Marcelo Veloso Maciel

Um estudo de caso do uso de mineração de dados e aprendizado de máquina no aprimoramento de inspeções de estações radio base

Brasil

Marcelo Veloso Maciel

Um estudo de caso do uso de mineração de dados e aprendizado de máquina no aprimoramento de inspeções de estações radio base

Trabalho de conclusão

Universidade de Pernambuco – UPE

Residência Tecnológica em Inteligência Artificial

Brasil

List of Figures

Figure 1 – Número de Abonados vis-à-vis Avaliados pré e pós balanceamento . . . 8
Figure 2 – Distribuição de acurácias. Acurácia mediana anotada em cada caixa. . 9
Figure 3 – Matriz de confusão num banco de teste de 60%. 1 é “Abonado”. 10

List of Tables

Contents

Introdução 5

1 Descrição do Caso 6

2 Solução proposta 7

Conclusão 11

Bibliography 12

Introdução

Nas últimas décadas a temática do impacto social da inteligência artificial vem tomando centralidade no imaginário prospectivo do cidadão médio, da comunidade científica e dos agentes estatais (1, 2, 3). A ascensão do assunto na opinião pública não é desconexa de mudanças no contexto econômico e político (4). A difusão da internet na sociedade, culminando nas tecnologias IoT (5), faz com que dados passem a ser consideradas pela The Economist [1] o novo petróleo.

Esse papel dos dados pressupõe a capacidade dos agentes econômicos de extrair valor deles. É essa a seara de inserção dos algoritmos de inteligência computacional, particularmente os de aprendizado de máquina. Algoritmos de aprendizado de máquina são aqueles que aprendem com uma experiência com relação a alguma tarefa e uma medida de performance se a performance na tarefa melhora com a experiência (6). Se os dados são o novo petróleo então os algoritmos utilizados para extrair informação e aprender com esses dados podem ser considerados os novos motores da economia.

O presente estudo apresenta um caso de sucesso da aplicação de sistemas inteligentes de recuperação e análise de informação de relativa simplicidade no aprimoramento de um processo rotineiro na indústria de telecomunicações: a inspeção de estações rádio base. O restante do trabalho está estruturado da seguinte forma

[1] Fonte: <https://tinyurl.com/y39u52kk>. Acessado em 1 de Novembro de 2019 .

1 Descrição do Caso

Como referenciado anteriormente o sistema alvo de interesse do nosso estudo está inserido no âmbito da indústria de telecomunicações. Na rede de celulares a mediação entre o celular dos usuários e as companhias telefônicas é feita pelas Estações Rádio Base (doravante ERB ou sítio celular). São nesses sítios que estão instalados os equipamentos necessários para a comunicação entre aparelhos celulares e as centrais de comunicação das agências telefônicas. Nesses ambientes são realizadas vistorias frequentes tendo em vista sua relevância para a qualidade do serviço de telefonia. Nessas vistorias são checados itens referentes às chaves do sítio, à rua de acesso, alarmes externos, aterramento, baterias, cabos, fontes de energia, antenas, dentre centenas outros. Essa vistoria é um trabalho conjunto entre técnicos que visitam os sítios e engenheiros de telecomunicação que analisam as informações. Atualmente essa troca de informação é feita da seguinte maneira: o técnico visita a ERB e para cada item de um *checklist*, que tem até 600 itens a depender da empresa de telefonia detentora do sítio, tiram fotos que são enviadas a um sistema, onde são aceitas ou rejeitadas pelos engenheiros na central. Contudo, nem todo item precisa ser checado a depender de condições particulares da ERB. Estes itens são, portanto, abonados.

Em conversas com técnicos e engenheiros responsáveis pelas inspeções foram identificadas ao menos duas possibilidades de aplicação de inteligência computacional no aperfeiçoamento do processo: a definição de quais itens são abonados e quais são aprovados ou rejeitados. O problema da dispensa do item, enfoque do presente trabalho, é que os técnicos não sabem de antemão quais itens devem ser abonados em um determinado sítio. Ao chegarem a ERB, desta forma, primeiro devem checar dentre centenas de itens em uma lista quais são dispensáveis e só então iniciam o trabalho da vistoria propriamente dita. Isso contribui drasticamente para a lentidão da atividade. Nossa contribuição para a redução do tempo despendido nessa checagem é descrita em seguida.

2 Solução proposta

Temos por problema a determinação de quais itens de um checklist são passíveis de abono. Isso pode ser modelado como um problema de classificação binária : dado um conjunto de características de um sítio e qual o item desejamos prever se ele é da classe "abonado" ou não (7). Especialistas apontaram a seguinte lista de características de um sítio que os próprios técnicos usam para abonar manualmente os itens:

- Tipo de site: 'RT', 'GF', 'RF', 'IN';
- Tipo de tecnologia: WCDMA, LTE, GSM;
- Frequência: 450Mhz, 700Mhz, 850Mhz, 1800Mhz, 2100Mhz, 2600Mhz;
- Equipamentos de radiofrequência (RF): Diplex, Triplex, Quadriplex, EHCU, Filtro, TMA, DTMA

Essas informações, contudo, não estão prontamente disponíveis. Uma fonte possível de informação são os Projetos Preliminares de Instalação (PPIs). Eles estão disponíveis em um sistema interno das empresas de telefonia, ao qual nos foi dado acesso, em formato pdf. Tivemos acesso também à base de checklists dos sítios. Identificamos 602 ERBs cadastrada nesse sistema das quais baixamos cerca de 150 PPIs e os checklists de fevereiro a setembro. Dentre os PPIs foram identificados 3 padrões de documento. Como um esforço inicial trabalhamos na extração de informação de um único padrão. Dado esse recorte de um único tipo de documento, a intersecção entre o grupo de sítios que tínhamos tanto o checklist quanto o PPI tem uma cardinalidade de 44.

As características das ERBs estavam de presentes de forma não estruturada em tabelas e textos nos ppis. A informação não contida nas tabelas, extraídas por meio de pacotes especializados, foi obtida por meio da tokenização dos textos. Desta forma geramos automaticamente uma base de características dos sítios. A partir da intersecção entre a base de características e a base de items geramos um banco de dados de 19000 observações. Na base temos 322 items únicos, com uma mediana de 243 items por ERB, e 19 atributos ('Items', 'Status', 'Operadora', 'Tipo de Site', 'WCDMA', 'LTE', 'GSM', '450Mhz', '700Mhz', '850Mhz', '1800Mhz', '2100Mhz', '2600Mhz', 'Diplex', 'Triplex', 'Quadriplexer', 'EHCU', 'Filtro', 'TMA', 'DTMA'), onde todos menos "Item" e "Tipo de Site" são variáveis binárias.

Uma inspeção inicial na base nos permitiu identificar um desbalanceamento no número de itens avaliados x os abonados, no "Status" do item. O desbalanceamento das

classes impacta na performance preditiva de modelos, na medida em que o modelo ganha um viés para a classe majoritária simplesmente pelo maior número de observações dessa classe, aumentando, portanto, o número de falso negativos (8). Como demonstrado na Figura 1 o número de itens avaliados era mais do dobro dos itens abonados, de forma que optamos pela sobreamostragem da classe minoritária por meio de um método de interpolação padrão: o SMOTE (Synthetic Minority Over-sampling Technique) (9).

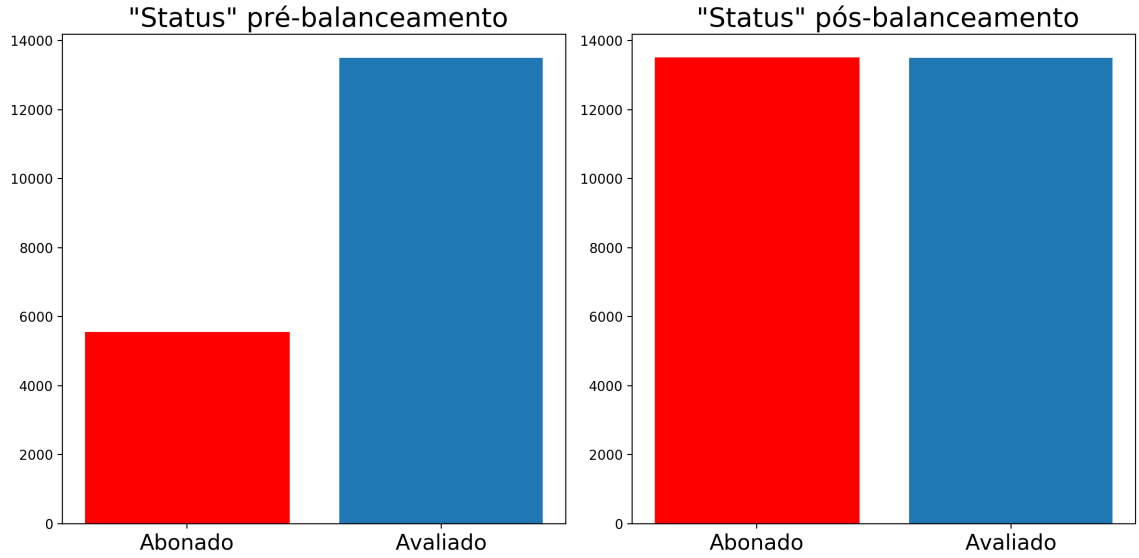

Figure 1 – Número de Abonados vis-à-vis Avaliados pré e pós balanceamento

Após o rebalanceamento codificamos os atributos “Item” e “Tipo de Site” por meio de one-hot-encoding. Uma outra opção de codificação seria atribuir um número inteiro a cada item, mas essa estratégia confundiria o modelo ao implicitamente atribuir ordem a uma variável nominal (8). A matriz de atributos após o rebalanceamento e a codificação tem 27002 linhas e 348 colunas.

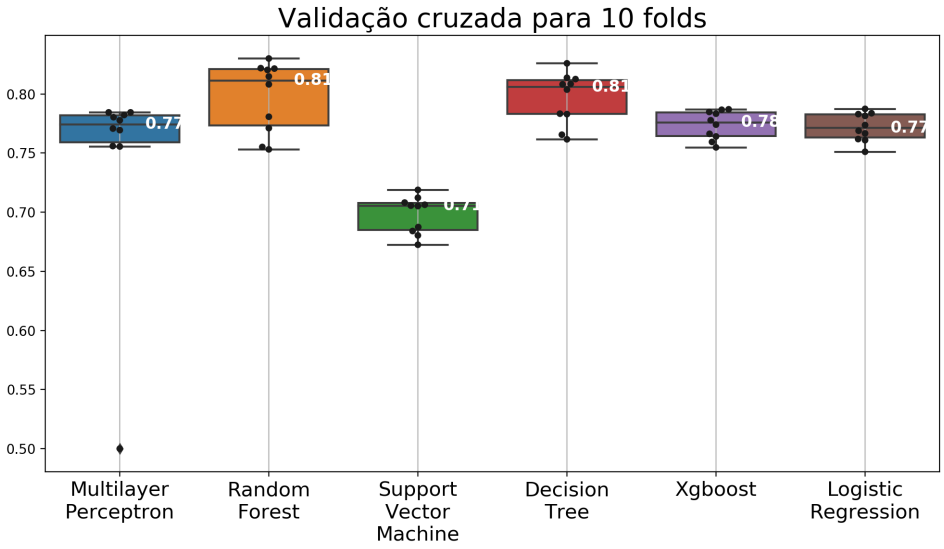

Figure 2 – Distribuição de acurácias. Acurácia mediana anotada em cada caixa.

Uma vez concluído o preprocessamento partimos para o uso de modelos preditivos de aprendizado de máquina. Fizemos um Grid Search[1], com validação cruzada (k-fold com 10 folds), dos seguintes modelos: Decision Tree, Multilayer Perceptron, Logistic Regression, Random Forest, Xgboost. A Figura 2 demonstra a distribuição de acurácias, (número de predições corretas) / (número total de predições) , dos melhores classificadores de cada tipo. A árvore de decisão (Decision Tree) e a floresta aleatória (Random Forest) foram os classificadores com melhor performance.

Nosso problema de interesse e banco de dados contêm características que 11 indica serem particularmente "tratáveis" por meio de árvores de decisão: os atributos têm poucos valores, o output assume valores discretos, descrições disjuntas são requeridas e os dados de treino podem conter erros.

A floresta aleatória, por sua vez, é simplesmente um conjunto de árvores de decisão, onde cada árvore de decisão é treinada em uma amostra sorteada da base (sorteia-se tanto linhas quanto colunas da base, com reposição). No nosso caso a floresta aleatória tem uma acurácia mediana um pouco maior do que as árvores de decisão (0.81 x 0.806), mas com um custo computacional muito maior. A árvore de decisão figura então como o algoritmo utilizado nesse estágio do projeto. A matriz de confusão, na qual utilizamos o melhor classificador da validação cruzada, na Figura 3 anos dá uma noção mais completa da performance do classificador: em torno de 16% das classificações como "avaliável" são falsos negativos.

[1] Modelos de aprendizado de máquina apresentam parâmetros, conhecidos como hiperparâmetros, que não são aprendidos internamente, eles são configurações do modelo. É considerada uma boa prática testar diferentes combinações de hiperparâmetros para identificar qual parametrização tem a melhor métrica de avaliação. O Grid Search é quando testamos exaustivamente as parametrizações estabelecidas (10).

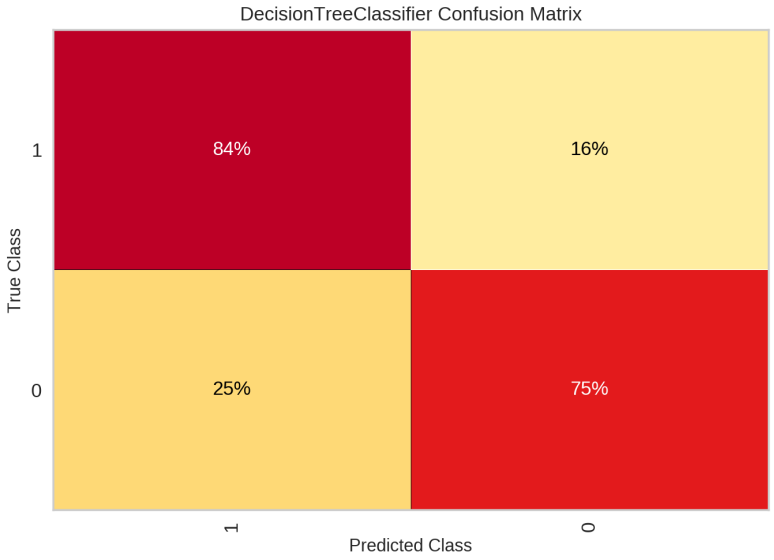

Figure 3 – Matriz de confusão num banco de teste de 60%. 1 é "Abonado".

Em conversas com especialistas, definiu-se que o "output" de interesse dos usuários seria quais itens teriam a maior probabilidade de serem abonados em determinado sítio. A árvore de decisão nos permite determinar a probabilidade, para o classificador, de uma instância ser de uma classe, no nosso caso a classe "Abonado". Sendo assim nossa solução é a seguinte:

1. o usuário indica qual a ERB de inspeção;
2. extrai-se da base construída qual as características do sítio;
3. as características preprocessadas são enviadas ao classificador treinado, a árvore de decisão, que retorna as probabilidades de pertencimento à classe "Abonado" de cada item do site;
4. retorna-se ao usuário a lista ordenada, pela probabilidade decrescente de pertencimento à classe, dos itens do site.

Conclusão

No presente trabalho apresentamos um caso de aperfeiçoamento do processo de inspeção de estacões rádio base por meio de mineração de dados e inteligência artificial. Embora grandes empresas de tecnologia como Google, Facebook e Amazon façam uso de grandes arquiteturas de redes neurais artificiais as quais necessitam de dezenas de horas de treinamento em unidades de processamento gráfico, a realidade da maior parte das empresas que buscam se inserir nessa nova era algorítmica difere em escopo (12). Se por um lado a inteligência artificial traz a possibilidade de uma riqueza de aplicações e otimizações no processo produtivo das empresas, por outro lado se faz necessária uma infraestrutura de dados que permita a aplicação dessas técnicas e uma "pipeline" de mineração e recuperação de informação (13). Ademais a restrição orçamentária e computacional e o imperativo da interpretabilidade[2] do funcionamento dos algoritmos nos direciona, nesses casos medianos, à algoritmos mais bem estabelecidos e simples em comparação aos de alta publicização (16).

No nosso caso a mineração de dados contidos em documentos contidos nos servidores internos de empresas de tecnologias em conjunção com um modelo simples e interpretável de aprendizado de máquina nos permitiu contribuir no processo produtivo. Há, contudo, muito a ser feito. Primeiramente, estender a mineração para todos os tipos de documentos contidos nos servidores é o próximo passo. Segundo, investigar como melhorar a acurácia dos classificadores, dado que temos um limiar máximo de aproximadamente 84%. Nossa hipótese, por meio de investigação dos servidores e conversas com usuários, é que o não cumprimento dos procedimentos de inspeção nas respostas aos itens gera ruído que confunde os classificadores. A despeito disso, ainda há a necessidade de investigar como aperfeiçoar os algoritmos independentemente da qualidade dos dados. Por fim, a determinaç de quais itens são abonáveis é somente a primeira tarefa, pois o problema de determinar, algoritmicamente, quais itens são aceitos ou rejeitados há de requerer uma pletora de estudos adicionais.

[2] No contexto de aprendizado de máquina a interpretabilidade é definida por 14, p.2 "como a habilidade de explicar ou apresentar em termos compreensíveis para humanos". Uma definição equivalente de interpretabilidade é: o grau no qual um humano pode compreender a causa de uma decisão (15).

Bibliography

1 CAMERON, J.; WISHER, W. *Terminator 2: Judgment Day*. [S.l.]: USA, 1991. Citado na página 5.

2 COCKBURN, I. M.; HENDERSON, R.; STERN, S. *The impact of artificial intelligence on innovation*. [S.l.], 2018. Citado na página 5.

3 MAKRIDAKIS, S. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, Elsevier, v. 90, p. 46–60, 2017. Citado na página 5.

4 KOGUT, B. M. *The global internet economy*. [S.l.]: MIT Press, 2003. Citado na página 5.

5 GUBBI, J. et al. Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems*, Elsevier, v. 29, n. 7, p. 1645–1660, 2013. Citado na página 5.

6 CARBONELL, J. G.; MITCHELL, T. M.; MICHALSKI, R. S. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer-Verlag, 1984. Citado na página 5.

7 JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. Citado na página 7.

8 FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. 2011. Citado na página 8.

9 CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 8.

10 GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. [S.l.]: O'Reilly Media, 2019. Citado na página 9.

11 MITCHELL, T. M. et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, p. 870–877, 1997. Citado na página 9.

12 CANZIANI, A.; PASZKE, A.; CULURCIELLO, E. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016. Citado na página 11.

13 SCHÜTZE, H.; MANNING, C. D.; RAGHAVAN, P. *An introduction to information retrieval*. [S.l.]: Cambridge University Press,, 2007. Citado na página 11.

14 DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. Citado na página 11.

15 MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, Elsevier, 2018. Citado na página 11.

16 DREISEITL, S.; OHNO-MACHADO, L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, Elsevier, v. 35, n. 5-6, p. 352–359, 2002. Citado na página 11.