Revista Brasileira de Computação Aplicada, Novembro, 2019

DOI: 10.5335/rbca.v11i3.9077
Vol. 11, Nº 3, pp. 122–132
Homepage: seer.upf.br/index.php/rbca/index

ISSN 2176-6649

ARTIGO ORIGINAL

Uma solução de mineração de dados para concessão de cupons de descontos em comércio eletrônico: um estudo de caso

A data mining solution for e-commerce discount coupons: a case study

Rosalvo Ferreira de Oliveira Neto[1], Ricardo Argenton Ramos[1], Cleidson Drummond da Silva[1]

[1]Universidade Federal do Vale do São Francisco

*rosalvo.oliveira,ricardo.aramos@univasf.edu.br; cleidsondru@gmail.com

Recebido: 01/02/2019. **Revisado:** 14/08/2019. **Aceito:** 26/09/2019.

Resumo

Este artigo pretende responder seguinte pergunta de pesquisa: "como construir uma solução eficiente de mineração de dados para um sistema de cupom de desconto?". Assim, neste artigo é proposto uma solução de mineração de dados para responder a essa pergunta. A solução é constituída por quatro componentes: 1) uso da técnica *Random Forest* como classificador, 2) tratamento dos valores ausentes, 3) enriquecimento da base de dados através da construção de novas variáveis e 4) uso do método de *Kolmogorov Smirnov* para a escolha do ponto de corte para tomada de decisão. Um estudo experimental foi realizado para validar a eficiência da solução proposta. Os resultados mostraram a adequação do método ao problema e que a estratégia de aquisição de conhecimento proposta aumentou o poder preditivo. Por fim, os resultados mostraram que a estratégia de tratamento de valores ausentes possui influência no poder discriminatório da solução. A contribuição deste estudo é um direcionamento para construção de soluções de mineração de dados em *web-shop*, dando diretivas sobre qual método de mineração de dados utilizar, qual a melhor estratégia para tratamento de valores ausentes, como melhorar o poder preditivo através da aquisição de conhecimento e ainda como escolher o melhor ponto de corte.

Palavras-Chave: Mineração de dados; Comércio eletrônico; Valores Ausentes; Random Forest; Kolmogorov Smirnov.

Abstract

This article aims to answer the following research question: "How to build an efficient data mining solution for a discount coupon system?". Thus, here a data mining solution is proposed to answer this question. The solution consists of four components: 1) use of the Random Forest technique as a classifier, 2) treatment of missing values, 3) enrichment of the database through the construction of new variables, and 4) use of the Kolmogorov Smirnov method to choose from the cut-off point for decision-making. An experimental study is conducted to validate the efficiency of the proposed solution. The results showed the acceptability of the Random Forest method to the problem and that the proposed knowledge acquisition strategy increased the predictive power. Finally, the results showed that the strategy of treating missing values has an influence on the discriminatory power of the solution. The contribution of this study is a guide to the construction of web-shop data mining solutions, giving guidelines on which data mining method to use, which is the best strategy for treatment of missing values, and how to improve predictive power through the acquisition of knowledge and how to choose the best cutting point.

Keywords: Data Mining; Electronic commerce; Missing Value; Random Forest; Kolmogorov Smirnov.

1 Introdução

As estatísticas mais recentes sobre o uso de Internet no mundo apontam que cerca de 4 bilhões de habitantes possuem acesso à Internet [1]. A quantidade de usuários de Internet tem influenciado o setor de comércio eletrônico. Setor que tem exercido um importante papel na economia mundial.

Este setor movimentou só no Brasil no natal de 2018 o valor de R$9,9 bilhões [2]. Uma das principais consequências deste crescimento é a migração do comércio de varejo físico para web-shop. Essa migração fornece aos comerciantes novas oportunidades para chegar a potenciais compradores em grande escala. No entanto, também traz novos desafios, tais como as relações entre comerciante e consumidor. O comerciante não conhece o cliente, seus desejos e intenções diretamente. Por isso, é mais complexo incentivar a compra dos clientes através de um web-shop (Hop, 2013). Uma grande quantidade de dados é coletada durante cada visita de um cliente a um web-shop (Midha and Singh, 2018). A disponibilização de grandes volumes de dados possibilita a aplicação de mineração de dados, que quando aplicado corretamente ao *e-commerce*, pode dar ao comerciante uma vantagem competitiva significativa, aumentando o volume de negócios e a satisfação do cliente. Uma aplicação de mineração de dados em *e-commerce* é a previsão de compra em uma loja on-line. Na prática, tal previsão proporciona algumas vantagens. Por exemplo, no caso de um cliente com elevada probabilidade para efetuar uma compra é possível recomendar produtos ao visitante a fim conseguir aumentar a quantidade de vendas. No caso de uma probabilidade menor, cupons de descontos podem ser oferecidos ao visitante para aumentar a motivação para a compra. Os métodos de mineração de dados podem ser aplicados para o cálculo apropriado das probabilidades de compra em um web-shop. No entanto, uma questão que surge é "como construir uma solução eficiente?". Este artigo propõe uma solução eficiente para previsão de compras em web-shop. O termo eficiente utilizado neste artigo é referente ao poder discriminatório. A solução é composta por quatro componentes: 1) aquisição de conhecimento através da construção de novas variáveis, 2) utilização da técnica *Random Forest* para estimativa de valores ausentes, 3) utilização da técnica *Random Forest* para estimativa das probabilidades de compras e 4) utilização do método *Kolmogorov Smirnov* para escolha do ponto de corte para tomada de decisão. Um estudo experimental foi realizado para validar a eficiência da abordagem proposta. O estudo utilizou a base de dados da competição internacional *Data Mining Cup* 2013. A solução proposta neste artigo venceria a competição, como será discutido no fim deste artigo.

O restante do artigo está dividido como segue. A seção 2 apresenta a definição do problema. A seção 3 descreve os conceitos importantes do Processo de Descoberta do Conhecimento. A seção 4 apresenta os trabalhos relacionados. A seção 5 detalha a solução proposta para estimativa das probabilidades de compras. A seção 6 exibe a metodologia experimental adotada no estudo. A seção 7 apresenta os resultados obtidos para validação da solução proposta. Por fim, a seção 8 conclui o trabalho e propõe trabalhos futuros.

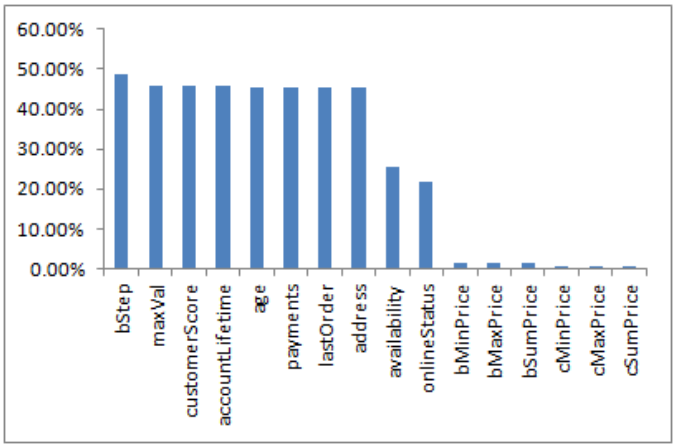

Figura 1: Percentual de valores ausentes em cada variável

2 Definição do Problema

O mecanismo de uma operação de compra na Internet funciona através das seções. Uma seção acontece quando uma visita a um site de compras é feita por um possível comprador. Durante a seção o visitante clica nos produtos a fim de ver os seus detalhes. Além disso, o comprador possivelmente irá adicionar ou remover produtos de sua cesta de compras. No final de uma sessão é possível que um ou vários produtos da cesta de compras sejam encomendados. As atividades do usuário dentro de uma seção são chamadas de transações. Uma transação é formada por uma série de variáveis. Essas transações são armazenadas e assim é possível trabalhar nesses dados para prever futuras compras.

Prever se o visitante faz uma compra ou não com base nos dados da transação coletados durante a sessão foi o desafio proposto pela *Data Mining Cup* (DMC) de 2013. A DMC é uma competição mundial que tem como foco tarefas de mineração de dados reais fornecidas por empresas (Team, 2013).

A DMC disponibilizou dois arquivos de transações: um arquivo de transações contendo as respostas, que deveria ser utilizado para construção da solução. Arquivo chamado de transact_train contendo aproximadamente 400,000 transações.

Outro arquivo de transações, que não continham as respostas, deveria ser utilizado para medir a qualidade da solução desenvolvida. Arquivo chamado de transact_class contendo aproximadamente 150,000 transações. Como cada sessão gera várias transações e o objetivo da competição era prever a probabilidade de compra de cada visitante foi necessário mudar a granularidade de cada arquivo de transação para que fosse considerada apenas a última transação. A lista com to-

[1] www.internetworldstats.com/stats.htm
[2] www.ebit.com.br

Tabela 1: Lista de variáveis

Nome da Coluna	Descrição
sessionNo	Número de execução da seção
startHour	Hora em que a sessão começou
startWeekday	Dia da semana em que a seção começou (1Mo; 2Tu;...;7Su)
duration	Tempo corrido em segundos desde o início da seção.
cCount	Número de produtos clicados
cMinPrice	O preço mais baixo de um produto clicado
cMaxPrice	O preço mais alto de um produto clicado
cSumPrice	Soma dos preços de todos os produtos clicados
bCount	Quantidade de produtos colocados no cesto de compras
bMinPrice	Preço mais baixo de todos os produtos colocados no cesto de compras.
bMaxPrice	Preço mais alto de todos os produtos colocados no cesto de compras
bSumPrice	Soma dos preços de todos os produtos colocados no cesto de compras
bStep	Etapa de processamento da compra
onlineStatus	Indicação se o cliente está on-line (y - sim; n - não)
availability	Status de entrega
customerID	Número do cliente
maxVal	Preço de compra máximo admissível para o cliente
customerScore	Avaliação do cliente do ponto de vista da loja
accountLifetime	Tempo de vida da conta do cliente em meses
Payments	Número de pagamentos efetuados pelo cliente
Age	Idade do cliente
address	Forma de endereço do cliente (1 - Sr.; 2 - Sra.; 3 - Empresa)
lastOrder	Tempo em dias transcorridos desde a última venda
order	Resultado da sessão (y - compra; n - não compra)

das as variáveis disponíveis é apresentada na Tabela 1.

Em aplicações reais de mineração de dados é comum a ocorrência de valores ausentes. Por exemplo, os visitantes da web-shop podem não ter uma conta de usuário registrada ou não ter indicado certas preferências. Como resultado, os dados de transação contêm muitos valores ausentes, o que requer uma estratégia de resolução para corrigir as imprecisões. A Fig. 1 mostra a ocorrência de valores ausentes no arquivo transact_train.

3 Processo de Descoberta do Conhecimento

O processo de descoberta do conhecimento em base de dados proposto por Fayyad and Stolorz (1997) é indicado como uma boa prática para construção de soluções de Mineração de Dados. O processo é composto por cinco etapas: 1) seleção, 2) pré-processamento, 3) transformação, 4) mineração de dados e 5) avaliação de desempenho. As etapas de pré-processamento e mineração de dados, em geral, consomem entre 50% e 80% do tempo de um projeto (Neto et al., 2017). A seguir serão descritas as principais tarefas realizadas nestas etapas:

3.1 Tratamento de valores ausentes

Dentre as atividades da etapa de pré-processamento está o tratamento de valores ausentes. A ocorrência de valores ausentes é comum em problemas reais de mineração de dados. A explicação para a ocorrência de valores ausentes pode ser falha na captura dos dados, informações que o usuário não se propôs a informar, variáveis que foram criadas após um determinado período, entre outras. Diversas estratégias para lidar com valores ausentes são encontradas na literatura (Gelman and Hill, 2006), dentre elas podemos destacar:

- Eliminação de variáveis: esta é a abordagem mais simples, uma vez que não obriga realizar nenhum tratamento, pois as soluções são construídas apenas com variáveis que possuem informação e as variáveis que possuem valores ausentes são excluídas;
- Substituição por um único valor: em vez de eliminar informação, esta abordagem completa os valores ausentes utilizando um valor único. Isso garante que não serão eliminadas informações para construção da solução. O valor ausente pode ser substituído pelo valor mínimo, valor máximo ou valor médio. Esta abordagem só pode ser aplicada para variáveis numéricas.
- Substituição por um valor estimado: esta abordagem substitui o valor ausente por um valor estimado a partir de uma técnica de mineração de dados como, por exemplo, Redes Neurais Artificiais, Regressão logística, *Random Forest*, entre outras;
- Criação de um novo valor: esta abordagem propõe a criação de um novo valor "categoria"para os valores ausentes. Esta abordagem só pode ser aplicada para variáveis nominais.

3.2 Construção de novas variáveis

De acordo com Pyle (1999), além do tratamento de valores ausentes outro objetivo da fase de pré-processamento é transformar os dados em um formato

que permita a aplicação de um algoritmo de mineração de dados. De acordo com Krogel (2005), em geral, esta tarefa representa o processo de *Feature Constructor*. Este processo é responsável por construir variáveis a partir da base de dados original. A construção de novas variáveis de entrada é uma forma sistemática de embutir conhecimento do domínio em um projeto de descoberta do conhecimento. Essas variáveis serão utilizadas como entrada por uma técnica de mineração de dados durante a etapa de *Data Mining*. O processo de construção de novas variáveis é um dos mais antigos e ainda desafiadores problemas (Hastie et al., 2009). De acordo com Witten et al. (2011), a melhor maneira de construir variáveis é manualmente, baseado no entendimento do problema de aprendizagem e no significado de cada atributo.

3.3 Seleção da técnica de mineração de dados

Dentre as técnicas de mineração de dados existentes na literatura, a *Random Forest* vem se destacando em diversas áreas como, por exemplo, análise de risco de crédito Neto et al. (2017), Reconhecimento Facial (Ribeiro and Neto, 2017), Reconhecimento de Digitais (Mendes and Neto, 2018) entre outras áreas. Esta técnica é baseada em árvore de decisão, que é um dos modelos de classificação mais utilizados na literatura devido à facilidade de compreensão de sua resposta, que é organizada na forma de uma árvore e a partir desta é possível extrair facilmente regras do tipo "Se-Então"(Polat and Gunes, 2007). Uma árvore de decisão utiliza a estratégia de dividir para conquistar. Um problema é decomposto em subproblemas e recursivamente a mesma estratégia é aplicada a cada subproblema. A simplicidade da árvore de decisão também traz desvantagens. A principal delas é a instabilidade provocada por ruídos nos dados (Hastie et al., 2009). A técnica de *Random Forest* melhora a estabilidade e precisão da árvore de decisão por incorporar um grande número de árvores em um único classificador (Diniz et al., 2013). *Random Forest* é um *ensemble* de árvore de decisões, no qual as variáveis que serão utilizadas em cada árvore são selecionadas de forma aleatória. A estratégia de *ensemble* utilizada pelo *Random Forest* é *Bagging*. *Bagging* é um acrônimo para *bootstrap aggregating*. A sua ideia central é construir vários classificadores individuais a partir de uma amostra *bootstrap* (amostra com reposição do mesmo tamanho do conjunto de treinamento, em que cada exemplo tem a mesma chance de ser escolhido). O objetivo é reduzir a variância do erro do classificador final utilizando o sistema de voto, onde cada classificador terá o mesmo peso "importância"no sistema de voto.

4 Trabalhos Relacionados

Saleem et al. (2019) apresentam estratégias para melhorar a taxa de conversão em sites de comércio eletrônico. Os autores definem taxa de conversão como uma visita ao site que resulta em uma compra. O estudo destaca quatro fatores essenciais para aumentar a taxa de conversão: 1) melhorias na logística e distribuição dos produtos, 2) fortalecimento da segurança no sistema de pagamento, 3) estratégias de *marketing* para captura de novos clientes e 4) métodos para melhorar a conversão de clientes durante a navegação do site. Embora o estudo destaque a importância de cada um destes itens, os autores apenas ilustram as oportunidades que melhorias nestes itens poderiam proporcionar, no entanto, o estudo não propõem soluções que implementem estas estratégias. É importante destacar que diversos trabalhos são encontrados na literatura que destacam oportunidades de melhorias para o comércio eletrônico, no entanto, poucos trabalhos propõem soluções práticas para sua melhoria, uma justificativa para ausência destes trabalhos é o fato de existirem poucas bases de dados de domínio público que possam ser utilizadas com esta finalidade. Dentre os estudos que propõem soluções, foram encontrados trabalhos que utilizam bases de dados de empresas privadas de comércio eletrônico da China. Dentre estes estudos, podemos destacar (Zhao et al., 2016, Zeng et al., 2019, Zaim et al., 2019).

Zhao et al. (2016) apresentam uma solução para prever compras em uma plataforma de varejo on-line para consumidores da China. Os autores utilizaram três grupos de variáveis, o primeiro relacionado aos cliques dos usuários, o segundo é referente a características das compras e o último relacionado ao carrinho de compras. Em seguida, foram realizados sete experimentos de previsão de compras do usuário com diferentes combinações dos três grupos, e o desempenho da previsão de compra foi observado. A técnica de mineração de dados utilizada foi *Random Forest*. De acordo com os autores, os resultados mostraram que os três grupos de variáveis apresentam um bom desempenho, mas que um modelo que utilize todos os grupos de variáveis juntos alcança um poder preditivo maior.

Zeng et al. (2019) utilizaram técnicas estatísticas para modelar o comportamento de usuários que fazem compras durante o dia do festival de compras na China. A base de dados foi composta por 31 milhões de logs gerados neste dia do festival. A validação da solução foi feita através da validação cruzada. Para a modelagem os autores relacionaram a eficácia de vários indícios de possíveis ações de compra (por exemplo, o efeito do tempo total de navegação, o número de cliques, categorias de produtos e a hora do dia em compras futuras). Com o resultado, os autores conseguiram testar a eficácia e identificar os comportamentos críticos dos usuários que determinam os indícios de futuras compras. Os autores chegaram à conclusão que as vendas do comércio eletrônico são fortemente estimuladas por descontos e promoções.

Zaim et al. (2019) propuseram uma abordagem dinâmica de avaliar a satisfação de compra de um cliente em um site de comércio eletrônico. A solução considera informações da navegação do usuário e também das revisões dos usuários referentes aos produtos da loja. Os autores não detalham todas as variáveis que foram utilizadas no estudo, mas citam algumas como, por exemplo, o tipo de ação do usuário (adicionar produto aos favoritos, adicionar produto ao carrinho de compras e clicar na descrição do produto) e o tempo de navega-

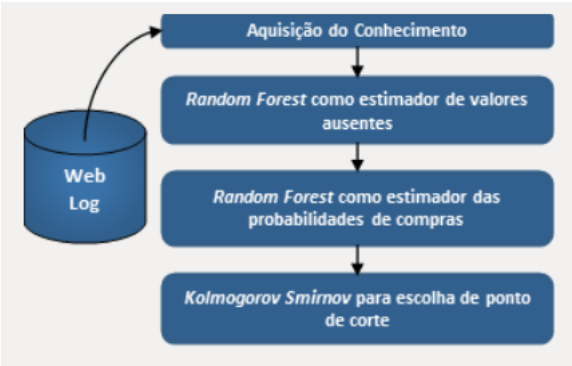

Figura 2: Arquitetura da abordagem proposta

ção do site. Os autores informam que a solução utiliza técnicas de Processamento de Linguagem Natural para criar variáveis a partir das revisões de seus produtos. A técnica de mineração de dados utilizada foi uma árvore de decisão. A base de dados utilizada no estudo foi disponibilizada pela TMALL, uma importante unidade do grupo Alibaba que é conhecida como uma das principais empresas de comércio eletrônico da China.

Os trabalhos apesentados nesta seção não abordam o tratamento de valores ausentes e não detalham as variáveis criadas, que depende da estrutura de cada base de dados disponibilizada.

5 Solução Proposta

A Fig. 2 mostra a arquitetura da abordagem proposta. Ela é composta por quatro camadas: na primeira camada, Aquisição do Conhecimento, são especificadas as variáveis que devem ser construídas para embutir conhecimento do domínio e maximizar o conteúdo estatístico da informação. A segunda camada utiliza a técnica de *Random Forest* para estimação de valores ausentes. A terceira camada propõe a utilização também do *Random Forest* como método principal de *data mining* para estimativa das probabilidades de compras. A quarta camada propõe a utilização do método *Kolmogorov Smirnov* para escolha do ponto de corte para tomada de decisão. A seguir cada camada será descrita em detalhes.

5.1 Aquisição do Conhecimento

De uma forma geral, a tarefa de construção de novas variáveis é muito mais dependente do conhecimento do domínio do que a construção de um classificador, por isso o conhecimento do domínio é um requisito (Zhao et al., 2009). Alguns estudos têm demonstrado que a utilização do conhecimento do domínio para construção de novas variáveis de entrada aumenta o poder preditivo do modelo (Neto et al., 2017). Soluções que atingiram boas colocações em importantes competições internacionais, utilizaram a estratégia de construção de novas variáveis para embutir conhecimento do domínio, como ser visto em Adeodato et al. (2008) e Adeodato et al. (2009). As variáveis propostas nesta abordagem são:

- qtdSession: quantidade de transações (número de linhas do arquivo de transação com o mesmo identificador da seção);
- Efetividade: razão entre as variáveis cCount (número de produtos clicados) e bCount (número de produtos colocados no carrinho da loja virtual);
- bStepLast: o valor da variável bStep na penúltima transação;
- max_bStep: o valor máximo da variável bStep no histórico de transação daquela sessão;
- availabilityLast: o valor da variável availability na penúltima transação.

5.2 *Random Forest* como estimador de valores ausentes

A abordagem de mineração de dados proposta neste artigo defende a utilização da técnica *Random Forest* como estimador de valores ausentes para variáveis numéricas, tendo em vista os bons resultados obtidos em outras áreas como apontados no trabalho de Shah et al. (2014). E também pelos seguintes pontos negativos encontrados nas demais estratégias: 1) a eliminação de variáveis reduz a quantidade de exemplos para construção da solução, 2) a substituição pelo valor mínimo ou máximo sofre influência de valores aberrantes e 3) a média reduz a variabilidade dos dados uma vez que as estimativas de variâncias tendem a ser subestimadas, pois a estimativa pela média não leva em consideração o relacionamento entre as variáveis (Eekhout et al., 2014).

Para as variáveis categóricas a abordagem proposta utiliza a criação de um novo valor como recomendado por Rud (2000).

5.3 *Random Forest* como estimador das probabilidades de compras

Existe uma grande variedade de técnicas de mineração de dados. Ngai et al. (2009) indicaram na literatura mais de 30 métodos utilizados em artigos científicos em sistemas de relacionamento com cliente. Por isso, a identificação de uma técnica adequada para o problema de previsão de compras em web-shop é uma contribuição relevante.

A solução proposta neste artigo sugere a utilização da técnica *Random Forest* como estimador para as probabilidades de compras. A escolha do *Random Forest* como técnica de mineração de dados para este problema é justificada por sua elevada capacidade de generalização para base de dados que possuem variáveis numéricas e categóricas que é o cenário deste estudo. Além de sua capacidade de evitar *overfitting* quando escolhido um correto número de árvores para composição da floresta.

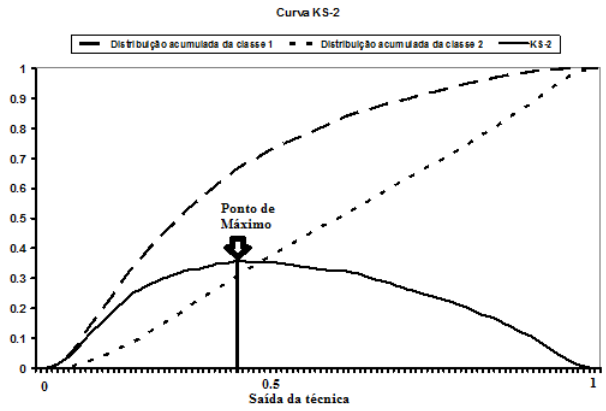

Figura 3: Ilustração do método KS para escolha do ponto de corte

5.4 *Kolmogorov Smirnov* para escolha do ponto de corte

A técnica adotada na solução proposta produz uma saída contínua entre zero e um. No entanto, para tomada de decisão de conceder ou não um cupom de desconto é necessário a escolha de um ponto de corte (limiar), abaixo do qual a decisão é feita para uma classe ou outra. A solução proposta sugere a utilização do método estatístico *Kolmogorov Smirnov* (KS) (Rezac and Rezac, 2011) para escolha do ponto de corte. Essa estratégia tem sido adotada com sucesso na indústria financeira para concessão de crédito.

O KS é um método estatístico não paramétrico utilizado para medir a aderência entre funções de distribuições acumuladas (Adeodato, 2015). Em problemas de classificação binária, a curva do KS é a diferença entre duas funções de distribuição acumuladas tendo a saída da técnica como variável independente. Uma distribuição contém a pontuação da classe 1 e, a outra, a da classe 2. O ponto de corte escolhido é o ponto de máximo da curva do KS. Ele maximiza a separação entre as classes. A Fig. 3 ilustra a escolha de um ponto de corte a partir do método do KS.

6 Metodologia Experimental

A metodologia experimental utilizada neste estudo está ilustrada na Fig. 4. Foram utilizadas as duas bases de dados descritas na seção de Definição do Problema. A primeira base de dados foi utilizada para construção das soluções e a segunda para mensurar o poder preditivo. A metodologia foi projetada para mostrar a eficiência da abordagem proposta, por isso os experimentos foram conduzidos em pares: utilizando *Random Forest* e utilizando árvore de decisão. Além disso, a metodologia objetiva mensurar importância do pré-processamento, por isso os experimentos foram conduzidos utilizando diferentes estratégias de tratamento de valores ausentes e também com e sem aquisição de conhecimento (construção de novas variáveis).

A métrica de avaliação de desempenho utilizada foi a

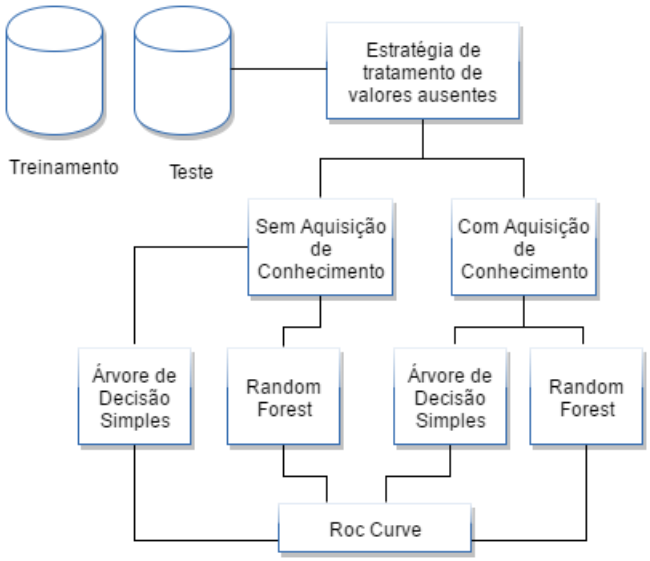

Figura 4: Metodologia Experimental

área sob a Curva ROC (*Receiver Operating Characteristic*). A Curva ROC representa o compromisso entre a taxa de verdadeiros positivos e os falsos positivos com base em uma saída contínua ao longo de todos os seus valores de ponto de corte. A Curva ROC é construída a partir da Sensibilidade contra 1-Especificidade para os diferentes valores de pontos de corte do modelo (Provost and Fawcett, 2001).

Por fim, a metodologia avalia a importância da escolha correta do ponto de corte utilizando o método do KS. Para isso, é realizada uma análise de qual seria o desempenho da solução proposta na competição do DMC 2013.

7 Resultados e Discussões

As simulações foram realizadas conforme especificado na seção Metodologia Experimental. A quantidade de árvores utilizadas nos experimentos com *Random Forest* foram 600. Este número foi escolhido após uma investigação preliminar para verificar a quantidade mais adequada ao problema investigado. As variáveis horário e dia da semana foram removidas da construção da solução porque os valores contidos no arquivo de treinamento são diferentes dos valores contidos no arquivo de teste. As estratégias investigadas de aquisição de conhecimento e tratamento de valores ausentes foram utilizadas conforme especificado na seção Abordagem Proposta. Além do *Random Forest*, os experimentos foram executados com Árvore de Decisão por ser uma das técnicas mais populares na literatura e por estar presente nos estudos encontrados como trabalhos relacionados. Nas Figs. 5 a 9, as siglas DT_AUC e RF_AUC significam área sob a curva ROC na Árvore de Decisão e *Random Forest* respectivamente. O símbolo * identifica o experimento que utiliza Aquisição de Conhecimento. A análise dos resultados está dividida em três subseções para uma melhor interpretação dos resultados.

7.1 Influência da Estratégia de Aquisição do Conhecimento

Os resultados mostram que a estratégia de aquisição de conhecimento proposta neste estudo aumenta o poder preditivo da solução uma vez que os resultados obtidos com essa abordagem superam os resultados obtidos sem a estratégia de aquisição de conhecimento independente da técnica de mineração de dados utilizada e da abordagem de tratamento de valores ausentes. Na Fig. 5, que utiliza apenas variáveis sem valores ausentes, a solução que utiliza árvore de decisão sem aquisição de conhecimento obteve uma área sob a curva ROC de 0,752 contra uma área sob a curva ROC de 0,91 para a árvore de decisão utilizando aquisição de conhecimento. Este aumento no poder preditivo proporcionado pela construção de novas variáveis é mantido quando substituímos a técnica de mineração de dados de árvore de decisão pelo *Random Forest*, como pode ser observado ainda na Fig. 5, o Random Forest com aquisição de conhecimento obtém uma área sob a curva ROC de 0,956 contra 0,79 sem a estratégia de aquisição de conhecimento. É importante destacar que esse comportamento se mantém para as demais estratégias de tratamento de valores ausentes como pode ser observado na Fig. 6, que mostra os resultados do valor mínimo para valores ausentes. Na Fig. 7, que mostra os resultados do valor máximo para valores ausentes. Na Fig. 8, que mostra os resultados do valor médio para valores ausentes. Na Fig. 9, que mostra os resultados do valor estimado do Random Forest para valores ausentes.

A justificativa para um maior poder preditivo quando uma solução utiliza a estratégia de aquisição de conhecimento é que a construção das novas variáveis maximiza o conteúdo estatístico da informação. Para ilustrar este procedimento, a Tabela 2 descreve o ganho de informação associado a cada nova variável proposta neste trabalho. O Ganho de Informação é a redução esperada da entropia devido a "classificação"de acordo com um determinado atributo de entrada (Witten et al., 2011). A Eq. (1) exibe a fórmula da entropia, onde P(Y) representa a probabilidade a priori da variável alvo. A Eq. (2) exibe o cálculo do Ganho de Informação para uma variável, onde H(X|Y) representa o valor da entropia da variável alvo Y apenas considerando a variável X.

$$Entropia\ H(Y) = -\sum p(y)\log p(y) \quad (1)$$

$$Ganho\ de\ informao\ GI(X,Y) = H(Y) - H(X|Y) \quad (2)$$

7.2 Influência da Estratégia de Tratamento de Valores Ausentes

Outro resultado relevante deste estudo é que ele mostra que a estratégia de tratamento de valores ausentes possui influência no poder discriminatório da solução. A estratégia de eliminação de variáveis é a que proporciona o menor poder preditivo, como pode ser observado na Fig. 5, seguido pela substituição do valor mínimo, conforme pode ser observado na Fig. 6. A explicação para isso é que a estratégia de eliminar variáveis reduz a quantidade de informação disponível, o que prejudica a inferência do modelo preditivo, principalmente quando existe um grande número de exemplos e variáveis com ocorrência de valores ausentes, que é o caso para o problema estudado. A substituição dos valores ausentes para o mínimo prejudica a solução porque tende a inferir que esses clientes não realizariam a compra. Por exemplo, a variável número de produtos no carrinho informa a quantidade de produtos que o cliente pretende comprar, utilizando esta estratégia o valor mínimo seria zero, o que só deve ocorrer em situações normais quando o cliente não realiza a compra. As demais estratégias de tratamento de valores ausentes produzem resultados semelhantes (Fig. 7, Fig. 8 e Fig. 9). No entanto, a estimativa utilizando o *Random Forest* tende a ser mais robusta conforme descrito na seção III.

Tabela 2: Ganho de Informação das variáveis construídas

Variável	Ganho de Informação
qtdSession	0.348
Efetividade	0.099
bStepLast	0.172
max_bStep	0.3966
availabilityLast	0.0172

Por fim, os resultados mostram a adequação do método *Random Forest* ao problema, uma vez que os resultados obtidos com esta técnica possuem maior área sob a curva ROC para todos os experimentos realizados no estudo como pode ser observado nas Figs. 5 a 9 que exibem os resultados obtidos.

7.3 Desempenho na Competição

Em 2013, a DMC teve 99 equipes inscritas na competição de 77 universidades e 24 países. Dessas, 60 conseguiram desenvolver soluções. De acordo com Hop (2013), a solução vencedora obteve 144 erros no conjunto de teste. Se fosse utilizada a solução proposta neste artigo: a estratégia de aquisição do conhecimento, o *Random Forest* como estimador de valores ausentes e como estimador das probabilidades de compras e o método do KS para escolha do ponto de corte, seriam encontrados 142 erros com um ponto de corte de 0,32. Esta solução venceria a competição. A solução proposta neste artigo reduz em 63% o erro obtido pela melhor equipe da América latina participante da competição, que obteve 224.12 erros.

8 Conclusões

Este artigo apresentou uma abordagem eficiente para previsão de compras em web-shop. A abordagem é composta por quatro componentes: 1) aquisição de conhecimento através da construção de novas variáveis,

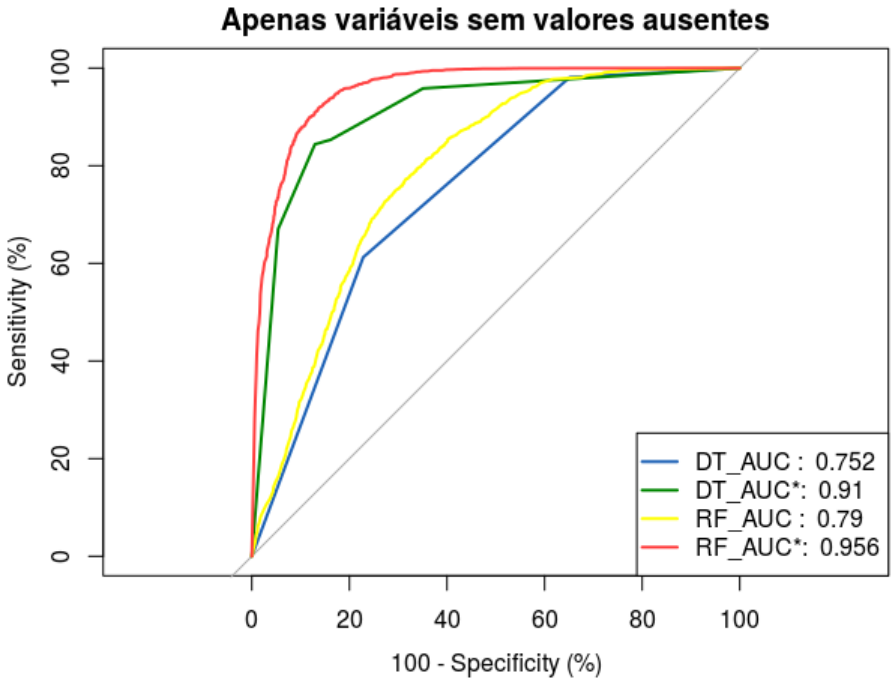

Figura 5: Curvas ROC dos modelos utilizando apenas as variáveis sem valores ausentes.

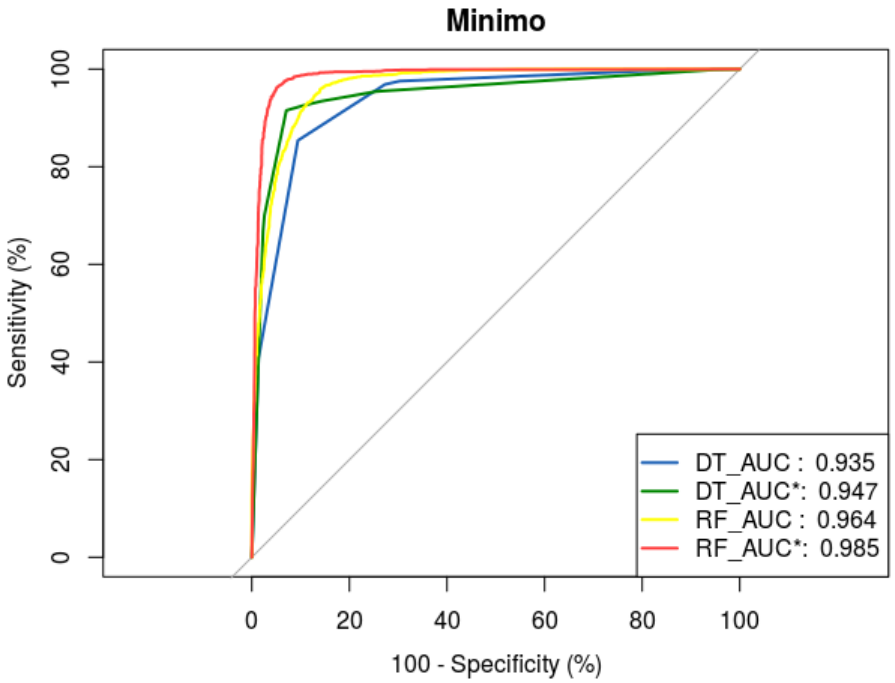

Figura 6: Curvas ROC dos modelos utilizando o valor mínimo como estratégia para os valores ausentes

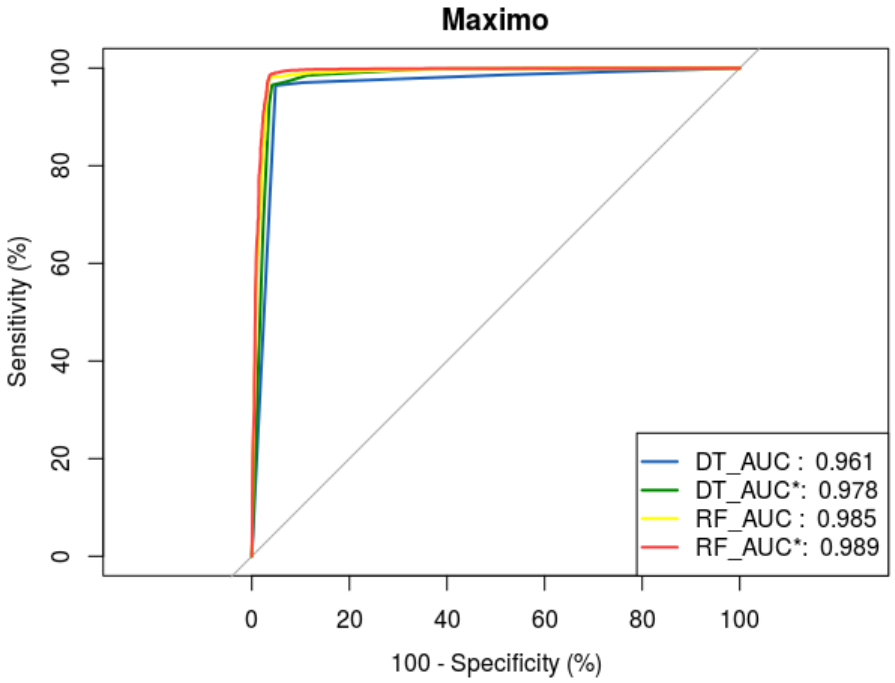

Figura 7: Curvas ROC dos modelos utilizando o valor máximo como estratégia para os valores ausentes

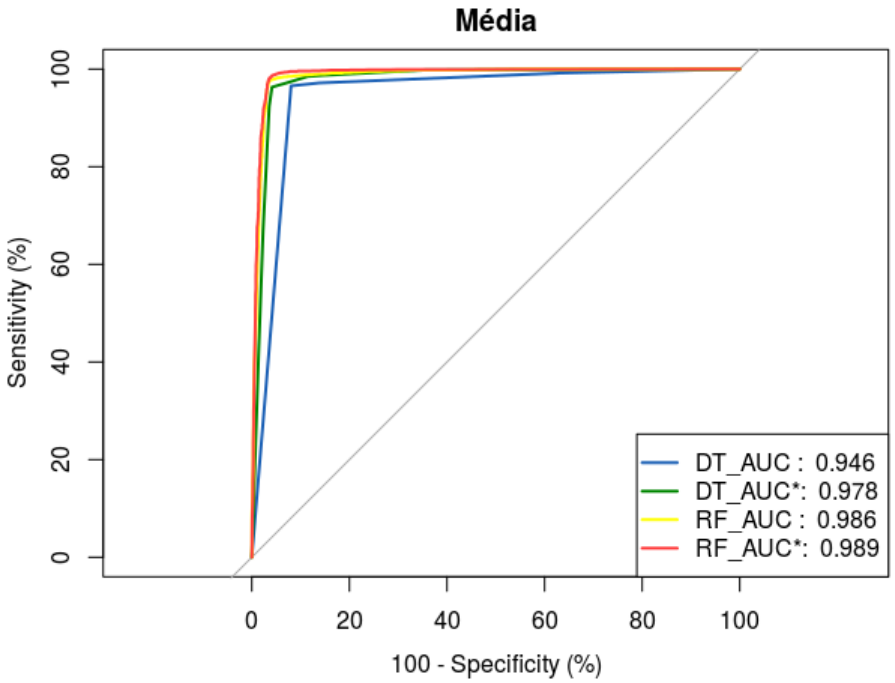

Figura 8: Curvas ROC dos modelos utilizando a média como estratégia para os valores ausentes

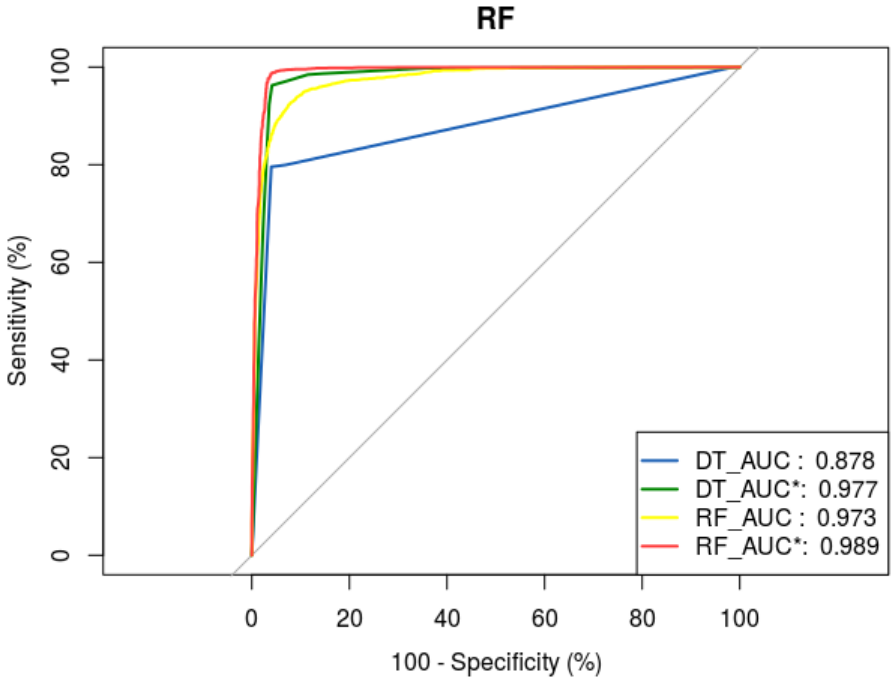

Figura 9: Curvas ROC dos modelos utilizando o *Random Forest* como estratégia para os valores ausentes

2) utilização da técnica *Random Forest* para estimativa de valores ausentes, 3) utilização da técnica *Random Forest* para estimativa das probabilidades de compras e 4) utilização do método *Kolmogorov Smirnov* para escolha do ponto de corte para tomada de decisão. O estudo experimental mostrou que a abordagem proposta produz um desempenho eficiente capaz de vencer a competição da DMC 2013. A principal contribuição deste estudo é um direcionamento para construção de soluções de mineração de dados em web-shop, pois no desenvolvimento de tais soluções diversos questionamentos surgem, tais como: qual método de mineração de dados utilizar? Qual a melhor estratégia para tratamento de valores ausentes? Como melhorar o poder preditivo através da aquisição de conhecimento? Como escolher o melhor ponto de corte? Este estudo fornece um direcionamento inicial que pode ser levado em consideração por empresas da área de comércio eletrônico, tendo em vista que os resultados obtidos neste estudo apontam um resultado competitivo. No entanto, cada empresa possui suas peculiaridades e devem adaptar a abordagem proposta para sua realidade. Como trabalho futuro, pretendemos aplicar a abordagem proposta a uma empresa brasileira do setor de comércio eletrônico através de um projeto de pesquisa.

Referências

Adeodato, P. (2015). Variable transformation for granularity change in hierarchical databases in actual data mining solutions, *The 16th International Conference Intelligent Data Engineering and Automated Learning*, pp. 146–155. https://doi.org/10.1007/978-3-319-24834-9_18.

Adeodato, P., Arnaud, A., Vasconcelos, G., Cunha, R., Gurgel, T. and Monteiro, D. (2009). The role of temporal feature extraction and bagging of mlp neural networks for solving the wcci 2008 ford classification challenge, *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pp. 57–62. https://doi.org/10.1109/IJCNN.2009.5178965.

Adeodato, P. J. L., Vasconcelos, G. C., Arnaud, A. L., Cunha, R. C. L. V., Monteiro, D. S. M. P. and Neto, R. F. O. (2008). The power of sampling and stacking for the pakdd2007 cross-selling problem, *International Journal of Data Warehousing and Mining* **4**: 22–31. https://doi.org/10.4018/jdwm.2008040104.

Diniz, F., Neto, F. M., Júnior, F. d. C. and Fontes, L. M. (2013). Redface: um sistema de reconhecimento facial baseado em técnicas de análise de componentes principais e autofaces, *Revista Brasileira de Computação Aplicada* **5**(1): 42–54. http://doi.org/10.5335/rbca.2013.2627.

Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R. and Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level, *Journal of Clinical Epidemiology* **67**(3): 335–342. https://doi.org/10.1016/j.jclinepi.2013.09.009.

Fayyad, U. and Stolorz, P. (1997). Data mining and kdd: Promise and challenges, *Future generation computer systems* **13**(2): 99 – 115. https://doi.org/10.1016/S0167-739X(97)00015-0.

Gelman, A. and Hill, J. (2006). *Missing-data imputation*, Analytical Methods for Social Research, Cambridge University Press, p. 529–544. https://doi.org/10.1017/CBO9780511790942.031.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, Springer.

Hop, W. (2013). *Web-shop order prediction using machine learning*, Master's thesis computational economics, Erasmus University Rotterdam, Rotterdam, Holanda.

Krogel, M.-A. (2005). *On propositionalization for knowledge discovery in relational databases.*, PhD thesis, Otto von Guericke University Magdeburg.

Mendes, R. and Neto, R. O. (2018). The power of ensemble models in fingerprint classification: A case study, *INFOCOMP* **17**(1): 1–10. http://www.dcc.ufla.br/infocomp/index.php/INFOCOMP/article/view/557.

Midha, N. and Singh, V. (2018). Classification of e-commerce products using reptree and k-means hybrid approach, *in* V. Aggarwal, V. B.and Bhatnagar and D. K. Mishra (eds), *Big Data Analytics*, Springer Singapore, Singapore, pp. 265–273. http://doi.org/10.1007/978-981-10-6620-7_26.

Neto, R., Adeodato, P. J. and Salgado, A. C. (2017). A framework for data transformation in credit behavioral scoring applications based on model driven development, *Expert Systems with Applications* **72**: 293 – 305. https://doi.org/10.1016/j.eswa.2016.10.059.

Ngai, E. W. T., Xiu, L. and Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Syst. Appl.* **36**(2): 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021.

Polat, K. and Gunes, S. (2007). Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast fourier transform, *Applied Mathematics and Computation* **187**(2): 1017–1026. http://dx.doi.org/10.1016/j.amc.2006.09.022.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments, *Machine Learning* **42**(3): 203–231. https://doi.org/10.1023/A:1007601015854.

Pyle, D. (1999). *Data preparation for data mining*, morgan kaufmann.

Rezac, M. and Rezac, F. (2011). How to Measure the Quality of Credit Scoring Models, *Czech Journal of Economics and Finance* **61**(5): 486–507. Disponível em https://ideas.repec.org/a/fau/fauart/v61y2011i5p486-507.html.

Ribeiro, J. and Neto, R. (2017). Eigenface vs random forest: A comparative study on face recognition, *XVII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg), Brasília, Distrito Federal, Brazil*.

Rud, O. (2000). *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*, Data Warehousing Data Mining, John Wiley & Sons.

Saleem, H., Uddin, M. K. S., ur Rehman, S. H., Saleem, S. and Aslam, A. M. (2019). Strategic data driven approach to improve conversion rates and sales performance of e-commerce websites, *International Journal of Scientific Engineering Research* **10**(4): 1 – 6.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study, *American Journal of Epidemiology* **179**(6): 764–774. http://doi.org/10.1093/aje/kwt312.

Team, D. (2013). *Data Mining Cup 2013*. Disponível em http://www.data-mining-cup.de/en/review/goto/article/dmc-2013.html.

Witten, I. H., Frank, E. and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zaim, H., Haddi, A. and Ramdani, M. (2019). A novel approach to dynamic profiling of e-customers considering click stream data and online reviews, *International Journal of Electrical and Computer Engineering* **9**(1): 602 – 612. http://ijece.iaescore.com/index.php/IJECE/article/download/12080/11091.

Zeng, M., Cao, H., Chen, M. and Li, Y. (2019). User behaviour modeling, recommendations, and purchase prediction during shopping festivals, *Electronic Markets* **29**(2): 263 – 274. https://doi.org/10.1007/s12525-018-0311-8.

Zhao, H., Sinha, A. P. and Ge, W. (2009). Effects of feature construction on classification performance: An empirical study in bank failure prediction, *Expert Systems with Applications* **36**(2, Part 2): 2633 – 2644. https://doi.org/10.1016/j.eswa.2008.01.053.

Zhao, Y., Yao, L. and Zhang, Y. (2016). Purchase prediction using tmall-specific features, *Concurrency and Computation: Practice and Experience* **28**(14): 3879 – 3894. https://doi.org/10.1002/cpe.3720.