadas as facilidades propostas pela tecnologia, já eram conhecidos há cerca de 30 anos, contudo, os computadores não estavam à altura da teoria desenvolvida, por isso não eram tão disseminados. São hoje conhecidos como “Machine Learning” (Aprendizado de Máquina), e fazem parte das técnicas utilizadas na Inteligência Artificial (área que entrega soluções, produtos e/ou serviços pautados em “aprendizagem” de algoritmos matemáticos e/ou estatísticos)⁸⁰.

Modelos de Machine Learning surgem com a necessidade de automatizar tarefas pautadas pelo processamento e análise orientados por dados, e que, podem simular o comportamento humano, de forma a entregar valores fundamentados em um código de ética, conforme supracitado em seção anterior⁸¹.

A intencionalidade destes modelos deve permitir que humanos façam inferências ou previsões, a partir de novos conjuntos de dados, sempre buscando padrões que expliquem relacionamentos entre os dados, e que tenham como propósito principal, o “aprendizado”, ou seja, a capacidade de adaptar, modificar e melhorar o “comportamento” desses dados, sem que sejam infringidas as regras de compliance (que devem ser pré-estabelecidas antes da reunião da presidência, conforme supracitado). Ainda, treinar, construir, formular ou induzir um modelo de conhecimento (ou seja, fazer com que “ele aprenda”), orientado por dados, e buscar padrões, para se fazer uma estimativa, teste ou predição de valores desconhecidos

⁸⁰ ESCOVEDO, Tatiana e KOSHIYAMA, Adriano, Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise. Casa do Código, 2020.

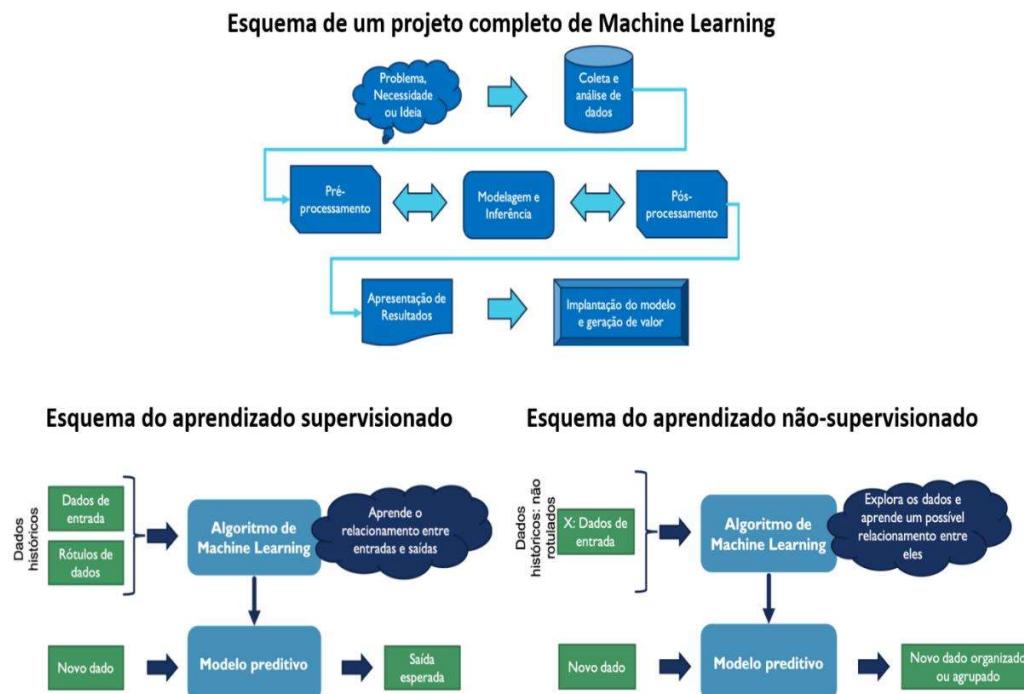
⁸¹ ESCOVEDO, Tatiana e KOSHIYAMA, Adriano, Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise. Casa do Código, 2020.

para atributos de conjuntos de dados. Esses modelos podem ser supervisionados e não-supervisionados⁸² (**Figura 4**).

Os modelos ditos supervisionados (caixa branca), devem ser construídos a partir de dados de um dataset (dados de entrada) em geral, apresentados como pares ordenados, ou seja, entrada oferecida e saída desejada. E então, esses dados podem ser rotulados, dado que a saída pode ser esperada. Ou seja, o DS deve apresentar para o algoritmo a ser desenvolvido, um número suficiente de situações (registros ou instâncias) de entradas e saídas desejadas (rotuladas à priori), cujo objetivo deve ser aprender uma regra geral que possa mapear as entradas em saídas; de forma técnica, são exemplos, os problemas de classificação e de regressão⁸³.

Para os modelos não-supervisionados, não existe a informação dos rótulos históricos (não são esperadas saídas à priori), e, o algoritmo não recebe treinamento de resultados esperados durante o treino. O algoritmo nesse caso, deve “descobrir sozinho”, pautado na exploração dos dados “imputados”, os possíveis relacionamentos entre eles. Este processo então, pretende identificar regularidades entre os dados a fim de agrupá-los ou organizá-los em função de similaridades encontradas entre os diferentes atributos dos dados; de forma técnica, são exemplos, problemas de clusterização (agrupamento) e de associação⁸⁴.

Figura 4: Workflow de ML e especificações de ML supervisionado e não supervisionado



Fonte: Modificado de Escovedo e Koshiyama pelos autores.

E então, continuando na jornada do DS, no episódio do chamado à aventura (**Figura 3**), deve-se considerar o desenvolvimento desses modelos, tanto os supervisionados como os não

⁸² ESCOVEDO, Tatiana e KOSHIYAMA, Adriano, Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise, Casa do Código, 2020.

⁸³ ESCOVEDO, Tatiana e KOSHIYAMA, Adriano, Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise, Casa do Código, 2020.

⁸⁴ ESCOVEDO, Tatiana e KOSHIYAMA, Adriano, Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise, Casa do Código, 2020.

supervisionados, de forma transparente, ou seja, de forma que possa ser compreendido e que tenha valores explícitos, conforme descrito nas seções anteriores.

E o episódio da ajuda necessária para o DS, pode vir de plataformas, ambientes e softwares livres e abertos, que, contam com muitas vantagens, quais sejam, a melhoria continua de todo o processo desenvolvido e disseminado por comunidades mundiais que são compostas por mentores com disponibilidade para auxílio necessário, e para o desenvolvimento de códigos, métodos e técnicas como um movimento de educação permanente; promovendo a cultura de estabilidade e confiança na utilização de códigos fontes estáveis; contudo, para que tudo isso ocorra, é necessário que o usuário (o unicórnio, o DS, em sua jornada), saiba escolher com qualidade e fundamentação teórica, o que usar e o que deve fazer, e, que em parte pode ser iluminado e facilitado pelos mentores das comunidades. Além disso, a empresa deve ter sempre um time com expertise e que seja sênior, para auxiliar cada novo unicórnio que chega. Para além de tudo isso, alguns cuidados devem ser tomados. Pois, é, sempre necessário ter cautela com possíveis riscos, por exemplo, com soluções ou produtos que sejam de forma legal adquiridas por empresas proprietárias, e que, embora abertos, não sejam livres de pagamentos, além disso, o que já foi largamente discutido nesta nota, acerca da responsabilidade legal, de como desenvolver os modelos com valores reais para a sociedade, e, colaborar com essas comunidades, deve sempre ser prioridade, e, cada membro deve evoluir, e tornar-se mentor, colaborando com a educação permanente de cada novo integrante, bem como, com a produção de conteúdo e conhecimento.

Desta forma, a travessia, a provação e a superação do unicórnio em sua jornada poderão trazer para a sociedade/comunidade uma solução, um produto, um serviço de IA (inteligência artificial interpretável e explicável, promovendo uma grande virada na situação real em que iniciou sua trilha, otimizando o processo e promovendo uma cultura de aprendizagem permanente, integrativa e colaborativa.

Alguns modelos de ML ainda são muito dificilmente explicáveis, então, novas técnicas e tecnologias podem ser utilizadas, como painéis (dashboards) mais ágeis como os de propriedade da Microsoft® ou mesmo da Google ®. Além de técnicas que facilitem a compreensão dos modelos, mesmo que desenvolvidas em linguagens mais complexas como por exemplo, o R e o Python.

Com este intuito, estes autores se candidataram na comunidade em que frequentam, **R-Ladies-SP**, quando a Escola de Modelos de Regressão (EMR), um evento científico na área de Estatística, de repercussão nacional, em sua 17ª edição convidou a comunidade para apresentação de um trabalho nos dias 29,30 de novembro e 1,2 de dezembro, p.p.⁸⁵.

A comunidade **R-Ladies-SP** é parte de um projeto maior, que foi concebido pela estatística brasileira Gabriela de Queiroz em 2012, nos EUA, onde ela atua até hoje. "Ela queria retribuir à comunidade depois de ir a vários encontros e aprender muito de forma gratuita." A comunidade tem vários capítulos funcionando no mundo todo em muitos Estados no Brasil, que cresce exponencialmente, conta com cerca de 50 países, mais de 170 cidades (capítulos), mais de 62k membros, que já realizaram mais de 2k eventos oficiais, além de tantos locais e/ou dentro das comunidades⁸⁶.

R-Ladies é uma organização mundial cuja missão é promover a diversidade de gênero na comunidade R. A comunidade R sofre de uma sub-representação de gêneros minoritários (incluindo, mas não se limitando a mulheres cis/trans, homens trans, não binários, genderqueer, agender) em todas as funções e áreas de participação, seja como líderes, desenvolvedores de pacotes, palestrantes/participantes de conferências, educadores ou

⁸⁵ <https://eventos.ibge.gov.br/emr2021>

⁸⁶ <https://rladies.org/>

usuários. Como uma iniciativa de diversidade, a missão da **R-Ladies** é alcançar uma representação proporcional ao encorajar, inspirar e capacitar pessoas de gêneros atualmente sub-representados na comunidade R, construindo uma rede global colaborativa de líderes R, mentores, alunas e desenvolvedores para facilitar o progresso individual e coletivo em todo o mundo⁸⁷.

R-Ladies dedica-se a fornecer uma experiência livre de assédio para todos. Não são tolerados assédio de participantes de nenhuma forma. Existe um código de conduta que se aplica a todos os espaços **R-Ladies**, incluindo encontros, Twitter, Slack, listas de e-mail, tanto online quanto offline. Qualquer pessoa que violar este código de conduta pode ser sancionada ou expulsa desses espaços a critério da Equipe de Liderança Global. Alguns espaços **R-Ladies** podem ter regras adicionais em vigor, que serão claramente disponibilizadas aos participantes. Os participantes são responsáveis por conhecer e cumprir essas regras.

E neste ambiente, de forma democrática, colaborativa e integrada, esses autores desenvolveram o minicurso de mesmo título desta nota, a partir de vivências próprias de cada um⁸⁸, na “vida como ela é nas empresas”.

E, para finalizar esta nota, será apresentado um roteiro breve fundamentado para quem quiser compreender como “a mágica é feita”, e, para maiores interessados, só procurar qualquer um dos autores no linkedin.

Roteiro da conversa entre R e Power BI: a vida como ela é nas empresas

- 1. PASSO ZERO:** Para iniciar um trabalho orientado por dados é importante compreender o ciclo de dados (**Figura 5**).
- 2. CENÁRIO⁸⁹:** A base de dados contém informações fictícias dos consumidores da Telco comunicações e está disponível de forma aberta.
- 3. OBJETIVO** desta análise: Prever o comportamento do consumidor (ações em prol de manter os consumidores)
- 4. CONTEXTUALIZAÇÃO:** CHURN: neste caso, carrega a ideia de deixar a empresa. Assim, o objetivo deve ser a questão: o que a empresa pode fazer para reduzir a *probabilidade* do cliente sair da empresa? Quais áreas podem ajudar nesse serviço?
- 5. PREPARAR O DATASET:** limpar e organizar os dados recebidos
- 6. EXPLORAR:** Análise descritiva exploratória inicial (**Figuras 6 e 7**).
 - 6.1.** Validar evidências de poder preditivo (IV E WOE)⁹⁰: técnicas que permitem a interpretação e compreensão do modelo (**Figuras 8 e 9**)
 - 6.2.** Usar o Power BI para fazer as análises descritivas exploratórias de forma mais rápida e objetiva (**Figuras 10 e 11**)

⁸⁷ <https://rladies.org/>

⁸⁸ Um dos autores é professora universitária e estuda e aprende na comunidade sobre BI e aplicação da estatística nessa área.

⁸⁹ https://www.kaggle.com/blastchar/telco-customer-churn?select=WA_Fn-UseC_-Telco-Customer-Churn.csv

⁹⁰ <https://towardsdatascience.com/model-or-do-you-mean-weight-of-evidence-woe-and-information-value-iv-331499f6fc2 - e - https://medium.com/mlearning-ai/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building-3b3b98efe0e8>

7. ANÁLISE INICIAL: Perfil de Churner=> a análise descritiva exploratória permite apresentar à instituição hipóteses sobre “o que esperar do data set”, e, com as medidas de validação de evidências, é possível discutir alguns insights sobre potenciais fraquezas que facilitem o churn (desistência da assinatura), tais como (análise real feita com esse data set fictício, depois de todas as análises realizadas pelos autores e discutidas em equipe), lembrar que, todas essas hipóteses devem ser largamente discutidas com profissionais com expertise na área de telecom:

- 7.1. Contrato mês a mês
- 7.2. Pagar com cheque eletrônico
- 7.3. Ter pagamento superior a \$58,4
- 7.4. Serviço de internet mais caro: Fibra ótica
- 7.5. O cliente que não se interessa por serviços adicionais

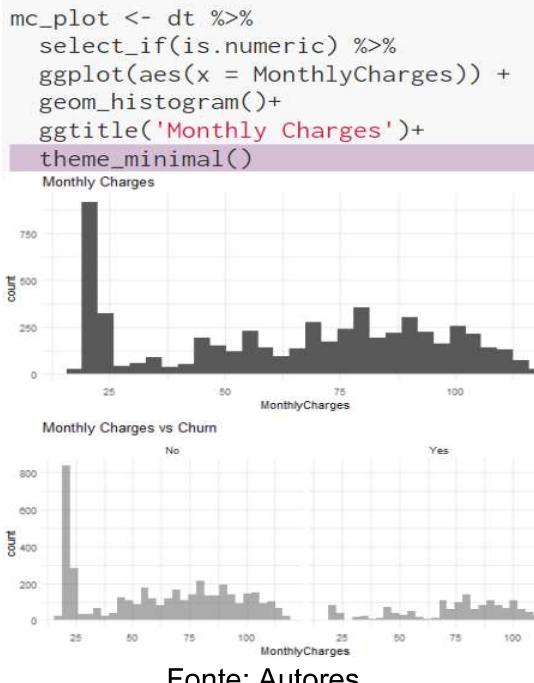
8. CONSTRUIR MODELOS COM AS VARIÁVEIS DE INTERESSE: Construir modelos até que se encontre o modelo “campeão, com as variáveis entendidas como importantes na fase exploratória

- 8.1. Preparar as variáveis para o modelo (**Figura 12A**) técnica de dummyficação
- 8.2. Modelar: Testar o modelo xgboost (ML com RL) (**Figura 12B**)
 - 8.2.1. Regularizar o modelo (**Figura 12C**)
- 8.3. Avaliar a performance do modelo - usar curva ROC, para verificar verdadeiros positivos e falsos positivos; Lift Metric, para avaliar o planejamento de ações; expectde maximum profit measure, para verificar a estimativa para valor do consumidor, essa métrica ajuda a “entregar” o threshold (limiar) ótimo e o lucro esperado no momento do estudo(**Figuras 13, 14 e 15**) – o ideal seria que cada vez mais as cores estivessem separadas, tal que o vermelho mais à esquerda e o verde mais à direita – a probabilidade de churn para essas variáveis é muito alta – para a curva ROC pode-se compreender como critério estrito, os valores com evidência forte, aqueles com uma pequena fração de falsos positivos e, com uma relativa pequena fração de verdadeiros positivos, no canto no canto inferior esquerdo da curva ROC; critérios menos estritos, conduzem a maiores frações de ambos os tipos, no canto superior direito da curva ROC; ou ainda, quanto maior a área da curva ROC, maior a capacidade discriminante do modelo; quanto mais próxima da linha diagonal mais aleatório é este modelo, ou seja, não se sabe se pode ser considerado discriminante ou não, e, o índice de Youden indica o melhor cutoff, aquele que otimiza a classificação dos verdadeiros positivos e dos verdadeiros negativos da curva no espaço ROC.
- 8.4. Interpretar modelos técnica SHAP => alocação ótima do crédito, e análise combinatória (Teoria dos jogos, Lloyd Shapley, Nobel de economia) - Esta técnica avalia a contribuição de cada “jogador”, dado um jogo cooperativo e uma aliança de jogadores, sendo capaz de estimar uma “distribuição justa” dos ganhos para cada participante (**Figura 16**)- indica a alta probabilidade do cliente feliz “dar churn” quando avaliadas as variáveis de interesse.

Figura 5: Ciclo de Ciéncia de Dados com tidyverse⁹¹

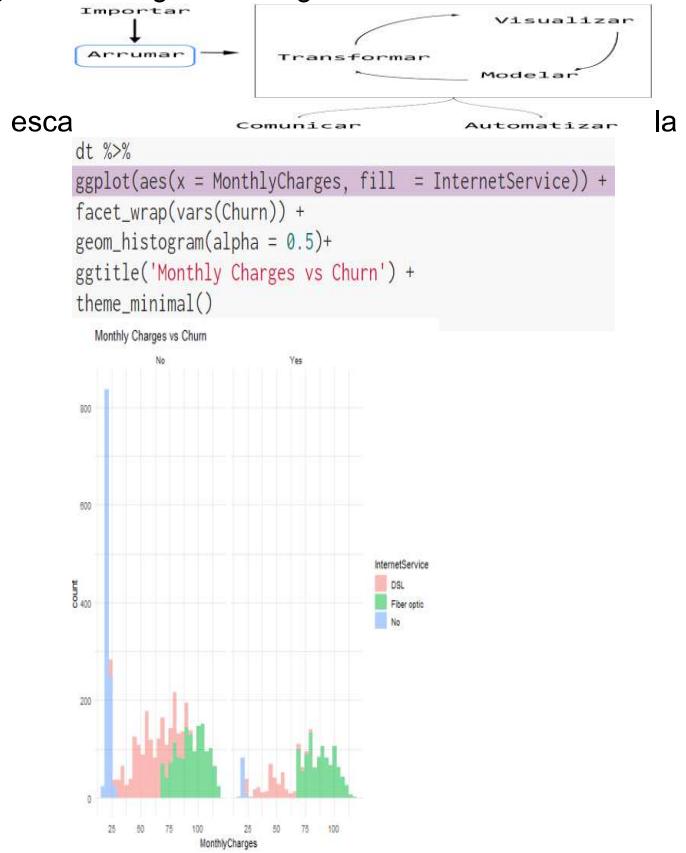
Fonte: Curso-R⁹² modificado pelos autores

Figura 6: Código em R e gráficos iniciais



Fonte: Autores

Figura 7: Código em R e gráficos destacados com mesma l

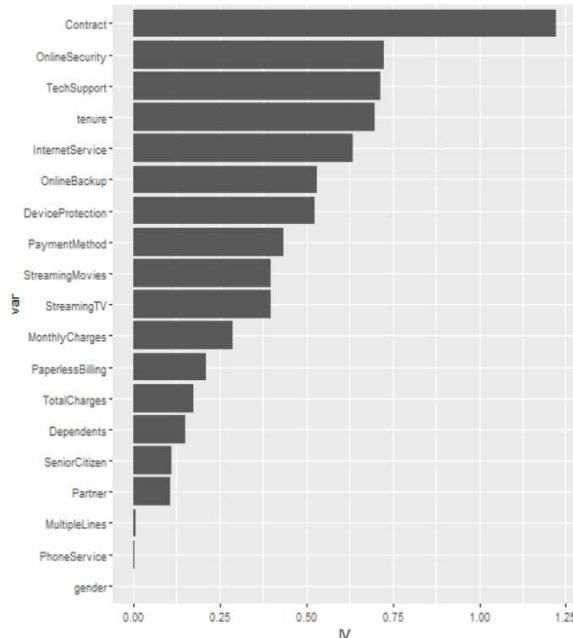


Fonte: Autores

⁹¹ <https://www.tidyverse.org/>

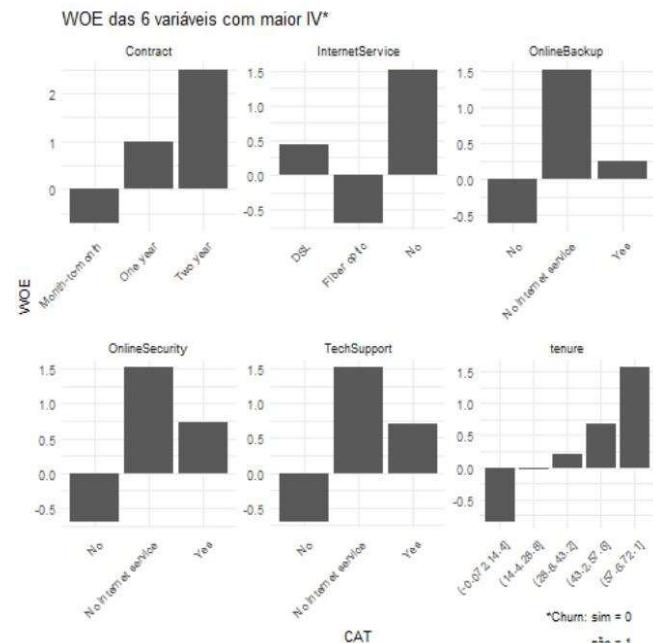
⁹² <https://livro.curso-r.com/7-manipulacao.html>

Figura 8: IV para os atributos do DataSet



Fonte: Autores

Figura 9: WOE apresenta o impacto dos atributos na taxa de churn



Fonte: Autores

Figura 10: Análise descritiva exploratória no Power BI





Fonte: autores

Apresentar um dashboard no Power BI vai agilizar o processo todo, e facilitar as análises mais detalhadas, por ser feito com gráficos mais simples e visuais que todos comprehendem e por ser mais rápido de construir.

Figura 12:Código para Dummyficação

```

leng_vars <- sapply(dt,
                      function(x) ifelse(is.character(x),n_distinct(x),NA))

leng_vars <- leng_vars[!is.na(leng_vars)]

categoricas <- dt %>%
  select_if(is.character) %>%
  select(names(leng_vars[leng_vars > 2])) %>%
#  select(-customerID, -Churn)
  select(-customerID)

dummy <- dummyVars(" ~ .", data=categoricas)
newdata <- data.frame(predict(dummy, newdata = categoricas))

numericas <- dt %>%
  select_if(is.numeric)

entrada_xgb <- bind_cols(newdata,numericas)
entrada_xgb <- janitor::clean_names(entrada_xgb)

```

Fonte: Autores

Figura 12A: Código para modelar XGBOOST

```

index_treino <- sample(1:nrow(entrada_xgb), nrow(entrada_xgb)*0.75)

treino <- entrada_xgb[index_treino,]
validacao <- entrada_xgb[-index_treino,]

treino <- xgb.DMatrix(as.matrix(treino)
                      , label = ifelse(dt$Churn[index_treino] == 'Yes', 1, 0)
                      )

validacao <- xgb.DMatrix(as.matrix(validacao)
                           , label = ifelse(dt$Churn[-index_treino] == 'Yes', 1, 0)
                           )

```

Fonte: Autores

Figura 12B: Código para regularizar modelo XGBOOST

```

dt %>%
  summarise( sum(Churn == 'No')/sum(Churn == 'Yes'))

##   sum(Churn == "No")/sum(Churn == "Yes")
## 1                         2.760516

watchlist <- list(train = treino, eval = validacao)
parametros = list(
  max_depth = 3,
  eta = 0.15,
  colsample_bytree = 1,
  subsample = 1,
  objective = "binary:logistic",
  eval_metric = 'auc',
  scale_pos_weight = 2.76)

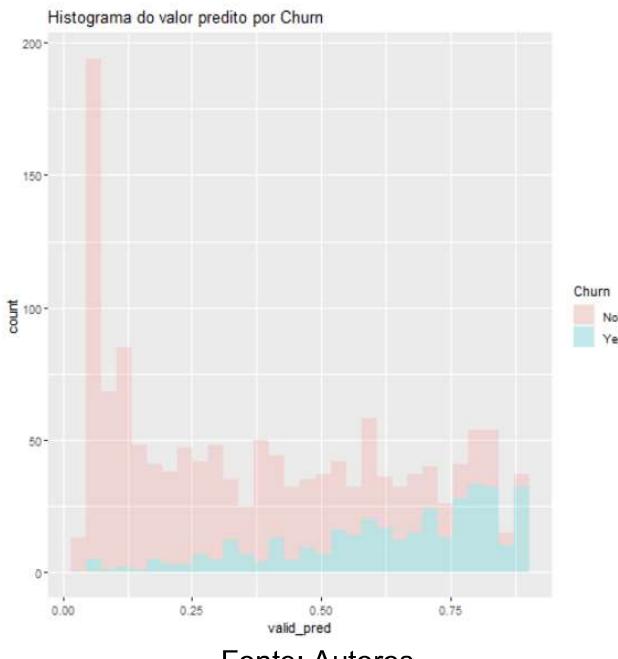
modelo <- xgb.train(data = treino,
                     nrounds = 20,
                     params = parametros,
                     watchlist)

## [1]  train-auc:0.824866    eval-auc:0.805865
## [2]  train-auc:0.830431    eval-auc:0.812493
## [3]  train-auc:0.842469    eval-auc:0.822613

```

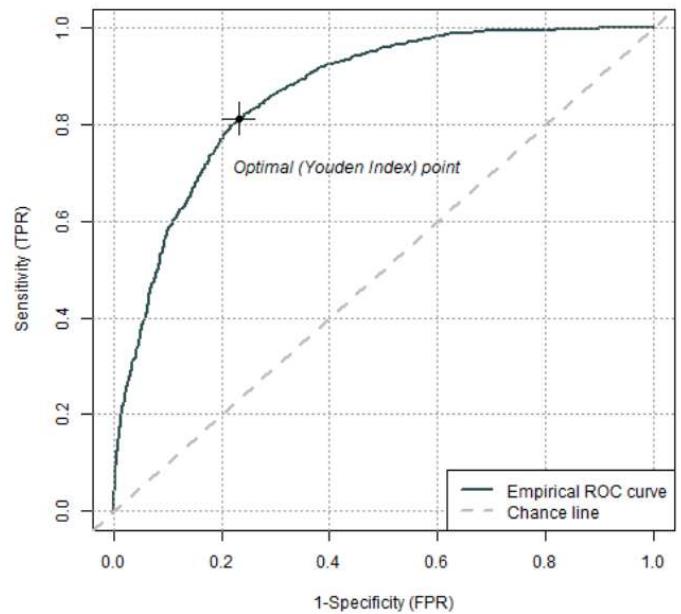
Fonte: Autores

Figura 13: Avaliar performe com histograma do valor predito



Fonte: Autores

Figura 14: Avaliar performance com curva ROC/AUC e Youden index



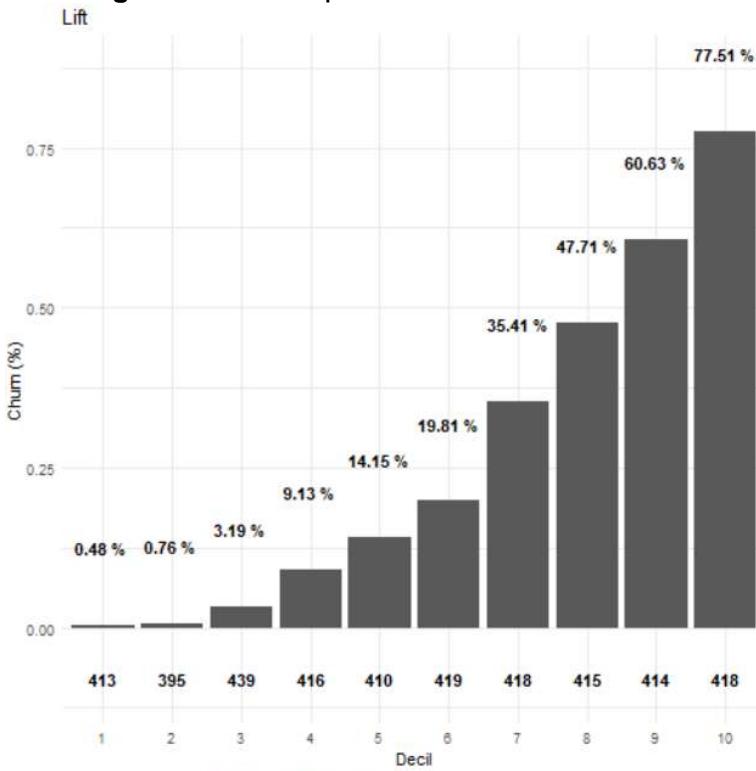
Índice de melhor trade off entre sensibilidade $P(\hat{y} = 1|y = 1)$ e especificidade $P(\hat{y} = 0|y = 0)$

$$J = \max_c(Sensibilidade_c + especificidade_c - 1)$$

```
library(ROCIt)
ROCIt_obj <- rocit(score=treino_pred, class=dt[index_treino,]$Churn)
```

Fonte: Autores

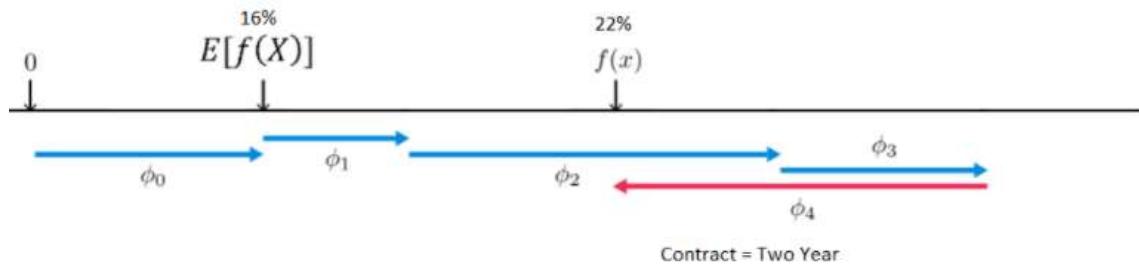
Figura 15: Avaliar performance com lift metric



```
## # A tibble: 2 x 2
##   Churn `mean(MonthlyCharges)`
##   <chr>      <dbl>
## 1 No          60.9 
## 2 Yes         74.3
```

Fonte: Autores

Figura 16:Técnica Shap



Fonte: Autores

Este gráfico indica a alta probabilidade do cliente feliz com a assinatura “dar chum” quando são avaliadas as variáveis de interesse (supracitadas). Este gráfico deve ser analisado de forma global e local para a compreensão e validação do modelo, de forma “interpretável e explicável”.

Referências Bibliográficas

- ✓ Download do R
 - <https://cran.r-project.org/>
- ✓ Download do RStudio
 - <https://rstudio.com/products/rstudio/download/#download>
- ✓ Passo a passo da instalação (livro Curso-R)
 - <https://livro.curso-r.com/1-instalacao.html>
- ✓ Dowload do Microsoft Power BI
 - <https://powerbi.microsoft.com/pt-br/>
- ✓ Power BI
 - <https://community.powerbi.com/t5/Community-Blog/Some-Ideas-and-Techniques-for-Customer-Churn-Analysis-in-Power/ba-p/991121>
 - <https://docs.microsoft.com/pt-br/power-bi/connect-data/desktop-r-scripts>
 - [Power BI Desktop Installer Changes & WebView2 | Blog do Microsoft Power BI | Microsoft Power BI](#)
- ✓ R & Power BI
 - Executar scripts do R no Power BI Desktop - Power BI | Microsoft Docs
- ✓ R-Studio
- ✓ Vídeos com Prof Samuel Macêdo
 - https://www.youtube.com/watch?v=-f_xB4mDxOc&t=631s
- ✓ Livros, dicas, blog e comunidades em Curso-R
 - <https://curso-r.github.io/zen-do-r/index.html>
- ✓ R-studio gráficos usados no power BI
 - <https://www.r-graph-gallery.com/circular-barplot.html>
 - <https://statisticsglobe.com/graphics-in-r>
 - <https://r4ds.had.co.nz/>
 - <https://r4ds.hadley.nz/>
 - <https://bookdown.org/ndphillips/YaRrr/where-did-this-book-come-from.html>
- ✓ Cores nos gráficos (e daltonismo)
 - <https://marcusnunes.me/posts/como-alterar-as-cores-em-um-grafico-ggplot2/>
 - <https://karloguidoni.com/post/paletas-de-cores-disponiveis-no-r/>
- ✓ Pacote XGBOOST
 - <https://medium.com/applied-data-science/new-r-package-the-xgboost-explainer-51dd7d1aa211>
 - <https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27>
 - <http://dx.doi.org/10.1145/2939672.2939785>
 - <https://medium.com/@gabrieltseeng/gradient-boosting-and-xgboost-c306c1bcfaf5>
 - T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System , 2016
 - J. Friedman, Greedy Function Approximation: A Gradient Boosting Machine 1999
 - Ihler, Ensembles: Gradient Boosting Youtube video , 2012
- ✓ Shap
 - <https://medium.com/@gabrieltseeng/interpreting-complex-models-with-shap-values-1c187db6ec83>
 - <https://www.r-bloggers.com/2019/03/a-gentle-introduction-to-shap-values-in-r/>

- <https://towardsdatascience.com/one-feature-attribution-method-to-supposedly-rule-them-all-shapley-values-f3e04534983d>
- <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>
- S. Lundberg, S Lee, A Unified Approach to Interpreting Model Predictions , 2017
- <https://www.r-bloggers.com/2019/03/a-gentle-introduction-to-shap-values-in-r/>
- ✓ Estatística
- ✓ Regressão Logística - livro eletrônico aplicado a área de negócios
 - <https://smolski.github.io/livroavancado/index.html>
- ✓ Modelagem - livro teórico
 - https://www.ime.usp.br/~giapaula/texto_2013.pdf
- ✓ Ciência de dados (aplicada)
 - <https://cienciadedados.github.io/dados/>
- ✓ Machine Learning
 - <https://christophm.github.io/interpretable-ml-book/>
- ✓ Termos da área de software engineering
 - IEEE. Ieee standard glossary of software engineering terminology, 1990
- ✓ Área de negócios
 - <https://github.com/IBM/telco-customer-churn-on-icp4d>