ANÁLISE DE IMPACTO DE INVESTIMENTO EM MÍDIAS DIGITAIS EM CONVERSÕES ONLINE UTILIZANDO APRENDIZADO DE MÁQUINA

MARCELO NUNES NOGUEZ[1]; MARILTON SANCHOTENE DE AGUIAR[1]

[1]*Universidade Federal de Pelotas – {mnnoguez, marilton}@inf.ufpel.edu.br*

1. INTRODUÇÃO

Empresas de diferentes nichos e mercados estão cada vez mais a procura de soluções digitais automatizadas, que agreguem inteligência nas decisões de negócios. Personalização da jornada do usuário, classificação de perfis e integração de dados online e offline são termos e expressões em voga. Também nunca se houve tantas formas de investimentos para um negócio, e um formato comumente utilizado hoje em dia é o da mídia paga. Para disputar posição num mercado de alta concorrência, as empresas e marcas apostam cada vez mais nas novas mídias digitais para evidenciar seu produto ou marca em detrimento ao do concorrente. Diferentes estratégias online são utilizadas, tais como: anúncios em redes sociais, motores de busca e portais através de mídia programática, um formato de compra e venda de espaços publicitários na *web*; mídia display, formato de anúncios em vídeos e banners online; e, *addressable media*, formato de publicidade que conecta marcas a consumidores individuais, permitindo estratégias personalizadas para cada consumidor.

A medida em que a quantidade de informações e dados disponíveis sobre o negócio de uma dada empresa cresce ao longo do tempo, a capacidade humana para a geração de valor neste mesmo ramo tende a diminuir. Para alcançar a excelência na tomada de decisões e se manterem competitivas no mercado, empresas são levadas a serem *data-driven* e a utilizarem inteligência de negócios. No contexto de análise de dados, *Business Intelligence* é o processo de entrega de decisões lógicas e executáveis de negócios através da manipulação analítica de dados e apresentação de informações dentro de um determinado ambiente de negócios (SHERIF, 2016). Para estas tarefas, diferentes métodos de análise podem ser aplicados e, dentre eles, *Machine Learning*.

Machine Learning é um ramo da Inteligência Artificial e Ciência de Dados, bem como um campo da ciência da computação, no qual visa o aumento de performance de um algoritmo para uma determinada tarefa através da experiência. Por definição, diz-se que um computador aprende através da experiência E realizando alguma classe de tarefas T e com a mensuração de performance P, caso sua performance nas tarefas T, mensuradas por P, melhore com a experiência E (MITCHELL, 1997). Diferentes métodos no campo de aprendizagem de máquina podem ser utilizados para gerar valor e inteligência de negócio e, dentre deles, a Regressão se destaca como uma técnica que nos permite utilizar diferentes fontes de dados para predizer resultados futuros.

Este trabalho propõe verificar a existência de uma relação causal entre dados de mídia e visitas online para um determinado automóvel A, de uma empresa do setor de automóveis, aplicando a técnica de Regressão Linear Múltipla. Esta análise se torna útil para saber se determinados investimentos de marketing retornam de fato para o *awareness* da marca na forma de consultas e visitas ao seu negócio online. E, caso seja verificada a influência dos canais de mídia nas visitas ao site, validar também se existe aproximação entre o resultado apresentado pelo modelo e os dados esperados. Por serem dados sensíveis a marca, para efeito de análise eles serão representados de maneira genérica.

A Regressão Linear Múltipla é uma técnica de regressão no qual se assume uma dependência linear entre a variável dependente (a ser predita) e o conjunto de variáveis independentes. O resultado é uma simplificação dos dados reais para uma função linear, com cada variável independente sendo atrelada a um coeficiente, que denota a sua significância para variável observada (dependente) de acordo com o modelo. Já o conjunto de dados utilizado representa diferentes investimentos em canais de mídia digital, tais como *video on demand*, programática e anúncios em *display*, pesquisas online redes e em sociais, e TV ao longo de dois anos, coletados mensalmente. Os dados a serem preditos são o número de visitas ao website da marca para o mesmo intervalo de tempo.

O artigo organiza-se da seguinte maneira: na Seção 2, segue o panorama geral da solução utilizando Regressão Linear Múltipla e a metodologia utilizada para atingir os resultados esperados. Os resultados e as discussões relacionadas ao trabalho são apresentados na Seção 3 e, por fim, a conclusão é apresentada na Seção 4.

2. METODOLOGIA

Para realizar a validação da relação entre as visitas a página do veículo A no website e os investimentos em mídia para ele, foi desenvolvida a abordagem proposta na plataforma Google Colaboratory, um serviço baseado em nuvem para a implementação de *notebooks* (similar a arquivos de código) Python, versão 3.6.9. De acordo com BISONG (2019), o Google Colaboratory é um recurso para prototipar modelos de *machine learning* com opções potentes de *hardware* como GPUs e TPUs, provendo um ambiente Jupyter *notebook* sem servidor para desenvolvimento interativo. Todos os dados coletados ao longo de 2 anos foram compilados em um *dataset*, contendo 14 colunas com diferentes canais de mídia, 1 coluna com as visitas e 1 coluna com a data das coletas. Os canais de mídia, para efeitos de análise, serão retratados de forma genérica, devido à natureza de seu significado. O tratamento dos dados foi realizado através das bibliotecas Pandas e NumPy. Pandas é uma biblioteca de código livre especializada em tratamentos de dados comumente voltados a ciência de dados para a linguagem Python. Já NumPy é uma biblioteca que contém estruturas de dados e funções para computação numérica. A biblioteca Scikit-learn (PEDREGOSA et al., 2011), para a linguagem Python, foi responsável por disponibilizar todas as técnicas de estatística e aprendizado de máquina utilizadas no desenvolvimento deste trabalho.

Para a validação de relação entre os canais de mídia e as visitas, foi utilizada a técnica de RFECV (*Recursive Feature Elimination with Cross-validation*) que nos permite criar um *ranking* dentre as características (canais de mídia), de maior a menor relevância para a variável a ser observada (visitas ao site). Resumidamente, a técnica RFECV permite verificar o número ótimo de características para uma determinada variável, dado um método de validação cruzada, um modelo de aprendizado e uma métrica de validação. Para o método de validação cruzada, foi utilizada a técnica de *K-fold cross-validation* (KFCV), que permite estimar o erro de predição de maneira mais fidedigna para conjuntos menores de dados, que é o escopo desde trabalho. O funcionamento completo da técnica pode ser encontrado em (HASTIE et al., 2009), porém, resumidamente, o conjunto de dados é dividido entre K partes de igual tamanho. Para um dado subconjunto K do conjunto de dados divididos em K partes, os K-1 subconjuntos restantes são utilizados para treinar o modelo preditivo, enquanto o erro de predição é calculado utilizando-se como referência o subconjunto restante, K.

Para o modelo de aprendizado, foi utilizada a técnica de Regressão Linear Múltipla, já citada anteriormente. A métrica de validação utilizada (*score*) foi R^2, onde a melhor pontuação é 1,0, podendo ser negativa (o modelo pode ser arbitrariamente ruim). Um modelo constante que desconsidera as características de entrada tem uma pontuação de 0,0.

Após a aferição dos canais de mídia que possuem maior relação com as visitas ao site, foi realizada a exclusão de características inadequadas ao modelo do veículo. Utilizou-se o subconjunto de canais de mídia com a maior pontuação (*score*) e foram testados todos os subconjuntos possíveis sem permutação.

Para o objetivo desde trabalho, após a fase de seleção dos canais, foi realizada a construção do modelo ideal, treinado com o subconjunto de dados dos canais escolhidos pela etapa anterior, utilizando Regressão Linear Múltipla. Por fim, com o modelo pronto, 4 cenários foram avaliados: 1) orçamento real mantido com redistribuição de gastos, dando maior peso aos canais mais bem ranqueados pelo RFECV; 2) cortes de 50% e 3) 75% do orçamento real, mantendo a redistribuição de gastos do cenário 1; e, 4) aumento de 15% do orçamento real, também mantendo a redistribuição de orçamento. A análise dos cenários foi feita através da biblioteca Matplolib. A Matplotlib é um pacote de geração de gráficos 2D utilizado para desenvolvimento de aplicações e scripts interativos em Python (HUNTER, 2007).

3. RESULTADOS E DISCUSSÃO

A técnica RFECV elencou 9 das 14 características (ou canais) originais como tendo o melhor valor de R^2, obtendo um resultado de -2,3 como *score* mais alto. Coeficientes negativos indicam contribuição negativa as visitas ao site. Já o *score* negativo indica que o modelo não teve boa performance correlacionando os canais de mídia com as visitas ao site. Mesmo com a melhor configuração de canais, ou seja, que melhor beneficiam as visitas, o modelo elencou os canais C, D, G, H como afetando negativamente a visitas ao site, tendo destaque negativo para o canal D, com coeficiente de -0,27. Os valores negativos podem estar relacionados ao fato do *dataset* utilizado ser esparso, possuindo muitos valores faltantes. Porém, treinando o modelo com os canais escolhidos por RFECV, se constatou uma aproximação fidedigna aos dados de visitas reais, como pode ser observado na Figura 1.

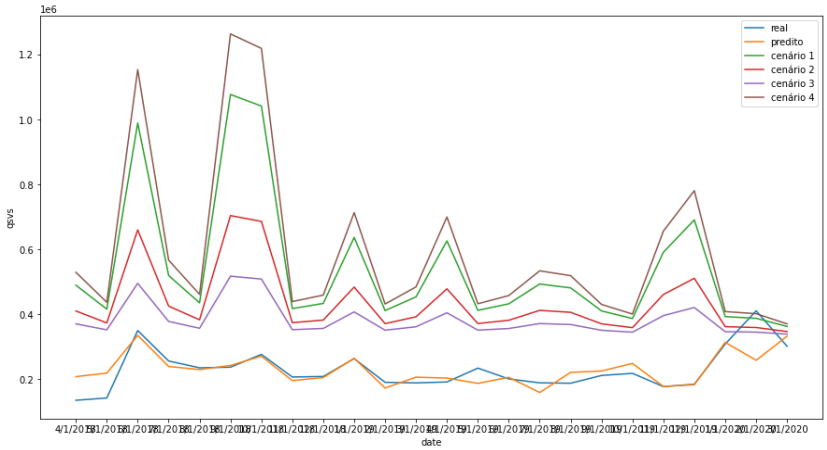

Figura 1. Resultados do modelo e cenários

A Figura 1 retrata valores de visitas ao longo do tempo, tanto reais quando preditos. É possível notar que os valores preditos para as visitas (linha laranja) se

mantêm fiéis aos resultados de visitas reais (linha azul), não se aproximando apenas de valores mais deslocados da média, nos primeiro e penúltimo meses. Para todos os cenários executados, o modelo prediz valores maiores de visitas, comparado aos valores reais. Todos os cenários analisados têm a mesma distribuição de orçamento entre as *features*, dando maior peso para os canais mais bem ranqueados. Os melhores resultados foram produzidos nos cenários 1 e 4 (linhas verde e marrom). Mesmo os resultados dos cenários 2 e 3 (linhas vermelha e roxa) que preveem cortes consideráveis de orçamento de 50 e 75% respectivamente, verificou-se um resultado com maior número de visitas que o real, para os meses observados. No cenário mais otimista (cenário 4), verificou-se um aumento de até 5,51x mais visitas do que a média de valores reais, relativamente expressivo levando em conta que visitas são nas casas dos milhares.

4. CONCLUSÕES

Este trabalho apresentou uma análise do impacto dos investimentos de mídia na propriedade digital da marca, utilizando *machine learning*. Como resultado, verificou-se que o modelo treinado, por mais próximo que fosse dos valores preditos aos reais, produz *score* negativo para o método de regressão utilizado, mesmo com a utilização de técnicas que buscavam elevá-lo.

Ainda, verificou-se a dificuldade do modelo de regressão linear em relacionar os dois conjuntos de dados, possivelmente pelo fato de o conjunto de dados para os canais de mídia possuir muitos valores nulos, e baixo alcance de dados. Notou-se que o modelo também tem a tendência a estimar sempre valores reduzidos para as visitas, dado o *score* negativo, mesmo para o melhor caso (apenas utilização de *features* que mais contribuem para a variável observada).

Apesar do *score* obtido levar o modelo a uma tendência de subestimar a predição, isto é, rebaixando os valores preditos, o modelo garante uma cota inferior de visualizações, produzindo resultados consideráveis na prática. Como trabalho futuro, pretende-se utilizar diferentes técnicas de regressão e estender a análise para os diversos modelos de veículo da marca, explorando suas relações com os diferentes canais de mídia digital.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BISONG, E.O. Google Colaboratory. In: BISONG, E.O. **Building Machine Learning and Deep Learning Models on Google Cloud Platform**. Berkeley: Apress, 2019. Cap.7, p.59-64.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning (2nd edition)**. Springer-Verlag, 2009.

HUNTER, J. D. “Matplotlib: A 2D Graphics Environment”. **Computing in Science & Engineering**, vol. 9, no. 3, p. 90-95, 2007.

MITCHELL, T.M. **Machine Learning**. McGraw-Hill: New York, 1997.

OLIPHANT, T.E. **A guide to NumPy**. EUA: Trelgol Publishing, 2006.

PEDREGOSA, et al. “Scikit-learn: Machine Learning in Python”. **JMLR**, v.12, p. 2825-2830, 2011.

SHERIF, A. **Practical Business Intelligence**. Birmingham: Packt Publishing Ltd, 2016.