Previsões de venda usando anúncios das redes sociais

F. A. Author
Glevson da Silva Pinto

B. Author, Jr.
Priscilla Amarante de Lima

Resumo - A assimilação da inteligência artificial pelo marketing Leads (digital) torna-se mais uma ferramenta importante na busca por possíveis compradores, contribuindo para o aumento das vendas através da exploração e potencialização dos anúncios presentes em sites de compras online. O presente artigo tem por objetivo apresentar um modelo preditor com uso de Aprendizagem de máquina e o algoritmo Naive Bayes aplicada ao dataset de marketing digital para prever as vendas a potenciais compradores. Por meio de abordagem estatística descritiva e relato da construção do modelo, analisou-se os dados e foi aplicado método de teste Cross Validation em que o melhor desempenho foi obtido com K = 5, uma acurácia de predição satisfatória maior que 90%. Prova que a abordagem com algoritmo é satisfatória para prever um conjunto de potenciais clientes.

⬛ **Justificativa** As intenções de compra são informações coletadas por meio de pesquisas de intenção, nas quais produtos e serviços são medidos para avaliar a aceitação dos mesmos pelo consumidor. São informações que toda empresa deve ter e acompanhar para coletar dados e desenvolver um plano estratégico para entender possíveis necessidades ou ajustar as condições necessárias para tornar o produto atraente para os clientes.

Com relação às pesquisas sobre a intenção de compra online, a maioria tem considerado os principais construtos do modelo de aceitação da tecnologia (technology acceptance model – TAM), o qual teoriza que as percepções de utilidade e de facilidade de uso ajudam a determinar a adoção do comércio eletrônico (GEFEN e STRAUB, 2000).

A Leads é uma estratégia conhecida no marketing digital que busca dados de possíveis clientes que estão com suas características mapeadas e com possibilidade de comprarem produtos que foram publicados na página da propaganda (Barbeito, 2017). Mas como a estratégia de Lead ainda não é suficiente para prever se um cliente vai comprar o produto ou não, então aplicou-se no projeto o classificador Naive Bayes.

OBJETIVO

Descobrir se um usuário compra um produto clicando no anúncio no site com base em seu salário, idade e sexo.

BASE DE DADOS

Nossa base de dados pode ser acessada no Kaggle, em https://www.kaggle.com/rakeshrau/social-network-ads contendo um conjunto de dados categóricos para determinar se um usuário comprou um produto específico em sites de compra online.

ANÁLISE EXPLORATÓRIA DE DADOS

Nessa etapa importa-se os dados em seguida exibe os dados para iniciar o pré-processamento de dados, é importante compreender os dados para realizar o pré-processamento. Através das análises podemos identificar possíveis problemas nos dados, verificar como os dados estão distribuídos e realizar uma melhor transformação. Durante esse processo realiza-se verificação dos tipos de dados conforme ilustrado na Figura 1.

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15600575	Male	25	33000	0
9	15727311	Female	35	65000	0

Figura 1. Dataset social-network-ads

As colunas representam informações dos possíveis clientes e são compostas de quatro atributos:

Sexo: valor categórico,
Idade: valor inteiro,
Faixa salarial: valor inteiro
Compra: atributo classificador [0,1] (não comprou ou comprou).

Romero et al (2013) técnica de exploração de dados é utilizada para tornar um conjunto de dados adequado ao algoritmo que vai aplicar. Essas técnicas podem ser listadas com as seguintes tarefas: Eliminação de atributos, junção dos dados, identificação de quantitativos de linhas e colunas e seus atributos,Identificar se existem dados faltantes (missing values), Analisar a distribuição de frequência das colunas quantitativas e analisar possíveis correlações entre as variáveis.

Um volume de informações importantes foram extraídos na exploração, com estatística descritiva pode-se verificar o quantitativo e as características do conjunto de dados. Nesse caso é ilustrado na Figura 2 pode-se ver o conjunto de informações referentes a idade, salário e ao sexo dos clientes contidos no dataset.

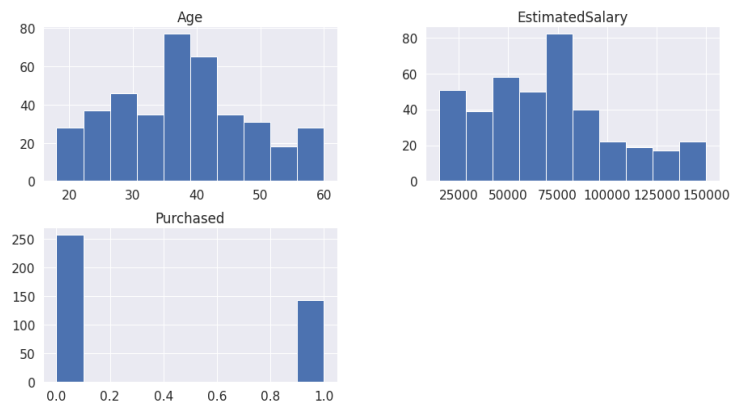

Figura 2. Histrograma do dataset idade, salário e sexo dos clientes

Por meio da medida de frequência é possível medir a proporção de vezes que um atributo assume no dataset. Pode ser observado na Figura 3 a faixa etária que mais clica em anúncios em sites que estão entre 25 - 45 anos.

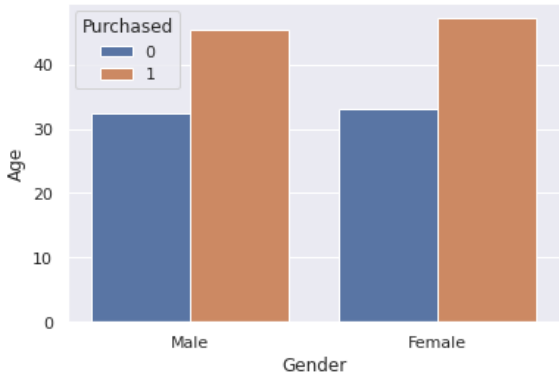

Figura 3. Faixa etária que mais clica em anúncios

Ainda nessa etapa realizou-se exploração dos dados multivariados por meio de correlação. A correlação pode ser feita entre variáveis numéricas e é de extrema importância quando se quer saber qual dado influencia mais em um certo resultado ou em outro dado. Assim pode-se definir de forma mais fácil e prática quais dados escolher para analisar e chegar a uma determinada conclusão. Na Figura 4 pode-se concluir que as quatros variáveis selecionadas do dataset social-network-ads, por meio da estratégia Leads feita por uma equipe de marketing digital, contém correlação forte o que indica que juntando os dois métodos Leads e classificação feita por meio de um algoritmo os resultados são ricos para a área de marketing e vendas de uma empresa.

Figura 4. correlação forte entre as variáveis: sexo, idade e estimativa salarial.

Por último na Figura 5. têm-se aplicação do método SMOTE (Synthetic Minority Oversampling Techniques). Neste método, são gerados mais dados das classes de minoria através da adição de instâncias em segmentos de linhas que juntam os k membros de uma determinada minoria. A partir disso, essa pesquisa terá 250 instâncias para cada classe de “1” representando quem compra e 250 instâncias para cada classe “0” representando quem não comprou (Márquez-Vera et al, 2013).

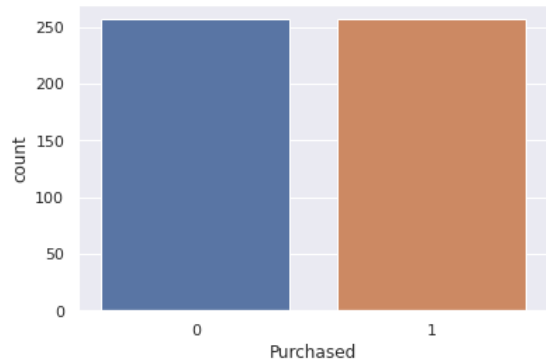

Figura 5. balanceamento do dataset usando o método SMOTE

CLASSIFICADOR NAIVE BAYES - GAUSSIAN

No projeto assumimos features que obedecem a uma distribuição de probabilidade gaussiana. O classificador multinomial Naïve Bayes é um dos modelos mais populares no aprendizado de máquina. Tomando como premissa a suposição de independência entre as variáveis do problema, o modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas.

Sendo um modelo adequado para classificação de atributos discretos, o Naïve Bayes tem aplicações na análise de crédito, diagnósticos médicos ou busca por falhas em sistemas mecânicos.

Baseado em teoria probabilística, o Naive-Bayes trabalha com a ideia de independência de atributos. Sendo considerado como ingênuo, desconsidera a associação entre os atributos e os analisa como condicionalmente independentes. Naive-Bayes oferece bons resultados quando se tem disponível um conjunto de treinamento médio ou grande. Já o Processo Gaussiano é um método que utiliza distribuições de probabilidade para estimar a classe de um ponto através de inferência Bayesiana (Rasmussen 2003).

EXPERIMENTOS

Nessa etapa foi dividido o conjunto de dados em conjunto de treinamento e conjunto de testes. Fornecendo o tamanho do teste representando 20% do dataset, o que significa que nossa amostra de treinamento contém 320 conjuntos de treinamento e a amostra de teste contém 80 conjuntos de teste. O cálculo do desempenho preditivo é exibido em termos de acerto e erro encontrado no experimento (Toussaint, 1974). Os dados de treinamento foram levados em indução e se ajustou bem ao modelo do Naive Bayes. Os dados de teste simularam conjunto de dados novos ao preditor e que não foram vistos antes. Esse experimento a medida de desempenho foi exibida por meio da acurácia que é obtida por meio da operação de soma da diagonal principal da matriz, dividida pela soma dos valores de todos os elementos da matriz nesse experimento foi igual a 0.92.

Também foi aplicado o método Cross Validation K = 5. O Cross Validation envolve a divisão dos dados em vários conjuntos (partes), um dos quais é usado para treinamento e o outro é usado para testar e avaliar o desempenho do modelo.

Na Figura 6 tem-se a saída do experimento com cross validation o que justifica os valores apresentadores com variação são as mudanças dos objetos e prova também a sensibilidade do conjunto de dados. Com K = 5 partições tem-se mudanças no resultado de acordo com cada partição onde neste experimento obteve-se os seguintes valores de [0.81, 0.96, 0.92,0.81,0.90].

```
1 from sklearn.model_selection import cross_val_score
2 pred_gnb = cross_val_score(GaussianNB(), X, y, cv=5)
```

```
1 pred_gnb
```

```
array([0.8125, 0.9625, 0.925 , 0.8125, 0.9   ])
```

Figura 6. Cross Validation, k=5.

Por fim nessa etapa pode-se concluir que o cross validation apresentou um bom desempenho para o experimento provando que o modelo é viável e pode ser utilizado para um cenário em produção.

ANÁLISE DOS RESULTADOS

Nesta etapa obteve-se matriz de confusão para o problema apresentado contendo duas classes 0 = “não Compra”, 1 = “compra”. Na matriz tem-se o número de verdadeiros positivos da classe positiva classificado corretamente (VP). Verdadeiros negativos (VN), exemplares da classe negativa classificados corretamente. (FP) falsos positivos cujo a classe é verdadeira é negativa, porém foi classificado errado como

pertencendo a classe positiva e por último tem-se o valor de (FN) Falsos negativos que pertencem a classe positiva foram classificados de forma errada como negativo (Monard e Baranauskas, 2003). Dessa forma tem-se na Figura 7, o resultado da matriz referente a esse modelo na qual apresentou resultado satisfatório com boa acurácia e baixo valor de FN e FP.

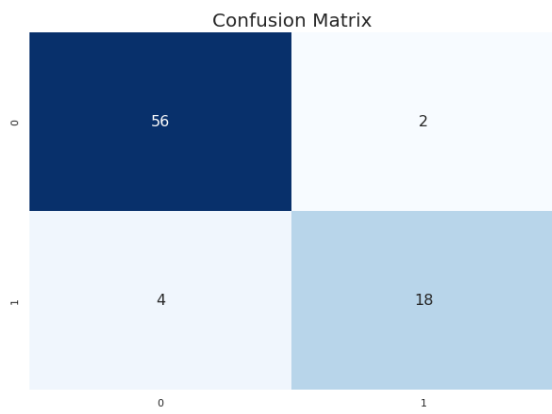

Figura 7. Matriz de confusão modelo Naive Bayes 0 = não compra , 1 = compra.

Podemos avaliar usando a matriz de confusão e a pontuação de precisão, comparando os valores de teste previstos e reais.

Observamos que 56 representa usuários que não são possíveis compradores, 18 representa usuários que são compradores. Com relação ao 2 representa o falso positivo, ou seja, foi classificado de forma incorreta e 4 representa o falso negativo, pois foi novamente classificado incorretamente. Tivemos uma acurácia satisfatória acima de 90%.

CONCLUSÃO E DISCUSSÃO

Os dados estatísticos nos mostram que a maioria dos usuários não são possíveis compradores clicando em anúncios presente em site de compras online; Provando que a aplicação do método de Marketing Leads é satisfatório para construção do dataset de um modelo de predição com uso do algoritmo Naive Bayes. A faixa etária que mais clica em anúncios está entre 25 - 45 anos.

Acurácia satisfatória acima de 90% com uso de validação cruzada para o conjunto de treinamento e teste com k = 5, mostrou-se eficiente para o treinamento e teste.

Para trabalhos futuros pode-se ampliar o conjunto de dados usando um rastreador de dados na web aplicado a um novo conjunto de dados ou ampliar o número de instâncias e atributos do dataset. Também é possível aplicar outras técnicas de aprendizagem de máquina para obter preditores com melhor desempenho.

⬛ REFERÊNCIAS

1. Barbeito, Diego. O que são Leads e por que são tão importantes?. Disponível em: https://sejamais.io/o-que-sao-leads-e-por-que-sao-tao-importantes/. 2017

2 GEFEN, D.; STRAUB, D. W. The relative importance of perceived ease-ofuse in IS adoption: a study of e-commerce adoption, Journal of the Association for Information Systems , v. 1, artigo 8, p. 1-30, 2000.

3. Kaggle. dataset social network. Disponível em: https://www.kaggle.com/rakeshrau/social-network-ads

4 LORENA, Ana Carolina; GAMA, João; FACELI, Katti. Inteligência Artificial: Uma abordagem de aprendizado de máquina. Grupo Gen-LTC, 2000.

5. MÁRQUEZ-VERA, C.; Morales, C. R.; Soto, S. V. Predicting School Failure and Dropout by Using Data Mining Techniques. IEEE Journal of Latin American Learning Technologies, Vol. 8, no. 1, February, 2013.

6. Rasmussen, C. E. (2003). Gaussian processes in machine learning. In Summer School on Machine Learning, pages 63–71. Springer.

7. ROMERO, Cristobal; ESPEJO, Pedro G.; ZAFRA, Amelia; VENTURA, Sebastian. Web usage mining for predicting final marks of students that use Moodle courses. Computer Applications in Engineering Education, v. 21, n. 1, p. 135-146, 2013