Trabalho de Aprendizado de Máquina

Victor Luiz Fortes Rivelo[1], Lucas Henrique Mantovani Jacintho[1], and Vinicius Henrique Borges[1]

Universidade de São Paulo, São Carlos , Brasil
{victor.rivelo, lucasmantovani, vinicius.henrique.borges}@usp.br
https://www.icmc.usp.br

Abstract. Este artigo tem como objetivo comparar técnicas utilizadas para resolução de problemas de predição de compra em tempo real para e-commerces, apoioado no grande aumento no número de empresas do ramo e o impacto das vendas no mercado mundial, realizaremos os comparativos utilizando algoritmos de Machine Learning como Multi Layer Perceptron, Support Vector Machine, Random Forest, para comparação analisaremos as métricas de acurácia, precisão, recall, e f1-score, contrastando-as para desta forma garantirmos análises realistas e concretas. seção 2.1,

Keywords: online shoppers · prediction · MLP · SVM · comparison of methods · Decision tree · Random forest · shop intention · real time

1 Introdução

Segundo a pesquisa feita por TIC Domicílios [10] pela primeira vez 69,8% da população brasileira possui conexão com a internet, e este número vem crescendo (vide Figura 1). Note que o processo, por natureza, detém de características naturais de expansão acelerada dado a fatores como globalização e evolução tecnológica.

Segundo o EcommerceBrasil [13], o ranking de meios de pesquisa e impacto nas futuras decisões de compra é:

- Internet – 66 %
- Conselhos de amigos e parentes – 61%
- Jornais – 43%
- Televisão – 42%
- Mala direta tradicional – 37%
- Revistas – 28%
- Rádio – 28%.

Ambos os estudos citados anteriormente nos mostram que com o aumento da quantidade de pessoas que utilizam internet no Brasil vide figura 1 e no Mundo, maior será o impacto da mesma nas relações pessoais e interpessoais. Focaremos neste artigo na compra e intenção de compra de produtos em tempo real, e neste quesito a internet assume posição de destaque com 66% explicando

o grande crescimento do marketing digital no mundo do comércio eletrônico [13]. Seguindo a lei de oferta e demanda natural dos mercados podemos notar também que, segundo Olhar Digital [12], o número de páginas web cresceu 1114% nos últimos 10 anos.

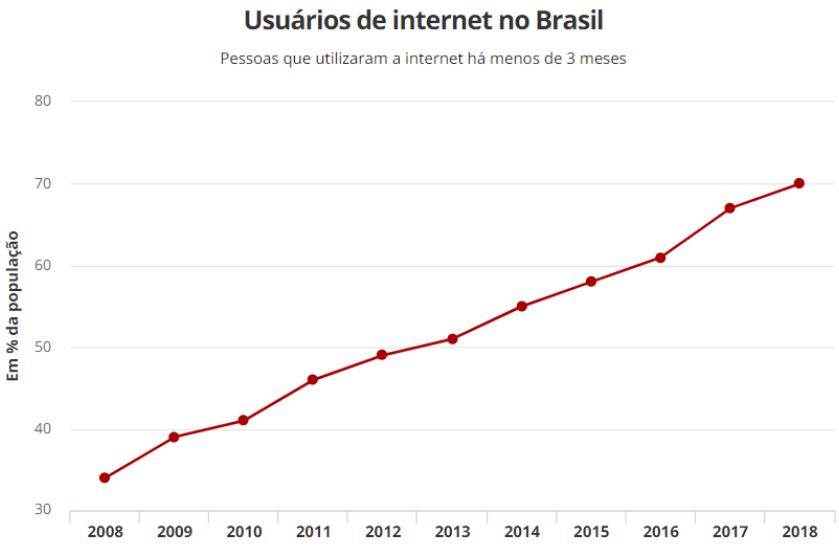

Fig. 1. Usuários de internet no Brasil
https://g1.globo.com/economia/tecnologia/noticia/2019/08/28/uso-da-internet-no-brasil-cresce-e-70percent-da-populacao-esta-conectada.ghtml

Sobre todo o montante de usuários na internet, segundo o G1 [14], em 2018, 43,7 milhões de pessoas compraram pela internet no país, o que representa 34% do total de usuários online. Analisando a correlação entre os dados podemos concluir que a gestão sobre as vendas no mundo tecnológico é um diferencial competitivo. Se analisarmos a notícia feita por Portal Ecommerce [15] veremos que o consumidor brasileiro não tem um tempo médio para realizar uma compra que se aplique de modo geral, e portanto, para cada categoria de produto existe um número de horas. Em adição, a mesma notícia nos mostra que para comprar itens de beleza a média de tempo gasto é de aproximadamente 3 dias, 12 horas e 14 minutos. Tempo mais do que suficiente para direcionar propagandas de marketing ao usuário e oferecer promoções para persuadi-lo a comprar o produto de sua loja. Para ser assertivo e obter boas conversões das divulgações em compras, é interessante inferir com a maior taxa de acerto possível, qual a intenção de compra do usuário tendo como resultado o aumento das vendas e do impacto da marca sob seus clientes.

Uma das formas de predizer a intenção de compra é utilizando algoritmos de Machine Learning como descrito em Real-time prediction [1]. Como a quantidade de dados gerada é expressiva, a análise manual não é efetiva a longo prazo, e por isso decidimos neste artigo investir esforços nesta metodologia.

O desenvolvimento desse artigo foi motivado além das indagações citadas acima, mas também devido a participação do grupo na disciplina Aprendizado de Máquina de código SCC0276_1Sem_2020. O artigo Real-time prediction [1] serviu de inspiração para o grupo. Nos referenciaremos a ele como artigo original durante o estudo, dado que todo o processo de revisão sistemática e análise inicial foi baseado no mesmo. Há de se notar que o artigo decorre o estudo sobre a mesma base que será analisada na seção 3.1, assim como o próprio artigo será aprofundado na seção 2.1.

Apoiados no artigo original desse conjunto de dados Kaggle [1], utilizaremos também as técnicas de MLP, SVM, Decision Tree e Gaussian Naive Bayes. Faremos a análise e também pré-processamento dos dados, como *oversampling* e *undersampling*. Também utilizaremos técnicas feature selection para sele]ao dos atributos utilizados na classificação.

Este artigo está organizado de forma a, primeiramente, revisar e apresentar quais as propostas, técnicas ou metodologias que foram utilizadas para identificação da intenção de compra de um visitante navegando em um website, utilizando como referência artigos publicados que fazem uso da mesma base, ou de bases similares, porém, com o mesmo objetivo. A seção seguinte contém uma breve explicação textual do conjunto de dados, assim como uma exploração mais aprofundada e pré-processamento. Também expõe as técnicas e modelos utilizados para a predição. Em seguida, exibimos os resultados obtidos, Por último, apresentamos uma conclusão do trabalho desenvolvido, assim como direcionamentos para trabalhos futuros.

2 Revisão Bibliográfica

2.1 Artigos Selecionados

A análise do problema, retirado da plataforma Kaggle, mostrou a existência de um trabalho expressivo sobre a base. Após análise do grupo de pesquisa, concordou-se na relevância do mesmo e em utilizá-lo como artigo inicial para o processo de revisão sistemática.

O artigo original utiliza as técnicas de MLP, SVM e Decision Tree com alguns modelos para cada tipo de técnica. Para MLP, testaram modelos com 1 camada escondida variando a quantidade de neurônios: 10, 20 e 40. Para a técnica SVM, utilizaram dois tipos de kernel: linear e rbf. E para a Decision Tree, utilizaram o C4.5, que é uma extensão do ID3 e permite atributos numéricos, e Random Forest [3]. Devido à criticidade do sistema utilizado como objeto de estudo pelos autores - o qual necessita de classificações em tempo-real para customizar a experiência dos visitantes - eles decidiram não utilizar algoritmos *lazy-learning* como o k-NN, pois estes tendem a apresentar respostas mais lentas dado o fato de analisarem toda a base de dados a fim de obterem uma classificação [4].

Em uma outra, porém semelhante, abordagem, um grupo composto por cientistas da empesa Alibaba e da Universidade de Sydney propuseram fazer a predição de conversão analisando o comportamento do usuário após o clique em um anúncio [5]. Em sistemas anteriores a tentativa era de prever a conversão apenas na sequência de ações Clique/Compra. No artigo citado, os pesquisadores incluem outras ações intermediárias, como Adicionar ao Carrinho e Adicionar à Lista de Desejos. Para isso, coletaram dados do Alibaba e realizaram experimentos utilizando Gradient Boosting Decision Tree, Deep Neural Network, Deep Neural Network com Over-Sampling e com o próprio modelo proposto pelo grupo, chamado de ESM^2.

Com uma abordagem muito semelhante ao artigo citado anteriormente, porém com diferença nas técnicas utilizadas, é valido citar o trabalho de Bigon et al [6]. Neste trabalho, o grupo também analisa o comportamento do usuário no e-commerce acompanhando a navegação e registrando eventos do tipo "view", "detail", "add", "remove", "buy" ou "click". E utilizando a sequência que o usuário realizou no website tentam prever se a sessão será finalizada em compra ou não. O diferencial desse trabalho é o uso de uma LSTM Seq2Label, que analisa uma sequência e retorna um rótulo. Como comparação utilizaram as técnicas Naive Bayes, Markov Chain e LSTM - Language Model.

Em um outro artigo Pîrvu et al[7], explorando também o nosso problema de predição através da sessão de log do usuário no browser, tendo foco na utilização de informações de sessões anteriores do mesmo usuário, para então analisar os dados, os dados são retirados do Google Analytics Reporting API v4, aqueles de maior interesse são separados em um vetor de ações comuns composto por: visita regular a uma página, visita aos detalhes de um item em específico, adicionar um item ao carrinho de compras, visitar a página de checkout, e realizar o pagamento da transação. No artigo citado os pesquisadores utilizaram uma metodologia de rede neural com recorrência dupla, implementada através da utilização de células de LSTM contendo 30 hidden states; as estratégias de atribuição utilizadas foram linear e decaimento por tempo, combinadas com normalização standard ou min-max.

2.2 Desempenho das Técnicas Identificadas na Revisão sistemática

Table 1. Resultado dos métodos aplicados no artigo [3]

Model	Accuracy(%)	True-positive rate	True negative	F1-score
C4.5	82.34	0.79	0.85	0.82
MLP	87.24	0.84	0.92	0.86
RBF SVM	84.11	0.75	0.94	0.82

	CVR@top0.1%			CVR@top0.6%			CVR@top1%		
Method	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
GBDT	4.382%	14.348%	6.714%	16.328%	9.894%	12.322%	27.384%	7.384%	11.631%
DNN	4.938%	15.117%	7.445%	17.150%	10.495%	13.021%	28.481%	8.196%	12.729%
DNN-OS	5.383%	15.837%	8.034%	17.38%	10.839%	13.353%	29.032%	8.423%	13.058%
ESMM	5.813%	16.295%	8.570%	18.585%	11.577%	14.267%	29.789%	8.961%	13.777%
ESM^2	**6.117%**	**17.145%**	**9.017%**	**23.492%**	**10.574%**	**14.584%**	**30.032%**	**9.034%**	**13.890%**

Fig. 2. Resultado dos métodos aplicados no artigo [5]

	CTCVR@top0.1%			CTCVR@top0.6%			CTCVR@top1%		
Method	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
GBDT	2.937%	0.701%	1.132%	4.870%	0.649%	1.145%	8.894%	0.531%	1.002%
DNN	3.168%	0.851%	1.341%	5.269%	0.768%	1.340%	9.461%	0.643%	1.204%
DNN-OS	3.382%	0.871%	1.385%	5.369%	0.801%	1.395%	9.863%	0.673%	1.260%
ESMM	3.858%	0.915%	1.479%	5.504%	0.828%	1.439%	10.088%	0.691%	1.294%
ESM^2	**4.219%**	**1.001%**	**1.618%**	**5.987%**	**0.900%**	**1.566%**	**10.991%**	**0.753%**	**1.410%**

Fig. 3. Resultado dos métodos aplicados no artigo [5]

Table 2. Resultado dos métodos aplicados no artigo Bigon et al[6]

Model	Accuracy
Naive Bayes	0.821
Markov Chain	0.882
LSTM - Language Model	0.909($\pm$0.004)
LSTM - S2L('avg' pooling)	0.927($\pm$0.003)
LSTM - 2SL('Last')	**0.932**($\pm$**0.002**)

Table 3. Resultado do métodos aplicados no artigo Pîrvu et al[7]

Attribution Model	Normalization	Loss	Accuracy(0.5/0.2)	Accuracy(0.8/1.25)
Linear	Min-Max	**0.0056**	**47.75**%	**40.72**%
Linear	Min-Max	**0.0056**	47.10%	39.50%
Time decaying	Min-Max	0.0065	44.73%	13.46%
Time decaying	Min-Max	0.0065	44.68%	13.39%

3 Material e Métodos

3.1 DataSet

O conjunto de dados utilizado se trata de registros de navegação de usuários em um website durante o período de 1 ano, a fim de evitar tendência a uma campanha específica. O conjunto pode ser encontrado online em diversos repositórios, como Kaggle [1] e UCI [2].

O dataset consiste em:

- 12,330 linhas;
- 18 atributos sendo eles 10 numéricos e 8 categóricos;
- Não possui atributos nulos;

Os atributos categóricos são:

- Sistema Operacional;
- Browser utilizado;
- Região do usuário;
- Tipo de tráfego (como o usário chegou ao website): banner, SMS, e outros;
- Tipo do visitante (Novo, Retornando, Outro);
- Fim de semana ou não;
- Mês do acesso;
- Houve compra ou não

Já os numéricos são:

- Administrativo: quantas páginas administrativas foram visitadas;
- Duração em páginas administrativas: quantos segundos o visitante permaneceu em páginas administrativas;
- Informativa: quantas página informativas foram visitadas (comunicação, endereço, outras informações);
- Duração em páginas informativas: quantos segundos o visitante permaneceu em páginas informativas;
- Produto: quantas páginas relacionadas ao produto foram visitadas;
- Duração em páginas de produto: quantos segundos o visitante permaneceu em páginas relacionadas ao produto;
- Taxa de saída com visita de apenas uma página - *Bounce Rate*;
- Taxa de saída - *Exit Rate*;
- Valor da página;
- Proximidade de dia especial (Dia das mães, dia dos namorados e outros);

4 Implementação

A implementação do código estará disponível em: Colab Network Code
`https://colab.research.google.com/drive/1twRahf06w3gItbKvf48UeSvkGNYmCONt`

4.1 Exploração e Pré-processamento

A exploração do dados e os métodos de pré-processamento utilizados estão disponíveis no notebook, junto com respectivas explicações necessárias.

Para o método de normalização dos dados utilizamos duas funções distintas da biblioteca sklearn: *StandardScaler* e *Normalizer*. Ambos foram comparados e apresentaram medidas de desempenho próximas, porém na média o método *StandardScaler* apresentou resultados superiores, por conta disso escolhemos utilizar os seus resultados para serem exibidos na tabela 5.

5 Tabelas de Resultados

Table 4. Tabela de resultados com dados puros

Modelo do classificador	Accuracy	Precision	Recall	F1-score
MLP	81.97%	70.56%	80.24%	73.14%
SVM	84.30%	92.14%	50.51%	46.75%
Decision Tree	85.27%	72.44%	72.66%	72.55%
Random Forest	89.75%	82.47%	76.24%	78.81%
Gaussian Naive Bayes	84.72%	71.37%	71.34%	71.36%

Table 5. Tabela de resultados com dados normalizados (StandardScaler)

Modelo do classificador	Accuracy	Precision	Recall	F1-score
MLP	87.64%	77.25%	73.74%	75.28%
SVM	88.55%	81.72%	70.38%	74.12%
Decision Tree	85.50%	72.86%	73.13%	73.00%
Random Forest	89.75%	82.47%	76.24%	78.81%
Gaussian Naive Bayes	79.37%	67.19%	75.21%	69.21%

Table 6. Tabela de resultados com seleção de atributos

Modelo do classificador	Accuracy	Precision	Recall	F1-score
MLP	88.00%	78.79%	71.88%	74.55%
SVM	84.30%	92.14%	50.51%	46.75%
Decision Tree	84.69%	71.63%	73.31%	72.41%
Random Forest	89.69%	82.55%	75.70%	78.47%
Gaussian Naive Bayes	84.88%	71.62%	71.19%	71.40%

Table 7. Tabela de resultados com Oversampling

Modelo do classificador	Accuracy	Precision	Recall	F1-score
MLP	87.38%	76.38%	81.22%	78.40%
SVM	73.92%	64.39%	74.13%	65.08%
Decision Tree	85.99%	73.70%	72.26%	72.94%
Random Forest	89.26%	79.90%	79.85%	79.87%
Gaussian Naive Bayes	81.54%	68.80%	75.51%	70.95%

Table 8. Tabela de resultados com Undersampling

Modelo do classificador	Accuracy	Precision	Recall	F1-score
MLP	71.91%	72.68%	71.68%	71.52%
SVM	62.89%	63.13%	63.01%	62.84%
Decision Tree	80.08%	80.10%	80.11%	80.08%
Random Forest	85.32%	85.45%	85.25%	85.29%
Gaussian Naive Bayes	76.52%	76.55%	76.46%	76.48%

Para comparação dos modelos o grupo decidiu por utilizar como objeto de estudo de maior peso a métrica recall, pois para a natureza do problema, o interesse se dá pela quantidade de amostras que realmente tem como resultado a compra e foram classificadas corretamente. Além disso, setamos os algoritmos para um estado fixo de random state = 42, desta forma garantimos que em toda execução dos modelos serão obtidos os mesmos resultados, para o conjunto de treino e teste aplicados aos modelos, que posteriormente auxiliam na manutenção do princípio de replicabilidade e avaliação transparente.

Tendo em vista a perspectiva citada, os modelos assumiram em ordem o ranking decrescente a seguir:

Table 9. Ranking dos resultados

Modelo do classificador	Pré-Processamento	Recall
Randon Forest	Undersampling	85.25%
MLP	Oversampling	81.22%
MLP	Dados puros	80.24%
Decision Tree	Undersampling	80.11%
Randon Forest	Oversampling	79.85%

6 Conclusão

Neste trabalho realizamos a classificação da intenção de compra utilizando diversos tipos de técnicas e modelos, a fim de compará-los. Os modelos foram MLP, SVM, Decision Tree, Random Forest e Gaussian Naive Bayes. Para validação empregamos a técnica Holdout, que separa o conjunto de dados em dois: um para o treino e um para o teste. Realizamos também o pré-processamento dos dados de duas formas: a primeira remove a média e transforma a variância para 1; a outra normaliza cada atributo para a escala 1, ou seja, os valores vão de 0 a 1. Além disso, aplicamos o Select K-Best uma técnica de seleção de atributos, retornando os 10 melhores atributos para estudo em nossa base, e também duas técnicas para balancear a quantidade de classes (oversampling e undersampling). Vale ressaltar, que as técnicas foram utilizadas de forma independente, portanto, não combinamos seus resultados.

6.1 Trabalhos Futuros

Como trabalhos futuros, propomos avaliar algoritmos de agrupamento. É necessário, porém, levar em consideração que a proposta original era construir um sistema de tempo real, o que não é viável nessa situação, devido à necessidade de processar o conjunto de dados completo.

Além disso, é possível também combinar os métodos de pré-processamento realizados. Assim é possível montar um conjunto de dados único e avaliar se apresenta melhores resultados quando comparado com os datasets isolados.

References

1. C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, Yomi Kastro, : Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks (2018)
2. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset
3. Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput & Applic 31, 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0
4. Atkeson C.G., Moore A.W., Schaal S. (1997) Locally Weighted Learning. In: Aha D.W. (eds) Lazy Learning. Springer, Dordrecht
5. Wen, H., Zhang, J., Wang, Y., Bao, W., Lin, Q., & Yang, K. (2019). Conversion Rate Prediction via Post-Click Behaviour Modeling. arXiv preprint arXiv:1910.07099. .
6. Bigon, L., Cassani, G., Greco, C., Lacasa, L., Pavoni, M., Polonioli, A., & Tagliabue, J. (2019). Prediction is very hard, especially about conversion. Predicting user purchases from clickstream data in fashion e-commerce. arXiv preprint arXiv:1907.00400.
7. Pîrvu, C., Mihai , & Anghel A. (2019). PREDICTING NEXT SHOPPING STAGE USING GOOGLE ANALYTICS DATA FOR E-COMMERCE APPLICATIONS. arXiv preprint arXiv:1905.12595.
8. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
9. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
10. Notícia Canaltech, `https://canaltech.com.br/internet/pesquisa-do-ibge-revela-que-aumentou-o-numero-de-usuarios-de-internet-no-brasil-129545/`. Último acesso em 9 Maio 2020
11. LNCS Homepage, `http://www.springer.com/lncs`. Last accessed 4 Oct 2017
12. Olhar Digital Notícia, `https://olhardigital.com.br/noticia/dados-mostram-o-crescimento-impressionante-da-internet-em-10-anos/85914`. Último acesso em 9 Maio 2020
13. EcommerceBrasil Artigo, `https://www.ecommercebrasil.com.br/artigos/a-internet-influencia-na-sua-decisao-de-compra/`. Último acesso em 9 Maio 2020
14. Notícia G1, `https://g1.globo.com/economia/tecnologia/noticia/2019/08/28/uso-da-internet-no-brasil-cresce-e-70percent-da-populacao-esta-conectada.ghtml`. Último acesso em 10 Maio 2020
15. Portal Ecommerce, `http://portaldoecommerce.com/2017/08/17/quanto-tempo-as-pessoas-levam-para-fazer-compras-online/`. Último acesso em 10 Maio 2020