Assessing the performance of classifying Algorithms by predicting success outcome of crowd-funding campaign

Msc Data Analytics

Ravikeerthi Teliegi
Student ID: 10503966

School of Computing
Dublin Business School

Supervisor: Basel Magableh

Assessing the performance of classifying Algorithms by predicting success outcome of crowd-funding campaign

Ravikeerthi Teliegi
10503966

Contents

1 Introduction **2**

2 Related Work **3**
2.1 Insights on working of Crowdfunding campaigns 3
2.2 Attributes leading to a successful projects 4
2.3 Data Mining methodologies in crowdfunding campaigns 5
2.4 Data mining methodologies 6

3 Methodology **6**
3.1 Business understanding 7
3.2 Data Understanding 7
3.3 Data cleaning and preparation 7

4 Implementation **9**
4.1 Attributes selected for analysis 9
4.2 SVM 10
4.3 LightGBM vs XGBoost 10
4.4 Random Forest 13
4.5 Precision and Recall 14

5 Evaluation **14**
5.1 SVM 15
5.2 Random Forest 15
5.3 Training time 15
5.4 Backers vs categories 15
5.5 Discussion 16

6 Conclusion and Future Work **16**

Abstract

Crowdfunding has enabled many creators, founders and entrepreneurs to directly appeal to the public in order realise their funding goals. Such websites remove a lot of barriers for those who don't typically need a large amount of funding. There are a lot of crowdfunding platforms like Kickstarter, Indiegogo, etc., which are used by many to run campaigns in order to reach their funding target. There are many

such campaigns running on a lot of these websites. Not all of these campaigns meet their funding goals. It is not always clear why a campaign succeeds or fails on these platforms. In this paper we look at data for campaigns on Kickstarter. Knowing what factors influence whether a campaign succeeds or fails would be an interesting question. In this study, the author looks data for campaigns which have already been completed on kickstarter. The objective of this study is to determine which factors play a major part in the success of a campaign. Five types of classifiers are used to predict campaign success viz., the Support vector machine (SVM) classifier, LightGBM algorithm, Random Forest and XGBoost . Additionally we test the effect that the number of backers have on the success of a campaign. The results of such a study could be very useful for potential creators and entrepreneurs when thinking of launching a campaign. The study also offers an insight to the reasons why certain campaigns succeed and others fail.

1 Introduction

Businesses are often classified on the basis of the methods by which they raise capital. Thus we have sole proprietorships, partnerships and joint stock companies as some of the major forms of business organisation. Entrepreneurs who are just beginning often find it difficult to obtain funding for their ventures. Where the funding requirement is small, the entrepreneur may be able to fund their venture themselves. But this is not often the case. There are many situations where people have put their entrepreneur aspirations on hold because of lack of funds. In these cases they may look elsewhere for their funding requirements by entering into a partnership with someone or trying to obtain venture capital form an investor. There is an alternate way called crowdfunding. Crowdfunding has existed in various forms for a long time. The basic idea here is to split the funding goal into small amounts which is then fulfilled by a large number of people. Thus a large amount of capital can be raised. In today's era many would be entrepreneurs have taken this route. With the rise of the internet it has become much easier to obtain funding this way. Inventors, founders and entrepreneurs to directly appeal to the public in order realise their funding goals. This has been made possible through the use of websites like KickStarter, Indiegogo etc. Such websites remove a lot of barriers for those who don't typically need a very large amount of funding. They are used by many to run campaigns in order to reach their funding target. There are many such campaigns running seeking funding for various projects. This is an interesting concept which has been made easier through the use of technology.

Not every such campaign reaches its funding goals. There may be many reasons for a such a campaign to fail. Maybe the product is not interesting enough. There may be a lack of awareness about a campaign which would have fared better with a little promotion. It is not always clear why a campaign succeeds or fails on these platforms. Knowing reliably the reasons behind success or failure of a campaign would be very useful.

Kickstarter is crowdfunding website that was launched in 2009. People can launch campaigns for their creative product ideas. These campaigns are aimed towards obtaining funding from individual backers for these ideas. These backers are general users who might be interested in funding the idea put forth in the campaign. Each backer pledges some money towards the campaign. In return for this backers may receive rewards. These rewards differ from campaign to campaign. The reward could some form of acknowledgement or involvement in developing the product. The campaigner has a particular funding

target which has to be reached. There also is a deadline for the campaign to end. The amount pledged collectively by all the backers needs to meet the target set by the campaigner within that deadline. The money that is pledged by the backers is only available to the campaigner if the funding target has been met before the deadline. If the target is not met, no money changes hands. Of all kickstarter projects ever launched, only 36% have been successful (*Kickstarter: Project Funding Success Rate 2018 — Statista*; 2018). Successful projects have obtained over 3.6 billion U.S. Dollars in funding till date. Being able to know the likelihood of success of a Kickstarter campaign is a very interesting problem.
Campaigners can benefit from knowing the factors which have an influence on the success or failure of a campaign. Knowing this helps the campaigner prepare and take measures to improve the chances of his project succeeding. The campaigners could launch online promotions, improve the product offering, make modifications to the campaign etc. The backers can also benefit from this information. Backers could be more judicious when deciding to back a certain campaign. Knowing the probability of success of a campaign also makes it easier to mobilise support for a campaign.
The purpose of this study is to explore a possible model to determine the success of a project. The other main focus is to build a classification model to determine the success of a campaign. For this we propose to compare the performance of three alternative classifiers. Firstly the SVM classifier is fitted to the model. A XGBoost is then used.Thirdly a random forestis used. The forth and final model used is are LightGBM classifier . LightGBM is a relatively new type of boosting algorithm which is an alternative to the more popular XGBoost algorithm. It much less resource intensive and more efficient. Another aspect of this research is to check the influence that the number of backers have on the success of a kickstarter campaign. In this study we look at data for campaigns which have already been completed on Kickstarter. The objective of this study is to determine whether the number of backers have a significant influence in the success of a campaign. The results of such a study could be very useful for potential creators and entrepreneurs when thinking of launching a campaign.

2 Related Work

2.1 Insights on working of Crowdfunding campaigns

Kickstarter is not the not the only crowd funding website. There are many others, with two of the next most popular being Indiegogo.com and Gofundme.com. The terms differ for all the sites. But the idea of collecting funds from a large number of people remains the same.There has been a lot of research on crowdfunding. There are several studies which specifically look at the underlying factor responsible for the success of a campaign. Lins, Fietkiewicz and Lutz (Lins et al.; 2016) looked at how successful a campaigner was at convincing the crowd to fund his idea. In a sample of 264 campaigns, they looked at linguistic behaviours affecting the likelihood of raising funds. They discovered success is related to the use of positive language patterns and promoting innovation. Wang, Zhu and others (Wang et al.; 2017a) looked at campaign descriptions and their effect on campaign success.
Kickstarter is a reward based crowdfunding platform. As a consequence, the success of a campaign is associated strongly with the rewards offered by the campaigner. Lin, Lee

and Chang (Lin et al.; 2016), analyses approximately 3k projects and 30k rewards. Using various statistical techniques, they found that projects with more rewards, with limited offerings and late-added rewards are more likely to succeed. (Greenberg et al.; 2013) has predicted the relation between ratings of the crowdfunded products. Here the author has also compared the traditional product and crowdfunded product on e-commerce website amazon and has analyzed the characters of the products sold.
Many projects which have collected more than the required funding for example Pebble and e- paper watch had put up a sum of $100k dollars, surprisingly more than fifteen thousand people pledged an amount of 2.6 million in a span of three days.
There is also a study between the fan base of jolla and Nokia when released in 2014.Nokia and jolla used different strategies i.e. Jolla (Jussila et al.; 2016) used the crowdfunding campaign to be successful and then to launch their product. Whereas nokia on the other hand released the product and is stuck to its old practices of releasing the product first. (Wang, Li, Liang, Ye and Ge; 2018) has conducted research over 1000 projects to find the relation between bakers and creators.(Greenberg et al.; 2013) thinks there could be hidden factors i.e past experiences, location and network that could increase the accuracy of data mining algorithms.
(Ahmad et al.; 2017) feels more than half the projects are rejected because of insufficient funds gathered. A data-set containing more than 26 thousand projects were studied to understand the factors that affect the of getting a successful funding. (Butticè et al.; 2018) has discussed about green funding and how its encouraged in different countries and used the data from Kickstarter from 2009 to 2012. The user interaction has attracted many user centered innovation and business ideas. (Lee and Sohn; 2019) has proposed a framework to discover business ideas and their user centre attraction using topic modelling on kickstarter websites.

2.2 Attributes leading to a successful projects

Authors (Jussila et al.; 2016) explains the effect of fan base and their interaction on the environment of kickstarter funding. The authors (Wang et al.; 2017a) employed text mining techniques in order to study the sentiment factor. They discovered that positive sentiment in the description contributed to the campaign's success, on the other hand, such sentiment should not be part of the campaign title. By suggestion of (Etter et al.; 2013) category, duration,images and videos are the key factors of an successful project. (Wang, Li, Liang, Ye and Ge; 2018) confirmed that comments and replies are positively related to the crowd funding success. (Sharma and Lee; 2018) explains how the creators of unsuccessful projects made more money than the successful projects , which shows rising more money doesn't make a project successful.
The authors(Greenberg et al.; 2013) explains the importance of number of backers. Positive sentiment in the description contributes to a better accuracy of predicting success (Wang et al.; 2017b) also the number of active backers and positive sentiment stimulates the users psychologically.(Kromidha; 2015) discusses about the big five personality traits and their correlation in kickstarter campaign success they are namely 1)openness which describes expression like intellect and liberalism.2)Conscientiousness which describes orderliness, self-discipline, self efficiency.3)Extra-version which describes energy friendliness and cheerfulness. 4)Agreeableness which describes modesty and morality. 5) Neurotic-ism which describes anger and anxiety.
The characters of the crowdfunded project on e commerce websites can be compared

used by the backers who support the crowd funding campaigns(Sharma and Lee; 2018) . Kickstarter users asked more questions on the products than the other products on the e-commerce websites.
The author has discussed about the different types of crowding platforms in different countries (Gera and Kaur; 2018) Kickstarter was developed in USA, crowdfunder.co.uk and Croedo was developed in singapore.The author (Gera and Kaur; 2018) has also seen that in usa there is a higher entrepreneur skills compared to europe and asia in terms of crowdfunding platform. (Sharma and Lee; 2018) finds that the creators of unsuccessful projects are like to have less number of backers.
(Ahmad et al.; 2017) while assigned weights to individual trees and found 13 different predictors that were more suitable for predicting accuracy. He also found that these predictors will serve both the creators and the backers the author (Butticè et al.; 2018) differs from other campaigns and only diffused in few countries which have are not environmental sustained. (Lee and Sohn; 2019) has discussed many businesses such as tutorials for math and seeking job opportunities in an emerging platform for campaign success on kickstarter. Also (Lee and Sohn; 2019) has explained the ideas preferred by the us were very different from other countries.

2.3 Data Mining methodologies in crowdfunding campaigns

Social set theory along with association rule was used to display the backers who were involved before and after the campaign(Jussila et al.; 2016). (Lin et al.; 2016)Using random forests they were able to obtain a prediction accuracy of over 85%. (Etter et al.; 2013) proposes a model using text mining on the campaign's
Description and the social features, SVM was used to predict the text mining and had an accuracy of 76 percent. (Wang, Li, Liang, Ye and Ge; 2018) has used BosonNLP to find the length, comment quality. Logistic regression was used to check the multicollinearity. The sentiment analysis resulted more than 95 percent accuracy for positive and negative comments.
The author performs a (Wang et al.; 2017b) sentiment analysis to see the personality, level of education, status in the society on the description of the campaign. Lexical based analysis,machine learning techniques including random forest and svm were used with 10 fold validation. Here the (Wang et al.; 2017b) uses SVM and POS and displayed 84 percent of precision and 92 percent of recall while performing sentiment analysis .(Sharma and Lee; 2018) has performed Chi-square statistical test on the product ratings on the e-commerce website.
(Greenberg et al.; 2013) has performed various algorithms to predict the success rate of crowdfunding campaigns. The author has proposed various models i.e ADA booster classifier, random forest, svm and other tree based models and performed using Wika and achieved an accuracy in predicting campaign success. (Gera and Kaur; 2018) have done web scraping is done from crowdfunding platforms to extract the data also the author has compared the live projects of the year 2017 to perform results.
The author has used random forest as a learning model to classify the Kickstarter program (Ahmad et al.; 2017). The authors have assigned optimal weights to the individual classifiers in the random forest. The learnt algorithms were tested against the new datasets and found to have a more than 90 percent accuracy with more than 90 percent of precision and recall. The (Ahmad et al.; 2017) has also calculated the accuracy of campaign success on 15 categories and 12 features and has found an average of more

than 90 percent. Also, the author has performed a state of art on many machine learning algorithms including MLP classifier, Random forest, Bayes Net, LWL Decisions ump, Naive Bayes and logistic regression and confirmed that random forest performed better. The authors (Butticè et al.; 2018) has used two hypothesis they are about country's environmental sustainability and the diffusion of the green campaigns. (Butticè et al.; 2018) have used random forest and presented that about 10 percent of the campaigns are green. Author (Lee and Sohn; 2019) has used conjoint analysis to analyse the most preferred topic In terms of funding they received. To achieve this the author has used Latent Dirichlet Allocation for topic modelling with 10-fold cross validation to discover innovative topics hidden in the project sets by focusing on perplexity.

2.4 Data mining methodologies

Internet finance P2P funding, which is caused a great loss for investors in china as the borrowers have a more transparent way of transaction were both sides can reap(Ma et al.; 2018) benefits in P2P funding..In another (Wang, Wang, Qin and Xia; 2018)work where data mining models where LightGBM and XG Boost was find the relation between the trajectory of GPS devices and the mode of transportation in major countries like USA and China. Both(Wang, Wang, Qin and Xia; 2018) (Ma et al.; 2018) have used machine boosting methods i.e LIGHTGBM and XG BOOST. Author (Eskandarpour and Khodaei; 2018) has used SVM to increase the margin to find the accuracy of grid operations. SVM was used for feature selection to eliminate those features which were irrelevant for prediction also the author explains how good SVM is at handling class imbalance(Maldonado and López; 2018). (Wang et al.; 2017a) Has used Random forest, XGBoost and LightGBM on miRNA molecules in classifying breast cancers. The author has used 10 fold cross validation on the data set and found that LightGBM performed best with an accuracy of 99 percent which was 2 percent higher than Random Forest and XGBoost.
Author(Greenberg et al.; 2013)has used random forest which is a boosting algorithm has created a branch mark for predicting campaign success. Also according to(Ma et al.; 2018) (Wang, Wang, Qin and Xia; 2018) (Maldonado and López; 2018) multi-observation model performed better than multidimensional method. As our project is dealing with massive dimension and class imbalance, as our project deals with the campaign success. so in this project we have considered SVM and LightGBM to perform our analysis on kickstarter campaigns.

3 Methodology

A research work is best to built with a road-map.For our analysis, data mining is divided Into different phases i.e. plan of action at each stage. For our research the author has implemented CRISP-DM approach. It's a hierarchical model which process with small phases(Wirth; 2000). The phases of CRISP DM are noted down they are:

1. Business Understanding
2. Data Understanding
3. Data Preparation

4. Evaluation

5. Deployment

3.1 Business understanding

Business understanding is the first step of CRISP -DM approach. The first step and the foremost step is to choose and understand the motive of the research. The work is motivated by the issues faced by the entrepreneurs in the crowd funding society i.e. Kick starter. our business objective is to help entrepreneurs reach their campaign success. our another motive behind this work is to find out how backers are a necessary part of campaigns and also help business users find the best categories to invest on. This research will help entrepreneurs take better decisions on categories and the backers to invest on the right campaign. Business understanding is considered as one of the most important phases of data mining. Business understanding is phase where the problem faced in data mining from a business perspective is observed and further a plan to achieve its objectives is made. Under the business understanding phase there are certain vital steps that are required to be taken. They include decisive business objectives, understanding the critical situations, looking in to the goals of data mining goals and finally formulating a decisive project plan

3.2 Data Understanding

At this phase raw data is collected and an analyst is provided this data who has to understand and find the problems with this data and also see what is useful for the business and what is not required. The analyst should also look for data that is otherwise hidden while analysing this data. This phase has four vital steps ie collection of raw data, creating a brief summary of the data, discovering the data and cross referring the validity of the data.

This research is based on data from web robots.io. Webrobots.io has been collecting data on kickstarter campaigns since 2014. This data is publicly available on their website. The data is collected by the webrobot.io scraper which navigates through kickstarter campaign pages on a monthly basis. The results are available in the JSON and CSV formats. This scraper collects a variety of publicly available information from campaign pages on kickstarter. This is a very comprehensive dataset consisting of 30 variables. The variables collected are a mix of numeric and text columns. Variables measuring the campaign blurb, the location, url of the creator are text. There are a total of 205091 campaigns in the dataset. This dataset was obtained by concatenating multiple individual monthly datasets generated at webrobots.io.

3.3 Data cleaning and preparation

Data preperation is the phase where the data collected is arranged according the business needs and prepared making it ready for further use to be put into the tool which is being used for modelling. The available data was cleaned using python programming language. Initially the columns having categorical textual value i.e. currency type, country,category and similar rows were given a unique number based on the categorical text.

Sl.no	Attribute
1	Backers_count
2	Blurb
3	Category
4	Converted_pl amnt
5	Country
6	Created_at
7	Creator
8	Currency
9	Currency_symbol
10	Currency_trail
11	Current_currency
12	Deadline
13	Disable_commun
14	Friends
15	Fr_rate
16	Goal
17	Id
18	Is_backing
19	Is_starrable
20	Is_starred
21	Launched_at
22	Location
22	Name
23	Permissions
24	Photo
25	Pledged
26	Profile
21	Slug
22	Source_url
23	Spotlight
24	Staff_pick
25	State
26	State_changed_at
27	Static_usd_rate
28	Urls
29	Usd_pledged
30	Usd_type

Table 1: Attributes

Attributes which were dichotomous i.e yes or no was encoded with one and zeros. The created date and the end date was converted from date format and subtracted to find the number of the days the project was live on kickstarter platform. The Goal and pledged amount were converted to numbers from exponential format. The attribute category was cleaned by removing unwanted text and Finally the name attribute was cleaned and classified on the basis of length of the name, existence of exclamation mark, number of words and seen if the words were written in upper case.
Finally the data was split with with a ration 0f 75 percent test data and 25 percent train data.

4 Implementation

Implementation is the fourth stage of our crisp DM approach.All the methods applied are in this phase and is executed . Before we apply models to our data set we select our attributes.

4.1 Attributes selected for analysis

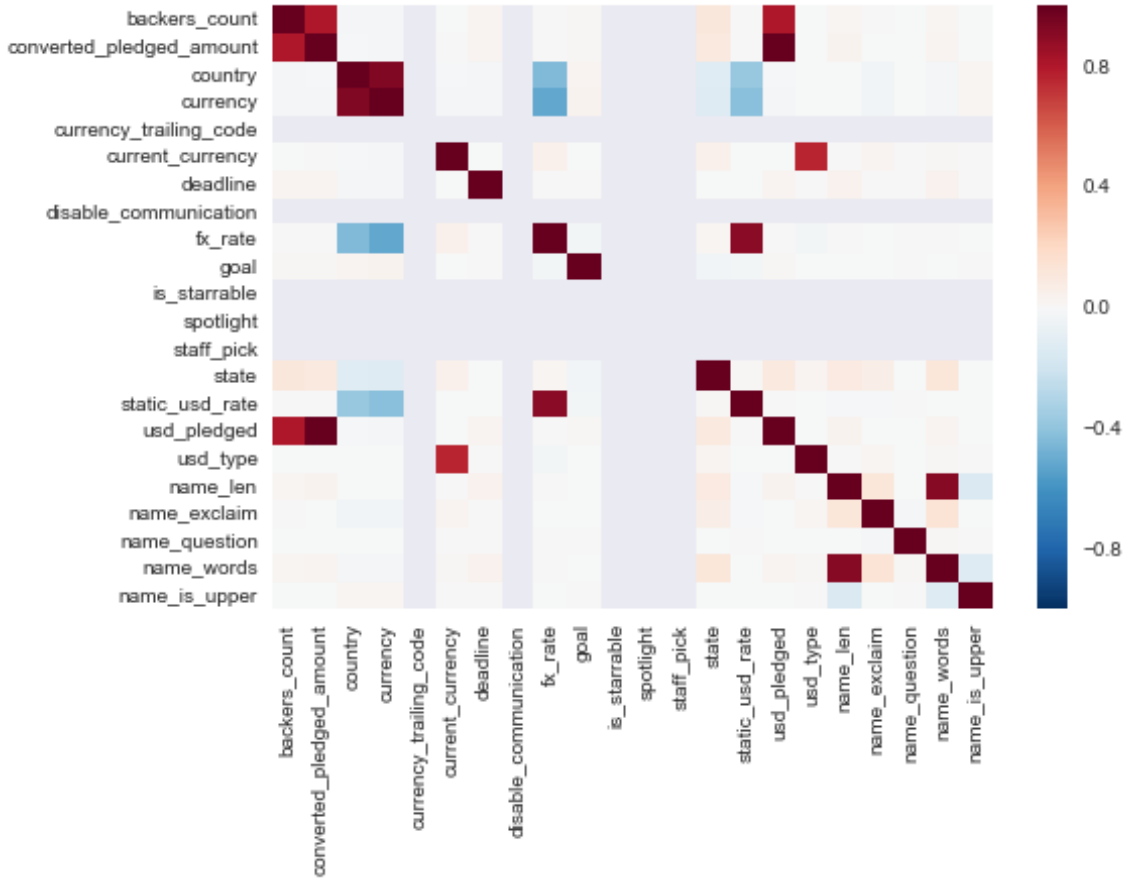

Figure 1: Heat Map

A correlation matrix (heat map) is constructed to visualize the necessary factors for our analysis. when analysis the co-relation the pledged amount and backers were high. Also pledged amount,USD pledged, converted pledged has a high co-relation so only the

usd pledged was kept for our analysis. Also the sate of the project i.e weather the project will be successful or is got co-relation with name of the project, category backers and USD pledged.

4.2 SVM

In order to build a classification model for kickstarter projects three different methods were considered. The first method that was considered was the use of a Support Vector Machine(SVM) classifier. The SVM classifier works well as a classifier by mapping points to higher dimensional feature spaces. Through this mapping separations which may not have been possible in two dimensions maybe become apparent in higher dimensions. By using hyperplanes to classify data points higher dimension space this classifier also works well with non linear data.

Different Kernel Functions
Linear
Nonlinear
Polynomial
Radial basis function (RBF)
sigmoid

We have used the RBF kernel.

Think of the Radial Basis Function kernel as a transform or processor to generate new features by measuring the distance between all other dots to a specific dot or dots centers. The most popular/basic RBF kernel is the Gaussian Radial Basis Function:Gamma () controls the influence of new features (x, center) on the decision boundary. The higher the gamma, the more influence of the features will have on the decision boundary, more wiggling the boundary will be.
But the RBF kernel has a number of other properties. it's important to also draw contrasts to other methods to develop context.
It is a stationary kernel, which means that it is invariant to translation. Suppose you are computing K(x,y).K(x,y). A stationary kernel will yield the same value K(x,y)K(x,y) for K(x+c,y+c)K(x+c,y+c), where cc may be vector-valued of dimension to match the inputs. For the RBF, this is accomplished by working on the difference of the two vectors. For contrast, note that the linear kernel does not have the stationarity property.
The single-parameter version of the RBF kernel has the property that it is isotropic, that is the scaling by yy occurs the same amount in all directions. The moral of the story is that kernel-based methods are very rich, and with a little bit of work, it's very practical to develop a kernel suited to your particular needs. But if one is using an RBF kernel as a default, a reasonable benchmark for comparison.

4.3 LightGBM vs XGBoost

The next method that was considered was the use of the LightGBM library to build a binary classification model. LightGBM is a relatively new algorithm in the world of gradient boosting algorithms. Adaboost and XgBoost are other popular boosting algorithms. These algorithms are collectively called ensemble techniques. These techniques work by building multiple weak learners and then using decision trees to choose the optimal classifier. LightGBM works by growing leaf-wise as opposed to the tree-wise growth employed

by the XGBoost algorithm. This offers advantages in accuracy due to the complexity of the trees produced. Overfitting may pose a problems due to the depth of the tree that is produced. This can be overcome by controlling the max-depth parameter. LightGBM offers many advantages over other boosting algorithms. It has a faster training speed and better efficiency as compared to alternatives.
LightGBM can be controlled and optimized by the following features:
Num_leaves: The number of leaves per tree
Learning_rate: The learning rate of algorithm.
Max_depth: maximum depth of the tree which is our case is kept default of 10
min_data: The number of data in the leaf to control the fitting.

The gradient boosting decision tree (GBDT) is outstanding amongst other performing classes of algorithms in AI applications. One usage of the gradient boosting decision tree xgboost is one of the most mainstream calculations in data science field.
The subtleties of both xgboost and lightGBM and what makes them so powerful. By understanding the underlying algorithms, it's obvious that with even a single parameter, which is simple but leads to a powerful hyperparameter tuning.
There is no strategy to the best split. Hence, the different strategies that xgboost and lightGBM present are techniques for finding the approximate best split. One thing that can be confounding is the contrast between xgboost and lightGBM . Xgboost and light-GBM are both subtypes or explicit occurrences of the GBDT calculation. In spite of the fact that the two of them execute based on basic
calculation, they each introduce various tricks to make training more efficient or to improve performance.
Background: Gradient Boosting Decision Trees So as to comprehend GBDTs, we have to comprehend the decision tree(growth of the trees). decision trees are a strategy for parting the data. Each branch in a decision tree partitions the data into one multiple trees based on the requirement.
Decision trees are adaptable and interpretable. Be that as it may, a solitary decision tree is inclined to overfitting.There are different methods for confining the adaptability of a decision tree, for example, by constraining its profundity. In this manner, decision trees are commonly not utilized alone: rather, various decision trees are utilized together. Slope boosting decision trees are one strategy (among many) of consolidating the forecasts of different decision trees to make expectations that sum up well.
The thought behind GBTDs is extremely straightforward: join the expectations of different decision trees by including them together. For example, on the off chance that we were attempting to foresee lodging costs, the anticipated cost for any data point would be the total of the forecasts of every individual decision tree.
GBDTs are prepared iteratively for example one tree at once. For example, a GBDT that endeavors to foresee lodging costs would initially prepare a basic, feeble decision tree on the data and crude lodging costs. The decision tree is capable of dealing with costs for example, the mean squared blunder that is by recursively parting the data in a manner that boosts some rule until some point of confinement for example, the profundity of the tree is met. The rule is picked with the goal that the misfortune capacity is limited by each split.
Xgboost and lightGBM are by all account not the only usage of GBDTs. The motivation behind why xgboost and lightGBM are dealt with like they are the agents of GBDTs is that they
(1) both have simple to-utilize open source usage

(2) are quick and precise.
(3) is especially significant; however the GBDT is actualized in sklearn, it is much more slow than xgboost and lightGBM.

2. Developing the Tree

Both xgboost and lightGBM utilize the leaf-wise development procedure when developing the decision tree. When preparing every individual decision tree and parting the data, there are two procedures that can be utilized: level-wise and leaf-wise.
The level-wise technique keeps up a reasonable tree, though the leaf-wise procedure parts the leaf that diminishes the misfortune the most.
An outline exhibiting the contrast between level-wise and leaf-wise development Level-wise preparing can be viewed as a type of regularized preparing since leaf-wise preparing can build any tree that level-wise preparing can, while the inverse does not hold. Accordingly, leaf-wise preparing is progressively inclined to overfitting yet is increasingly adaptable. This settles on it a superior decision for enormous datasets and is the main alternative accessible in lightGBM.
Already, leaf-wise development was a restrictive element of lightGBM, however xgboost has since actualized this development technique . This methodology is accessible for the histogram-based strategy (which I will clarify beneath), so as to utilize it, clients should set the tree method parameter to hist and develop approach parameter to lossguide.
3. Finding the Best Split
The key test in preparing a GBDT is the way toward finding the best split for each leaf. At the point when innocently done, this progression requires the algorithm to experience each component of each datum point. The computational multifaceted nature is therefore
.

Present day datasets will in general be both enormous in the quantity of tests and the quantity of highlights. For example, a tf-idf network of a million records with a jargon size of 1 million would have a trillion passages. Along these lines, a credulous GBDT would take always to prepare on such datasets. There is no strategy that can locate the best split while abstaining from experiencing all highlights of all data focuses. In this manner, the different strategies that xgboost and lightGBM present are techniques for finding the estimated best split.

3.1 Histogram-based techniques (xgboost and lightGBM)

The measure of time it takes to fabricate a tree is corresponding to the quantity of parts that must be assessed. Frequently, little changes in the split don't make a big deal about a distinction in the exhibition of the tree. Histogram-based strategies exploit this reality by gathering highlights into a lot of canisters and perform parting on the receptacles rather than the highlights. This is identical to subsampling the quantity of parts that the model assesses. Since the highlights can be binned before structure each tree, this technique can extraordinarily accelerate preparing, diminishing the computational intricacy to .

The contrast among xgboost and lightGBM is in the points of interest of the improvements. Beneath, we will experience the different manners by which xgboost and lightGBM enhance the essential thought of GBDTs to prepare exact models productively. In spite of the fact that adroitly straightforward, histogram-based strategies present a few decisions that the client must make. Right off the bat, the quantity of canisters makes an exchange off among speed and precision: the more receptacles there are, the more exact the algorithm is, however the more slow it is too. Besides, how to isolate the highlights into discrete canisters is a non-paltry issue: partitioning the containers into equivalent

interims (the least complex technique) can regularly bring about a lopsided assignment of data. Despite the fact that the subtleties are past the extent of this post, the "most adjusted" strategy for isolating the canisters really relies upon the slope insights. Xgboost offers the =approx, which figures another arrangement of canisters at each split utilizing the gradient statistics. LightGBM and xgboost with the tree technique set to hist will both figure the containers toward the start of preparing and reuse similar receptacles all through the whole preparing procedure.

3.2 Overlooking meager sources of info (xgboost and lightGBM)

Xgboost and lightGBM will in general be utilized on forbidden data or content data that has been vectorized. Accordingly, the contributions to xgboost and lightGBM will in general be meager. Since a large portion of the qualities will be 0, glancing through every one of the estimations of a scanty element is inefficient. Xgboost proposes to overlook the 0 highlights when processing the split, at that point designating every one of the data with missing qualities to whichever side of the split decreases the misfortune more. This lessens the quantity of tests that must be utilized when assessing each split, accelerating the preparation procedure.

Despite the fact that lightGBM does not empower overlooking zero qualities as a matter of course, it has a decision called zero as missing which, whenever set to Genuine, will view each of the zero qualities as absent. lightGBM will treat missing qualities similarly as xgboost as long as the parameter use missing is set to Genuine (which is the default conduct).

3.3 Subsampling the data:Gradient-based One-Side Sampling (lightGBM)

This is a technique that is utilized only in lightGBM. The basic perception behind this strategy is that not all data focuses contribute similarly to preparing; data indicates with little angles tend be all the more very much prepared (near a nearby minima). This implies it is progressively effective to focus on data focuses with larger gradients The most direct approach to utilize this perception is to just overlook data focuses with little slopes when processing the best split. Nonetheless, this has the danger of prompting one-sided testing, changing the appropriation of data. For example, if data that had a place with the "youthful" age gathering would in general be less all around prepared, the inspected data will have an a lot more youthful age circulation. This implies the split is probably going to be more youthful than the ideal worth.

So as to moderate this issue, lightGBM additionally arbitrarily tests from data with little inclinations. This outcomes in an example that is as yet one-sided towards data with enormous angles, so lightGBM expands the heaviness of the examples with little inclinations when figuring their commitment to the adjustment in misfortune (this is a type of significance examining, a strategy for effective testing from a discretionary appropriation).

4.4 Random Forest

Random forest is based on a small idea i.e wisdom in the crowd. The aggregation of various results of various predictors gives a wise and a better prediction compared to a single tree. Combining of different trees is called ensamble learning. The random forest chooses a different subsets of various features and therefore build as many of decisions trees as possible.The parameters can be changed to generalise the prediction. Tuning a model a risky and tedious task.There are many combinations between the parameters. One of the alternative to allow the machine to decide the best combo for the prediction.

	Precision	Recall	f1-score	Support
0	0.99	0.76	0.86	19818
1	0.85	0.99	0.92	26472
Total/avg	0.91	0.90	0.89	46290

Table 2: Light GBM with backers

They are:
Grid Search
Grid Search definition The model can be used for various combination possible for the functions using cross validation.

For instance, The model is tested with 10,20,30 no of trees and every tree can be rested over a various mtry to 1,2,3,4,5. There for there are total of 15 different models which are formed.
Random Search definition

The main difference of random search and grid search, so the search including random searches cannot evaluate the combinations of various hyper meters id the searching space. Ins-ted they choose combination at each and every iteration. The main advantage us that it lowers the computational cost. Set the control parameter

4.5 Precision and Recall

Precision and Recall are two of the most important measures used to measure the performance of classifiers. Precision measures the ratio of true positive cases to the total number of cases identified as positives. Recall on the other hand is the ratio of true positives to the number of all cases which are relevant. Both these measures are often at tug of war with each other. Increasing precision often leads to lower recall and vice versa. Another metric called the F1 score makes use of both the precision and recall in order to gauge classifier performance. The F1 score is the harmonic mean of the precision and recall. This value ranges between 0-1 with 1 indicating perfect precision and recall. A higher F1 score indicates better model performance.
In order to gauge the relative performance of these models precision, recall and the F1 score was considered for each model

5 Evaluation

The last phase of our CRISP DM approach is deployment, where all the models are built and implemented. They are also evaluated to find weather the results generated are relevant or not. Every model in this research was built based on the requirements of finding the best output and factors leading to accuracy.

The SVM classifier that was fit to the data yielded an average precision value of 0.85 and a recall of 0.84. The F1 score for the model was 0.85. In this case the precision and recall values are very close to each other. As a result the F1 score is also very similar.
The XGboost model performed much worse than SVM and LightGBM and was hence not considered in the final evaluation.

	Precision	Recall	f1-score	Support
0	0.77	0.90	0.83	19818
1	0.91	0.80	0.86	26471
Total/avg	0.85	0.84	0.85	46290

Table 3: SVM

	Precision	Recall	f1-score	Support
0	0.99	0.69	0.81	33601
1	0.81	0.99	0.89	43671
Total/avg	0.90	0.84	0.85	77272

Table 4: Random forest

The results clearly show that LightGBM is superior to the other classifiers and that the number of backers has a very strong influence on the success of a Kickstarter campaign.

5.1 SVM

The SVM classifier that was fit to the data yielded an average precision value of 0.85 and a recall of 0.84. The F1 score for the model was 0.85 as shown in table 3. In this case the precision and recall values are very close to each other. As a result the F1 score is also very similar.

5.2 Random Forest

The Random Forest classifier yielded a average precision value of 0.90 and recall of 0.84. This is also close to SVM and their F1 score is similar.

5.3 Training time

SVMs are much slower than mlp and LightGBM. The reason behind this is that is the quadratic optimization because of which number of variables are high. In our case the number variables are the number of training instances. LightGBM was primarily built for speed with many cores and high frequency. LightGBM outperformed both the algorithm in terms of training time.

5.4 Backers vs categories

From Figure 2.we can see There are more number of supporters for Video games , product Design and less number of supporters for Non fiction and Technology. From this we

	Precision	Recall	f1-score	Support
0	0.99	0.27	0.43	31758
1	0.65	0.99	0.79	42305
Total/avg	0.80	0.69	0.63	74063

Table 5: LightGBM without Backers

1.png 1.png

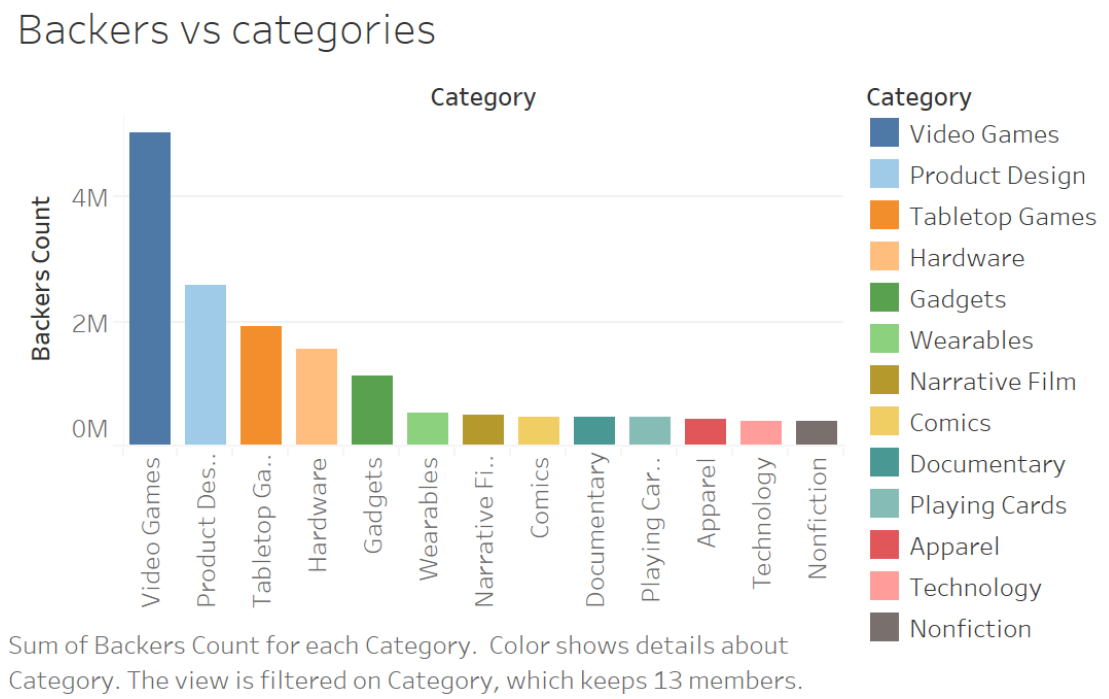

Figure 2: Backers Vs Categories

can infer that who are investing in Video game category start up projects will be more successful rather than Technology start ups.

5.5 Discussion

The LightGBM algorithm performed much better with a higher average precision and recall. This algorithm was run twice. The first time was by excluding the backers_count variable and the second time by including the same. There was a marked difference in the performance of the classifier with the presence of that one variable. from table 1 and 4 The precision and recall scores for the model without the inclusion of backers_count resulted in a precision of 0.81 and a recall of 0.71 . With the inclusion of that one variable the precision jumped up to 0.91 and the recall to 0.90. The F1 scores were also markedly different for both. The model without the backers_count variable had an F1 score of 0.66 whereas after inclusion the F1 score went up to 0.89. This indicates that backers_count is a very important feature influencing the success of a kickstarter campaign. The results of XG boost is not included as the the data overfit the algorithim due to high skewed data class imbalance.
Also we measured the accuracy of live projects and out of 6 thousand odd live projects we could anlyse and success of 2000 odd projects using our XG Boost algorithm.

6 Conclusion and Future Work

Based on the results obtained it can seen that LightGBM outperforms the other classification methods. This is because of the way it operates by growing the decision tree leafwise as opposed to tree wise. This leads to greater accuracy. This accuracy dropped

when excluding the backers as a variable in the data. It is very apparent that taking the number of backers into consideration has a great influence on the accuracy, precision and recall scores. LightGBM is clearly a superior classifier as compared to SVM and Random Forest. It is also quite fast and efficient as compared to other boosting techniques.

Kickstarter is not the only crowdfunding website around. There are other websites like Indiegogo and kiva which are alternatives. Indiegogo operates under slightly different rules where the amount pledged by backers is dispersed enough if the campaign fails to meet its funding goals. Kiva on the other hand aims to provide funding for people in poor and developing countries. While not exactly the same these websites operate on the same basic principles i.e., funding campaigners by outsourcing the financing to a large number of people. A comparative analysis of these three websites can yield a general sense of the factors which influence people to fund a project. This can help campaigners on these platforms better frame their campaigns and maximize their chances of success.

References

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

Butticè, V., Colombo, M. G., Fumagalli, E. and Orsenigo, C. (2018). Green oriented crowdfunding campaigns: Their characteristics and diffusion in different institutional settings, *Technological Forecasting and Social Change* .

Eskandarpour, R. and Khodaei, A. (2018). Leveraging Accuracy-Uncertainty Tradeoff in SVM to Achieve Highly Accurate Outage Predictions, *IEEE Transactions on Power Systems* **33**(1): 1139–1141.

Etter, V., Grossglauser, M. and Thiran, P. (2013). Launch Hard or Go Home!: Predicting the Success of Kickstarter Campaigns, *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, ACM, New York, NY, USA, pp. 177–182.

Gera, J. and Kaur, H. (2018). Influence of Personality Traits on Campaign Success, *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Noida, India, pp. 14–15.

Greenberg, M. D., Pardo, B., Hariharan, K. and Gerber, E. (2013). Crowdfunding Support Tools: Predicting Success & Failure, *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, ACM, New York, NY, USA, pp. 1815–1820.

Jussila, J., Menon, K., Mukkamala, R. R., Lasrado, L. A., Hussain, A., Vatrapu, R., Kärkkäinen, H. and Huhtamäki, J. (2016). Crowdfunding in the Development of Social Media Fanbase – Case Study of Two Competing Ecosystems, *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 4495–4504.

Kickstarter: Project Funding Success Rate 2018 — Statista (2018). https://www.statista.com/statistics/235405/kickstarter-project-funding-success-rate/.

Kromidha, E. (2015). A comparative analysis of online crowdfunding platforms in USA, Europe and Asia, *eChallenges E-2015 Conference*, pp. 1–6.

Lee, W. S. and Sohn, S. Y. (2019). Discovering emerging business ideas based on crowdfunded software projects, *Decision Support Systems* **116**: 102–113.

Lin, Y., Lee, W. and Chang, C. H. (2016). Analysis of rewards on reward-based crowdfunding platforms, *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 501–504.

Lins, E., Fietkiewicz, K. J. and Lutz, E. (2016). How to Convince the Crowd: An Impression Management Approach, *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 3505–3514.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, *Electronic Commerce Research and Applications* **31**: 24–39.

Maldonado, S. and López, J. (2018). Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification, *Applied Soft Computing* **67**: 94–105.

Sharma, V. and Lee, K. (2018). Predicting Highly Rated Crowdfunded Products, *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 357–362.

Wang, B., Wang, Y., Qin, K. and Xia, Q. (2018). Detecting Transportation Modes Based on LightGBM Classifier from GPS Trajectory Data, *2018 26th International Conference on Geoinformatics*, pp. 1–7.

Wang, N., Li, Q., Liang, H., Ye, T. and Ge, S. (2018). Understanding the importance of interaction between creators and backers in crowdfunding success, *Electronic Commerce Research and Applications* **27**: 106–117.

Wang, W., Zhu, K., Wang, H. and Wu, Y. J. (2017a). The Impact of Sentiment Orientations on Successful Crowdfunding Campaigns through Text Analytics, *IET Software* **11**(5): 229–238.

Wang, W., Zhu, K., Wang, H. and Wu, Y. J. (2017b). The Impact of Sentiment Orientations on Successful Crowdfunding Campaigns through Text Analytics, *IET Software* **11**(5): 229–238.

Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining, *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29–39.