Predictive modelling for startup and investor relationship based on crowdfunding platform data

IOP Conf. Series: Journal of Physics: Conf. Series **971** (2018) 012002 doi:10.1088/1742-6596/971/1/012002

Predictive modelling for startup and investor relationship based on crowdfunding platform data

Andry Alamsyah, Tri Buono Asto Nugroho
School of Economics and Business, Telkom University, Bandung, Indonesia

andrya@telkomuniversity.ac.id, tribuonoasto@student.telkomuniversity.ac.id

Abstract. Crowdfunding platform is a place where startup shows off publicly their idea for the purpose to get their project funded. Crowdfunding platform such as Kickstarter are becoming popular today, it provides the efficient way for startup to get funded without liabilities, it also provides variety project category that can be participated. There is an available safety procedure to ensure achievable low-risk environment. The startup promoted project must accomplish their funded goal target. If they fail to reach the target, then there is no investment activity take place. It motivates startup to be more active to promote or disseminate their project idea and it also protect investor from losing money. The study objective is to predict the successfulness of proposed project and mapping investor trend using data mining framework. To achieve the objective, we proposed 3 models. First model is to predict whether a project is going to be successful or failed using K-Nearest Neighbour (KNN). Second model is to predict the number of successful project using Artificial Neural Network (ANN). Third model is to map the trend of investor in investing the project using K-Means clustering algorithm. KNN gives 99.04% model accuracy, while ANN best configuration gives 16-14-1 neuron layers and 0.2 learning rate, and K-Means gives 6 best separation clusters. The results of those models can help startup or investor to make decision regarding startup investment.

1. Introduction

Crowdfunding platform is a place where startup exhibit and promote their project idea to public to get funded. The most funded crowdfunding platform for creative project is Kickstarter. Since its launch on April 28,2009, over 14 million people have commit to invest on many categories project, $3.3 billion has been invested, and 132,039 projects have been successfully funded [1]. In Kickstarter, startup promote their project by explain their idea, set project goal, deadline, and rewards. As consequence, people show their interest on the project, and finally fund the project. Crowdfunding platform is an efficient way for startup to get funded without traditional liabilities, such as in the conventional bank. There is various category project that startup can participate. A safety procedure is provided to ensure achievable low-risk environment. The startup promoted project must accomplish their funded goal target. If they fail to reach the target, then there is no investment activity take place. Regarding the crowdfunding platform rules, in this paper, we proposed models to predict and map the project successfulness to get funded.

There have been several references about crowdfunding platform and Kickstarter. Aslam [2] investigate the pattern of the relationship between startup and its investors based on network topology context using the small world characteristics. This study explores deeper about startup – investor

relationship on the model construction context using data mining framework. Greendberg et al [3] construct a classification models to predict project successfulness using static project data such as goal, category, pledged, video available or not, and some others. Etter et al [4] also built prediction model on Kickstarter with the model performance is 85% of correct predication, and propose Support Vector Machine model adopted on social attributes such as tweets with accuracy of 72%. Stam [5] also doing study based on Etter's work using static data project and social attributes on Twitter. He extends the study by adding the duration for a project to achieve their funding goal. This study adopts some of static attributes project data that have been studied before, with addition of comments and updates data attributes.

Based on the references above, we construct 3 models: First, the classification model using K-Nearest Neighbor (K-NN) to predict whether the project will be successful or failed. Second, the prediction model using Artificial Neural Network (ANN) to predict the number of successful project on each project category. Third, the model to map investor trend on investing the project using clustering methods based on K-means algorithm. The choice of each data mining model depends on data characteristics and the questions that investors and startups most asked. Through interviews, we investigate several scenarios, and narrowing down the results into three models above.

2. Theoretical Background

2.1. Crowdfunding

Crowdfunding is an initiative undertaken to raise money for new project proposed by someone by collecting small to medium-size investment from several other people. In the process of asking the public for donation that provide startup capital for new ventures, with this entrepreneurs and small business owners can bypass venture capitalist and angel investor entirely and instead pitch ideas straight to everyday internet users, who provided financial backing [6]. For those who invest their money will get rewards as appreciation for supporting their project.

2.2. Data Mining

Data mining is a scientific method in Knowledge Discovery in Database (KDD) process consisting of the application of data analysis and discovery of algorithm that, under acceptable computational efficiency limitation, to produce unknown data pattern [7].

2.3. Classification

Classification is the processing of finding set of models or function which describe and distinguish data classes or concepts [8]. For example, in the intrusion detection application, the classification algorithm gathers sufficient "normal" and "abnormal data, then produce classifier that can learn from those data classifier to label new unseen audit data whether belong to the "normal" or "abnormal" class [9]. One of classification methods based on non-parametric is KNN.

2.4. Prediction

Prediction is a model to predict the future output from data input based on certain criteria or formula [10]. One of the best blackbox algorithm that can be used is ANN, which imitate human brain system to simulates learning proses [11]. ANN prediction is useful when facing highly dimensional non-linear system.

2.5. Clustering

Clustering is model to group the objects of a dataset into meaningful cluster or subclass [12]. There are several methods to measure separation between data, some of them based on their distance, hierarchical, and data density. K-Means algorithm is clustering methods based on distance in vector space.

3. Methodology

3.1 Theoretical Framework

The research idea is based on the data exploratory analysis, thus construct models based on the data patterns. However, we also validate our findings with the most asked questions about business aspect by interviews the startups and investors. We come to conclusion that the previous 3 models we proposed answer the real-world problem. We explore crowdfunding data project, construct models and validate it to show models accuracy. Our research framework can be seen in Fig 1.

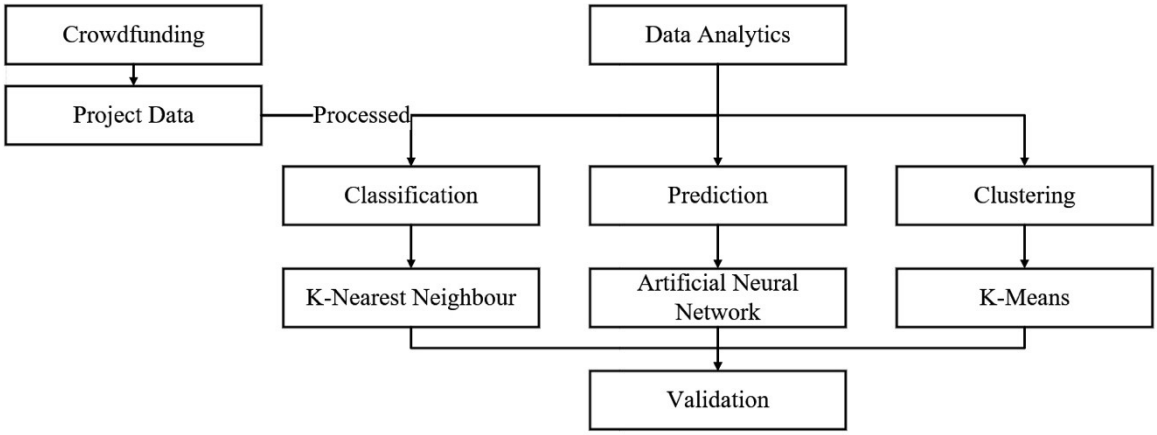

Figure 1 Research Framework

3.2 Dataset

We collect Kickstarter data through their open Application Programming Interface (API). The duration of data collection is from January 2016 to February 2017 which contains 45498 data projects. Table 1 shows dataset profile of Kickstarter.

Table 1 Kickstarter profile dataset

Attribute	Description	Type
Backers	Count backers that have backing on project	Features
Category	Category that project belongs to	Features
Comments	Total comment on project page	Features
Duration	Length of project campaign in day	Features
Goal	Goal of project in dollar	Features
Pledged	Total money of gathered in dollar	Features
State	Project state in the end of deadline	Target
Updates	Total updates on project page by creator	Features

3.3 Classification Process

The objective is to classify whether the project belong to success or fail class to get investment. We use state attribute as prediction target and the rest of the attributes as classifier set. We apply K-nearest neighbor (KNN) algorithm to classify which class a project belongs to, base on their distance closeness. We assign k=1 to define the project class of its nearest neighbor either “successful” or “failed” class. We validate the model based on 10 folds cross validation statistical performance.

3.4 Prediction Process

The objective is to construct the prediction model of number successful project per category. By using ANN algorithm, we collect sample from a multivariate value series dataset by windowing the input data per week. The windowing process is to transforms a given sample set containing series data into new example set containing single valued example. We run the methods on each project category to predict the number of successful project in each category. Dolezel et al [13] use same kind of dataset where X(k-1) and Y(k-1) as input and the output is Y(k). We train and test the data to estimate the performance of prediction model through validation process. ANN algorithm construct model by a feed-forward neural network trained by back propagation algorithm. Trial and error procedure applied to find the least error configuration by 1000 iterations. Model error rate measured using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Table 2, show the attributes that we use for prediction process.

Table 2 Dataset profile in time-series

Attribute	Description	Type
Historical successful project	Number of successful project per week per category	Features
Successful project	Number of successful project per category	Prediction

3.5 Clustering Process

The Objective is to map investor trend in investing on project. We construct clustering methods of pledged data that show total of money of investor to fund the project, and the backers data to show number of investor who fund the project. K-means algorithm is used to determine the similarity by means of the distance of objects data to each other, and assign the which cluster an object belongs to.

4. Result and Analysis

4.1. Classification Model

Figure. 2. shows the result of KNN classification model to classify whether a project belongs to "successful" or "failed" class. To classify those result, use the following attributes: pledged, backers, duration, goal, category, comments, and updates for social aspect. It also shows the result prediction of the actual data, in bottom left with blue dot shows prediction and actual project is failed. In bottom right, it shows prediction project is failed and the actual project is successful. In upper left, it shows prediction project is successful and the actual project data is failed. In upper right, it shows prediction and actual project is successful. The model accuracy can be seen in Table 3. It shows that prediction model has a high accuracy, class recall and class precision. It indicates that model has successfully predict most of the project entity.

By using this model for real world application, we have the information what category a project belongs to. To avoid a failed project, model characteristics inform us what aspect or data attributes needs to be improved. In real-world terms, we can call this model to predict a project likely to be successful or failed, but actually this model is classification problem.

Table 3 Classification Model Accuracy

Accuracy 99.04%	True Failed	True Successful	Class Precision
Pred. Failed	28389	214	99.25%
Pred. Successful	221	16673	98.69%
Class Recall	99.23%	98.73%	

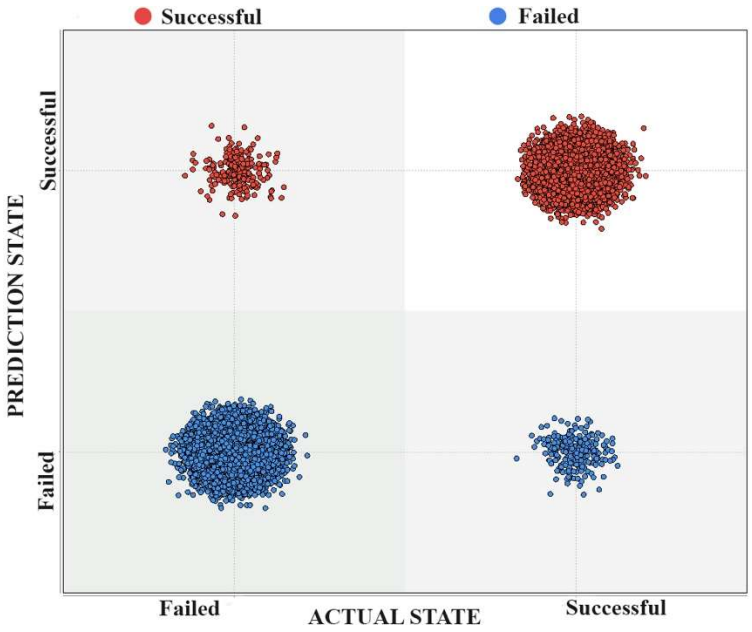

Figure 2 Classification Model Visualisation

4.2. Prediction Model

ANN algorithm predicts the number successful project per category. ANN contains Input Layer, Hidden Layer, and Output Layer of connected artificial neurons. The successful project data summarized based on 16 project categories. The output layer is one of the 16 project categories. We do trial and error to find the lowest error in hidden layer configuration. Table 4 shows the 4 lowest error. The result show that the lowest error and highest accuracy is in is Art category.

Table 4 Layer Configuration Error Test Result

Layer	RMSE	MAE
16-4-1	4.001	1.941
16-8-1	4.559	2.090
16-10-1	5.749	3.481
16-14-1	3.243	2.574

In Table 4 show that 16-14-1 configuration have the lowest error rate, we focus on building neural network based on this result. To measure the accuracy, we use three different value of Learning Rate 0.1, 0.2 and 0.3 while more than those values shows higher error. Table 5 shows the accuracy result of each learning rate. The 0.2% learning rate has the highest accuracy with 83.3%. It means predicting the number of successful project per category using ANN has high prediction rate. The ANN configuration of 16-14-1, can be seen in Figure 3.

Table 5 Learning Rate Test Result

Learning Rate	Accuracy
0.1	66.7%
0.2	83.3%
0.3	75%

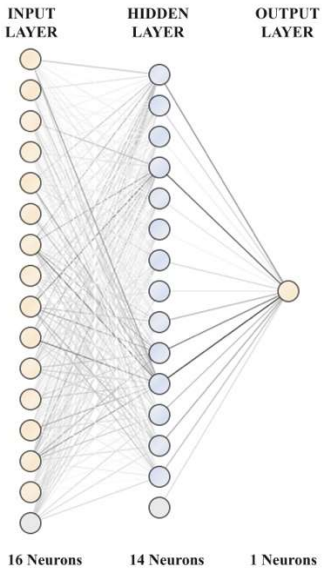

Figure 3 Neural Network Model

4.3. Clustering Model

To support investment activities exploratory analysis, we map number of investors and total investment values per project in Euclidean space. K-means algorithm is applied to the space. Based on elbow methods, number of clusters k=6 is the optimal configuration, thus we obtain 6 group / clusters. In Figure 4, it shows that cluster 1, cluster 3, cluster 4 and cluster 6 are normal scenario, where number of money invested and number of investor in project are both low. Cluster 5 has unique result, it shows total of invested money on projects is huge, while the number of investors are fairly low, which mean there are some investors that have high expectation on those projects. Cluster 2 shows different story, there are relatively small amount of total investment even though it attracts many investors.

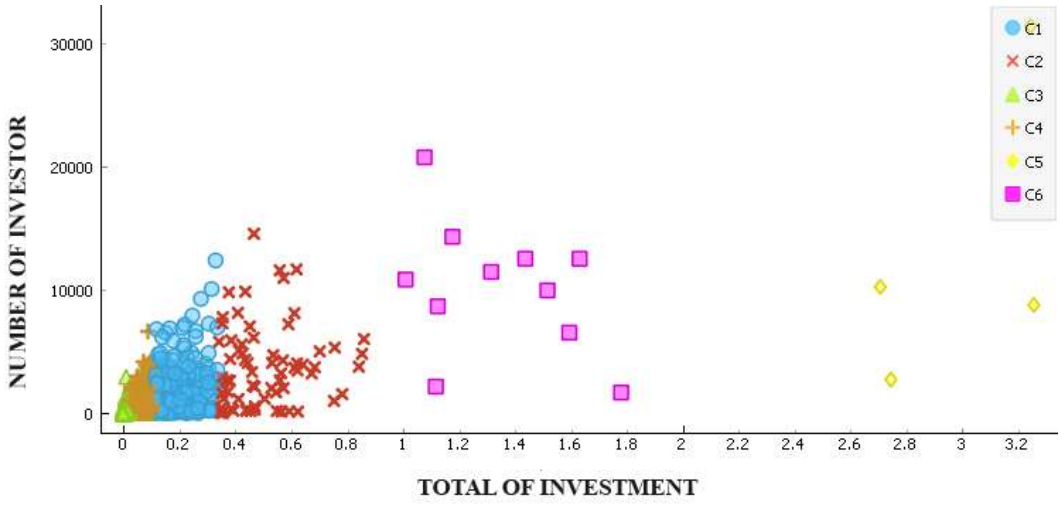

Figure 4 Clustering Model Visualization

Conclusion

In general, there are many possibilities of model construction to solve case scenario based on data pattern. It is not always guarantee that we can construct the model, since the dataset itself is not always give meaningful pattern. If we can find several patterns, thus we can construct many model, where each model answer different question. In this paper, our models answer most asked question real-world problem face by startups and investors.

This study presents how a dataset answer different questions or scenarios. First is about predicting project successfulness, where from classification model, we predict whether a project belong successful or failed class. The classification model has high accuracy 99.04%. From the prediction model, we predict the number of successful project on each project category. The prediction model has quite high accuracy learning rate of 83.3%. At last, the clustering model help us to map investor trend in investing their money. There are some projects that has fairly the same number of investor but have very high total of money invested. It shows some investor has high expectation in those projects.

Overall, the models can help startup and investor to understand comprehensively about their business nature. We found that Kickstarter represents crowdfunding platform business model, thus the model can closely represent real-world startup-investor relationship.

References

[1] Kickstarter, "Kickstarter About," [Online]. Available: https://www.kickstarter.com/about?ref=nav. [Accessed 30 September 2017].

[2] F. N. Aslam and A. Alamsyah, "The Small World Phenomenon and Network Analysis of ICT Startup Investment in Indonesia and Singapore," *The 7th Smart Collaboration for Business in Technology and Information Industry,* 2016.

[3] G. Michael D, P. Bryan, K. Hariharan and E. Gerber, "Crowdfunding support tools: predicting success & failure," *CHI'13 Extended Abstracs on Human Factors in Computing Systems,* pp. 1815-1820, 2013.

[4] V. Etter, M. Grossglauser and P. Thiran, "Launch Hard or Go Home! Predicting the Success of Kickstarter Campaigns," *Proceedings of the first ACM Conference on Online Social Networks (COSN'13),* pp. 177-182, 2013.

[5] M. Stam, *Crowdfunding Success Prediction: From Classification to Survival Regression and Back,* Amsterdam: University of Amsterdam, 2016.

[6] S. Steinberg, The Crowdfunding Bible: How to Raise Money for Any Startup, Video, Game, or Project, San Francisco: READ.ME, 2012.

[7] U. Fayyad, P.-S. G. and S. P, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," *Proc. 2nd Int. Conference on Knowledge Discovery and Data Mining,* pp. 82-88, 1996.

[8] H. J. Parashar, S. Vijendra and N. Vasudeva, "An Efficient Classification Approach for Data Mining," *International Journal of Machine Learning and Computing,* vol. 2, no. 4, pp. 446-448, 2012.

[9] W. Lee, S. J. Stolfo and K. W. Mok, "A Data Mining Framework for Building Intrusion Detection Model," *Proceedings of the 1999 IEEE Symposium on Security and Privacy ,* pp. 120-132, 1999.

[10] D. Olshon and S. Yong, Intoduction to Business Data Mining, United States: McGraw-Hill Irwin, 2007.

[11] K. M, Data Mining: concept, models, methods, and algorthms, New Jersey: John Wiley, 2011.

[12] T. Sajana, C. M. Rani and K. V. Narayana, "A Survey on Clustering Techniques for Big Data Mining," *Indian Journal of Science and Technology,* vol. 9, no. 3, 2016.

[13] P. Dolezel, P. Skrabanek and L. Gago, "Weight Initialization Possibilities for Feedforward Neural Network with Linear Saturated Activation Functions," *AFAC-PaperOnline,* vol. 49, no. 25, pp. 049-054, 2016.