CLASSIFICAÇÃO DE USUÁRIOS POR *DATA MINING* EM SOFTWARE FINANCEIRO

Natan Hermano de Maman[1]

Rodrigo Dalla Vecchia[2]

Daiane Rossi[3]

Marcelo Nogueira Cortimiglia[4]

Abstract: *The present study is an action research on the application of data mining techniques on classification of user data during the trial period of a software for financial market operations. This aims to identificate and select variables that characterize user subscribers or non-subscribers, through data mining techniques, to classify users in one of the two classes at the end of the trial period made available. Through the process of knowledge discovery in database were applied four different algorithms to extract useful knowledge through the data. The applied algorithms were Logistic Regression, Support Vector Machine, Naïve Bayes and Random Forest. As result, the Random Forest algorithm was the better representative, compared to the other models tested, to the classification of users problem, and presented the best performance with an area under the curve of 0.85.*

Keywords: Data Mining; Classification; Software; User.

Resumo: *Este estudo é uma pesquisa-ação com aplicação de técnicas de data mining de classificação em dados de usuários no período de teste de um software para operações no mercado financeiro.Ooobjetivo principal consiste em identificar e selecionar variáveis que caracterizam os usuários assinantes ou não assinantes e, através da classificação via data mining, alocar os usuários em uma das duas classes ao fim do período de teste. Com análise de banco de dados, foram aplicados quatro diferentes algoritmos para a extração de conhecimento útil através dos dados. Os algoritmos aplicados foram os de Regressão Logística, Support Vector Machine, Naïve Bayes e Random Forest. Como resultado, o algoritmo de Random Forest representou o que melhor se adaptou ao problema de classificação dos usuários, apresentando o melhor desempenho, e área sob a curva de 0,85.*

Palavras-chave: Mineração de Dados; Classificação; Plataforma; Usuários.

[1] Departamento de Engenharia de Produção e Transportes – Universidade Federal do Rio Grande do Sul (UFRGS) Porto Alegre – Brasil. Correo electrónico: natandemaman11@gmail.com

[2] Programa de Pós-Graduação em Ensino de Matemática – Universidade Federal do Rio Grande do Sul (UFRGS) Porto Alegre – Brasil. Correo electrónico: rodrigovecchia@gmail.com

[3] Programa de Pós-Graduação em Engenharia de Produção – Universidade Federal do Rio Grande do Sul (UFRGS) Porto Alegre – Brasil. Correo electrónico: dai-rossi@hotmail.com

[4] Programa de Pós-Graduação em Engenharia de Produção – Universidade Federal do Rio Grande do Sul (UFRGS) Porto Alegre – Brasil. Correo electrónico: cortimiglia@producao.ufrgs.br

1 INTRODUÇÃO

Com a consolidação da Internet e da Era da Informação, empresas de diversos segmentos do mercado viram ali uma oportunidade de aumentar a competitividade de suas atividades por meio do comércio eletrônico (*e-commerce*). Em 2016, o faturamento de *e-commerce* no Brasil foi de R$ 44,4 bilhões, um aumento de 7,4% em relação a 2015, um forte contraste com relação ao varejo em lojas físicas, que encolheu mais de 10% nos períodos de 2015 e 2014 (E-bit, 2017). Acompanhando a tendência, o marketing digital pretende construir um relacionamento com o indivíduo de forma eficaz, tendo a tecnologia da informação como ponto central (Kotler, *et al*., 2010; Smith, 2011; Ryan & Jones, 2012).

Esse rápido crescimento no *e-commerce* criou uma nova situação para empresas e consumidores. Com a disponibilização de produtos ou serviços usando o meio digital para os consumidores, há diversas alternativas disponíveis nesse meio. Logo, percebeu-se a necessidade de estratégias mais eficazes de marketing e relacionamento com o consumidor (CRM) (Poongothai *et al*., 2011).

Um fundamento do marketing digital é a investigação e compreensão do comportamento do consumidor através da análise quantitativa de dados, devido a geração de grandes volumes de dados. Consequentemente, cresce também o número de métodos para análises de dados gerados nos meios digitais a fim de gerar conhecimento a partir destes. Em particular, ganha importância a extração de conhecimento útil em grandes bancos de dados, também chamada de Descoberta de Conhecimento em Banco de Dados (DCBD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Han, 2005; Aggarwal, 2015). Traçar as características do comportamento de usuários é uma das formas de criar conhecimento como classificações ou predições, com aplicações diretas em e-commerce, CRM, análise de web, *data mining* e sistemas de informação, possível pela larga escala de dados disponíveis nestes meios (Mudiraj, 2011).

Um dos estágios do processo de DCBA é o *Data mining* ou mineração de dados, que tenta identificar padrões em bancos de dados para descobrir e extrair conhecimento útil através da aplicação de algoritmos de descoberta, os quais podem ser agrupados em quatro diferentes tipos: *clustering*, classificação, associação de padrões e análise de *outliers* (Etzioni, 1996; Karuna *et al*., 1999; James *et al*., 2013; Aggarwal, 2015). As aplicações mais típicas de *data mining* na área de marketing podem ser encontradas com objetivos de definir segmentações e públicos-alvo, análise de cesta de mercado, análise de retenção ou

vulnerabilidade de clientes, *retargeting* e recomendação de produtos (Sen, 1998; Aggarwal, 2015).

Assim, o objetivo principal deste estudo consiste em identificar e selecionar variáveis que caracterizam os usuários assinantes ou não assinantes e, através da classificação via *data mining,* alocar os usuários em uma das duas classes ao fim do período de teste. O software a ser testado um banco de dados de usuários cadastrados para um período gratuito de testes de operações no mercado financeiro. Através da aplicação de data mining de classificação, é possível gerar conhecimento útil para apoiar na tomada de decisão estratégica na área de marketing.

2 REFERENCIAL TEÓRICO

Descoberta de conhecimento em banco de dados (DCBD) está no centro do processo de aplicação de um método específico para descobrir padrões e extraí-los. DCBD pode ser definido como o processo não trivial de extração de informações implícitas nos dados, não conhecidas anteriormente, e potencialmente úteis. As áreas de aplicação da descoberta de conhecimento em banco de dados incluem marketing, financeira, telecomunicações, manufatura e agentes de internet (Han *et al*., 2005, Aggarwal, 2015). Logo, DCBD refere-se ao processo completo de descoberta de conhecimento útil nos dados. Uma das etapas em particular desse processo de descoberta é a de *data mining*, cuja é definida como a aplicação de análise de dados e algoritmos de descoberta para extrair padrões existente nos dados. O processo de DCBD possui outras etapas além do *data mining*, as quais são a de extração dos dados, pré-processamento ou limpeza dos dados, transformação dos dados e um interpretação apropriada dos resultados do *data mining* (Frawley *et al*., 1992; Fayyad *et al*., 1996; Han *et al*., 2005, Aggarwal, 2015).

Uma variedade de métodos tradicionais de *data mining* pode auxiliar no processo de descoberta de conhecimento, os quais podem ser agrupados em tarefas distintas. Um dos métodos de descoberta é o de padrão de associação, que permite correlacionar vendas de distintos produtos de modo a subsidiar algoritmos de recomendação. Outro método aplicável em *data mining* é o de detecção de *outliers*, o qual é relacionado a problemas de identificação de dados dissimilares ao grupo ou comportamento principal, como por exemplo, aplicações na identificação de fraudes em cartões de crédito. Já o método de descoberta de classificação permite a criação de classes a partir de atributos comuns, como o comportamento de compra do consumidor, assim, com aplicações diretas em marketing para o público-alvo. O método de

análise de *clustering* permite agrupar dados em grupos de características similares, com resultados que podem ser utilizados para auxiliar na estratégia de marketing ou na mudança dinâmica do site para determinado cliente, a fim de aumentar a chance de transação e/ou melhorar a experiência deste usuário. Em cada método citado existem diversos algoritmos aplicáveis, sendo que cada algoritmo busca solucionar o problema adaptando-se à situação-problema da melhor forma possível, com diferentes níveis de acurácia dependendo dos objetivos específicos e das características dos dados (Cooley *et al*., 1997; Han *et al*., 2005; James *et al*., 2013; Aggarwal, 2015). Os algoritmos aplicados e as formas de obter os dados para aplicação do processo de DCBD podem variar conforme a necessidade e os critérios, sendo amplamente estudado e aplicados.

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa se classifica como pesquisa-ação, empírica, que segue os critérios levantados por Tharenou (2007), com o objetivo gerar conhecimento para solução de problemas específicos e práticos existentes, de natureza aplicada (Manson, 2006). A abordagem é essencialmente quantitativa, por ferramentas matemáticas para coleta e análise de dados. Sendo descritiva e prescritiva, pois objetiva compreender em detalhe uma situação problema e propor soluções a partir dessa compreensão (Tripp, 2005; Miguel 2012).

Este trabalho tem como foco a seleção e identificação de variáveis para aplicação de técnicas de *data mining* em usuários em teste de um software, classificando-os em assinantes ou não assinantes, deste modo, extraindo conhecimento útil dos dados. O processo de DCBD da literatura com sua estrutura de macro etapas foi adaptado para a problemática da pesquisa.

A aplicação destas técnicas está inserida no processo de descoberta de conhecimento em banco de dados (DCBD), o qual envolve as etapas de extração, pré-processamento dos dados, transformação dos dados, aplicação de algoritmos de *data mining* e a análise e validação (Fayyad *et al*., 1996; Han *et al*. 2005; Crone *et al*. 2005; Pachidi *et al.* 2014). Nas primeiras etapas de extração e pré-processamento, utilizou-se a base de dados interna existente na empresa, na qual são registrados os dados de cadastro dos usuários pelo *e-commerce.* Dados como o uso efetivo do software, tempo de uso da plataforma e ordens registradas durante o teste, assim obtendo os dados alvo do estudo, no período de primeiro de Julho de 2016 até Setembro de 2017. Após a extração, a seleção das variáveis para o modelo faz-se necessária e fundamental, através de métodos que permitem a identificação das variáveis mais significativas para o problema em questão (Han et al., 2005). Na fase de

limpeza de dados, valores faltantes, errados ou inconsistentes são removidos. No caso de valores inconsistentes, avalia-se a presença de eventos atípicos como promoções, cortesias ou outras situações que possam influenciar as observações.

Em seguida, a transformação dos dados trata da alteração de observações categóricas para numéricas, assim, gerando a binarização das categorias para que estas possam ser interpretadas pelos algoritmos na fase posterior de forma eficiente (Han *et al*. 2005; Aggarwal, 2015). A transformação dos dados envolve a normalização destes, convertendo-os para a mesma escala, tornando possíveis comparações de grandezas diferentes.

Após a fase de pré-processamento e transformação dos dados segue a etapa de aplicação dos algoritmos de descoberta de conhecimento para classificação dos dados, neste trabalho são aplicados e comparados os algoritmos de *Naïve* Bayes, Kernel SVM, *Random Forest* e Regressão Logística utilizando a linguagem de programação Python. Através destes algoritmos de classificação estimam-se os identificadores de classes, os quais são utilizados para definir em qual classe será alocada a observação (Han *et al*. 2005; Pachidi *et al*. 2014; Aggarwal, 2015). Desta forma, as classes formadas pelos algoritmos são os diferentes usuários segmentados de forma binária, no estudo em questão, em assinantes ou não assinantes ao fim da realização do teste (James *et al*., 2013; Aggarwal, 2015).

Com os resultados dos algoritmos aplicados na fase de *data mining* selecionou-se o que apresentou o melhor desempenho, para assim gerar o conhecimento a partir dos dados extraídos da base de dados de clientes em teste, auxiliando o direcionamento e tomada de decisão das estratégias de marketing.

A pesquisa-ação desenvolvida por este trabalho aplica-se em uma base de dados de clientes que testaram um software de operações no mercado financeiro. O processo de teste gratuito da plataforma pode ser realizado por qualquer usuário que realizar o cadastro pelo e-commerce, sendo que os períodos de testes variam de sete a quinze dias conforme o produto; após a realização do teste gratuito o cliente só poderá realizar o teste novamente após um ano do final do teste anterior. A empresa fica localizada em Porto Alegre, região sul do Brasil. Atualmente a empresa não faz uso de nenhuma técnica de classificação dos usuários em teste.

4 RESULTADOS

A pesquisa teve início com a primeira etapa do processo de DCBD. A extração dos dados internos da empresa. Estes ficam armazenados em SQL Server. A base utilizada contém as informações de teste do software pelos clientes no período de Julho de 2016 até Setembro

de 2017. Possui 25.829 linhas e 25 colunas, onde cada linha é uma solicitação de teste feito por um usuário. A partir dos dados da fase de extração, realizou-se a fase de seleção e limpeza dos dados alvo do estudo, na parte de seleção, classificou-se as colunas sem relevância ou significância para as fases posteriores. Fez-se uso de dois método para criação de um ranking das colunas que seriam selecionadas para a fase posterior, através da medida qui-quadrado das variáveis e do valor-F, para assim medir o impacto das variáveis no modelo. As colunas foram adicionadas de cinco em cinco até a utilização de todas presentes na base de dados. Cada iteração foi avaliada nos dois métodos de seleção. As colunas utilizadas para seleção tinham as informações de tempo de uso da plataforma pelo usuário, o produto testado, por qual canal o teste foi realizado, a idade do usuário, qual Estado está localizado, quantos dias de teste esse usuário teve, se houve contato por e-mail ou telefone durante o teste, quantas ordens de compra ou venda foram realizadas, se o usuário já havia sido cliente anteriormente e a classificação binária se, após o período de teste, houve a compra do software ou não.

Na fase de limpeza, todas as linhas com valores faltantes no tempo de uso ou no envio de ordens foram eliminadas, além de testes solicitados, porém não ativados pelo usuário. Assim, a base original foi reduzida para 9.006 registros (linhas). Outro procedimento de limpeza de dados foi a exclusão de *outliers* encontrados na base, como por exemplo, erros nas horas de uso ou usuários com dias de testes maiores que o padrão, os quais podem ser resultado de promoções ou cortesias praticados pela empresa. A base foi separada em duas classes balanceadas, contendo 4.477 registros de assinaturas após a realização do teste e 4.529 não assinaturas. Com acurácia aleatória de 51,34%, favorece o desempenho dos algoritmos de classificação (Pachidi *et al.,* 2014).

Após pré-processamento, os dados foram normalizados para que assim nenhuma diferença na grandeza dos números afete os algoritmos de descoberta ou crie viés em seus resultados. A normalização dos dados também contribui com aumento do poder de processamento, que pode ser muito importante ao se trabalhar com base de dados grandes (Aggarwal, 2015). Ao fim da transformação e do pré-processamento, dados categóricos como produto testado, localização e canal de teste foram transformados em valores numéricos; com isso o número de colunas na base de dados totalizou 63.

Para aplicação dos algoritmos de descoberta a base de dados transformada foi dividida em duas partes: 80% dos registros virara a base de treino, para a fase de aprendizado do algoritmo, e os 20% foram reservador para teste. Todos os algoritmos aplicados (Regressão Logística, Kernel SVM, *Random Forest* e *Naïve* Bayes) foram configurados para classificar os dados de entrada nas classes como **não assinante** ou **assinante** após a realização do teste.

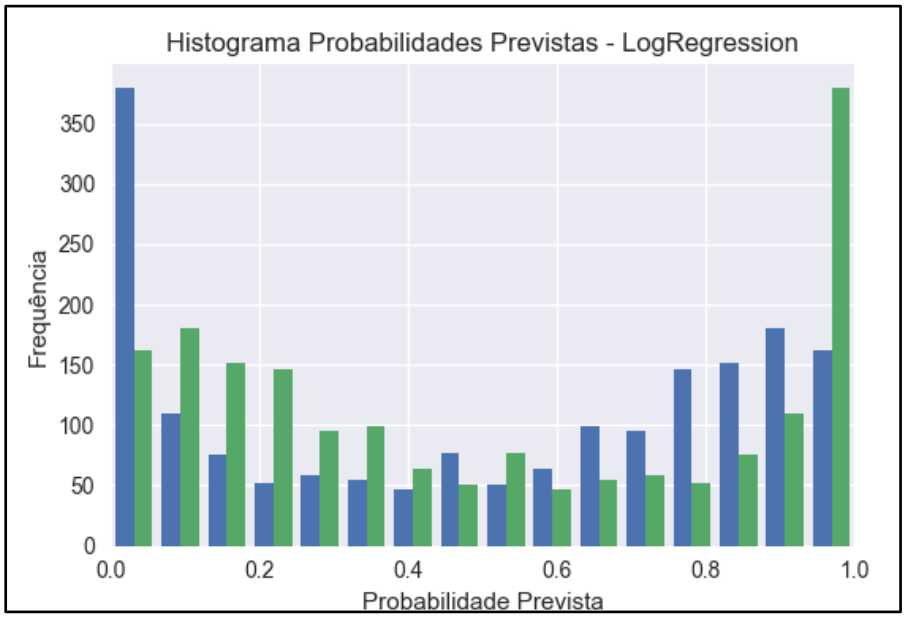
Na avaliação de cada modelo de classificação binário foram comparadas acurácias máximas, falsos positivos e falsos negativos através da matriz de confusão e a área sob a curva (AUC) de cada modelo pelo gráfico da curva ROC (*Receiver Operating Characteristics*). A matriz de confusão é a forma de representação da correlação de informações dos dados de referência (compreendido como verdadeiro) com os dados classificados.

Os valores são resultados, também, de métodos de otimização aplicados após a execução dos algoritmos. Dois métodos foram aplicados com esse intuito. Primeiro, o método de validação cruzada *K-Fold*, o qual aplica a base de treino e a base de teste em diferentes partes da base de dados. No caso deste estudo, foram 10 sub-bases, evitando assim possíveis variações na base. O segundo método de otimização aplicado foi o de *Grid Search*, em que variam-se determinados parâmetros de entrada dos algoritmos de descoberta a fim de buscar um melhor resultado em termos de acurácia do modelo.

4.1 REGRESSÃO LOGÍSTICA

No algoritmo de Regressão Logística, através do processo de seleção na fase de pré-processamento chegou-se ao número de colunas com a melhor performance para o modelo: a base de treino utilizada com 25 colunas classificadas pelo método do qui-quadrado resultando em uma acurácia máxima de 72,08%, com um desvio padrão de 1,37% com o limite de 0,5 na classificação de positivo ou negativo.

Figura 1 – Distribuição de probabilidades Regressão Logística



Fonte: elaborado pelos autores.

A distribuição de probabilidades do modelo é vista na Figura 1. A distribuição mostra uma concentração nos valores de 0 para valores classificados como não assinaturas e no valor de 1 para valores classificados como assinante, ambas as classes com um aumento de valores próximo a extremidade oposta.

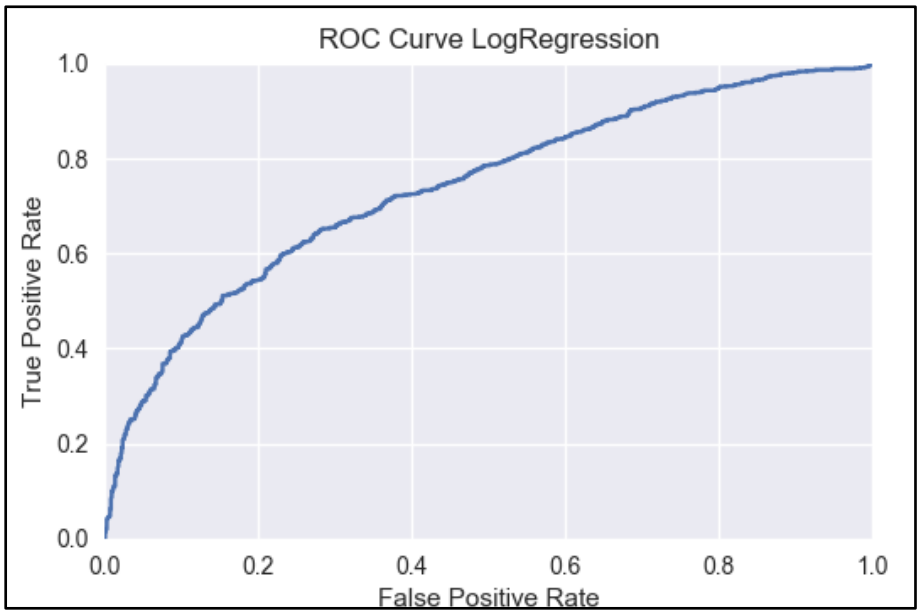
Na matriz de confusão, que mostra as frequências de classificação para cada classe do modelo, os maiores erros foram encontrados no quadrante de falsos negativos, isto é, o modelo indicou que o usuário não iria assinar o software mas na verdade o usuário realizou a assinatura, mas são encontrados tanto falsos positivos quanto falsos negativos, o que justifica a cauda de cada classe próximas a outra extremidade na distribuição de probabilidades, mostrados no Quadro 1.

Quadro 1 - Matriz de Confusão Regressão Logística

		Previsto	
		0	1
Real		763	161
		342	536

Fonte: elaborado pelos autores.

Figura 2 – Curva ROC Regressão Logística



Fonte: elaborado pelos autores.

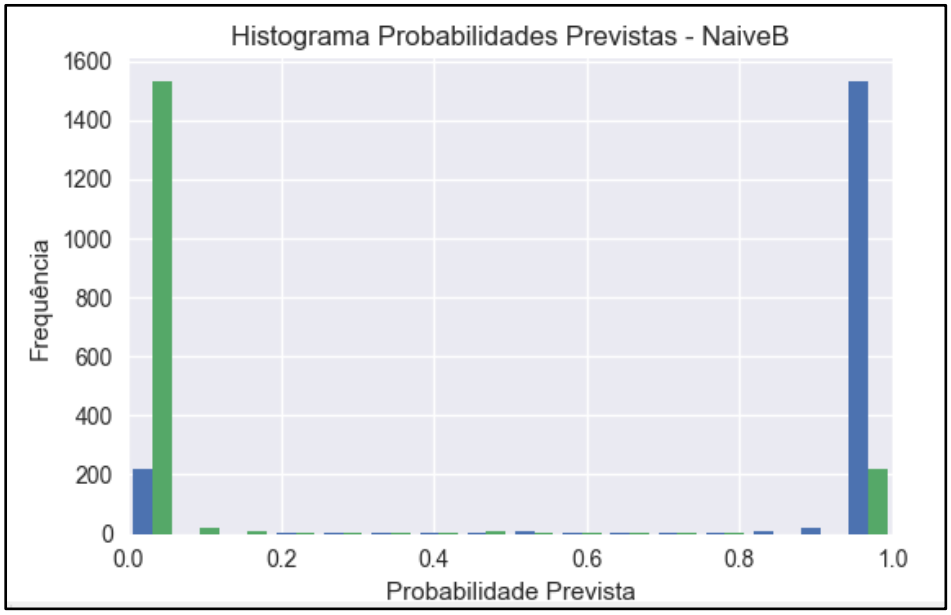
Na análise da curva ROC na Figura 2, a área sob a curva do modelo, que pode variar de 0 a 1, apresentou um valor de 0,7368, a qual leva em consideração todos os limites possíveis na classificação do modelo, utilizando a taxa de falso positivos e a taxa de positivos verdadeiros para determinar a habilidade geral de classificação do

modelo, método de avaliação mais indicado para classificações binárias (Ling *et al*., 2003; Han *et al*., 2005; James *et al*., 2013; Aggarwal, 2015).

4.2 NAÏVE BAYES

Para o algoritmo de *Naïve* Bayes foram realizados os mesmos procedimentos de seleção das variáveis, buscando o melhor desempenho possível para o modelo nas condições do estudo. A acurácia máxima do algoritmo de *Naïve* Bayes, com 0,5 de limite de classificação, foi de 69,08% e um desvio padrão de 5,86%. A distribuição das probabilidades do algoritmo pode ser observada na Figura 3. Os resultados apresentados foram resultado da seleção de 25 colunas através do método do qui-quadrado. Na distribuição, diferente da Regressão Logística, os valores ficaram concentrados em mais de 90% nos extremos, próximos a 0 e 1.

Figura 3 – Distribuição de probabilidades Naïve Bayes



Fonte: elaborado pelos autores.

Na matriz de confusão, no Quadro 2, os erros ficaram concentrados no quadrante de falsos negativos, isto é, casos em que o modelo havia previsto uma não assinatura que, na realidade, resultou em assinatura.

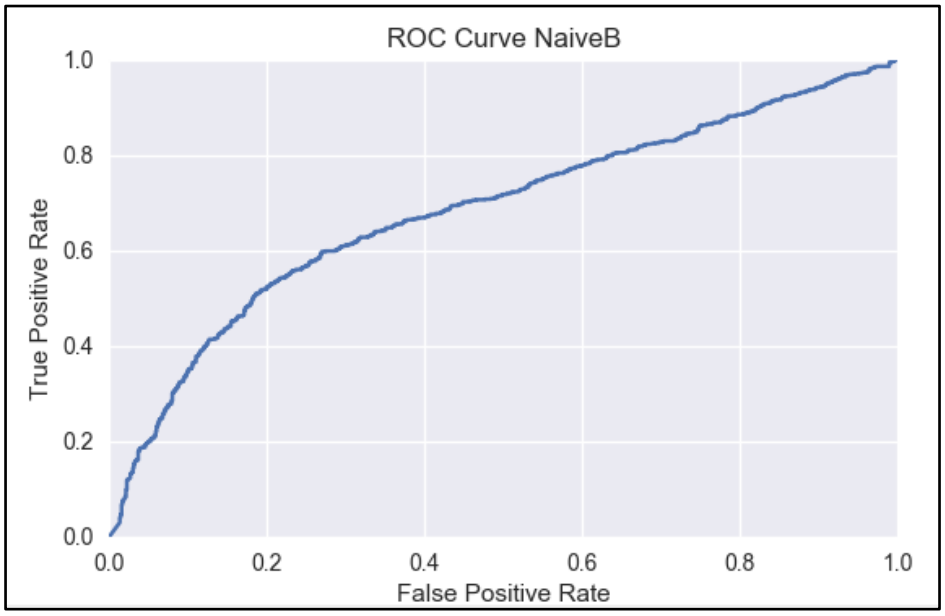
Quadro 2 - Matriz de Confusão Naïve Bayes

		Previsto	
		0	1
Real		855	46
		511	390

Fonte: elaborado pelos autores.

Na Figura 4, pode-se ver a curva ROC do modelo de *Naïve* Bayes; a área apresentada pelo modelo sob a curva foi de 0,6842.

Figura 4 – Curva ROC Naïve Bayes

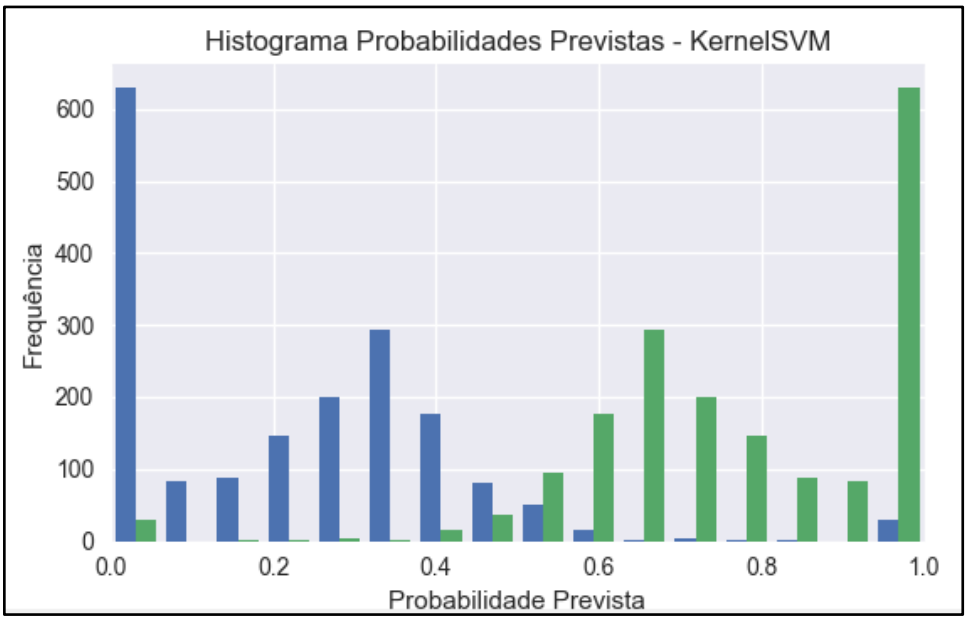


Fonte: elaborado pelos autores.

4.3 KERNEL SVM

Para o algoritmo de Kernel SVM foi utilizada a versão de classificação do vetor de suporte.

Figura 5 – Distribuição de probabilidades Kernel SVM



Fonte: elaborado pelos autores.

Com base no RBF Kernel, apontado pelo método de *Grid Search* como parâmetros mais apropriados para o modelo, resultando na seleção de 30 colunas pelo método da análise de significância do valor-F. Desta forma, a acurácia máxima do algoritmo ficou em 74,19%, com desvio padrão da acurácia de 3,26%. A distribuição da probabilidade ficou com os valores de probabilidade de não assinatura em torno de 0,3 e de assinatura em 0,7 e picos nos extremos, visto na Figura 5.

A matriz de confusão é apresentada no Quadro 3. Os resultados foram semelhantes ao modelo de Regressão Logística, com erros concentrados no quadrante de falsos negativos.

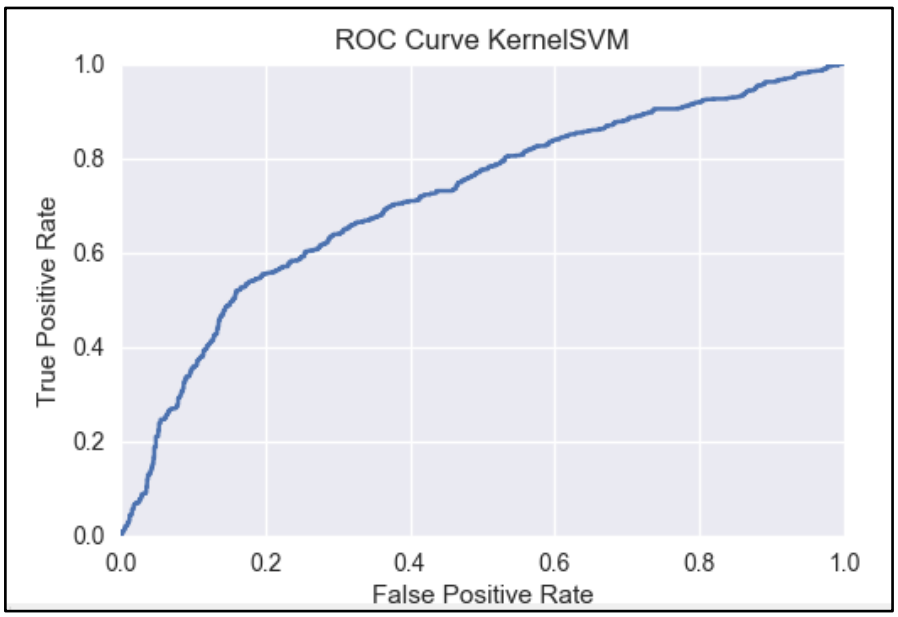
Quadro 3 - Matriz de Confusão Kernel SVM

		Previsto	
		0	1
Real		823	122
		343	514

Fonte: elaborado pelos autores.

O algoritmo Kernel SVM apresentou uma AUC de 0,7151; a curva pode ser vista na Figura 6.

Figura 6 – Curva ROC Kernel SVM

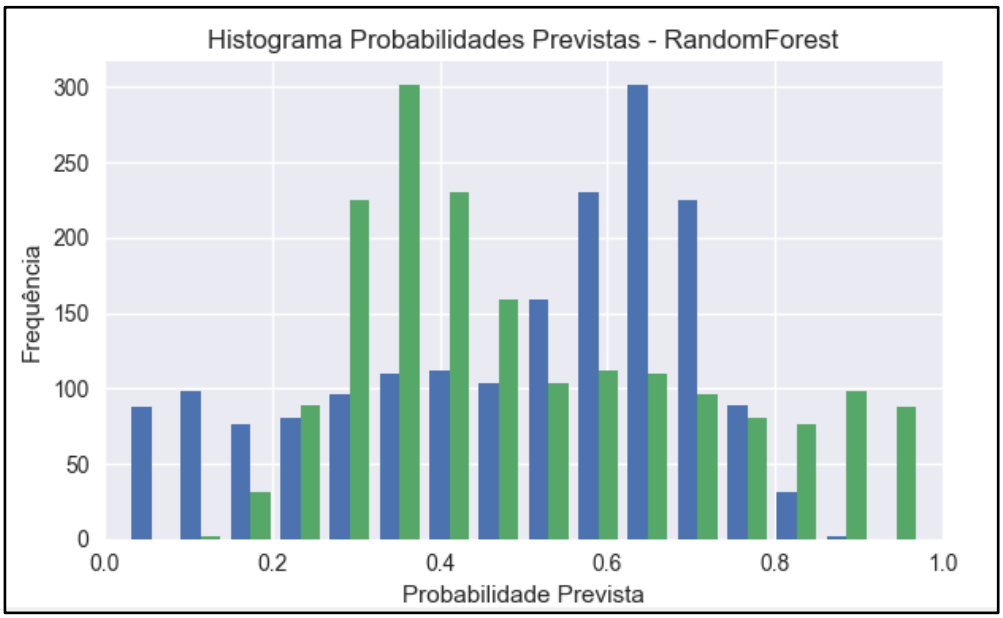


Fonte: elaborado pelos autores.

4.4 RANDOM FOREST

Com o algoritmo de *Random Forest*, a escolha de parâmetros otimizados pelo método de *Grid Search* indicou a utilização de 100 estimadores de árvore de decisão e critério definido como entropia, além da utilização de todas as colunas da base de treino. Esta aplicação teve uma acurácia máxima de 73,36%, com um desvio padrão de 1,69%, valores estes, encontrados com 0,5 de limite de classificação, com a distribuição observada na Figura 7. A distribuição de probabilidades do algoritmo resultou com os picos de cada classe de probabilidade próximas aos valores de 0,35 e 0,65.

Figura 7 – Distribuição de probabilidades Random Forest



Fonte: elaborado pelos autores.

Na matriz de confusão, os resultados estão mais distribuídos entre falsos positivos e falsos negativos do que os vistos nos outros modelos, como apresenta o Quadro 4.

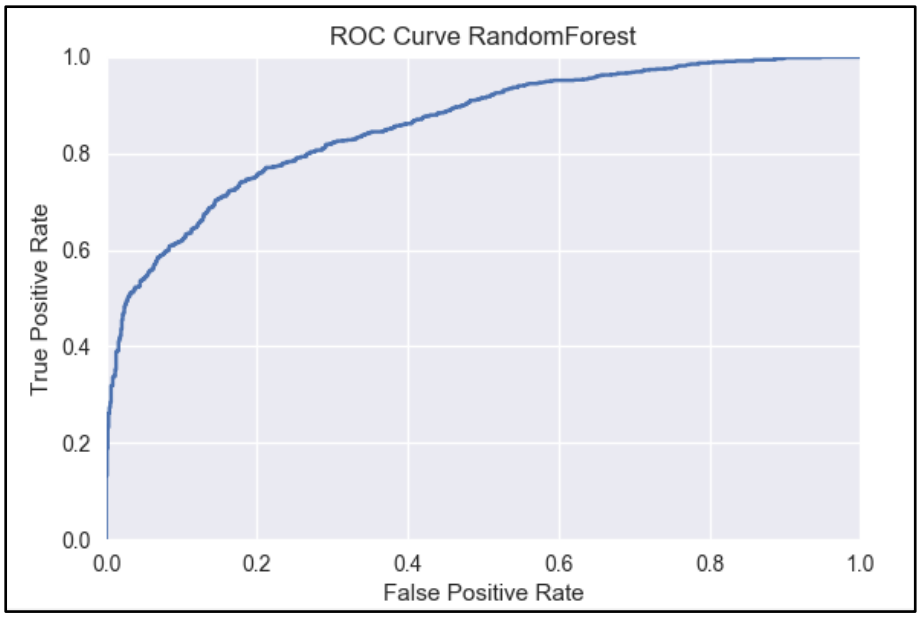
Quadro 4 - Matriz de Confusão Random Forest

		Previsto	
		0	1
Real		780	181
		299	542

Fonte: elaborado pelos autores.

Analisando a curva ROC na Figura 8 do modelo de *Random Forest*, a área sob a curva foi de 0,8592, apresentando o melhor valor dentro os modelos testados pelo presente estudo.

Figura 8 – Curva ROC Random Forest



Fonte: elaborado pelos autores.

Levando em consideração a natureza da pesquisa e os objetivos, a presença mais elevada de falsos negativos gera menos danos para o negócio que falsos positivos, pois nos falsos negativos seriam levados em consideração nas estratégias de marketing clientes que já iriam realizar a assinatura. Porém, para fins do objetivo do estudo este tipo de erro do algoritmo de classificação é prejudicial, já que o objetivo de gerar conhecimento útil para a área de marketing seria afetado por esforços desnecessários apontados pelo modelo como falsos positivos. A partir dos modelos testados e avaliados neste estudo indica-se pela utilização do algoritmo de *Random Forest*. Este apresenta uma AUC maior que os demais algoritmos e dentro de um nível considerado bom, segundo Batterham *et al.* (2017). Para modelos de forma geral, com a área abaixo da curva a partir de 0,84, e uma acurácia semelhante aos demais modelos, o conhecimento gerado serviria de maneira razoável como guia para as equipes direcionarem esforços para aqueles usuários com baixas probabilidades de compra. O desempenho razoável na acurácia do algoritmo pode ter relação com a falta de representatividade das variáveis utilizadas para explicar e prever a variável de saída, a escolha equivoca do modelo aplicado ou o nível complexidade do problema (Han *et al.*, 2005; Aggarwal, 2015).

5 CONCLUSÃO

Este estudo teve objetivo da identificação e seleção de variáveis para aplicação de *data mining* de classificação em uma base de dados de usuários, a fim de gerar um direcionamento de esforços para a equipe de marketing da empresa. Para atingir o objetivo, a metodologia de DCBD foi aplicada, para assim, gerar conhecimento através da base de dados extraída e auxiliar na classificação do usuário em teste. Quatro algoritmos de descoberta foram testados e avaliados, são eles, Regressão Logística, Kernel SVM, *Random Forest* e *Naïve* Bayes. Na avaliação dos modelos aplicados houve a análise da curva ROC de cada um, além da média AUC, e a análise da acurácia com auxílio da matriz de confusão.

Como resultado, tem-se a indicação da utilização do algoritmo *Random Forest*, o qual obteve o melhor resultado dentre os modelos testados (com o maior valor de AUC) de 0,8592, resultado considerado bom. Mesmo com o desempenho razoável por parte da acurácia, o modelo já pode servir de direcionador para o objetivo do estudo, visto que a empresa não utiliza atualmente, qualquer tipo de suporte. Outro ponto que favorece a escolha é a maior presença de falsos negativos, que gerariam esforços desnecessários pelas equipes, pois se tratam de clientes com probabilidade maior de assinatura, ao contrário do classificado pelo modelo.

REFERÊNCIAS

Aggarwal, C. C. (2015). *Data mining: The textbook*. Ed. Springer, New York.

Batterham, M., Neale, E., Martin, A., Tapsell, L., (2017). Data mining: Potential applications in research on nutrition and health. *Dietitians Association of Australia, Nutrition & Dietetics* 74, pp. 3–10.

Cooley, R., Mobasher, B., & Srivastava, J., (1997). *Web Mining: Information and Pattern Discovery on the World Wide Web*. University of Minnesota.

E-bit, (2017). *Relatório Webshoppers*, edição 35.

Etzioni, O., (1996). *The World Wide Web: Quagmire or gold mine*. Communications of the ACM, 39.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P., (1996). From data mining to knowledge discovery: An overview. *In Advances in knowledge discovery and data mining*. AAAI/MIT Press.

Frawley,W. J., Piatetsky-Shapiro, G. & Matheus, C., (1992). Knowledge Discovery in databases: An Overview. *AI Magazine*, vol. 13, number 3.

Han, J., Kamber, M., (2005). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers Inc.

James, G., Witten, D., Hastie, T., Tibshirani, R., (2013). *An Introduction to Statistical Learning with Applications* in R. Ed. Springer

Karuna, P. Joshi, Anupam Joshi, Yelena Yesha, and Raghu Krishnapuram, (1999). *Warehousing and Mining Web logs*, ACM.

Kotler, P., Kartajaya, H. & Setiawan, I., (2010). *Marketing 3.0: From Products to Customers to the Human Spirit*. s.l.:John Wiley & Sons, Inc..

Ling, X. C., Huang, J., Zhang, H., (2003). AUC: a Better Measure than Accuracy in Comparing Learning Algorithms. *Lectures notes Computer in Computer Science*, vol. 2671, Springer.

Manson, N.J., (2006). Is operations research really research? *Orion*, 22(5), 155-180.

Miguel, P.A., et al. (2012). *Metodologia de pesquisa em engenharia de produção e gestão de operações*. 2. ed. Rio de Janeiro: Elsevier: ABEPRO.

Mudiraj, P.V.G.S., Jabber, B., David, K. R., (2011). Web Mining: An Overview. International *Journal of Electronics Communication and Computer Engineering*. Volume 2, Issue 2.

Pachidi, S., Spruit, M., Weerd, I. Van de, (2014). Understanding users' behavior with software operation data mining. *Computers in Human Behavior* 30, pp 583–594, Elsevier.

Poongothai, K., Parimala, M. and Sathiyabama, S., (2011). Efficient Web Usage Mining with Clustering. *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 6, No 3.

Ryan, D. & Jones, C., (2012). *Understanding Digital Marketing: Marketing Strategies for Engaging the Digital Generation*, 2a Edição. s.l.:Kogan Page Limited.

Sen, S., (1998). An Overview of Data Mining and Marketing, Fordham University, Proceedings of the 1998 Academy of Marketing Science Annual Conference, Springer.

Smith, K. T., (2011). Digital Marketing Strategies that Millennials Find Appealing, Motivating, or Just Annoying. *Journal of Strategic Marketing*. V. 19, 2011, I6, pp. 489-499.

Tharenou, P., Donohue, R., Cooper, B., (2007). *Management Research Methods*. Cambridge University Press, Nova York.

Tripp, D., (2005). *Pesquisa-ação: uma introdução metodológica*. Educação e Pesquisa, São Paulo, 31, 443-466.