Modelo de avaliação de risco em campanhas *kickstarter* utilizando *machine learning*

Risk assessment model in kickstarter campaigns using machine learning

Gustavo Peixinho Cardoso Bacharelando, Universidade Nove de Julho (UNINOVE), Brasil – gustavopeixinhocardoso@gmail.com
Bruno Guimarães Sininbardi Bacharelando, Universidade Nove de Julho (UNINOVE), Brasil – brunogsininbardi98@gmail.com
Murilo Silva Sobral Bacharelando, Universidade Nove de Julho (UNINOVE), Brasil – murilo.sobral01@outlook.com
Edson Melo de Souza Mestre, Universidade Nove de Julho (UNINOVE), Brasil – souzaem@uni9.pro.br

RESUMO

O Crowdfunding tornou-se um fenômeno virtual por ser um modelo de negócio que visa arrecadar fundos para projetos coletivos de vários tipos. A Avaliação de Riscos com base nas características de tais projetos são um passo importante na minimização de perdas financeiras. Neste sentido, o objetivo deste trabalho é apresentar um modelo para avaliação de risco em projetos desta natureza utilizando técnicas Estatísticas e de Aprendizagem de Máquina (Machine Learning ou ML), considerando as etapas de preparação, processamento e análise dos resultados. Para o desenvolvimento do trabalho foi utilizada a ferramenta WEKA e a linguagem de programação Python com módulos específicos de ML. Por fim, este trabalho mostra que a utilização de técnicas de ML foi eficiente obtendo uma acurácia de 77%, superior aos 76% do modelo proposto por Etter, Grossglauser e Thiran (2013) e, por consequência, pode ser utilizada como um método na Avaliação de Risco em Campanhas Kickstarter, fornecendo condições aos investidores de mitigar os riscos associados aos projetos desta natureza por meio da análise de gráficos e dados numéricos.

Palavras-chave: Avaliação de Risco. Kickstarter. Machine Learning. Crowdfunding.

ABSTRACT

Crowdfunding has become a virtual phenomenon because it is a business model that aims to raise funds for collective projects of various kinds. Risk Assessment based on the characteristics of such projects is an important step in minimizing financial losses. In this sense, the objective of this work is to present a model for risk assessment in these projects using Statistical and Machine Learning (ML) techniques, considering the stages of preparation, processing and analysis of the results. For the development of the work the WEKA tool and the Python programming language with ML specific modules were used. The research showed that the use of ML techniques was efficient, obtaining an accuracy of 77%, higher to the 76% of the model proposed by Etter, Grossglauser e Thiran (2013) and can therefore be used as a method in Risk Assessment in Kickstarter Campaigns, providing conditions for investors to mitigate the risks associated with projects of this nature by analyzing graphs and numerical data.

Keywords: Risk Assessment. Kickstarter. Machine Learning. Crowdfunding.

Recebido em 12/02/2019. Aprovado em 24/02/2019. Avaliado pelo sistema *double blind peer review*. Publicado conforme normas da ABNT.
http://dx.doi.org/10.22279/navus.2019.v9n4.p66-79.914

1 INTRODUÇÃO

O financiamento coletivo ou Crowdfunding se tornou um fenômeno virtual por ter como objetivo a captação de recursos financeiros para realização de projetos dos mais variados tipos (COCATE; PERNISA JÚNIOR, 2012). A captação de tais recursos ocorre por meio da contribuição coletiva e informal de pessoas, denominadas backers, que estejam interessadas nos projetos ofertados por empresas, grupos ou pessoas físicas. Esse modelo de negócio surgiu na Europa no ano de 2006 com o site alemão Sellaband, que realizou captações para o desenvolvimento de projetos musicais (ETEMAD, 2017). Assim, bilhões de dólares passaram a ser movimentados anualmente ao redor do mundo, permitindo a viabilização de projetos em um curto prazo, além de contribuir para a criação de uma nova oportunidade de investimentos, remunerando os investidores em espécie ou com benefícios relacionados ao material produzido (MENDONÇA; MACHADO, 2015).

A difusão do modelo tomou grandes dimensões rapidamente, dando origem a diversas plataformas como o Prosper, JustGiving, Catarse, Kickante e, principalmente, a plataforma estadunidense Kickstarter (BRUTON *et al.*, 2015), focada em projetos culturais criativos e que acumula números expressivos: 16 milhões de backers, patrocinando 156.931 projetos, os quais acumulam cerca de 4.1 bilhões de dólares desde 2009 (KUPPUSWAMY; BAYUS, 2013; KICKSTARTER, 2019).

Nesta modalidade de negócio os backers realizam uma conexão direta com os empreendedores por meio das plataformas digitais que oferecem projetos inovadores e com grande potencial de êxito. Apesar de serem pequenos valores investidos, os projetos devem oferecer detalhes que permitam uma análise apurada, a fim de minimizar os riscos (VALIATI; TIETZMANN, 2012).

Neste contexto, avaliar os riscos envolvidos para esta modalidade de investimento é grande importância, uma vez que as características do projeto são essenciais para que ocorra o sucesso do mesmo e, por consequência, o retorno esperado seja alcançado.

O objetivo deste trabalho é apresentar uma metodologia que utiliza técnicas de Machine Learning (ML) para realizar avaliação de riscos em projetos de Crowdfunding da plataforma Kickstarter baseada em dados históricos. Para o desenvolvimento do trabalho foi utilizado o software WEKA (Waikato Environment of Knowledge Analysis) e a linguagem de programação Python com módulos Scikit-Learn (mineração e análise de dados), NLTK (processamento de linguagem natural) e Scipy (otimização numérica e álgebra linear).

2 TRABALHOS CORRELATOS

Segundo a literatura, a análise de riscos em projetos de Crowdfunding pode ser realizada considerando aspectos do projeto como descrição, duração, orçamento, número de colaboradores, entre outros (LICHTENBERG, 2011). É possível verificar na literatura que existe muitas técnicas utilizadas, as quais vão desde análises estatísticas baseadas simplesmente nos dados dos projetos, como também pela utilização da Inteligência artificial com a aprendizagem de máquina ou Machine Learning. Alguns trabalhos de grande relevância para o tema foram selecionados e são apresentados a seguir.

No trabalho de Zhu e Zhou (2016), foi utilizada a tecnologia blockchain para a avaliação e o controle de riscos em plataforma de Crowdfunding na China. Neste trabalho, os autores mostraram a importância da tecnologia para o setor financeiro, a qual permite verificar a integridade dos dados, além de ser um sistema distribuído que simplifica as transações financeiras entre investidores e empresários.

Dresner (2013) patenteou um sistema que permite coletar dados e realizar análises de risco, viabilidade de projetos, entre outros, permitindo a visualização por meio da internet. Tal sistema é capaz de realizar análises e disseminar os dados relacionados a transações financeiras por meio de uma interface de internet para minimizar o risco de investimento. Segundo o autor, "A interface do aplicativo permite que o usuário realize buscas no banco de dados em busca de informações relativas a um acordo financeiro em particular, incluindo um acordo financeiro, transação ou projeto e perfil do projeto."

A combinação de vários métodos empíricos como a mineração de textos, experimentos online e aprendizagem de máquina foram utilizados por Kim *et al.* (2017), que mediram o comportamento dos investidores e o impacto sobre os riscos dos projetos de Crowdfunding. Segundo os autores, "A assimetria de

informação entre criadores e apoiadores é uma das principais preocupações dos provedores de plataformas de crowdfunding." (KIM *et al.*, 2017, tradução nossa).

Já no trabalho de Li, Rakesh e Reddy (2016) foram utilizados métodos de regressão linear e logística para avaliar a potencialidade dos riscos de investimentos, entre outros aspectos relacionados aos projetos. Ainda, segundo os autores, tais métodos são importantes para avaliar a previsibilidade de projetos, utilizando características como o tempo necessário para atingir o sucesso de um projeto, dados parciais e recursos temporais obtidos no início dos projetos.

Por fim, Etter, Grossglauser e Thiran (2013) realizaram um estudo similar ao proposto neste trabalho, mas com uma abordagem diferenciada, onde os autores utilizaram dados de redes sociais, como o número de comentários e Tweets sobre a campanha, do valor já arrecadado, quantidade de contribuintes que já contribuíram para o projeto daquele usuário, entre outras informações, atingido uma acurácia de 76% na predição de sucesso de um projeto, enquanto neste trabalho foi obtido o valor de 77% com a técnica XGBoost.

A literatura mostra que as técnicas de Machine Learning são amplamente empregadas para a avaliação de riscos em projetos do tipo Crowdfunding. Portanto, este trabalho traz contribuições para um modelo de predição com características exclusivas dos dados provenientes do próprio Kickstarter.

3 CARACTERÍSTICAS DO KICKSTARTER

Desde sua fundação em 2009, o Kickstarter já captou cerca de 4.1 bilhões de dólares em financiamentos coletivos ou Crowdfunding, se tornando a maior plataforma deste modelo de negócio na internet (ETTER; GROSSGLAUSER; THIRAN, 2013). Este modelo de negócio permite que backers patrocinem projetos dos mais variados tipos, oferecidos pelos empreendedores ou founders (LI; RAKESH; REDDY, 2016).

Segundo Etter, Grossglauser e Thiran (2013), o Kickstarter é a maior plataforma de Crowdfunding da atualidade, o que transmite confiança para aos backers. Entretanto, qualquer investimento, por menor que seja, apresenta margem de risco. Nos projetos Kickstarter não é diferente, apesar de grandes sucessos terem sido alcançados, muitos projetos não obtiveram sucesso. Um dos riscos iminentes já parte do conceito de que todo projeto Kickstarter é *all or nothing* (tudo ou nada), ou seja, se um projeto arrecadar 99,99% de seu objetivo, o mesmo irá falhar, pois não atingiu a meta na sua completude (KUPPUSWAMY; BAYUS, 2013; MAROM; ROBB; SADE, 2016; LI; RAKESH; REDDY, 2016).

Dentro deste ambiente virtual, os projetos são cadastrados com grande riqueza de detalhes, com o objetivo de atrair o maior número de backers possível, viabilizando então o projeto proposto. Este detalhamento inclui, além da descrição, imagens, vídeos, objetivo financeiro a ser alcançado e posicionamento de como está o andamento do projeto em relação às captações financeiras e desenvolvimento (KICKSTARTER, 2019).

Os dados oferecidos são de responsabilidade do proponente, ou seja, existem orientações de como estes devem ser preenchidos para que tenham mais visibilidade, mas isso não garante o êxito. Uma descrição pobre em detalhes ou escrita incorretamente pode não deixar transparente o objetivo, assim como o uso de palavras-chave que não estão alinhadas ao objetivo (QIU, 2013). Diversos projetos são lançados sem um planejamento bem realizado, o que desmotiva o investimento por parte dos *backers*, devido à falta de transparência nos objetivos do projeto ou então qual a rentabilidade que será obtida por eles.

Outra característica importante é que podem existir outros projetos com o mesmo fim, mas com características diferentes, o que torna a necessidade de analisar o risco de investimento ainda mais.

3.1 Dataset do Kickstarter

A base de dados utilizada neste trabalho foi disponibilizada pelo criador do Kickscraper (SAMANCI; KISS, 2014), que é uma API (*Application Programming Interface*) ou Interface de Programação de Aplicação, que fornece condições de coletar dados públicos do Kickstarter. Após a extração da base de dados, foram selecionadas as colunas a seguir que obedecem ao layout mostrado no Quadro 1.

Quadro 1 – *Features* originais do dataset *Kickstarter*

Nome do campo	**Significado**
id	Identificador do registro
kickstarter_id	Identificador do projeto
name	Nome do projeto
slug	Palavras-chave do projeto
blurb	Descrição do projeto
main_category	Categoria principal
sub_category	Categoria secundária
launched_at	Data de lançamento
deadline	Data de encerramento
state	Estado final (*Successful*, *failed*, *canceled, live*)
goal	Objetivo de arrecadação
pledged	Arrecadação obtida
percentage_funded	Porcentagem arrecadada
backers_count	Número de doadores
currency	Moeda (USD, BRL etc.)
created_at	Data de criação
updated_at	Data em que mudou o Estado final

Fonte: Elaboração própria (2019)

Conforme mostra o Quadro 1, o conjunto de atributos ou *features* do *dataset* permitem realizar uma análise detalhada dos projetos utilizando métodos estatísticos e técnicas de *Machine Learning*, as quais são apresentadas a seguir.

4 MACHINE LEARNING

O Aprendizado de Máquina ou *Machine Learning* (ML) é um ramo da Inteligência Artificial que permite que computadores tenham a habilidade de aprender sem a necessidade explícita de serem programados para tal fim, utilizando conjuntos de dados históricos como base (SIMON, 2013). Para trabalhar com ML, existem conjuntos de ferramentas e algoritmos que podem ser incorporados em programas como o WEKA ou com a linguagem de programação Python, entre outros (OLIVEIRA; GONZAGA; ZAMBALDE, 2016).

O WEKA (*Waikato Environment of Knowledge Analysis*) é um pacote de algoritmos criado pela Universidade de Waikato, Nova Zelândia, para ML (RUBIANO; GARCIA, 2016), o qual contém ferramentas para pré-processamento, visualização de dados, classificação, regressão, séries temporais, associações, agrupamentos e *Deep Learning* (YOUNG *et al.*, 2018).

Um outro conjunto de ferramentas muito utilizado para o desenvolvimento de modelos para ML é a biblioteca da linguagem de programação Python *Scikit-Learn* (LILLEBERG; ZHU; ZHANG, 2015), que possui diversos algoritmos como: árvores de decisão, redes neurais, vetorização de campos textuais, entre outros.

4.1 Técnicas de Machine Learning

Diversas técnicas são encontradas na literatura para a aplicação em ML, entre as quais este trabalho destaca o TF-IDF (*Term Frequency-Inverse document frequency*) que, de acordo como Grossman e Cormack

(2014), é uma técnica estatística de extração de dados textuais, onde cada palavra de um documento recebe um score baseado no número de vezes que ela se mostra presente (*Term Frequency* ou Frequência do Termo) e em quantos registros essa palavra apresenta (*Inverse document frequency ou* Inverso da frequência do documento).

Outra técnica utilizada é o *Gradient Boosting Machine* (FRIEDMAN, 2018), que constrói modelos de regressão, ajustando uma função de forma sequencial e parametrizada utilizando o Método dos Mínimos-Quadrados (MMQ) em cada iteração, permitindo que haja um ajuste simples da função, o que determina então o aprendizado.

Em complemento ao *Gradient Boosting*, o *XGBoost* é uma técnica que tem como principal característica a robustez para trabalhar com um grande volume de dados, obtendo um desempenho melhor com valores ausentes, como colunas em branco, além de ajustar valores de acordo com a situação, o tornando muito flexível e interessante (SILVA *et al*., 2017).

Também muito utilizada é a técnica da Regressão Logística, que é um modelo linear generalizado que classifica variáveis independentes, ajustando-as a uma variável de resposta. Esta técnica é considerada não paramétrica, uma vez que não exige suposições sobre o comportamento probabilístico dos dados de entrada, permitindo a estimação direta da probabilidade de ocorrência de um evento (BITTENCOURT; 2003, MAYRINK; HIPPERT, 2017).

O algoritmo J48 se apresenta como um poderoso classificador baseado em árvores e tem como característica principal selecionar o melhor atributo presente no conjunto de dados para dividir a amostra em subconjuntos caracterizados por classes (BHASKARAN; LU; AALI, 2015). Este algoritmo é uma evolução do C4.5 que constrói árvores de decisão a partir de um conjunto de treinamento, assim como o algoritmo ID3. Trabalhando sobre o conceito de atributos contínuos e discretos, além de trabalhar com atributos incompletos (SALZBERG, 1993).

Já o *CfsSubsetEval* é um avaliador que considera a capacidade preditiva individual de cada dado em relação à redundância entre eles, destacando o nível de relacionamento com a classe gerada (SELVAKUBERAN; INDRADEVI; RAJARAM, 2008).

Após uma breve contextualização sobre as técnicas e algoritmos empregados neste trabalho, a seção 5 descreve a metodologia proposta neste trabalho, detalhando os procedimentos utilizados para a análise dos dados.

5 METODOLOGIA

Para construção do modelo proposto, foi necessária a realização de algumas etapas para preparação do *dataset*, uma vez que os dados estão despadronizados para aplicação dos algoritmos descritos na seção 4.

5.1 Preparação dos Dados

O *dataset* utilizado, descrito na sessão 3, para o pré-processamento possui um total de 231.602 projetos (registros) com estados finais *successful* ou *failed*, ou seja, se obtiveram ou não sucesso.

Entre as *features* constantes, as denominadas *pledged*, *backers_count*, *updated_at*, *percentage_funded*, *id* e *Kickstarter_id* foram descartadas, uma vez que seus valores não possuem relevância para o modelo, podendo causar ruídos para o objeto do estudo realizado.

Além dessas, as *features live* e *canceled* também foram descartadas, pois representam campanhas em andamento ou interrompidas antes de sua conclusão.

A seguir são apresentadas as etapas do pré-processamento sobre o *dataset* do Kickstarter.

5.2 Extração das Características

Foi realizado o processo de *Feature Engineering*, que é a extração de características existentes no conjunto de dados, a fim de remover valores que não são úteis para o modelo e que podem atrapalhar no desempenho do processamento (HARRIS, 2017).

A ferramenta computacional de extração e transformação de dados *Pentaho Data Integration* foi utilizada para este fim, uma vez que em seu pacote estão disponíveis ferramentas de análise, integração, transformação e mineração de dados (BOUMAN; VAN DONGEN, 2009).

Os dados foram carregados no *software Pentaho* onde foi realizada a filtragem para a extração dos registros de texto inconsistentes, uma vez que há caracteres que podem quebrar o registro, conforme mostra a Figura 1.

Figura 1 – Exemplo de linha inconsistente no *dataset*

...	...	...
11	101	813209633
12	This UGLY-Shirts are for the FEW & PROUD with \"Under Guard, Like You\" ATTITUDE	"Irwin Madriaga"
13	111	18039350
...	...	...

Fonte: Elaboração própria (2019)

É possível verificar que o caractere (\") indica uma marcação por aspas dentro do registro, realizando um destaque para a frase *Under Guard, Like You*, o que irá causar uma quebra no registro, ou seja, os algoritmos não considerarão o destaque da palavra, causando um ruído na interpretação do texto "puro".

5.3 Processamento Textual

Posteriormente as informações textuais obtidas foram compiladas e transformadas em *features* extras para o modelo por meio da concatenação dos atributos *name, slug e blurb*, que deram origem ao atributo "Descrição".

Sobre este atributo foram extraídos os seguintes dados: quantidade de caracteres totais, ocorrências de letras maiúsculas, minúsculas, interrogações e exclamações. Além deste procedimento, os campos que não possuem valores numéricos no *dataset* foram mapeados por meio da técnica *One-Hot Encoding* (PIECH *et al.*, 2015), que consiste em um processo pelo qual as variáveis categóricas são convertidas em formato numérico, a fim de facilitar a sua manipulação pelos algoritmos de ML.

5.4 Criação das Variáveis Quantitativas

Existem diversas possibilidades de análises numéricas sobre o conjunto de dados do *dataset*, e, por este motivo, os autores optaram em desenvolver variáveis quantitativas baseadas nas razões entre a arrecadação que pudessem melhorar a qualidade das análises realizadas com as técnicas de ML, sendo, portanto: categoria, objetivo e arrecadação diária.

As médias de doações foram calculadas para cada categoria por meio da Equação 1, a fim de separar projetos que buscam arrecadar mais do que a maioria das campanhas que aquele grupo costuma pleitear.

$$m_c = \frac{\sum d}{qtd} \quad (1)$$

Na qual:
m_c = média da categoria;
d = doações da categoria;
qtd = quantidade de projetos da categoria.

Também foi extraída a razão entre o objetivo e a média de sua categoria, conforme a Equação 2.

$$r_{mc} = \frac{o}{m_c} \quad (2)$$

Na qual:
r_{mc} = razão do objetivo pela média da categoria;
o = objetivo;
m_c = média da categoria.

Outro passo é calcular o valor diário necessário para atingir a meta, como mostra a Equação 3.

$$vl_d = \frac{o}{t_d} \quad (3)$$

Na qual:
vl_d = valor diário a ser atingido;
o = objetivo;
t_d = tempo de duração da campanha;

Após a realização dos passos anteriores, o *dataset* ficou com as seguintes colunas (*features*): Sucesso, Objetivo, Categoria secundária, Categoria principal, Descrição, Tamanho da descrição, Moeda, Número de exclamações, Número de interrogações, Número de letras minúsculas, Número de letras maiúsculas, *Diff Laun/Dead Diff Create/Laun*, Dia semana Lançamento, Dia semana Encerramento, Razão Objetivo/Dias, Razão Objetivo/Média da categoria secundária, Razão Objetivo/Média da categoria principal, Objetivo maior que a categoria principal e Objetivo maior que a categoria secundária.

Por fim, os dados obtidos na etapa anterior foram analisados por meio de métodos estatísticos e processados com as técnicas de ML apresentadas na seção 6, uma vez que os procedimentos metodológicos para os experimentos foram incorporados aos resultados para melhor compreensão e visualização.

6 EXPERIMENTOS E RESULTADOS

Os procedimentos apresentados nesta seção abordam uma análise estatística realizada no *dataset*, a qual não necessita a utilização de ML para geração de resultados.

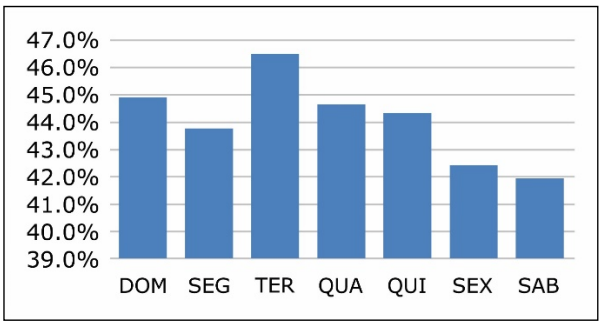
Após a análise, são mostrados os procedimentos e resultados obtidos com a aplicação das técnicas e algoritmos descritos na seção 4.

6.1 Análise do Dataset

Do total de projetos analisados, 129.101 (56%) não obtiveram êxito, contra 102.501 (44%) de campanhas sucedidas. Esta distribuição mostra que a acurácia de linha de base para este problema é de 56%, obtida por meio da hipótese de inferir todos os exemplos para a classe majoritária.

De acordo com as análises realizadas, os projetos que obtiveram maior de taxa de sucesso foram lançados às terças-feiras representando 46,4%, enquanto o sábado mostrou menor taxa com 42,0%, conforme mostra a Figura 2.

Figura 2 - Taxa de lançamento por dias da semana



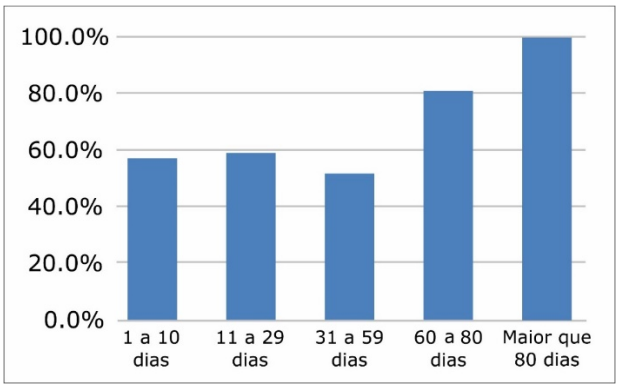
Fonte: Elaboração própria (2019)

Alguns projetos têm como objetivo a arrecadação de valores maiores que a média de doações de sua respectiva categoria principal, comportamento constatado em 3.529 registros, onde apenas 6% obtiveram resultado positivo, isso faz com que esta informação seja importante para o modelo acertar suas previsões em casos extremamente limítrofes.

A diferença, em dias, entre a data de criação e de lançamento possui um alto impacto para o sucesso da campanha. Para projetos criados e abertos ao público exatamente no mesmo dia, apresentam 64% de chance de sucesso, entretanto, caso a diferença for superior a 20 dias, a probabilidade de êxito sobe para 77%.

Para períodos de duração superiores a 60 dias, que correspondem a 17.158 campanhas, todas foram bem-sucedidas. Já no período entre 1 e 59 dias de duração a taxa de sucesso não varia de forma significativa. Não obstante, quando o projeto dura mais que dois meses, a probabilidade de êxito aumenta potencialmente, conforme mostra a Figura 3.

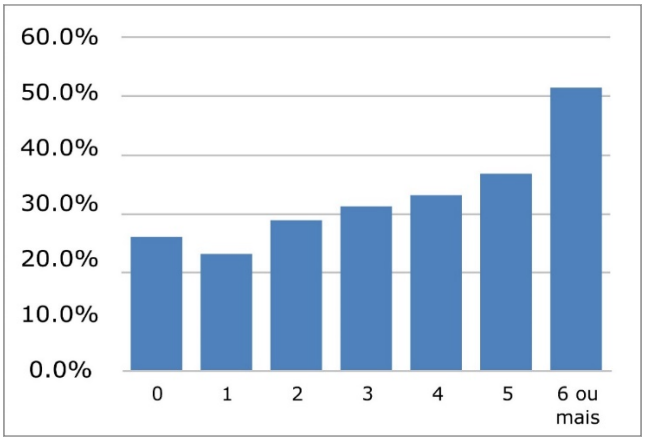
Figura 3 – Intervalo em dias para o lançamento de uma campanha



Fonte: Elaboração própria (2019)

Outra constatação sobre os dados é que uma campanha com menos de 50 caracteres no campo "Descrição do Projeto" tem uma probabilidade de êxito de apenas 23%, ao passo que entre 51 e 100 caracteres, o percentual se eleva para 42%. Quando há ocorrência de mais de 200, a chance de sucesso é de 50%. Ainda em relação ao campo Descrição do Projeto, a presença de apenas um caractere maiúsculo aponta êxito para o projeto de 26%, ao passo que cinco ou mais caracteres maiúsculos elevam este número para 52%, conforme mostra a Figura 4.

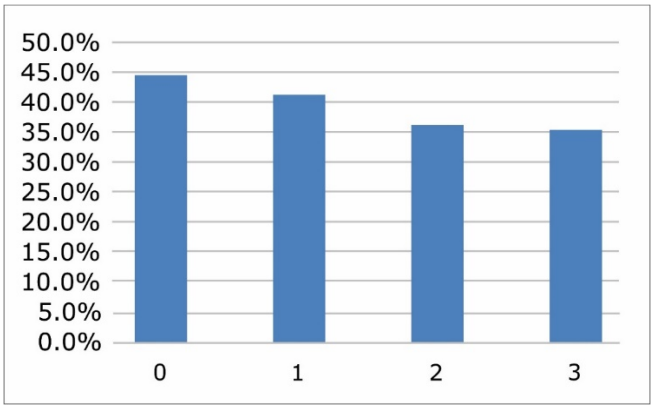
Figura 4 – Impacto da frequência de caracteres na descrição do projeto



Fonte: Elaboração própria (2019)

Já o número de interrogações mostrou ter um efeito diferente de todas as análises anteriores, apontando que as descrições que não têm esse caractere possuem taxa de sucesso de 44%, e quanto maior a recorrência, menor são as possibilidades de êxito do projeto, conforme mostra a Figura 5.

Figura 5 – Impacto do caractere interrogação na descrição do projeto



Fonte: Elaboração própria (2019)

6.2 Modelagem Preditiva com o WEKA

O mesmo *dataset* utilizado anteriormente foi carregado no software WEKA e o filtro da ferramenta *StringToWordVector* foi aplicado para transformação do campo descrição em um vetor de palavras, transformação TD-IDF.

A seguir foi realizada a 'sistemização', redução da palavra para o seu tronco raiz, com os algoritmos *AlphabeticTokenizer* e o *SnowballStemmer*. Com essa transformação, o tamanho da matriz esparsa totalizou 231.602 linhas por 1.022 colunas. Para reduzir o custo computacional e evitar a possibilidade de que os modelos generalizem de forma inapropriada para novos dados, foram selecionados os atributos com maior correlação para o sucesso ou fracasso do projeto, utilizando o mecanismo de seleção de *Features* *CfsSubsetEval*. E, para buscar os melhores *Subsets*, foi aplicada a busca *Best-First*, reduzindo a dimensionalidade da matriz para 15 colunas.

Posteriormente foi aplicado o algoritmo Zero-R do WEKA para estabelecer a linha de base para este problema e, como mencionado acima, este valor é de 56%, o que torna qualquer valor abaixo deste inviável para aplicações em cenários reais.

Por fim, foi aplicado o algoritmo J48 com um fator de confiança de 1%, o qual gerou na validação cruzada com dez subconjuntos uma árvore de informações. A Figura 6 mostra a árvore gerada pelo algoritmo.

Figura 6 - Árvore gerada pelo algoritmo J48

```
Diff Create/Laun <= 60
|   Diff Create/Laun <= 0
|   |   Objetivo maior que a media da categoria? <= 0
|   |   |   websit <= 2.571372
|   |   |   |   trump <= 3.41245: 1 (69990.0/25660.0)
|   |   |   |   trump > 3.41245
|   |   |   |   |   featur <= 0
|   |   |   |   |   |   by <= 1.303551: 0 (90.0/25.0)
|   |   |   |   |   |   by > 1.303551
|   |   |   |   |   |   |   NUMERO_UPPERCASE <= 8: 0 (3.0)
|   |   |   |   |   |   |   NUMERO_UPPERCASE > 8: 1 (8.0)
|   |   |   |   |   featur > 0: 1 (7.0/1.0)
|   |   |   websit > 2.571372: 0 (222.0/82.0)
|   |   Objetivo maior que a media da categoria? > 0: 0 (1339.0/330.0)
|   Diff Create/Laun > 0
|   |   Objetivo maior que a media da categoria? <= 0
|   |   |   fring <= 2.884182
|   |   |   |   Diff Create/Laun <= 1
|   |   |   |   |   album <= 0
|   |   |   |   |   |   danc <= 2.217123
|   |   |   |   |   |   |   th <= 2.238791
|   |   |   |   |   |   |   |   burn <= 3.314703
|   |   |   |   |   |   |   |   |   titanium <= 3.727831: 0 (64973.0/21832.0)
|   |   |   |   |   |   |   |   |   titanium > 3.727831
|   |   |   |   |   |   |   |   |   |   of <= 0
|   |   |   |   |   |   |   |   |   |   |   titanium <= 4.470754: 1 (15.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   titanium > 4.470754
|   |   |   |   |   |   |   |   |   |   |   |   titanium <= 4.626639: 0 (6.0)
|   |   |   |   |   |   |   |   |   |   |   |   titanium > 4.626639: 1 (9.0/3.0)
|   |   |   |   |   |   |   |   |   |   of > 0: 0 (2.0)
|   |   |   |   |   |   |   |   burn > 3.314703
|   |   |   |   |   |   |   |   |   NUMERO_UPPERCASE <= 5: 0 (33.0/8.0)
|   |   |   |   |   |   |   |   |   NUMERO_UPPERCASE > 5: 1 (161.0/63.0)
|   |   |   |   |   |   |   th > 2.238791
|   |   |   |   |   |   |   |   NUMERO_UPPERCASE <= 11: 0 (362.0/173.0)
|   |   |   |   |   |   |   |   NUMERO_UPPERCASE > 11: 1 (339.0/121.0)
|   |   |   |   |   |   danc > 2.217123: 1 (863.0/342.0)
|   |   |   |   |   album > 0
|   |   |   |   |   |   danc <= 2.479073
|   |   |   |   |   |   |   NUMERO_UPPERCASE <= 5
|   |   |   |   |   |   |   |   NUMERO_UPPERCASE <= 3: 0 (125.0/48.0)
|   |   |   |   |   |   |   |   NUMERO_UPPERCASE > 3
|   |   |   |   |   |   |   |   |   want <= 0: 1 (363.0/173.0)
|   |   |   |   |   |   |   |   |   want > 0: 0 (32.0/10.0)
|   |   |   |   |   |   |   NUMERO_UPPERCASE > 5: 1 (3594.0/1397.0)
|   |   |   |   |   |   danc > 2.479073: 0 (27.0/9.0)
|   |   |   |   Diff Create/Laun > 1: 0 (44230.0/6250.0)
|   |   |   fring > 2.884182: 1 (372.0/78.0)
|   |   Objetivo maior que a media da categoria? > 0: 0 (5830.0/156.0)
Diff Create/Laun > 60: 1 (15447.0)
```

Fonte: Elaboração própria (2019)

Analisando os valores obtidos, o algoritmo J48 apresentou um desempenho e acurácia satisfatórios para o problema, conforme mostra a Tabela 1. O atributo com maior entropia, ou seja, o nó inicial da árvore, é

o *Diff Create/Laun*, em que todos os projetos com mais de dois meses entre a criação e o lançamento foram bem-sucedidos.

O modelo também fez grande uso do fato do objetivo de a campanha ser maior que a média de doações da categoria daquele projeto, a árvore tem vários nós que são desenhados com base no número de letras maiúsculas que a descrição do projeto possui, ligando isso ao *score* que as palavras mais relevantes apresentam.

Tabela 1 – Resultado do Desempenho do Algoritmo J48

Modelo	**Acurácia**	***Recall***	**Área ROC**	**Precisão**
J48	72,65%	72,70%	77,77%	72,60%

Fonte: Elaboração própria (2019)

Os resultados mostram que a metodologia usada para extrair conhecimento da base de dados foi o diferencial para a obtenção deste desempenho. Entretanto, como o WEKA não possui alguns algoritmos de ML, lançou-se mão da linguagem Python, utilizando os pacotes XGBoost, Scikit-Learn, NLTK e Scipy, a fim de melhorar a análise dos dados e as predições e que é descrito a seguir.

6.3 Modelagem Preditiva com Python

Utilizando a linguagem Python com o NLTK, foi realizado o carregamento das funções de 'stemização' e 'tokenização' aplicadas no campo descrição, a fim de obter as palavras-chave.

Em seguida foi aplicado o algoritmo TFIDFVectorizer, resultando na dimensionalidade de 1.000 atributos para cada linha do dataset. Como esta função retorna uma matriz esparsa composta apenas pelos termos do campo, foi necessário concatenar os dados obtidos com as informações categóricas (duração do projeto, categoria, entre outros) por meio do método hstack da biblioteca Scipy, totalizando 1.017 atributos, e assim, como na concepção do algoritmo no WEKA, foram selecionadas as features mais significativas para a compreensão do problema pelo modelo de predição.

Após a seleção, foi empregado o método SelectFromModel, que tem a capacidade de recuperar os atributos considerados mais importantes de modelos baseados em conjuntos de árvores. O modelo utilizado para esta seleção foi o RandomForest, utilizando o coeficiente de Gini, que é um índice que representa uma medida de desordem (NASCIMENTO JUNIOR, 2017), para mensurar a relevância das features do modelo.

Com a matriz retornada por esta técnica, contendo 30 atributos, foram executados quatro algoritmos, dos quais todos passaram pela GridSearch, método de Busca Exaustiva que testa todas as combinações de parâmetros fornecidos, testados também por validação cruzada de dez subconjuntos. As melhores combinações são mostradas na tabela 2.

Tabela 2 – Melhores Parâmetros por Modelo

Modelo	***N_estimators***	***Max_depth***	***Learning_rate***
Extra Trees	500	30	-
*Adapative Boostin*g	500	-	0.5
XGBoost	400	-	0.4
Regressão logística	-	-	-

Fonte: Elaboração Própria (2019)

Com os parâmetros obtidos, os quatro algoritmos foram executados e apresentaram os resultados das acurácias, como mostra a Tabela 3.

Tabela 3 – Resultado dos Modelos de Predição

Modelos	Acurácia geral	Desvio padrão
Extra Trees	63,00%	10%
Adapative Boosting	76,10%	1%
XGBoost	77,00%	1%
Regressão logística	67,00%	5%

Fonte: Elaboração Própria (2019)

7 CONCLUSÃO

O modelo proposto neste trabalho se mostrou eficiente e satisfatório na avaliação do Risco de Investimentos em projetos de Crowdfunding no Kickstarter, além de obter uma acurácia de 77%, superior aos 76% do modelo proposto por Etter, Grossglauser e Thiran (2013).

Sobre a análise dos dados é possível verificar que as chances de sucesso de um projeto estão associadas diretamente a qualidade do texto informado no campo Descrição do Projeto, onde uma campanha com menos de 50 caracteres no campo descrição do projeto tem uma probabilidade de êxito de apenas 23%, ao passo que entre 51 e 100 caracteres, o percentual atinge 42% e, esse número aumenta para 50% quando há ocorrência de mais de 200 caracteres.

Por outro lado, a presença de apenas um caractere maiúsculo apresenta êxito de 26% para o projeto ao passo que cinco ou mais caracteres maiúsculos elevam este número para 52%.

A taxa de sucesso de um projeto se mostrou oscilante em relação ao dia da semana em que um projeto é lançado, mostrando que as terças-feiras são o melhor dia, apresentando uma taxa de sucesso de 46,4%, ao passo que aos sábados este valor decai para 42%.

As técnicas de ML mostraram que a qualidade dos dados é fundamental para que o modelo possa realizar o processamento dos dados e apresentar resultados visuais que permitam a tomada de decisão de forma mais direta, como mostra a árvore de atributos e valores na seção 6, Figura 6.

Por fim, este trabalho mostrou que a utilização de técnicas de ML pode ser utilizada como um método na Avaliação de Risco em Campanhas Kickstarter, fornecendo condições aos investidores de mitigar os riscos associados aos projetos desta natureza e, por este motivo, em trabalhos futuros os autores pretendem explorar a correlação dos atributos do dataset, a fim de fornecer ainda mais parâmetros de análise. Ademais, vislumbram a criação de uma ferramenta online que possa realizar a predição de uma campanha em tempo real.

REFERÊNCIAS

BHASKARAN, Subhashini; LU, Kevin; AALI, Al Mansoor. Student Performance and Time-To-Degree Analysis Using J48 Decision Tree Algorithm. *In*: MANAGING INTELLECTUAL CAPITAL AND INNOVATION FOR SUSTAINABLE AND INCLUSIVE SOCIETY: MANAGING INTELLECTUAL CAPITAL AND INNOVATION, 2015, Bari. **Proceedings of the MakeLearn and TIIM Joint International Conference 2**. Bari: ToKnowPress, 2015. p. 2029-2029. Disponível em: https://ideas.repec.org/h/tkp/mklp15/2029.html. Acesso em: 14 jul. 2019.

BITTENCOURT, Hélio Radke. Regressão logística politômica: revisão teórica e aplicações. **Acta Scientiae**, v. 5, n. 1, p. 77–86, 2003.

BOUMAN, Roland; VAN DONGEN, Jos. **Pentaho Solutions:** Business Intelligence and Data Warehousing with Pentaho and MySQL. Indianapolis, Indiana: Wiley Publishing, 2009.

BRUTON, Garry *et al.* New financial alternatives in seeding entrepreneurship: Microfinance, crowdfunding, and peer-to-peer innovations. **Entrepreneurship: Theory and Practice**, v. 39, n. 1, p. 9–26, 2015.

COCATE, F. M. Flávia Medeiros; PERNISA JÚNIOR, Carlos. Estudo sobre crowdfunding: fenômeno virtual em que o apoio de uns se torna a força de muitos. *In*: SIMPÓSIO NACIONAL ABCiber, 5., 2011, Florianópolis.

Anais [...]. Florianópolis: UDESC/UFSC, 2011. p. 1-14. Disponível em: http://abciber.org.br/simposio2011/anais/Trabalhos/artigos/Eixo%206/17.E6/226-353-1-RV.pdf. Acesso em: 11 jul. 2019.

DRESNER, Steven. System and Method of Data Collection, Analysis and Distribution. **U.S. Patent Application**, US 2013/0185228 A1, n. 19, filed July 18, 2013. v. 1, p. 1-28, 2013. Disponível em: https://patentimages.storage.googleapis.com/66/a5/ab/e5529cbc15e6ae/US20130185228A1.pdf. Acesso em: 14 jul. 2019.

ETEMAD, Hamid. The emergence of online global marketplace and the multilayered view of international entrepreneurship. **Journal of International Entrepreneurship**, v. 15, n. 4, p. 353–365, 2017.

ETTER, Vincent; GROSSGLAUSER, Matthias; THIRAN, Patrick. Launch hard or go home! Predicting the Success of Kickstarter Campaigns. *In*: THE FIRST ACM CONFERENCE ON ONLINE SOCIAL NETWORKS (COSN'13), 2013, Boston. **Proceedings of the First ACM conference on Online Social Networks (COSN'13)**. Boston: ACM, 2013. p. 177-182. Disponível em: https://infoscience.epfl.ch/record/189675. Acesso em: 14 jul. 2019.

FRIEDMAN, Jerome H. Greedy Function Approximation: a Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2018.

GROSSMAN, Maura R.; CORMACK, Gordon V. Grossman-cormack glossary of technology-assisted review, the. **Fed. Cts. L. Rev**., v. 7, p. 85, 2014.

HARRIS, Michael. **Feature Engineering for Algorithmic and Machine Learning Trading**. [2017]. Disponível em: https://medium.com/@mikeharrisNY/feature-engineering-for-algorithmic-and-machine-learning-trading-d0326305ac7b. Acesso em: 20 jan. 2019.

KICKSTARTER (EUA) (ed.). **About Kickstarter**. 2019. Disponível em: https://www.kickstarter.com. Acesso em: 20 jan. 2019.

KIM, Keongtae *et al.* Information Disclosure and Crowdfunding: An Empirical Analysis of the Disclosure of Project Risk. **Academy of Management Proceedings**, v. 2017, n. 1, p. 12360, jan. 2017.

KUPPUSWAMY, Venkat; BAYUS, Barry L. **Crowdfunding Creative Ideas:** The Dynamics of Project Backers in Kickstarter. [2018]. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2234765. Acesso em: 20 jan. 2019.

LI, Yan; RAKESH, Vineeth; REDDY, Chandan K. Project success prediction in crowdfunding environments. *In*: WSDM '16, 2016, San Francisco. **Proceedings of the Ninth ACM International Conference on Web Search and Data Mining**. New York: ACM, 2016. p. 247-256. Disponível em: https://dl.acm.org/citation.cfm?id=2835791. Acesso em: 14 jul. 2019.

LICHTENBERG, Frank R. Entrepreneurial Risk-Taking in Crowdfunding Campaigns. **Small Bus Econ**. January, p. 843–859, 2011.

LILLEBERG, Joseph; ZHU, Yun; ZHANG, Yanqing. Support vector machines and word2vec for text classification with semantic features. *In*: COGNITIVE INFORMATICS AND COGNITIVE COMPUTING, 2015, Beijing. **Proceedings of the 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)**. Beijing: IEEE, 2015. p. 136-140. Disponível em: https://ieeexplore.ieee.org/abstract/document/7259377. Acesso em: 14 jul. 2019.

MAROM, Dan; ROBB, Alicia; SADE, Orly. Gender dynamics in crowdfunding (Kickstarter): Deals, and taste-based discrimination. **NBER working papers**, n. 430, p. 1–75, 2016.

MAYRINK, Victor; HIPPERT, Henrique S. A hybrid method using Exponential Smoothing and Gradient Boosting for electrical short-term load forecasting. *In*: LATIN AMERICA CONGRESS ON COMPUTATIONAL INTELLIGENCE (LA-CCI), 2016, Cartagena. **Proceedings of the 2016 IEEE Latin American Conference on Computational Intelligence (LA-CCI)**. Cartagena: IEEE, 2016. p. 1-6. Disponível em: https://ieeexplore.ieee.org/abstract/document/7885697. Acesso em: 14 jul. 2019.

MENDONÇA, Rafael Uttempergher de; MACHADO, Luiz Henrique Mourão. Análise do crowdfunding no empreendedorismo brasileiro – características e tendências. **South American Development Society Journal**, v. 1, n. 3, p. 37–53, 2015.

NASCIMENTO JUNIOR, Luiz Antônio Ferreira. Aplicando método do gradiente ótimo na otimização do cálculo do grau de cobertura das regras em árvores de decisão Fuzzy. **Revista Brasileira de Computação Aplicada**, v. 9, n. 3, p. 31-43, 2017.

OLIVEIRA, Paulo Lima Júnior; GONZAGA, Luiz de Castro Júnior; LUIZ ZAMBALDE, Andreé Luiz. Applying Textmining to Classify News About Supply and Demand in the Coffee Market. **IEEE Latin America Transactions**, v. 14, n. 12, p. 4768–4774, 2016.

PIECH, C. *et al*. Deep Knowledge Tracing. *In*: CORTES, C.; LAWRENCE, N. D.; LEE, D. D.; SUGIYAMA, M.; GARNETT, R. (ed.). **Advances in Neural Information Processing Systems 28**. Montreal: Curran Associates, Inc., 2015. p. 505–513.

QIU, Calvin. Issues in Crowdfunding: Theoretical and Empirical Investigation on Kickstarter. **SSRN Electronic Journal**, 2013. Disponível em: http://www.ssrn.com/abstract=2345872. Acesso em: 1 jul. 2019.

RUBIANO, Sandra Milena Merchan; GARCIA, Jorge Alberto Duarte. Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance. **IEEE Latin America Transactions**, v. 14, n. 6, p. 2783–2788, 2016.

SALZBERG, Steven L. C4. 5: Programs for machine learning by J. Ross Quinlan Morgan Kaufmann Publishers, Inc., 1993. **Machine Learning,** v. 16, n. 3, p. 235-240, 1994.

SAMANCI, Murat; KISS, Gabor. **Exploratory Study on Technology related Successfully Funded Crowdfunding Projects. Post Online Market Presence.** 2014. 44 f. Master Thesis (Entrepreneurship), Department of Business Administration, Lunds University, Lund, 2014.

SELVAKUBERAN, K; INDRADEVI, M; RAJARAM, R. Combined Feature Selection and classification – A novel approach for the categorization of web pages. **UK Journal of Information and Computing Science**, v. 3, n. 2, p. 83–89, 2008.

SIMON, Phil. **Too big to ignore:** the business case for big data. Roboken, New Jersey: John Wiley & Sons, 2013. 256 p.

SILVA, Italla Dayanna da *et al*. Non-Stationary Demand Forecasting Based on Empirical Mode Decomposition and Support Vector Machines. **IEEE Latin America Transactions**, v. 15, n. 9, p. 1785–1792, 2017.

VALIATI, Vanessa Amália Dalpizol; TIETZMANN, Roberto. Crowdfunding: O Financiamento Coletivo como Mecanismo de Fomento à Produção Audiovisual. *In*: CONGRESSO DE CIÊNCIAS DA COMUNICAÇÃO NA REGIÃO SUL, 13., 2012, Chapecó. **Anais** [...]. Chapecó: Pontifícia Universidade Católica do Rio Grande do Sul, 2012. v. 2, n. 6, p. 1-13. Disponível em: http://www.intercom.org.br/papers/regionais/sul2012/resumos/R30-1090-1.pdf. Acesso em: 14 jul. 2019.

YOUNG, Tom *et al*. Recent trends in deep learning based natural language processing [Review Article]. **IEEE Computational Intelligence Magazine**, v. 13, n. 3, p. 55–75, 2018.

ZHU, Huasheng; ZHOU, Zach Zhizhong. Analysis and outlook of applications of blockchain technology to equity crowdfunding in China. **Financial Innovation**, v. 2, n. 1, p. 29, 2016.