International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

http://10.12785/ijcds/110147

Predicting Success of Campaigns on Membership based Patreon Crowdfunding Platform

Partha Mukherjee[1], Youakim Badr[1] and Srushti N Karvekar[1]

[1]*The Pennsylvania State University, Malvern, USA*

Received 1 June. 2021, Revised 17 Nov. 2021, Accepted 4 Jan. 2022, Published 20 Jan. 2022

Abstract: Crowdfunding platforms, such as the Patreon platform, are a means of regular financial support to entrepreneurs and artists who create independent content in the form of images, videos, podcasts, comics, games, or any media that supporters enjoy. Entrepreneurs leverage their potential base of patrons by using various social media platforms. Even though this collaboration has proved to be a practical approach to raising funds, it is difficult to predict the success rates of new projects. In this paper, we consider Patreon as the membership-based platform and our empirical analysis shows that half of proposed projects turn out to be successful. In this research, we build a data analytics approach to predict the rate of success of Patreon projects based on a dataset containing details of various features and historical information about previous projects. We employed a family of supervised classifiers that includes Naïve Bayes, Logistic Regression, Random Forest, and Boosting algorithms to predict the success of a given project. Currently, the Gradient Boosting classifier has achieved an average accuracy of more than 74%. Such results could help creators to define a path to better promote their content and improve monthly pledges.

Keywords: Crowdfunding, Patreon, Graphtreon, XGBoost, Social media, Gradient Boosting

1. Introduction

Since 2007, content creators are using crowdfunding platforms increasingly to raise funds for their content [1]. In crowdfunding platforms like Kickstarter and Indiegogo, content creators can get only one-time funding. They can collect either the total fund amount, or the sum raised by the fulfillment of the campaign. Such a business model works when creators plan to launch products in limited time. Content creators who publish their work on regular basis need recurrent funding. Thus, membership-based platforms like Patreon serve as a support to raise funding.

The Patreon platform aims to support content creators from various fields for the longer haul. Content creators collect funds as a recurring payment from their supporters, known as patrons, who can access content early along with special benefits and merchandise [2]**.** It is very difficult for creators who do not have any established reputation to find fans and supporters to fund them. Before crowdfunding platforms like Patreon, they used to make money using an advertising-based campaign where they would promote their content. They would get a part of the revenue earned in return for their investment.

The business model that Patreon follows is to ask creators for 10 percent of their earnings. That includes the fees for platform and processing of payment of 5 percent each [3]. The platform fee varies with the functionalities and services that content creators opt for. Overall, creators earn most of the revenue generated on the Patreon platform.

Graphtreon is an independent platform that collected data from the Patreon website using its API. According to Graphtreon, out of the total of 220,000 creators on Patreon, over 132,000 creators have at least one sponsor to keep up their work. This means that almost half of content creators have been successful in raising funds. In the past year, the number of project initiators on Patreon has risen by 37 percent. In the past year, there is 42 percent increase in the number of sponsors or patrons with an increase in average monthly earnings by 40 percent [4].

Project creators' turnover rate is strongly correlated with the income. People may attribute the success of the project to the platform, but it is highly dependent on content developers' patrons' base. Thus, if they do not obtain sufficient financial backing, the content developers begin to churn after a while. As very few content developers have a very high monthly income, the earnings graph is also skewed. A crowdfunding campaign is a success when the campaign obtains the necessary funds to meet the budget to keep the project alive on the crowdfunding platform, and is considered very successful when the funding exceeds the budget [5].

E-mail address: pom5109@psu.edu , yzb61@psu.edu, snk5287@psu.edu

Thus, we are exploring the Patreon platform to understand how content creators could ensure regular earnings and support from their patrons. We believe there is no known prior research on the prediction of success of the Patreon membership-based platform. The work presented in this paper is a significant extension of the work published in [6]. The current work aims to explore the various factors that contribute towards the growth of any crowdfunding project and help creators know how to improve their content to increase their patronage and success rate.

We have organized the paper as follows. In section 2, we provide the prior research on crowdfunding, particularly on the Kickstarter platform. Section 3 delineates the context of our research with the research questions that we address in this study. Section 4 and 5 explain the data exploration and the experimental settings respectively. Section 6 exhibits the results, while section 7 discusses the result and the implication. Section 8 concludes the work and provides research directions.

2. Related work

Crowdfunding refers to an internet fundraising activity where a founder issues an open invitation to investors, fans or followers to raise funds through contributions or in return for incentives or equity for a project or venture. The crowdfunding platform generates valuable information about the product or service specific to the campaign, in addition to the financial aspect. Hemer [7] explained the accuracy of measuring the success of a crowdfunding campaign is the reason for the patron's contribution in terms of funding. Genevsky et al. [8] showed that success of the crowdfunding projects is perceived as an effect of individual patron's decision to fund. There has been an increase in Crowdfunding platforms that support funding raising over past years. Considering the popularity of such platforms, attempts have been made to anticipate the success of a project on ventures such as Kickstarter [9] and IndieGoGo [10]. Many studies have attempted to identify the factors associated with the successful projects using Kickstarter which is a reward-based crowdfunding with significantly lower risk for the content creators [11]. For example, Ullah and Zhou [12] examined with 27,117 crowdfunding campaigns and showed campaigns with realistic goals, a definite period of completion, and communication with potential funders turned to be the successful ones. They also established that the gender of campaigners influenced the generation of funding on crowdfunding platforms. Schraven et al. [13] evaluated the campaign success from the perspective of cognitive biases of the participants in w.r.t duration of assessment of positivity about the Kickstarter projects. Shane and Cable [14] and Zheng et al. [15] have showed that the influence of an individual's social media connections in raising funds successfully is significant. The social network serves as an early pool of supporters for a project campaign [16] and helps in endorsing the creator's content to more external patrons [17]. This means that a creator's social network would impact his /her project success positively. Crosetto and Regner [18] have also found that videos, images, and blog entries on various websites serve as determinants of success for the crowdfunding ventures. Thus, these studies serve as evidence that the crowd will reward communicating the project's aim via social media.

Chung and Lee [19] showed that the incorporation of social media features in the set of predictors improved the prediction accuracy of success of the crowdfunding projects. The predictive model shows a 76.4% success rate using static features, while the use of Twitter features along with it means the success prediction is increased to 78.9%. Their study basically collected the dataset of Kickstarter projects consisting of the predictors that include: the profiles of the projects and corresponding users in the platform, time varying data related to projects, and information from the Twitter account of each user. The authors analyzed the aspects of the projects from the perspective of user behaviors and project features on the Kickstarter platform.

A myriad of prediction algorithms from the data mining perspective have been developed and compared. Numerous models have been built using SVM, decision trees, KNN and all the algorithms are being evaluated and compared based on their advantages and disadvantages, depending on various contexts [20], [21]. Experiments show that these algorithms perform differently with respect to the sample size, sample features, and the application domain [22]. Therefore, we can say that choosing of an algorithm is essential based on the sample characteristics and the application contexts.

Diverse machine learning classifiers such as logistic regression, decision tree, SVMs, REPTree, etc. were used by Greenberg, Pardo, Hariharan, and Gerber [23] to classify successful and unsuccessful crowdfunding projects. Decision tree was the optimal classifier with success prediction of 68%.

Ahmad, Tyagi, and Kaur [1] used weighed random forests models (i.e. a variant of random forest model) to measure the success and it generated 94.2% accuracy in success prediction. This study proposed to apply the same classifier model over all projects or over each category of projects.

Etter [24] used project specific committed funding as time series data to categorize the active projects as a success or failure with certain confidence. This study involved social features too, to measure the success of prediction using the Support Vector Machine (SVM) model, and achieved an improvement of 4% accuracy.

3. Study context

The number of sponsors backing different projects on the Patreon platform has crossed over 3 million[1]. There has been a rise of 50 percent patrons each year which is quite an impressive retention rate for the platform [25]. In our research we aim to understand the success determinants and how the success of a Patreon project can be influenced

by social media platforms. The creator of projects on the Patreon platform generally gets backing from his or her followers or patrons who have been following the creator's content on social media sites. This research study aims to provide guidelines to patrons on how they can improve based on social media features, to raise more funds and ensure the success of their project proposals. The study is done using the data from Graphtreon [4] and our target is whether the project is a success.

In this research, we aim to explore three research questions as follows:

• What are the determinants to measure the success of creator of Patreon project?

• How does social media influence determine the success of a new Patreon project?

• What type of content becomes popular in Patreon projects that gets more patronage?

First the creator must build their profile on Patreon by registering to the corresponding website. The registration is free of cost and there are no geographical restrictions on membership. Once registered, the creators can set their creator page in motion by delineating their work. Content creators should solicit the money on a per month basis by choosing the sponsors or patrons from different tiers they belong. The content creators decide on the tiers of the patrons. Further, the project creators set up their goals based on the income and let their potential patron base know about the future in the project.

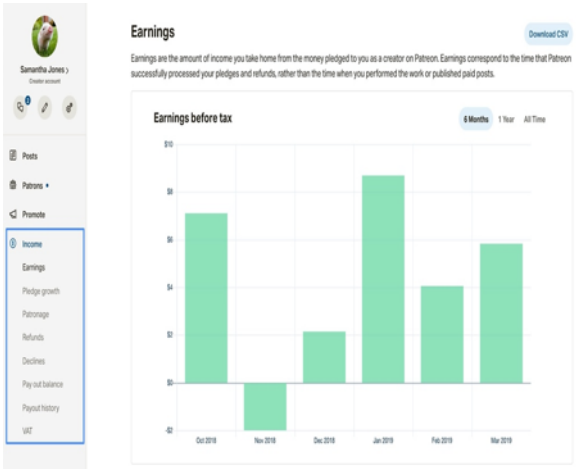

Figure 1. Patreon Creator Dashboard[1]

Every content creator has a dashboard which keeps track of funding growth, monthly earnings, patronage, payment details, and tax details as shown in Figure 1. It also displays the count of posts for patrons-only and public (non-patrons), the count of comments and likes made by the content maker and patrons on their page. This information is only available to the respective creator. Thus, due to the data privacy policy, one cannot acquire data directly from Patreon.

The Creator dashboard includes the following details, which are derived from https://support.patreon.com

1) **Posts** section gives a glimpse of the posts' engagement by showing terms like Views, Likes, Comments, Traffic sources, and Viewers.

2) **Income** section shows the details related to the earnings of the creator. The information displayed has the following values:

• Earnings - Summarizes income associated with the creator page.

• Pledge growth - Provides a month-wise summary of commitment to the page of the creator.

• Patronage - Summarizes the patronage that each creator receives.

TABLE I. Dataset Attributes / Features

Column Name	Description
Graphtreon	Graphtreon URL
Name	Content creator name
Category	The type of the content
Patrons	Count of sponsors or backers supporting the project
Earnings	Income on monthly basis
Range	If a project is either month long, or podcast or charged per published post
Is Nsfw	If the campaign is Not Safe For Work
Facebook Likes	Count of Facebook likes for a project
Twitter Followers	Count of Twitter followers of the project
YouTube Subscribers	Count of YouTube Subscribers
YouTube Videos	Number of YouTube Videos
YouTube Views	Count of views on the YouTube Videos
Status	Success=1, Failure =2

• Refunds – Summarizes detailed refunds to the patrons due to some fraudulent transactions or other technical issues due to Patreon website.

• Declines - Appraises the degree to which patrons' commitment is successfully exercised.

• Pay out balance - to set up payout information.

• Payout history - Summarizes monthly deductions from the creator balance.

• VAT – Outline of the Value-Added Tax (VAT) that Patreon accumulates from backers located in the European

Union (EU) countries.

By using the Graphtreon data, we deploy a data analytic process to predict a project's success or churn and what are the factors that contribute towards the growth of any project. By answer these questions, it becomes possible for a content creator to know how to improve their content to increase their patronage.

4. Data Collection and Processing

The dataset has been collected on daily basis from the Graphtreon website[2] and includes statistical information and rankings of about 250,000 projects that are available from March 2015 to June 2020 on the Patreon Platform. The data covers 28 different categories and includes the name of the content group, the number of patrons, their monthly income and the details on their profiles in different social media platforms, i.e., Facebook, Twitter, and YouTube. Some content creators do not own any social accounts connected to the Patreon site to upload and make the content sharable. There are 13 predictors to define the attributes of each maker as shown in Table I.

Out of content creators, 50% have been running their crowdfunding projects successfully, whereas the others have churned out. The successful and unsuccessful projects in terms of content creation in Patreon from our collected dataset are shown in Table Table II.

TABLE II. Projects on Patreon

Item	Count
Total number of projects	245,884
Projects with success	131,617
failed projects	103,019

Table IV shows the success rates and dates of the various creators whereasTable III gives the number of creators of the top 6 categories. We can see that 26% of creators fall in the video category. Games has the second highest number of creators. Music, Podcasts, Writing, and Painting have a similar number of creators in the range of 6 to 7 percent.

Figure 3 shows the success rate for the various categories. We can see that most of the categories have an average success rate 40%. Although the highest number of creators are in Video category, the success rate is 46%. Podcasts, dance, and theater categories have the highest success rate of 57% each. Creators in the music category

TABLE III. Creators per Category

Category	Creators
Video	64,839
Games	18,837
Music	16,902
Podcasts	15,096
Writing	14,459
Drawing & Painting	13,367

TABLE IV. Dataset Description Statistics

Parameters	Values
Total number of projects created	245,884
Number of successful projects	131,617
Number of failed projects	114,267
Success and failure ratio	0.535
Date of creation of first project	1^{st} March 2015
Date of creation of last project	30^{th} June 2020
Count of independent variables in dataset	13

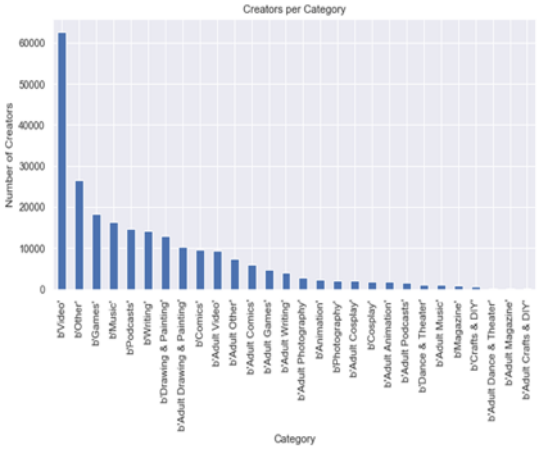

Figure 2. Creators per Category

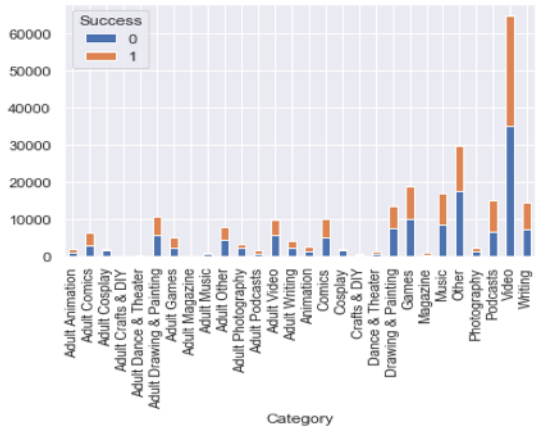

Figure 3. Success rate per Category

have a 50% chance of succeeding. However, the number of projects across categories vary. We believe that the prediction of fundraising for a crowdfunding project campaign should consider the category of the project content.

While exploring the data, there have been some creators who have not revealed their incomes. These creators were not included when the classifiers were being created, as earnings could not be imposed to each creator's profile. The earnings vary for every creator. Regular expressions with at least three characters are used to extract creators' information from the Graphtreon URL[3] connection such as "jam", "jame", "james" etc (i.e., for example, creators with first name James).

Data processing involved taking care of missing values in the data. Missing values are present in all the features and constitute over half of the data points, as a creator does not have accounts in all social media platforms. We encoded missing values of all the features using the Weight of Evidence (WOE) encoder. WOE methodology is extensively used in modeling the credit risk [26]. It is also named as default modeling probability [27]. The purpose of such encoding is to make the greatest variance between categories linked to the "Success" of the project, which is the target variable in our study. In each binned category, it calculates the number of successes and failures, then assigns each of the binned categories a logarithmic value [28], as shown in equation 1. In this transformation the information of the target variable has been utilized. The function WOE_Encoder() of the Python module category_encoders to use to calculate WOE. There are studies where the WOE is used in different applications [29], [30], [31].

$$WOE = ln\left(\frac{\%\ of\ Success}{\%\ of\ Failure}\right) \quad (1)$$

To demonstrate whether a function is associated with the target variable (i.e., success) leads us to use the Pearson correlation. Correlations between variables are described in Figure 4. The findings indicate that earnings are proportional to the number of backers. The remainder of the features display either very small or no correlations.

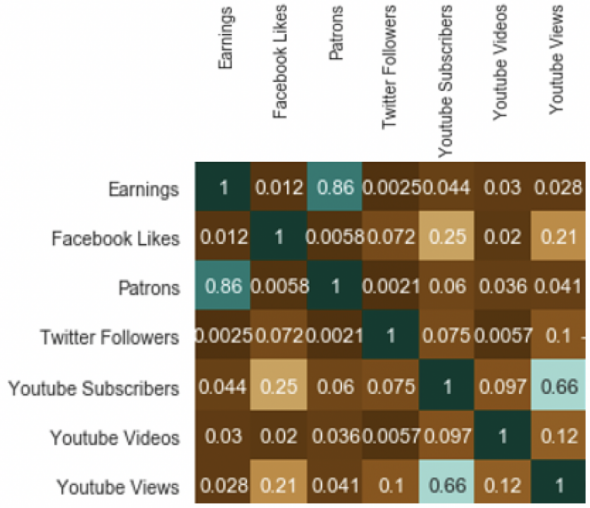

Figure 4. Correlations between variables

5. Experimental Settings

Contrary to traditional reward-based crowdfunding that characterizes a funding objective, Patreon is not straightforward from the perspective of a successful venture. Therefore, our research needs to incorporate a parameter related to "success" of a project. We believe in our analysis that if a creator has not yet churned off the site, it is marked as a success.

6. Model Selection

Since we need to predict the success and/or churn rate, this study is a classification problem and requires supervised machine learning algorithms. In our study we are interested in comparing the prediction of success in membership-based crowdfunding scenario between parametric and non-parametric classifiers in terms of different performance metrics. The parametric subset contains Naïve Bayes and Logistic, while the non-parametric subset contains random forest and two booster algorithms, such as XGBoost and Gradient Boosting Method (GBM). We have an intuition that distribution-free non-parametric classifiers will perform better to predict the success in crowdfunding set up [32].

To summarize, we select five different classification techniques:

- Logistic Regression,
- Naïve Bayes,
- Random Forest,
- XGBoost,
- GBM.

The pre-processed data is divided into a training and test set with a train-test ratio. The same training data is used for each classification model to get its highest performance accuracy. Mayr, Binder and Gefeller [33] in their study have explained the concept of boosting algorithms to improve the overall accuracy. Thus, Gradient Boosting and XGBoost algorithms were used in this study to improve the accuracy over the other methods.

1) Logistic Regression can be applied to this analysis as the financing may either result in success or failure. The logistic regression model is a non-linear regression relation between the target variable and multiple predictors. The coefficients ($\beta_i's$) can be used to determine how the presence of a set of attributes leads to an outcome. These coefficients can be used to rank the dependent variables [34]. P is the probability of an event to be a successful one. For the

logistic regression, the model would follow this equation,

$$\begin{aligned} &logit(\mathrm{P}) = \log[\mathrm{P(y = 1)}/(1 - \mathrm{P(y = 1)}] \\ &\quad = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \cdots + \beta_\mathrm{i} \mathrm{x}_\mathrm{i} \\ &For \mathrm{i} = 1, 2, \ldots, \mathrm{n} \end{aligned}$$

The Naïve Bayes classifier is a Bayesian probability-based model which assumes every attribute depends on the class. This classifier is helpful with high dimensional dataset as the probability of one attribute is not altered by other. The outcome of this classifier is a probability distribution function [35]. For Naïve Bayes,

$$P(y \mid x_1 x_2, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{3}$$

In our research the target variable y has two values: - "success" and "failure". In Naïve Bayes we need to determine the class 'y' with highest probability. Given the independent variables, we derive the label of the observation with maximum probability by means of equation 4.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{4}$$

A Random Forest consists of randomly generated decision trees which are independent of each other [36]. We need to adjust two parameters for the trees: - 1) the tree count and 2) count of features to be used for generating the trees. Gini impurity is used to determine the eigen values for splitting the nodes and thus the tree parses the features used to segment to minimize the impurity [37]. Gini impurity is calculated by equation 5.

$$Gini = 1 - \sum_j p_j^2 \tag{5}$$

where $p_j$is the probability of an item being in category j with $\sum p_j = 1$. So $1 - \sum_j p_j^2$is overall probability of misclassifying an event/item as it has been classified based on the probabilities of the overall distribution.

4) In Gradient Boosting algorithms, the weak classifiers are converted to strong classifiers. The weight for the weak classifiers is increased to classify the observations easily. In Gradient Boosting (GBM), the base-learner is fitted according to the negative loss function. Thus, the subsequent trees give better classification results [38]. Extreme Gradient Boosting (XGBoost) uses the gradient descent algorithm architecture. However, it is an improvement over GBM with its algorithmic enhancements and optimizations. XGBoost has certain features like tree-pruning, parallel processing, and regularization methods over GBM [39].

a) GBM is the optimization problem that works in two steps. The first one is to determine the step direction, and the second deals with optimization of the length of the step. We can express the GBM model using equation 6.

$$\begin{aligned} f(x) = \; & f^{(M)}(x) = \sum_{m=0}^{M} f_m(x) \\ = \; & f_0(x) + \sum_{m=1}^{M} \delta \rho_m \theta_m(x) \end{aligned} \tag{6}$$

Where $f_m(x) = \delta \, \rho_m \theta_m(x)$ is the step length ρ_m multiplied by a factor $\delta(0 < \delta < 1)$ and a base estimator $\theta_m(x)$ at each iteration m. f_0 is initialized using a constant involving the weight $\omega 0$ before iteration starts which can be further expressed as the loss function L.

b) On the other hand, XGBoost differs from GBM, and it solves the following equation to directly determine the step.

$$\frac{\partial L\left(y,\ f^{(m-1)}\right)(x) + f_m(x))}{\partial f_m(x)} = 0 \tag{7}$$

for each x in the dataset. L is the loss function w.r.t the current GBM estimate $f_m(x)$which can be written as shown in equation 8.

$$L(f_m) \propto \sum_{j=1}^{T_m} [G_{jm}\omega_{jm} + \frac{1}{2} H_{jm}\omega_{jm}^2] \tag{8}$$

For each region, the optimal weight $\omega_{jm} = -\frac{G_{jm}}{H_{jm}}$, $j = 1, \ldots T_m$. G_{jm} and H_{jm} respectively speak for the aggregation of gradient and hessian in region j.

7. Performance Metrics Used

With the given crowdfunding platform, we concentrate on the binary target variable that states whether a venture is a success or not. The target variable in our study is Success, which can take either of two possible outcomes: Success = 1 denotes successful campaign while Success = 0 denotes the campaign is an unsuccessful one.

We will be using the below metrics to evaluate the model performance.

Evaluation Metrics

- $Accuracy(Acc) = \frac{TP+TN}{TP+FP+FN+TN}$
- $Precision(Pr) = \frac{TP}{TP+FP}$
- $Recall(Re) = \frac{TP}{TP+FN}$
- $F - Measure = \frac{2*Precision*Recall}{Precision+Recall}$

Where TP = True Positives, TN = True Negatives, FP= False Positives, and FN = False Negatives.

For evaluating the classifiers, though we provide all the evaluation metrics, we mainly used accuracy and AUC (Area Under the Curve) under ROC curve (Receiver Operating Characteristic). ROC curve plots the predictive power of a classifier by providing true positive rate (TPR) and the false positive rate (FPR) on the y-axis and x-axis respectively. Any model is better because its ROC curve shows a low FPR or high TPR close to the top left corner. An AUC value equal to 0.5 denotes that 50 percent of the data was correctly guessed by the model.

8. Experimental result analysis

Upon testing the different models, the following values for the evaluation metrics shown in Table V were obtained.

We achieved the results shown in Table V, using the

TABLE V. Classifier performance in hold-out framework

Methods	Metrics				
	Acc	AUC	Pr	Re	F1-Score
Naïve Bayes	60	0.72	0.72	0.60	0.52
Logistic Regression	69	0.75	0.66	0.65	0.64
Random Forest	71.54	0.78	0.70	0.70	0.70
XGBoost	73.8	0.81	0.74	0.74	0.73
GBM	74.26	0.82	0.75	0.75	0.74

set of variables exhibited in Table II. From Table V , we compare the results of the different classification models we used in this study. It is observed that Gradient Boosting appears as the best performing classifier for the project success prediction on Patreon, from the perspective of all performance metrics. XGBoost stands as the second-best prediction model.

Figure 5. ROC for different classifiers in hold-out framework.

We plotted the ROC curve to visually illustrate the predictive power of the family of algorithms used in our study by showing TPR against the FPR at different threshold settings as shown in Figure 5.

In Figure 5, the AUC for the classifiers compute of separability between the outcomes of the target variable. The GBM classifier (the best one among the family of models), distinguishes between success and failure of content on the Patreon crowdfunding platform with 82% probability, while the XGboost stands second in the race with 81% probability.

We have further performed cross-validation on the dataset to ensure that every creator profile was trained and tested to get the best model. The data is segregated using 10-fold cross-validation. The same evaluation metrics are used to compare the performance of different classifiers. Values of all five-evaluation metrics are shown in Table VI as the performance metrics. In cross-validation too, Gradient Boosting (GBM) eclipses the other classifiers from all performance metrics standpoint.

TABLE VI. Classifier performance in cross-validation

Methods	Metrics				
	Acc	AUC	Pr	Re	F1-Score
Naïve Bayes	62	0.74	0.55	0.80	0.59
Logistic Regression	69	0.75	0.61	0.76	0.68
Random Forest	71	0.78	0.67	0.68	0.67
XGBoost	73	0.81	0.70	0.78	0.73
GBM	75	0.83	0.71	0.79	0.74

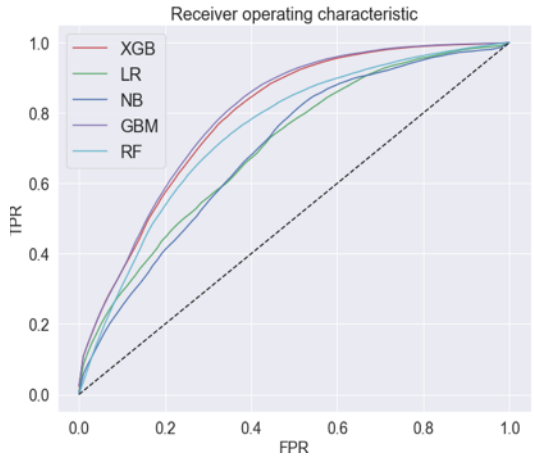

Figure 6. ROC for different classifiers in cross-validation framework

In the cross validation scenario, a similar pattern is observed and there is little change between the corresponding evaluation metrics for the two frameworks. Amongst the ensemble models, gradient boosting performs the best, while discriminating between the successful and unsuccessful projects. The AUC value of 83% shows that the data can be fit well by Gradient Boosting. Also, XGBoost shows almost similar AUC value of 81% (see Figure 6).

We chose a family of five algorithms where Naïve Bayes and Logistic Regression are parametric while the remaining three are non-parametric algorithms. It is observed that non-parametric algorithms perform better to distinguish between success and failure of content on the Patreon crowdfunding platform than the parametric classifiers considered in this study. We believe that the distribution of Patreon projects is uneven, like the Kickstarter projects [40], and free of assumption on the coefficients of logistic regression and the distribution on the input variables of Naïve Bayes classifiers. The last three classifiers are ensemble classifiers which are more powerful in the sense that it assimilates the predictions from multiple independent base classifiers (i.e. decision tree here) together compared to that from an independent classifier [41].

Overall, we conclude that our classifier's production was satisfactory. Regardless of how we split the dataset, our precision seems to touch an upper bound of 75 percent. This indicates that there is a probability that other variables exist that will help us to better identify the successful projects. We assume that possible additional variables could be the message postings on Patreon page, creators crowdfunding experience, demographic attributes of the creators and patrons, geo-location, creators' connectedness in the platform, and content analysis of the uploaded text and video as well. Our current dataset does not include these variables.

9. Discussion of results

Prediction of success of crowdfunding campaigns is becoming increasingly significant due to the growth in the use of different crowdfunding ventures by promising entrepreneurs, low-cost businesses, and individuals. We can ask the question "how are creators of crowdfunding projects using social media to activate their fans to create economic value"?

Using archival data from Patreon projects, we observe that boosting methods (GBM, and XGboost) appeared to be more efficient to predict the success of projects in comparison to the other three techniques, though Random Forest also reports more than 70% accuracy, the boosting algorithms further improve the prediction results in both the partitioning and cross validation set-up. In both set-ups we notice the non-parametric classifiers outperform the parametric classifiers. Furthermore, the non-parametric subset of the family of classifiers considered in this study is constituted with three ensemble classifiers. In this study the random forest is the basic ensemble classifier while XGBoost and GBM are the boosting variants that takes predictors in a sequential fashion instead of random manner. It makes the predictions better by boosting the weaker models, leveraging the residual patterns. Though XGBoost is theoretically more powerful than GBM, we believe GBM splits the base classifier (decision tree) leaf-wise while XGBoost splits the base classifier in level-wise fashion. The leaf wise split may result a little better accuracy than level-wise algorithm, by lowering more loss

We are interested to see which features contribute more to the prediction. The feature importance for the predictors is computed using the Gini impurity as shown in Figure 7. The list of features in Figure 7with importance responds to our first research question regarding the determinants of measuring the project success on Patreon. From Figure 7, we notice that incomes which depend on the number of patrons decides the success extensively. The content makers will churn if the number of refusals from sponsors or backers related to the campaigns grows. It is next to impossible to realize the projects with quality content with limited funds or resources. Additionally, the content makers do need the likes and following of the contents as a measure of appreciation as it is attributed to attract the sponsors and hence improves the chance of getting more sources of earnings.

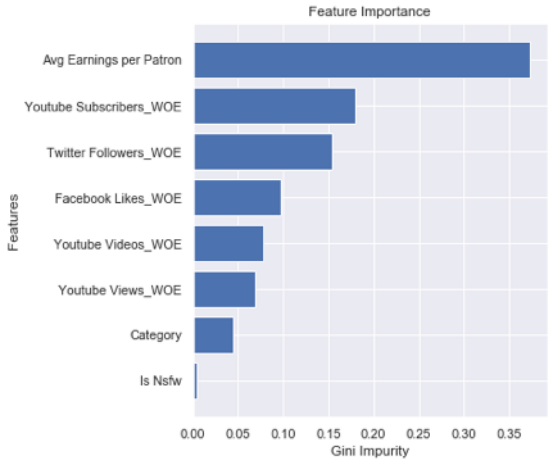

Figure 7. Feature Importance

Secondly, subscribers to YouTube also play an important part in the content's progress. One-fourth of all campaigners attached to Patreon platform belong to the video content creation category. The creators may display more videos to illustrate the process of content development to their fans or followers such as how to create comics, build online games, puzzles for psychological testing etc. This may explain why subscribers (followers) play a pivotal role in evaluating performance of the creator of the project. We also see from Figure 7, that Twitter and Facebook are effective in measuring the success of the projects as independent social media platforms. The aforementioned observations on the influence of social media platforms and type of content category to determine the success of the crowdfunding projects on Patreon answer our second and third research questions. We can also say that the creators should take steps to gain attention of the viewers with their projects on Patreon to increase content popularity as indicated by YouTube views in Figure 7.

We can also use this analysis for patronage forecasting since it can choose the successful projects to support in various categories. The projects which have failed could be analyzed and modified to improve the creator content. Along with the predictions of the success, we can also predict the other information on the ongoing projects such as the count of patrons, the earnings, and monetary progress etc.

From the experimental results we observe that the Gradient Boosting model accuracy reaches 75% while that for Naïve Bayes (the worst performing model) is only 60% to forecast the success of crowdfunded projects on Patreon. The result indicates that non-parametric models are more powerful to predict positive cases, which indicates the successfully financed projects.

10. Conclusion and future work

Prospective Patreon creators should have tools to analyze the features of their respective campaign profiles to predict the success of their projects before that are launched. We leveraged the supervised machine learning classifiers to facilitate them do so. We applied various models on Patreon projects dataset to determine whether the launched projects can be categorized as success or failures in this work. Our work in this area aims to assist new content creators with project planning so that they can deliver what the customer expects. The result of this study may facilitate the building of a forecasting engine that can advise potential creators in the creation of successful content for their projects, this may help the newcomers in the crowdfunding platforms for their campaigns. To support this forecasting tool, we employed a family of classifiers, ranging from naïve Bayes, to ensemble techniques. We observed that Gradient Boosting provided the best results, with the least runtime.

We have only studied the dummy results of success attribute of the project. We will further evaluate other aspects of crowdfunding by forecasting the number of patrons, the earnings for every category, and the growth in the creator profile in terms of contents and posts.

Though we are encouraged by this result, we look to explore additional features in the future for further future improvements on the present study. Although we have resorted to mainly supervised algorithms, we intend to explore the deep learning methodologies in our upcoming work.

In future study, we intend to include the dynamic features of the project campaign such as updates, comments, posts per day etc., and the social promotion attributes such as followers' counts, count of tweets, count of shares of YouTube videos, to comprehensively analyze how the outcome is affected by these attributes. We will carry out the experiment on how the addition and subtraction of features can improve the performance behavior of the classifiers, both in supervised and deep learning setups. The feature set will contain the present and the additional features explored in the near future.

We will apply our methodology on different crowdfunding platforms other than Patreon in future to compare the accuracy of classifiers to measure the success of projects across different crowdfunding platforms. As crowdfunding is a growing field, we hope to provide the creators with an enterprising framework to launch their project.

References

[1] F. S. Ahmad, D. Tyagi, and S. Kaur, "Predicting crowdfunding success with optimally weighted random forests," *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS). 2017. IEEE*, pp. 770–775.

[2] L. Wilson and Y. W. Wu, "A little bit of money goes a long way: Crowdfunding on Patreon by YouTube sailing channels," *Social Science Research Network (SSRN)*, vol. 2919840, 2019.

[3] C. Perrin, "The Innovative Business Model of Daft Punk," *Innovation in the Cultural and Creative Industries*, vol. 8, pp. 55–75.

[4] "Patreon Creators Statistics," *Graphtreon*, 2020.

[5] A. Fernandez-Blanco, J. Villanueva-Balsera, V. Rodriguez-Montequin, and H. J. S. Moran-Palacios, "Key factors for project crowdfunding success: An empirical study," *Sustainability*, vol. 12, no. 2, pp. 599–599, 2020.

[6] P. Mukherjee, Y. Badr, and S. Karvekar, "Prediction of Success in Crowdfunding Platforms," *International Conference on Decision Aid Sciences and Application (DASA)*, pp. 233–237, 2020.

[7] J. Hemer, "A snapshot on crowdfunding," *Arbeitspapiere Unternehmen und Region*, no. R2, 2011.

[8] A. Genevsky, C. Yoon, and B. Knutson, "When brain beats behavior: Neuroforecasting crowdfunding outcomes," *Journal of Neuroscience*, vol. 37, no. 36, pp. 8625–8634, 2017.

[9] D. Steinberg, "The kickstarter handbook: Real-life Crowdfunding success stories," 2012.

[10] C. Gallemore, K. R. Nielsen, and K. J. E. Jespersen, "The uneven geography of crowdfunding success: Spatial capital on Indiegogo," *EPA: Economy and Space*, vol. 51, no. 6, pp. 1389–1406, 2019.

[11] A. Schwienbacher, "Entrepreneurial risk-taking in crowdfunding campaigns," *Small Business Economics*, vol. 51, no. 4, pp. 843–859, 2018.

[12] S. Ullah and Y. Zhou, "Gender, anonymity and team: What determines crowdfunding success on Kickstarter," *Journal of Risk & Financial Management*, vol. 13, no. 4, pp. 80–80, 2020.

[13] E. Schraven, E. V. Burg, M. V. Gelderen, and E. J. Masurel, "Predictions of Crowdfunding Campaign Success: The Influence of First Impressions on Accuracy and Positivity," *Journal of Risk & Financial Management*, vol. 13, no. 12, pp. 331–331, 2020.

[14] S. Shane and D. Cable, "Network ties, reputation, and the financing of new ventures," *Management science*, vol. 48, no. 3, pp. 364–381, 2002.

[15] H. Zheng, D. Li, J. Wu, and Y. Xu, "The role of multidimensional social capital in crowdfunding: A comparative study in China and US," *Information Management*, vol. 51, no. 4, pp. 488–496, 2014.

[16] J. S. Hui, E. M. Gerber, and D. Gergle, "Understanding and leveraging social networks for crowdfunding: opportunities and challenges," *Proceedings of the 2014 conference on Designing interactive systems*, pp. 677–680, 2014.

[17] J. Huhtamäki, L. Lasrado, K. Menon, H. Kärkkäinen, and J. Jussila, "Approach for investigating crowdfunding campaigns with platform data: case Indiegogo," *Proceedings of the 19th International Academic Mindtrek Conference. 2015. Finland*, pp. 183–190.

[18] P. Crosetto and T. Regner, "It's never too late: Funding dynamics and self pledges in reward-based crowdfunding," *Research Policy*, vol. 47, no. 8, pp. 1463–1477, 2018.

[19] J. Chung and K. Lee, "A long-term study of a crowdfunding platform: Predicting project success and fundraising amount," *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 211–220, 2015.

[20] X. Chen, H. Wang, Y. Ma, X. Zheng, and L. Guo, "Self-adaptive resource allocation for cloud-based software services based on iterative QoS prediction model," *Future Generation Computer Systems*, vol. 105, pp. 287–296, 2020.

[21] W. Wang, G. Tan, and H. Wang, "Cross-domain comparison of algorithm performance in extracting aspect-based opinions from Chinese online reviews," *International Journal of Machine Learning & Cybernetics*, vol. 8, no. 3, pp. 1053–1070, 2017.

[22] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. Al-Shawakfa, "A comparison study between data mining tools over some classification methods," *International Journal of Advanced Computer Science Applications*, vol. 8, no. 2, pp. 18–26, 2011.

[23] M. D. Greenberg, B. Pardo, K. Hariharan, and E. Gerber, "Crowdfunding support tools: predicting success & failure," *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 1815–1820, 2013.

[24] V. Etter, M. Grossglauser, and P. Thiran, "Launch hard or go home! Predicting the success of Kickstarter campaigns," in *Proceedings of the first ACM conference on Online social networks*. ACM, 2013, pp. 177–182.

[25] J. Ciovacco, "Kickstarter and Patreon. The Rosen Publishing Group, Inc." 2016.

[26] D. S. Lade, "Estimation of credit risk for mortgage portfolios." MS Thesis, Nova School of Business and Economics," 2020.

[27] I. Wod, "Weight of evidence: A brief survey," *Bayesian Statistics*, pp. 249–270, 1985.

[28] D. L. Weed, "Weight of evidence: a review of concept and methods," *Risk Analysis*, vol. 25, no. 6, pp. 1545–1557, 2005.

[29] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International Journal of Computer Applications (IJoCA)*, vol. 175, no. 4, pp. 7–9, 2017.

[30] S. Baskoro and W. D. Sunindyo, "Predicting Issue Handling Process using Case Attributes and Categorical Variable Encoding Techniques," *International Conference on Data and Software Engineering (ICoDSE). 2019. IEEE*, pp. 1–5.

[31] C. Seger, ""An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing." B.S Thesis School of Electrical Engineering and Computer Science, Royal Institute of Technology (KTH)," 2018.

[32] F. Siddiqui and Q. M. Ali, "Performance of non-parametric classifiers on highly skewed data," *Global Journal of Pure and Applied Mathematics*, vol. 12, no. 2, pp. 1547–1565, 2016.

[33] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms-from machine learning to statistical modelling," *Methods of Information in Medicine*, vol. 53, no. 6, pp. 419–427, 2014.

[34] S. J. Lee, "Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data," *International Journal of Remore Sensing (IJoRS)*, vol. 26, no. 7, pp. 1477–1491, 2005.

[35] S. Mukherjee and N. Sharma, "Intrusion detection using naïve Bayes classifier with feature reduction," *Procedia Technology*, vol. 4, pp. 119–128, 2012.

[36] C. J. Mantas, J. G. Castellano, S. Moral-García, and J. Abellán, "A comparison of random forest based algorithms: random credal random forest versus oblique random forest," *Soft Computing*, vol. 23, no. 21, pp. 10 739–10 754, 2019.

[37] W. Wang, H. Zheng, and Y. J. Wu, "Prediction of fundraising outcomes for crowdfunding projects based on deep learning: a multimodel comparative study," *Soft Computing*, vol. 24, no. 11, pp. 1–19, 2020.

[38] J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3056–3064, 2015.

[39] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[40] E. Mollick, "The dynamics of crowdfunding: An exploratory study," *Journal of business venturing*, vol. 29, no. 1, pp. 1–16, 2014.

[41] Y. Liu, "eXtreme Gradient Boosting (XGBoost): Better than random forest or gradient boosting," *Statistics and Data Analysis using R*, 2018.

Partha Mukherjee Partha Mukherjee received his bachelor's degree in mechanical engineering in 1995 from Jadavpur University in India. He received his Master of Technology in Computer Science from Indian Statistical Institute in 2001. He earned his second graduate degree in computer Science from the University of Tulsa in 2008. He completed his Ph.D. from Penn State in information and technology with a minor in applied statistics in 2016. Previously he worked as post-doctoral research scholar in the management information systems department at Eller College of Management at the University of Arizona where he also taught a data mining for business analytics course. He served as a research consultant at the Indian Institute of Technology, (IITKGP) in designing and implementing a combinational unit of the ADSP 21020 microprocessor. Partha is a member of ACM, ACEEE, AIS, AiR and ASE. He has published papers in peer reviewed IEEE, Elsevier, and ACM Journals and conferences such as ACM Transaction on the Web, IEEE Intelligent Systems, Elsevier Electronics Commerce Research and Applications, and ACM SIGCHI. He is a reviewer of several journals such as Internet Research, Information Processing and Management, JSM Mathematics and Statistics etc. His research interests are in social computing, web analytics, data mining, e-commerce, and natural language processing with a focus of text simplification.

Youakim Badr Youakim Badr received his bachelor's and master's degrees in computer science from the Lebanese University and an additional master's degree in mathematical modeling and scientific software engineering from the Francophone University Agency (AUF). He earned his Ph.D. in computer science from the National Institute of Applied Sciences (INSA-Lyon), where he worked as an associate professor in the computer science and engineering department. Over the course of his research, Dr. Badr has worked extensively in the area of service computing (distributed systems) and information security. His current research strategy aims at developing a new software engineering approach for designing and deploying "smart connected devices" and building "smart service systems" for the Internet of Things. In addition, he conducts research activities on the integration of data analytic capabilities at the edge of the Internet of Things (Edge Machine Learning) and into connected devices (Built-in analytics) to make the Internet of Things smarter from the flood of data generated by connected devices. Dr. Badr authored or co-authored more than 95 publications in international journals and conferences, four edited proceedings, and two books published by Springer Verlag in the Studies in Computational Intelligence Series. In 2010, Dr. Badr spent his sabbatical leave at Cornell University and Pennsylvania State University. He has held short-term visiting scholar positions at the University of Sydney, the University of Namur in Belgium, and Zayed University in the UAE. Dr. Badr is vigorously involved in a series of international conferences. He is a professional member of IEEE, a lifetime member of ACM, and associate member of the ACM special interest group on knowledge discovery and data mining (SIGKDD).

Srushti N. Karvekar Srushti N. Karvekar is a graduate with Master's in Data Analytics from Pennsylvania State University. Working as a Graduate Research Assistant, she have published a couple of conference papers. Ms Karvekar demonstrated history of working in the software industry as a Programmer Analyst at Cognizant. She was involved in designing R, Python and SAS applications using Agile Methodologies and performing the analysis tasks associated with critical service management processes. She has experience in writing complex SQL queries to extract data for data analysis, report creation and visualization. Skilled in Python, Hadoop, Machine Learning, Deep Learning, Natural Language Processing, and R, she is currently working as an Analytics Engineer at CubeSmart Self-Storage as part of the Revenue Management department. Her work includes utilizing machine learning and data science methodologies to enhance the pricing management system.