TÉCNICAS DE *DATA MINING* NA CLASSIFICAÇÃO DE USUÁRIOS EM TESTE DE UM SOFTWARE DO MERCADO FINANCEIRO

Natan Hermano de Maman

Universidade Federal do Rio Grande do Sul (UFRGS)
natan.maman@ufrgs.br

RESUMO

O estudo apresentado aborda a aplicação de técnicas de *data mining* de classificação em dados de usuários no período de teste de um software para operações no mercado financeiro. O objetivo da pesquisa-ação realizada predizer a propensão dos usuário à assinatura do software a partir da aplicação de técnicas de data mining para a identificação e seleção de variáveis que permitam a classificação (alocação) nas categorias assinantes e não-assinantes. Através do processo de descobrimento de conhecimento em banco de dados foram aplicados quatro diferentes algoritmos, dessa forma, possibilitando a extração de conhecimento útil através dos dados. Os algoritmos aplicados foram os de Regressão Logística, *Support Vector Machine*, *Naïve* Bayes e *Random Forest*. A análise dos resultados aponta que o algoritmo de *Random Forest* apresentou o melhor desempenho na classificação dos usuários comparado aos outros modelos testados no estudo, com uma área sob a curva de 0,85, o que em parâmetros gerais, pode ser considerado um bom resultado.

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

ABSTRACT

The present study approaches the application of data mining techniques on classification of user data during the trial period of a software for financial market operations. The aim of this action research was to predict users' propensity to subscribe to the software through the application of data mining techniques to identify and select variables that allow the classification (allocation) of subscribers and non-subscribers. Through the process of knowledge discovery in database were applied four different algorithms, so, being possible 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 presented the best performance in the classification of users compared to the other models tested in this study, with an area under the curve of 0.85, which in general parameters is associated as a good result.

Keywords: Data Mining, Classification, Software, User

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 através 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 evoluiu para atender as necessidades oriundas com a utilização dessas novas formas de comércio, resultando no que hoje é referido como marketing digital. O marketing digital pretende, portanto, entender o comportamento das pessoas e a forma com que elas se relacionam com as tecnologias e, desta forma, 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 ambos, empresas e consumidores. No caso das empresas, destaca-se um movimento de disponibilização de produtos ou serviços usando o meio digital, com importantes reflexos na competitividade. Já para os consumidores, destacam-se as diversas alternativas disponíveis que surgiram rapidamente 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, visto que as interações digitais tipicamente geram 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 e, além disso, analisar e compreender os padrões de comportamento é uma das formas de criar conhecimento como classificações ou predições, além de se tratar de uma das áreas mais importantes da ciência da informação, 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).

Data mining, ou mineração de dados, é um dos estágios do processo de DCBD 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, este artigo tem como objetivo predizer a propensão à assinatura de usuários cadastrados para um período gratuito de teste de um software para operações no mercado financeiro. A partir da aplicação de *data mining* de classificação, busca-se alocar os usuários nas classes de assinantes ou não assinantes ao fim da realização do teste, de modo a gerar conhecimento útil e apoiar na tomada de decisão estratégica na área de marketing. O método de *data mining* é o de classificação a partir de dados relativos às características do usuário coletadas no momento de cadastro no *e-commerce*, etapa necessária para liberação do período de teste, e dados de uso do software instalado no computador do usuário durante o período de teste.

Justifica-se o tema pelo crescimento exponencial de usuários no meio digital no mundo e o nível de informação e conhecimento gerados pelas técnicas de *data mining*, além de possuir uma aplicação direta em e-commerce, CRM, análise da web e sistemas de informação da web, possibilitando entender e atender o usuário de uma melhor forma (Park, 2008; Mudiraj, 2011). Desta forma, pretende-se gerar informações que sejam consideradas úteis

para o negócio da empresa onde está sendo aplicado o processo de DCBD.

Este artigo segue a seguinte estrutura: na segunda seção é apresentado um referencial teórico sobre as definições de DCBD, *data mining* e *data mining* de classificação com suas aplicações. Na seção seguinte são descritas a situação-problema de uma empresa atuante no meio digital e os procedimentos metodológicos aplicados para a solução do problema. Na quarta seção são apresentados os resultados obtidos. Por fim, na quinta seção, são apresentadas as conclusões e considerações finais deste trabalho.

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*, a qual 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*, como apresenta a Figura 1, 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).

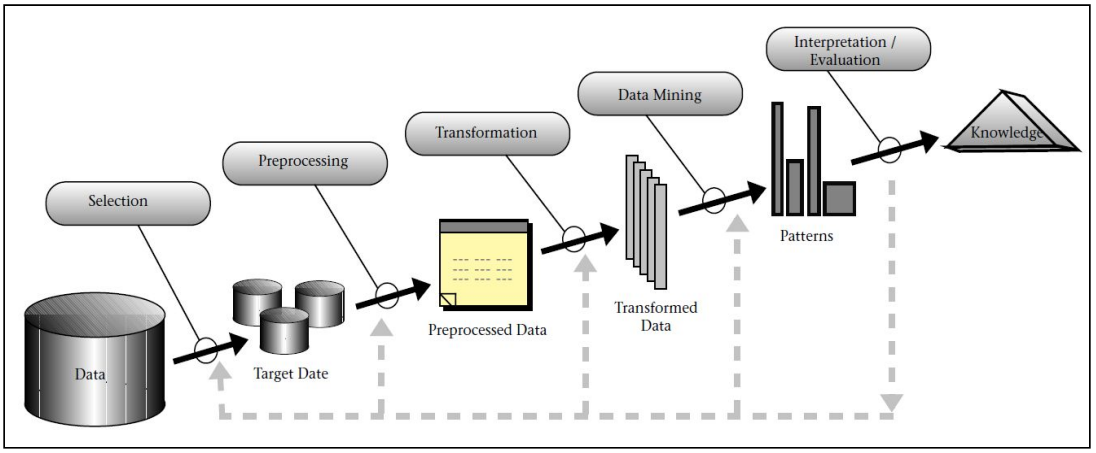

Figura 1 – Etapas do processo de DCBD. Fonte: Fayyad *et al*., 1996

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, cuja propósito é encontrar todas associações e correlações entre os dados em que a presença de certos itens implica na presença de outros, com um grau de confiança estabelecido. Aplicações deste método em *data mining*, por exemplo, 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. Um exemplo do resultado do método de classificação é o comportamento de compra do consumidor, assim, com aplicações diretas em marketing para o público-alvo. Por fim, 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. Oliveira *et al.* (2017), por exemplo, apresentaram a aplicação de *data mining* de classificação no apoio à tomada de decisão na escolha do ponto de pesca mais propício. Através da comparação dos resultados dos modelos de *Random Forest*, *Support Vector Machine* (SVM), Redes Neurais e Regressão Múltipla, o qual já era utilizado na escolha do local até então, o modelo classifica qual ponto de pesca deveria ser escolhido pelos pescadores, além de qual embarcação utilizar para aquela determinada espécie de peixes. O modelo utilizando o algoritmo *Random Forest* apresentou o melhor resultado, com um r^2 de 0,89, comparados com o modelo de regressão múltipla utilizado anteriormente, com r^2 de 0,79.

Outra forma de buscar maior acurácia nos resultados é utilizando métodos de aprendizado de conjunto, em que diferentes algoritmos trabalham em conjunto para aprimorar os resultados, como o caso do estudo de Tseng *et al.* (2017), no qual sugere-se um modelo para classificar por importância os fatores de risco e diagnosticar a recorrência do câncer de ovário em pacientes. Através de um banco de dados de pacientes, cinco algoritmos de *data mining* foram combinados e treinados para aumentar o desempenho do modelo: SVM, C5.0, *Random Forest*, *Extreme Learning Machine* (ELM) e MARS. Os resultados obtidos pelo método de aprendizado em conjunto superou os algoritmos testados de forma individual, gerando um modelo útil para diagnosticar câncer de ovário. De modo similar, Demigha (2016) propôs uma abordagem de *data mining* de classificação utilizando o algoritmo *Naïve* Bayes para detectar câncer de mama em pacientes, através de um banco de dados de pacientes cadastrados eletronicamente, treinou-se o modelo e obtiveram uma acurácia de 91,50% na predição e detecção do câncer.

Com o objetivo de testar diferentes formas de pré-processamento dos dados e analisar o impacto que esse fator implica na acurácia de algoritmos de *data mining* de classificação, Crone *et al.* (2005) aplicaram os algoritmos de classificação C4.5, Redes Neurais e SVM em dados de campanhas de marketing por e-mail, gerando resultados em dados sem pré-processamento e com diferentes níveis de pré-processamento. O modelo, após o pré-processamento, selecionava quais consumidores eram mais favoráveis à compra (e que, portanto, deveriam receber o e-mail da campanha); como resultado desse estudo, concluiu-se que o pré-processamento afeta diretamente no desempenho do modelo de descoberta, visto que melhorou os resultados dos três algoritmos testados, além do tempo de processamento.

Uma aplicação recorrente de *data mining* de classificação é na identificação de sinais fraudulentos em contas ou transações bancárias. Li *et al.* (2012) realizaram um estudo nessa área, aplicando técnicas de *data mining* para identificar e evitar golpes ou fraudes desse tipo. Através dos algoritmos de Redes de Bayes e regra de associação, utilizaram-se dados das contas dos clientes, como depósitos, saques, transações e relacionadas com contas classificadas como fraudulenta, obtendo resultados consistentes com o modelo.

Batterham *et al.* (2017), a fim de demonstrar a aplicabilidade de *data mining* na área da nutrição, além de comparar a utilidade e resultado dos diferentes modelos, desenvolveram seu estudo através de uma base de dados de paciente contendo níveis de glicose e gordura, IMC,

entre outros. Comparando resultados dos algoritmos de Árvore de Decisão, *Random Forest*, Regressão Logística, SVM e Redes Neurais, o modelo buscava classificar quais pacientes estavam aptos a realizar uma determinada atividade física. O desempenho do modelo foi avaliado pela área sob a curva (AUC) de cada algoritmo, e pela avaliação de erro, sendo que o melhor resultado foi encontrado pelo modelo Árvore de Decisão com 18% de erro total.

Com o objetivo de analisar o comportamento de clientes de um banco de pessoas físicas, Hu (2005) realizou um estudo utilizando métodos de *data mining* para identificar quais clientes do banco eram mais propensos a cancelar o serviço. A análise da taxa de cancelamento foi desenvolvida aplicando os algoritmos de Árvore de Decisão, *Naïve* Bayes e Redes Neurais, com algoritmos testados individualmente e na forma de aprendizado em conjunto. O modelo apresentou melhor acurácia na forma de aprendizado em conjunto na base de teste. Em seguida, uma aplicação prática foi realizada com 30 mil clientes com padrão de cancelamento selecionados pelo modelo de *data mining* de classificação, sendo metade contatada pelo banco e a outra metade não. Como resultado, a metade que recebeu contato teve taxa de cancelamento de 0,12% contra 5,6% da metade não contatada.

Outra aplicação de *data mining* de classificação na área médica pode ser vista em Ouali *et al.* (2006), cujo objetivo foi identificar a idade do início real da esquizofrenia em pacientes e outra variável relacionada à origem do desenvolvimento. Através dos dados clínicos e biológicos de pacientes, o estudo utilizou o algoritmo de Redes de Bayes para minerar os dados. Como resultado, permitiu propor limitações para a idade de início da doença nos pacientes com um embasamento mais preciso, além de auxiliar na determinação da variável contínua relacionada à origem do desenvolvimento neurológico da doença.

No estudo de Pachidi *et al.* (2014) sugere-se um modelo de mineração do uso de um usuário em um software, a fim de entender como o usuário final utiliza o software, além de identificar quais as funcionalidade mais utilizadas dentro do software, e dessas, quais podem ser expandidas ou descontinuadas. Com essas informações é possível melhorar a experiência do usuário, além de possibilitar inovação e diferenciação do produto. O modelo apresentado possui três passos. No primeiro, realiza-se uma análise de classificação com o algoritmo de Árvore de Decisão (CART) para identificar os fatores que influenciam as decisões dos consumidores. Na segunda etapa, através do algoritmo de *clustering,* define-se o perfil do usuário nos diferentes *clusters*. Por fim, utiliza-se as cadeias de Markov para analisar o

padrão sequencial dos usuários do software, isto é, quais caminhos são os mais percorridos. O modelo sugerido pelo estudo foi aplicado então em dados de usuários em teste de um aplicativo financeiro online. Após o pré-processamento, as três etapas do modelo e a avaliação dos resultados, concluiu-se que o modelo é válido para mineração do uso de usuários de um SaaS (*software as a service*); a classificação dos fatores que influenciam o usuário obteve 85% de acurácia, identificando cinco perfis de usuário e os caminhos mais percorridos pelos usuários em teste.

3. PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa, através da investigação da base empírica, intervenção e transformação de um sistema através da participação direta e ativa do pesquisador, pode ser classificada como pesquisa-ação, cuja abordagem na condução dos procedimento segue os critérios levantados por Tharenou (2007). Assim, este trabalho tem como objetivo gerar conhecimento para solução de problemas específicos e práticos existentes, logo, dito como de natureza aplicada (Manson, 2006). A abordagem é essencialmente quantitativa, em virtude da utilização de ferramentas matemáticas para coleta e análise de dados, e tem objetivo de cunho descritivo e prescritivo, 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. 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). A pesquisa-ação desenvolvida neste trabalho é conduzida seguindo o processo de DCBD encontrado na literatura e adaptado para a problemática da pesquisa, cuja estrutura das macro etapas é apresentada na Figura 2.

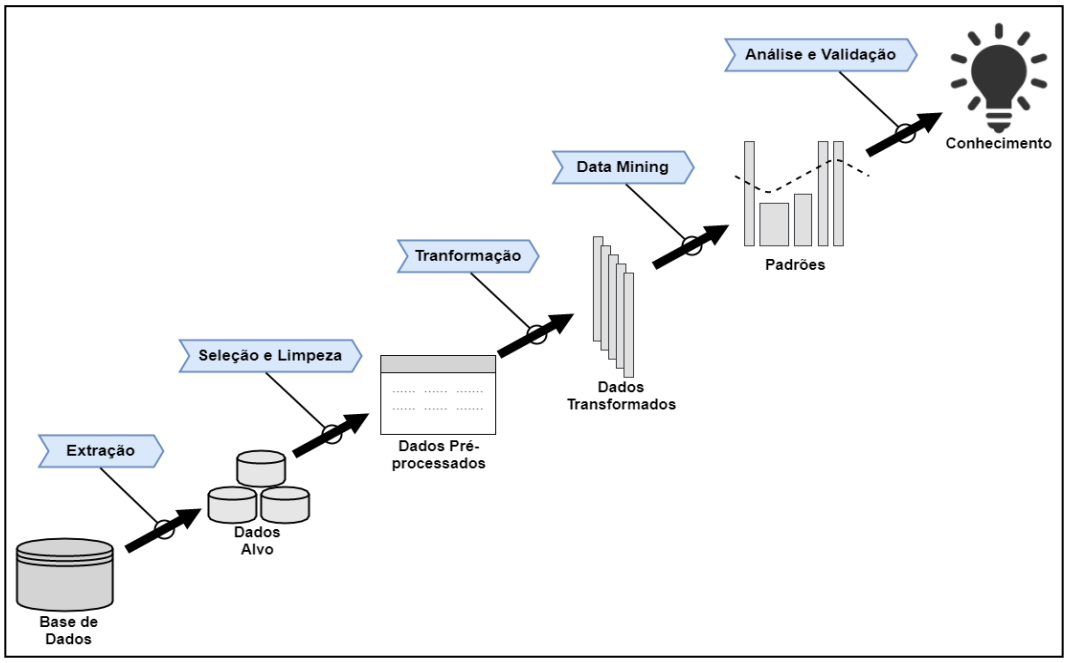

Figura 2 – Estrutura de macro etapas da pesquisa-ação.
Fonte: Adaptado de Han *et al*. 2005, Aggarwal, 2015.

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*, os quais precedem o uso efetivo do software, além de dados de tempo de uso da plataforma e ordens de compra ou venda de ativos registradas durante o teste, assim obtendo os dados alvo do estudo, os quais abrangem um período de primeiro de Julho de 2016 até Setembro de 2017. É importante ressaltar nessa etapa de extração a heterogeneidade dos dados, logo, o pré-processamento é fundamental (James *et al*., 2013; Aggarwal, 2015). 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). Outro ponto da transformação envolve a normalização dos dados, convertendo-os para a mesma escala, tornando possível 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, como por exemplo, campanhas direcionadas aos usuário indicados como mais propício a não contratar o software, buscando aumentar a chance de conversão.

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, e atua em um mercado de alta competitividade, onde a agilidade, qualidade e inteligência na gestão dos recursos tornam-se fundamentais na busca pelo resultado. Logo, destinar os recursos de forma correta aumenta a possibilidade de sucesso. Atualmente a empresa não faz uso de nenhuma técnica de classificação dos usuários em teste e acaba tratando todos da mesma forma estrategicamente.

4. RESULTADOS

A pesquisa teve início com a primeira etapa do processo de DCBD, a extração dos

dados internos da empresa em estudo, os quais 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, com 25.829 linhas e 25 colunas, onde cada linha é uma solicitação de teste feito por um usuário, representados pela Tabela 1.

A partir dos dados da fase de extração, realizou-se a fase de seleção e limpeza dos dados alvo do estudo, onde, 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 email 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.

Index	**Ativação**	**FimTeste**	**...**	**Ordens**	**CanalID**	**Venda.NaoVenda**
1	2016/07/01	2016/07/16	...	587	2	1
2	2016/07/01	2016/07/16	...	0	1	0
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
25.288	2017/09/30	2017/10/15	...	0	1	0
25.289	2017/09/30	2017/10/15	...	0	1	0

Tabela 1 - Representação Base de dados na fase de extração

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, grande parte devido ao elevado

número de teste com ativação porém sem tempo de uso por parte do usuário. Outro procedimento de limpeza 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 contém 4.477 registros de assinaturas após a realização do teste e 4.529 não assinaturas, com uma acurácia aleatória de 51,34%, apresentando assim duas classes balanceadas, o que favorece o desempenho dos algoritmos de classificação (Pachidi *et al.,* 2014).

Após o 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 reservados 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 de 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*).

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 bases de treino mais propensas a uma das classes de saída. 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. A distribuição de probabilidades do modelo é vista na Figura 3. 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.

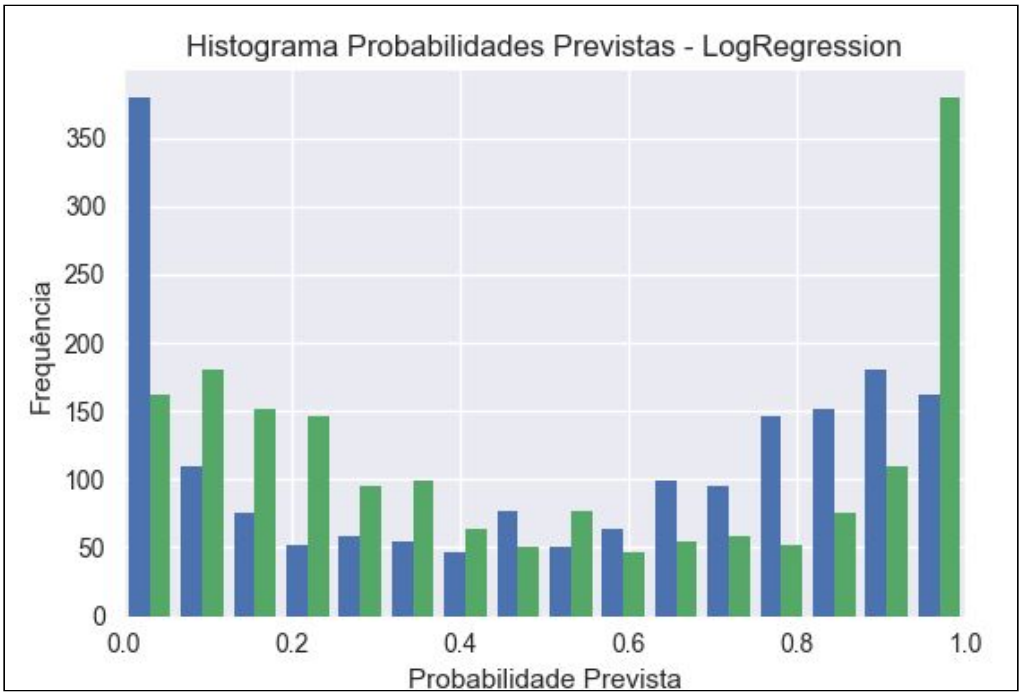

Figura 3 – Distribuição de probabilidades Regressão Logística

Na matriz de confusão 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.

	Previsto

		0	1
Real	0	763	161
	1	342	536

Quadro 1 - Matriz de Confusão Regressão Logística

Na análise da curva ROC na Figura 4, 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).

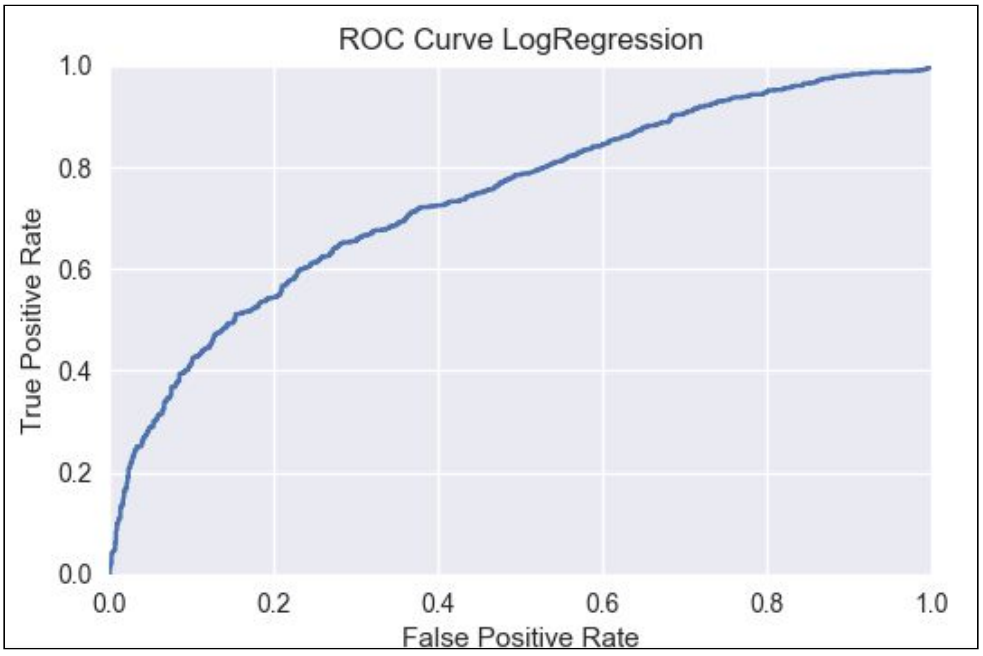

Figura 4 – Curva ROC Regressão Logística

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 5. 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.

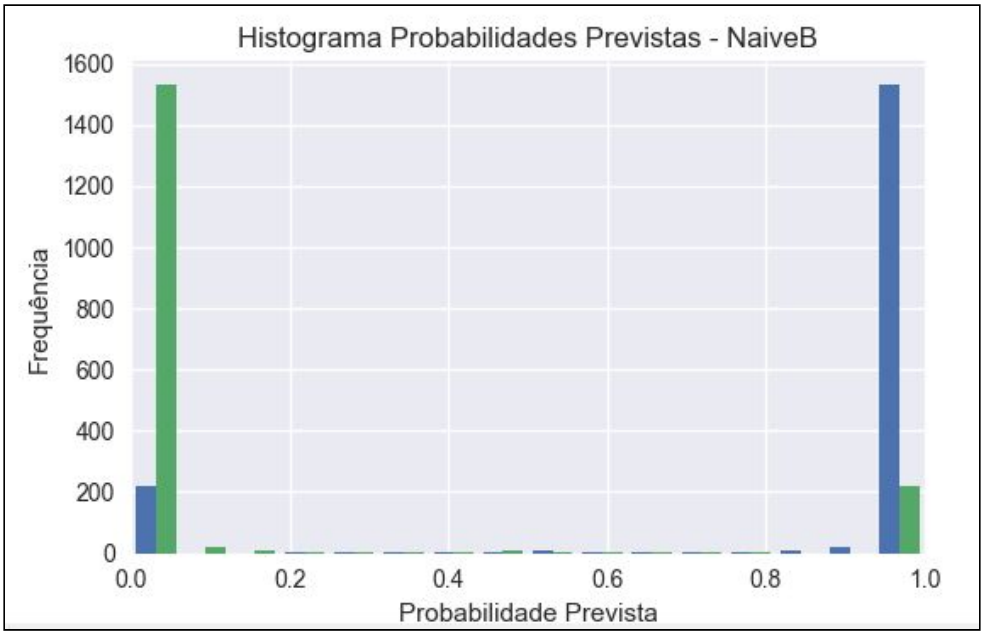

Figura 5 – Distribuição de probabilidades *Naïve* Bayes

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.

		Previsto	
		0	1
Real	0	855	46
	1	511	390

Quadro 2 - Matriz de Confusão *Naïve* Bayes

Na Figura 6, pode-se ver a curva ROC do modelo de *Naïve* Bayes; a área apresentada pelo modelo sob a curva foi de 0,6842.

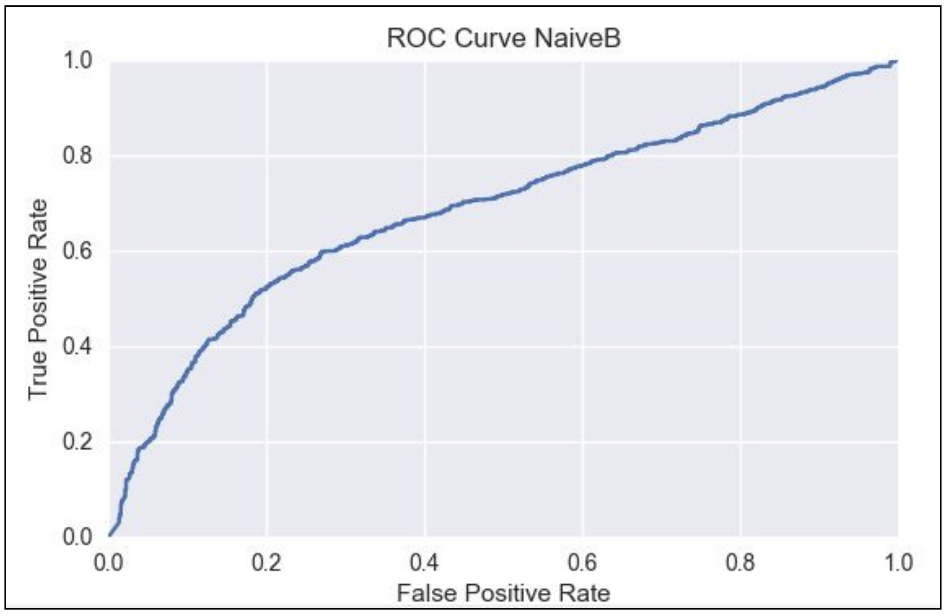

Figura 6 – Curva ROC *Naïve* Bayes

4.3 Kernel SVM

Para o algoritmo de Kernel SVM foi utilizada a versão de classificação do vetor de suporte, e 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 7.

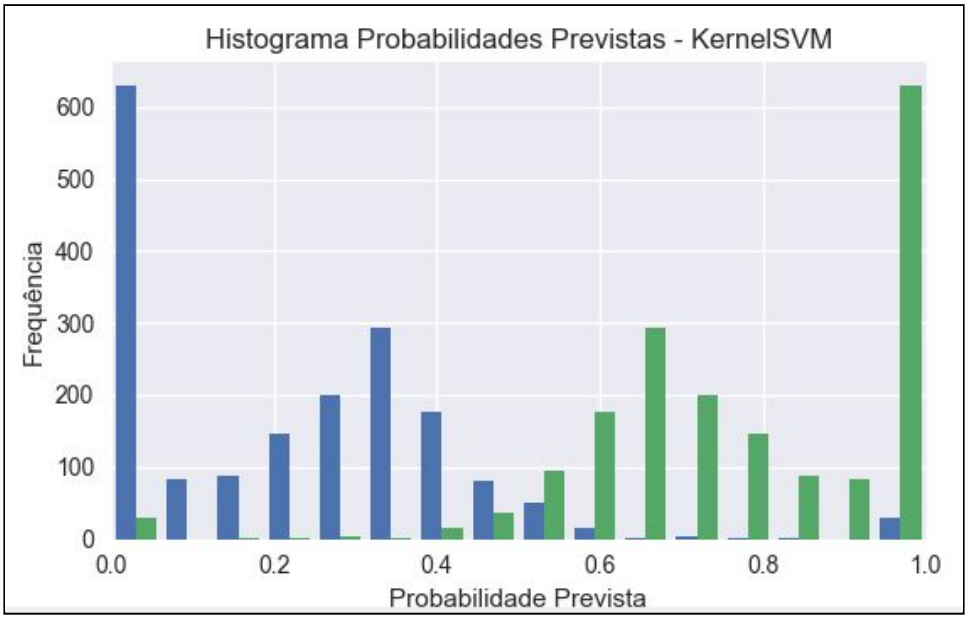

Figura 7 – Distribuição de probabilidades Kernel SVM

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.

		Previsto	
		0	1
Real	0	823	122
	1	343	514

Quadro 2 - Matriz de Confusão Kernel SVM

O algoritmo Kernel SVM apresentou uma AUC de 0,7151; a curva pode ser vista na Figura 8.

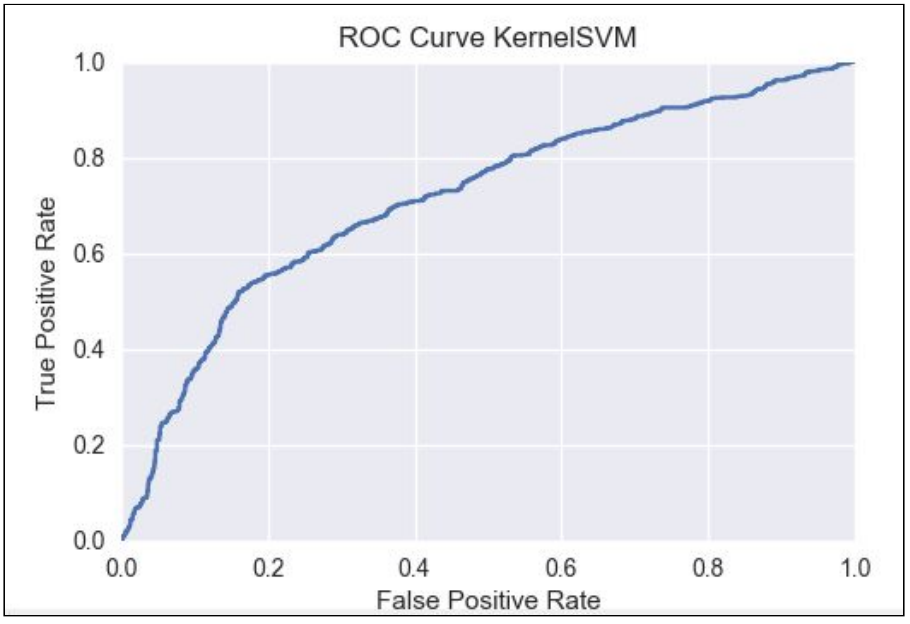

Figura 8 – Curva ROC Kernel SVM

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 9. A distribuição de probabilidades do algoritmo resultou com os picos de cada classe de probabilidade próximas aos valores de 0,35 e 0,65.

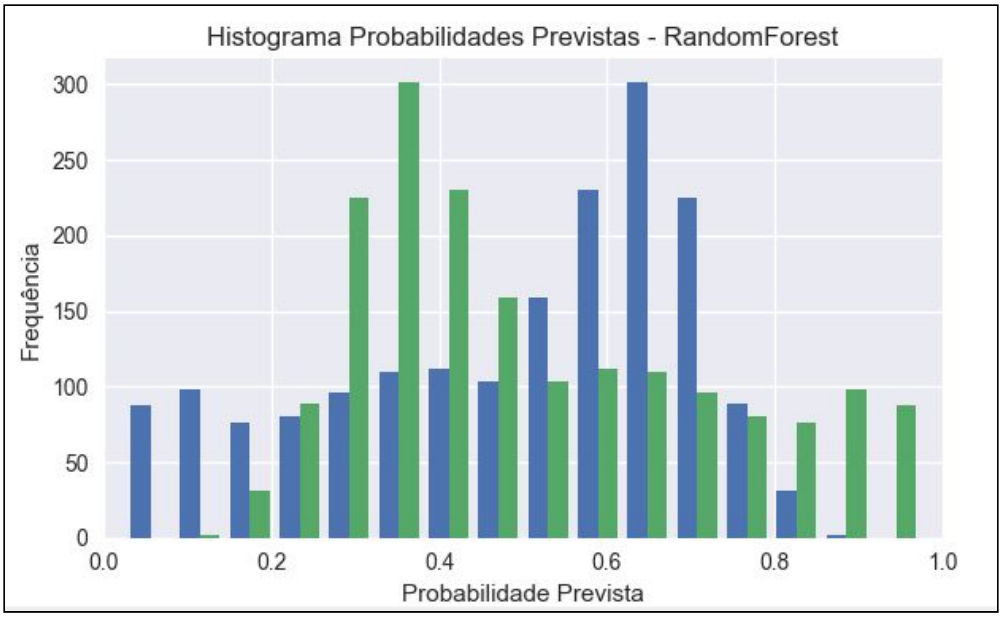

Figura 9 – Distribuição de probabilidades *Random Forest*

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.

		Previsto	
		0	1
Real	0	780	181
	1	299	542

Quadro 4 - Matriz de Confusão *Random Forest*

Analisando a curva ROC na Figura 10 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.

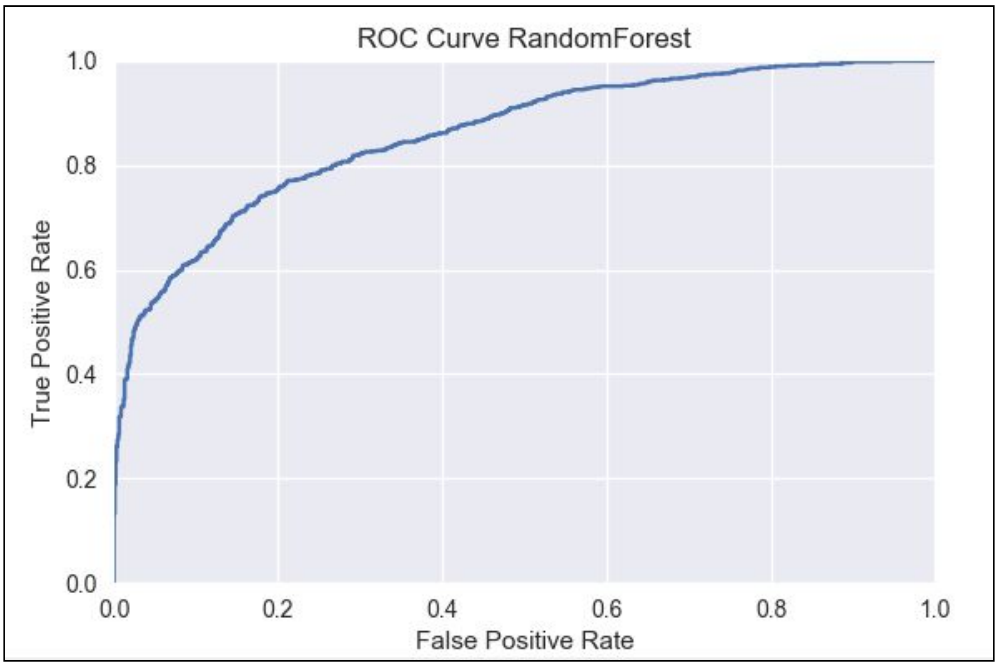

Figura 10 – Curva ROC *Random Forest*

Levando em consideração a natureza da pesquisa e os objetivos, a presença mais elevada de falsos negativos gera menos danos para o négocio 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*, por apresentar uma AUC maior que os demais algoritmos e dentro de uma nível considerado bom, segundo Batterham *et al.* (2017) para modelos de forma geral. Com a área de 0,8592, 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 como objetivo a identificação e seleção de variáveis para aplicação de *data mining* de classificação em uma base de dados de usuários que testaram um software, para classificá-los em assinantes ou não assinantes após o teste, a fim de gerar conhecimento útil a partir dos dados para a equipe de marketing da empresa em questão, direcionando os esforços para usuários com menos chance de adoção do software.

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 uma 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.

Para trabalhos futuros, sugere-se a aplicação de outro algoritmos não incluídos nesta pesquisa, como o de Redes Neurais ou o de K-NN, e até mesmo a tentativa de aprendizado em conjunto dos diferentes algoritmos (*ensemble learning*), para assim verificar se com a base atual é possível obter um desempenho maior do que apresentado pelo algoritmo *Random Forest*. Outras sugestões envolvem a utilização de outro método de seleção de variáveis, como o de Análise de Componentes Principais (PCA), ou inclusão de novas variáveis à base de dados que possam capturar fatores adicionais que possam influenciar na assinatura ou não de um usuário, assim, gerando novas pesquisas e comparando com os resultados encontrados neste estudo.

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.

Crone, S. F., Lessmann, S., Stahlbock, R., 2005. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research 173, pp. 781-800, Elsevier.

Demigha S., 2016. Mining knowledge of the patient record: The Bayesian classification to predict and detect anomalies in breast cancer. The Electronic Journal of Knowledge Management Volume 14, pp. 128-139.

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.

Hu, X., 2005. A Data Mining Approach for Retailing Bank Customer Attrition Analysis. Applied Intelligence 22, pp. 47–60, Springer.

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..

Li, S., Yen, D., Lu, W., Wang, C., 2012. Identifying the sings of fraudulent accounts using data mining techniques. Computers in Human Behavior 28, pp 1003-1013, Elsevier.

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., Fleury, A., Mello, C.H.P., Nakano, D.N., Lima, E.P., Turrioni, J.B., Ho, L. L., Morabito, R., Martins, R.A., Souza, R., Costa, S. E. G., Pureza, V., 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. raju, 2011. Web Mining: An Overview. International Journal of Electronics Communication and Computer Engineering. Volume 2, Issue 2.

Oliveira, M. M., Camanho, A. S., Walden, J. B., Miguéis, V. L., Ferreira, N. B., Gaspar, M. B., 2017. Forecasting bivalve landings with multiple regression and data mining techniques: The case of the Portuguese Artisanal Dredge Fleet. Marine Policy 84, pp. 110-118, Elsevier.

Ouali, A., Cherif, A. R., Krebs, M., 2006. Data mining based Bayesian networks for best classification. Computational Statistics & Data Analysis 51, pp. 1278 – 1292, Elsevier.

Park, S., Suresh, N. C., Jeong, B., 2008. Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. Data & Knowledge Engineering 65, Elsevier.

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.

Tseng, C., Lu, C., Chang, C., Chen, G., Cheewakriangkrai, C., 2017. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. Artificial Intelligence in Medicine 78, pp. 47-54, Elsevier.

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.