Aprendizagem de Máquina para Apoio à Predição de Vendas para Empresas do Ramo de Locação de Equipamentos para Eventos

Ítalo B. L. Carvalho[1], **Leonardo O. Moreira**[1]

[1]Instituto Universidade Virtual (UFC Virtual)
Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil

italoboss@gmail.com, leoomoreira@virtual.ufc.br

Abstract. *The growth in the number of events have favored the service sector and strategies that can increase the efficiency of companies in the industry can guarantee their balance. Thus, the use of Machine Learning techniques can help companies in decision making. For this, the present work, using the methodology of Data Mining, proposes to develop a tool, implemented in the language Java and using the tool WEKA, that helps in the prediction of sales to companies of equipment rental for events.*

Resumo. *O crescimento do número de eventos têm favorecido o setor de prestação de serviços e estratégias que possam aumentar a eficiência de empresas do ramo podem garantir seu equilíbrio. Sendo assim, o uso de técnicas de Aprendizagem de Máquina pode auxiliar as empresas em tomadas de decisão. Para isso, o presente trabalho, utilizando a metodologia de Mineração de Dados, se propõe a desenvolver uma ferramenta, implementada na linguagem Java e utilizando a ferramenta WEKA, que auxilie na predição de vendas para empresas de locação de equipamentos para eventos.*

1. Introdução

A todo instante, dados vêm sendo armazenados e agrupados em grandes conjuntos contendo informações ocultas de grande importância para o negócio. Em virtude dessa massividade de dados, extrair essas informações se torna uma tarefa incomum [Damasceno 2010]. Dentro deste contexto, métodos e técnicas para extração e análise de informações úteis são importantes na tomada de decisão [Bose and Mahapatra 2001]. É por se tratar de um processo de descoberta de padrões relevantes em banco de dados que a Mineração de Dados é uma técnica muito utilizada. [Bose and Mahapatra 2001]

Como dito por Bose e Mahapatra (2001), o avanço da tecnologia de armazenagem de dados permitiu que as empresas organizassem e armazenassem grandes volumes de dados empresariais de forma que pudessem ser analisados. Com o amadurecimento da área de Inteligência Artificial, criou-se um conjunto de técnicas de Aprendizagem de Máquina que se tornaram úteis na automação de atividades cruciais de descoberta de padrões em bancos de dados. [Bose and Mahapatra 2001] São esses fatores que têm mudado a forma de analisar dados, onde a Mineração de Dados integra técnicas de Aprendizagem de Máquina e Análise Estatística para que o analista de negócios possa descobrir padrões significativos nos dados empresariais.

De uma forma geral, a Aprendizagem de Máquina procura solucionar um problema real de relevância, mas para isso deve possuir bases de dados com informações que possibilitem atingir esse objetivo [Roza et al. 2016]. Para tal, como explicado por Roza et al. (2016), é importante conhecer os dados disponíveis para que seja possível determinar os problemas que poderão ser solucionados, além de abstrair os processos por meio da construção de modelos que representem as relações entre os dados e que sejam válidos para o contexto com que se deseja trabalhar.

1.1. Motivação e Questão de Pesquisa

O presente trabalho se insere no ramo de empresas de locação de equipamentos para eventos, tratando-se de um modelo de negócio onde a empresa contratada fornece equipamentos para a realização de eventos de variadas naturezas, como feiras, congressos, *shows*, eventos sociais, sendo o contrato fechado com empresas organizadoras ou clientes finais, esses últimos podendo ser pessoa física ou jurídica. De acordo com o SEBRAE (2017), o Brasil possui empresas de pequeno porte representando noventa porcento do setor de eventos, ofertando diversos tipos de produtos e serviços. Juntamente à ascensão do turismo de negócio, o setor de *shows* e espetáculos tem crescido e favorecido o mercado de locação de equipamentos para eventos.

Sendo assim, empregar técnicas que aumentem a eficiência de empresas desse setor é de importância econômica para o Brasil, oferecendo possibilidade de crescimento para empresas e garantindo empregos. Para isso, o presente trabalho pretende responder, em parceria com uma empresa do ramo baseando em seus dados anteriores, a seguinte questão: como Aprendizagem de Máquina pode auxiliar na predição de conversão de venda para empresas de locação de equipamentos para eventos?

1.2. Objetivos

O presente trabalho tem como objetivo principal desenvolver uma ferramenta que auxilie na predição da conversão de vendas para empresas de locação de equipamento para eventos, o que consiste em informar se o orçamento realizado será convertido em venda ou se será cancelado.

Para isso, será necessário selecionar dados, pré-processar e criar modelos de classificação, através de técnicas de Aprendizagem de Máquina, utilizando o cadastro de clientes e o histórico de vendas da empresa de locação de equipamentos para eventos. Com isso, por fim, avaliar a aplicação utilizando dados de testes.

2. Referencial Teórico

Para se chegar aos padrões úteis, é necessário passar por um processo de treinamento e isso se faz válido para toda técnica de Mineração de Dados. Essa fase de treinamento ocorre após o pré-processamento dos dados, onde estes são apresentados para o algoritmo de Mineração de Dados, que se torna o responsável por identificar os padrões [Damasceno 2010]. A Figura 1 representa a visão geral do processo de Mineração de Dados. [Bose and Mahapatra 2001].

Como citado por Roza et al. (2016), os algoritmos de aprendizado de máquina podem ser dividido em dois importantes grupos: os de aprendizagem não-supervisionada e a aprendizagem supervisionada. Já em seu trabalho, Schneider (2016) coloca que do ponto

Figura 1. Visão geral do processo de mineração de dados (Adaptada de Fig. 1 de [Bose and Mahapatra 2001])

de vista das entradas e natureza de aprendizado, podemos dividí-los em aprendizagem supervisionada, aprendizagem não supervisionada e aprendizado por reforço. Schneider (2016) também faz uma revisão e classificação de literatura sobre aplicações de técnicas de Apredizagem de Máquina na construção de modelos preditivos de perda ou retenção de clientes. Também cita que, em relação ao problema de análise preditiva de perda ou retenção de clientes, comumente são aplicados os modelos de aprendizagem supervisionada.

Do ponto de vista da saída desejada, as técnicas de Aprendizagem de Máquina podem ser categorizadas por outros grupos, como, associação, classificação, clusterização, entre outros. Para esse mesmo problema, os modelos de classificação são adotados [Schneider 2016]. Para a aprendizagem supervisionada, a instância analisada deve conter um atributo classe que irá classificá-la. São técnicas geralmente utilizadas em predição, pois tentam predizer para uma nova instância qual classe esta pertence, baseando-se no treinamento [Damasceno 2010]. Como dito por Roza et al. (2016), após se obter um modelo de classificação que descreva o conjunto de dados, espera-se que ele possa prever a classificação para novas entradas.

Para avaliar os resultados dos modelos obtidos, a prática mais utilizada consiste em separar os dados pré-processados em dados para treinamento e dados de testes, e comparar os valores de saída obtidos após o treinamento com os exemplos de teste sendo os valores de entrada [Roza et al. 2016].

2.1. WEKA

A WEKA (*Waikato Enviroment for Knowledge Analysis*) é uma ferramenta formada por um conjunto de implementações de algoritmos de várias técnicas de Mineração de Dados, escrita na linguagem de programação Java [Damasceno 2010]. O WEKA[1] é um software livre, sob domínio da licença GPL.

Como dito por Damasceno (2010), para a aplicação de técnicas de mineração de dados é preciso que os dados estejam estruturados, seja em uma planilha, banco de

[1]WEKA. Disponível em: `http://www.cs.waikato.ac.nz/ml/weka`. Acessado em: 20 de novembro de 2018.

dados ou alguma outra estrutura. No caso da ferramenta WEKA, devemos usar arquivo seguindo um formato chamado de ARFF. Podemos dividir esse arquivo em duas partes, uma contendo a listagem dos atributos, definindo o tipo e valores que podem representar, e outra composta pelas as instâncias, com seus valores definidos para cada atributo. A seguir, serão descritas as técnicas utilizadas no presente trabalho.

2.2. Árvores de Decisão

O objetivo dessa técnica é de criar um modelo que prevê, com base nas instâncias de entrada, o valor de uma variável de uma nova instância, sendo uma representação simples de método de classificação [Schneider 2016].

Classificadores de árvore de decisão são populares por, em geral, não requerer conhecimento avançado da técnica e são apropriados para descoberta de conhecimento exploratória [Schneider 2016]. Para Roza et al. (2016), a simplicidade em se montar as árvores, com algoritmos eficientes que escalam bem com o aumento do tamanho da base de dados utilizada, se torna outra razão para a popularidade dessa técnica. No WEKA, a classe que representa técnica de árvore de decisão é a weka.classifiers.trees.J48.

2.3. Regressão Logística

Essa é uma técnica de regressão onde a variável dependente é categórica, normalmente assumindo valores binários (como 0 ou 1) [Schneider 2016]. Essa técnica determina a probabilidade de ocorrência de uma classe específica com base nos valores das variáveis de entrada, utilizando uma função logística (Distribuição Acumulada Logística) [Schneider 2016]. No WEKA, a classe que representa a técnica de regressão logística é a weka.classifiers.functions.SimpleLogistic.

2.4. Naive Bayes

Essa técnica de classificação faz parte de um conjunto de classificadores probabilísticos simples baseados na aplicação do Teorema de Bayes, propondo uma forte independência entre as características [Schneider 2016]. Isso se deve ao fato de que o Teorema de Bayes fornecer um mecanismo formal para atualizar probabilidades [Morettin and Bussab 2017]. Para Schneider (2016) , classificadores Naive Bayes, em alguns tipos de modelos de probabilidade, podem ser utilizados com muita eficiência em um ambiente supervisionado.

Para essa técnica, porém, um número grande de características ou uma possuir uma característica com muitas possibilidades de valores faz com que se torne inviável de utilizá-lo. Contudo, alguns estudos apontam que uma vantagem desse método é que requer uma quantidade pequena de dados de treino para construção do modelo [Schneider 2016]. No WEKA, a classe que representa a técnica de Naive Bayes é a weka.classifiers.bayes.NaiveBayes.

3. Abordagem Proposta

Para este trabalho, usaremos como base a metodologia de Mineração de Dados descrita por Borges (2015) citando Chapman et al. (2000) conhecido por CRISP-DM (*Cross-Industry Standard Process for Data Mining*), metodologia a qual a comunidade empresarial frequentemente utiliza. Em meados de 1999 surgiu a versão 1.0 de CRISP-DM, com o intuito de descrever as principais etapas do processo de Mineração de Dados, desde a

análise do negócio até à entrega do produto final [Borges 2015]. Segundo Chapman et al. (2000), essas etapas constituem-se de:

1. Aprendizagem do negócio: identificar objetivos de negócio, verificar capacidade de acesso a recursos, determinar objetivos da Mineração de Dados e produzir planificação do projeto.
2. Análise dos dados: recolher dados, descrevê-los, explorá-los e verificar a qualidade dos dados.
3. Preparação dos dados: selecionar, limpar, construir, integrar e formatar os dados.
4. Modelagem: selecionar técnica de modelagem, gerar esquema de testes, construir e consultar o modelo.
5. Avaliação: avaliar resultados e rever processos.
6. Distribuição: planejar distribuição, manutenção e monitorização, produzir relatórios e rever o projeto.

Importante informar que a última etapa, a distribuição, não será realizada para esta pesquisa, pois trata-se de um processo no qual se faz necessário disponibilizar a ferramenta para diversos usuários de forma controlada e observável, não cabendo no escopo do presente trabalho.

3.1. Aprendizagem do Negócio

A primeira etapa do método foi realizada em parceria com o diretor executivo da empresa de locação de equipamentos, o qual já possuíra o conhecimento do negócio e saberia o que seria capaz de ser realizado. Após analise, foi decidido que seria utilizado os dados relacionados às vendas e clientes com o objetivo de prever qual a probabilidade de um novo orçamento ser convertido em venda para a empresa.

3.2. Análise e Preparação dos dados

Para estas etapas, foram realizadas reuniões com a equipe de desenvolvimento de software da empresa de locação de equipamentos para que pudesse ser realizada a coleta e análise dos dados.

Foram verificadas duas importantes características dos dados. Primeiro, apesar da empresa atuar desde 1999, devido a mudança nas tecnologias e regras de negócio utilizadas só seria possível utilizar dados gerados a partir do ano de 2013 e das vendas que foram inseridas na categoria de eventos, excluindo os contratos de duração acima de um mês. Por fim, possuímos dados não estruturados, fazendo-se necessário a realização da valorização de forma manual.

Devido ao limite de informações disponíveis e estruturadas em banco de dados e para manter privacidade de informações, foram selecionados os seguintes parâmetros para coleta:

- Forma de pagamento do serviço;
- Categoria do serviço, podendo ser este Evento ou Gerador, pois se diferenciam em itens, preparação e execução do serviço;
- Se o cliente é pessoa física ou juridica;
- Perfil do cliente, podendo ser este Organizadora de Eventos, Hotel, Pequenas Empresas, entre outros;

- Data do primeiro contato com o cliente;
- Valor total do orçamento para o serviço;
- Valor e tipo de desconto dado;
- Se o orçamento foi fechado ou negado.

Para preparação, foram selecionados os dados, filtrando as condições apresentadas acima e formatados em um arquivo de planilha para que pudesse ser realizada a valorização manual de registros que não foram possíveis de serem valorizados automaticamente. No total, foram selecionados 15225 instâncias para base de treinamento e teste.

3.3. Modelagem

Para esta etapa foram selecionadas três técnicas de modelagem, apresentadas na fundamentação teórica (árvore de decisão, Naive Bayes e regressão logística), para que possa ser realizada uma avaliação de qual técnica apresentou melhores resultados para o conjunto de dados apresentado.

Após finalizada a preparação dos dados, estes foram divididos em duas bases, uma de treinamento e outra de teste. A primeira, representando 95% (noventa e cinco porcento) dos dados, para ser utilizada em cada um dos algoritmos selecionados na ferramenta WEKA, para que sejam construídos os modelos de classificação. E a segunda, representando 5% (cinco porcento) dos dados, para que seja utilizada para testar a eficiência dos modelos de classificação, comparando os valores de resultados obtidos com o os valores de resultado já conhecidos.

3.4. Avaliação

Para a avaliação foi desenvolvido uma ferramenta utilizando a linguagem Java e a ferramenta WEKA com a responsabilidade de ler os dados de treinamento e teste, aplicar o treinamento aos algoritmos e utilizar o dados de teste para informar a porcentagem de acerto de cada modelo.

Além disso, foi desenvolvido um meio para que o usuário possa preencher, com valores possíveis, os parâmetros necessários para que se possa realizar a predição de uma venda.

4. Avaliação

A ferramenta desenvolvida[2] foi divida em 2 telas, onde para a primeira tela, o usuário é capaz de carregar um arquivo em formato CSV (*Comma Separated Value* ou Valores Separados por Vírgula) com o conjunto total de dados para a execução do treinamento dos algoritmos e teste dos mesmos, ilustrada na Figura 2.

Para a Figura 3, o usuário tem a possibilidade de preencher um formulário com os campos referentes a cada um dos parâmetros que compõe uma instância, esses definidos na etapa de análise e preparação dos dados, para que possa realizar a predição sobre uma venda.

[2]Disponível em: `https://github.com/italoboss/saleforecasting`.

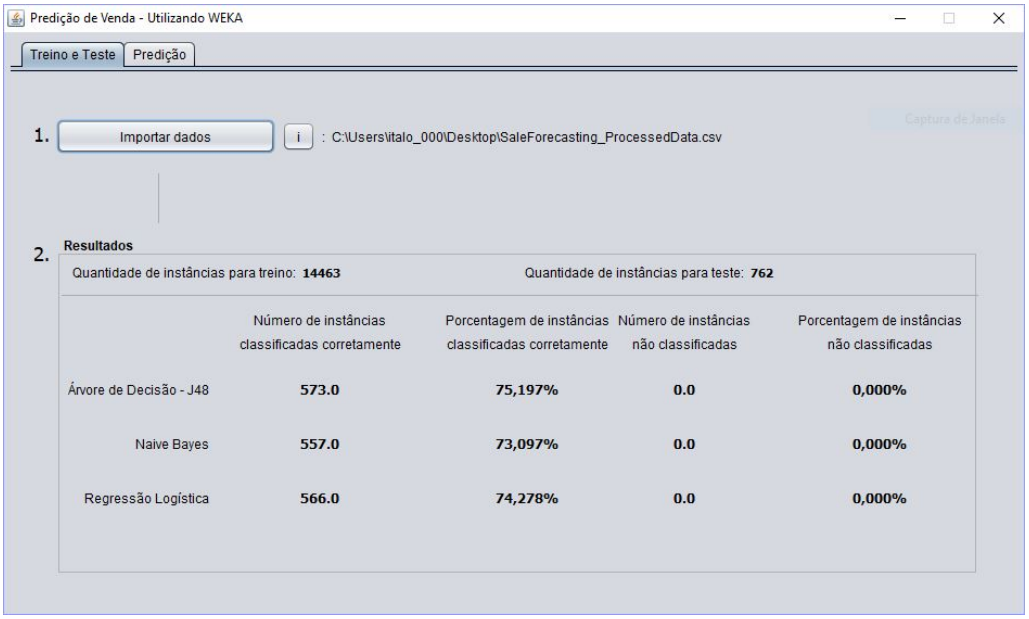

Figura 2. Captura da tela principal do sistema.

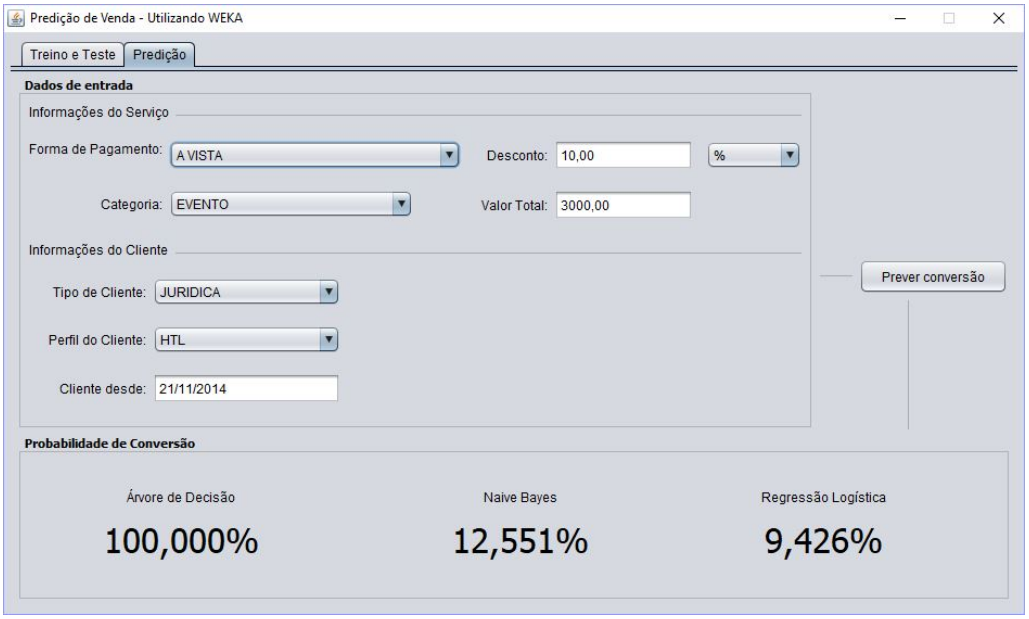

Figura 3. Captura de tela de predição do sistema.

4.1. Funcionalidades da Ferramenta

Após realizada a leitura do arquivo CSV, a ferramenta proposta faz a transformação para o formato aceito pela ferramento WEKA, o ARFF, e após isso é realizado um sorteio na ordem das instâncias e estas são separadas em dois conjuntos, um para treino, onde 95% do número total de instâncias são inseridas e as outras restantes adicionadas no outro conjunto para realização do teste. Para o presente trabalho, o número total de instâncias obtidas foi 15225, sendo aleatoriamente selecionadas 14463 instâncias para treino dos algoritmos e 762 para o teste.

Ao final da separação dos dados, a ferramenta proposta, utilizando a ferramenta WEKA, realiza o treinamento dos algoritmos e constrói os modelos referentes aos de Regressão Logística, Árvore de Decisão e Naive Bayes. Logo em seguida, utilizando do conjunto de teste realiza a avaliação dos modelos, apontando, por fim, os Número de instâncias classificadas corretamente e a Porcentagem destas referente ao total de instâncias de teste. Bem como a número de instâncias não classificadas, caso haja, e porcentagem destas em relação ao total de instâncias de teste.

A segunda função da ferramenta, em consequência do treinamento dos algoritmos, é a predição de uma nova instância. Para isso, há um formulário para ser preenchido e após isso, ao se clicar no botão "Prever conversão", os valores referente a probabilidade da venda ser fechada para cada um dos algoritmos é colocado em tela. Dessa forma, o usuário pode verificar qual a probabilidade de conversão da venda dentre os algoritmos que mais foram acertivos nos testes.

4.2. Testes dos Algoritmos

Devido a quantidade e aleatoriedade da distribuição das instâncias para treino e teste, foram realizadas 20 rodadas de treino e teste dos modelos para comparar a taxa de acerto de cada um desses. O resultado obtidos podem ser visualizados na Tabela 1.

Tabela 1. Resultado das rodadas de treinamento e teste

Média das Porcentagens de acerto		
Árvore de Decisão	Naive Bayes	Regressão Logística
75,485%	73,071%	75,039%

Pode-se observar que o modelo construído, através da ferramenta WEKA, para o algoritmo de Naive Bayes, em todas as rodadas, foi o com menor taxa de acerto. Contudo, a diferença para os demais modelos foram de, no máximo, 3 ponto percentuais.

Da mesma forma, podemos observar que as técnicas de Árvore de Decisão e Regressão Logística revezaram-se entre o com maior porcentagem de acerto, com diferença em pontos percentuais muito pequena, abaixo de 1.

Quanto a variação na taxa de acerto, a técnica de Regressão Logística teve a menor, com 2,1 pontos percentuais de diferença entre o menor e maior valor. Seguido da técnica de Árvore de Decisão, com 2,49 e de Naive Bayes com 3,15 pontos percentuais de variação.

Acredita-se que o porquê da técnica de Árvore de Decisão ter sido mais eficiente, dentre as abordadas, deve-se a simplicidade em se montar árvores e pela eficiência do algoritmo ao escalar com o aumento do tamanho da base de dados.

5. Conclusão

Neste artigo, utilizando-se de uma ferramenta desenvolvido na linguagem Java e a ferramenta WEKA, podemos observar o uso de três algoritmos de Aprendizagem de Máquina, Árvore de Decisão (representado pela classe J48), Naive Bayes e Regressão Logística, para predição de venda em empresas que prestam serviço de locação de equipamentos para eventos.

De forma geral, todos modelos abordados no presente trabalho tiveram uma taxa de acerto média de 74,53% nos testes executados. O que permite entender que há um espaço de 25 pontos percentuais para que possa ser aprimorado.

Para tanto, há duas características importantes, que seriam a quantidade de instâncias com variados exemplos e a quantidade e a relevância dos parâmetros selecionados para compor uma instância. Para a primeira característica, acredita-se que foi satisfatório o número de instâncias, pois somente para treino foram utilizados mais de quatorze mil registros.

Porém, para a segunda característica, apenas 8 parâmetros foram selecionados na etapa de análise dos dados e alguns com pouca variação de valor. A limitação do banco de dados disponibilizado não possibilitou uma melhor qualidade dos dados. Dessa forma, as combinações de parâmetros e resultados diminuiram as possibilidades de maiores variações para que pudessem ser modelados com maior precisão.

Para trabalhos futuros, entende-se que realizar a melhoria da limitação de parâmetros disponíveis supracitada poderia trazer resultados com maior precisão. Além disso, pensa-se na possibilidade de envolver análise dos modelos criados, a partir dos dados preparados, para que seja possível avaliar relevância dos parâmetros de entrada. Bem como avaliar o uso de outras técnicas de Aprendizagem de Máquina.

Referências

Borges, H. D. (2015). *Exploração de séries temporais em processos de previsão de vendas*. PhD thesis.

Bose, I. and Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & management*, 39(3):211–225.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Damasceno, M. (2010). Introdução à mineração de dados utilizando o weka. In *Anais do V Congresso Norte-Nordeste de Pesquisa e Inovação (CONNEPI)*, CONNEPI '10, pages 1–14.

Morettin, P. A. and Bussab, W. O. (2017). *Estatística básica*. Editora Saraiva.

Roza, F. S. d. et al. (2016). Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas.

Schneider, P. H. (2016). *Análise preditiva de Churn com ênfase em técnicas de Machine Learning: uma revisão*. PhD thesis.

SEBRAE (2017). *Como montar um serviço de locação de equipamentos para eventos*. Disponível em: `http://www.sebrae.com.br/sites/PortalSebrae/ideias/como-montar-um-servico-de-locacao-de-equipamentos-para-eventos,eec87a51b9105410VgnVCM1000003b74010aRCRD`. Acessado em: 28 de Novembro de 2017.