Crowd Funding Based Business Project Classification Using Data Mining

Md Ashraf Uz Zaman Shahria

2014–2–60–001

Rulia Islam

2014–2–60–022

Sk. Sabit Faisal

2014–2–60–014

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering
East West University
Dhaka-1212, Bangladesh

January, 2019

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Dr. Shamim H. Ripon, Associate Professor, Department of Computer Science and engineering, East West University. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

....

Dr. Shamim H. Ripon

Supervisor

....

Md. Ashraf Uz Zaman Shahria

2014–2–60–001

....

Rulia Islam

2014–2–60–022

....

Sk. Sabit Faisal

2014–2–60–014

Letter of Acceptance

This project entitled "**Crowd Funding Based Business Project Classification Using Data Mining**" submitted by Md Ashraf Uz Zaman Shahria (2014–2–60–001), Rulia Islam (2014–2–60–022) and Sk. Sabit Faisal (2014–2–60–014) to the Computer Science and Engineering Department, East West University, Dhaka-1212, Bangladesh is accepted as satisfactory for partial fulfillment of requirements for the degree of Bachelors of Science(B. Sc.) in Computer Science and Engineering.

Board of Examiners

....

Dr. Shamim H. Ripon
Associate Professor
Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh

....

Dr. Ahmed Wasif Reza
Associate Professor and Chairperson
Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh

Abstract

We can simply define data mining as a process that involves searching, collecting, filtering and analyzing the data. It is important to understand that this is not the standard or accepted definition. But the above definition caters for the whole process.The utilization of machine learning in each area has gone to an incredible level amid this decade. In business part, a huge measure of new organizations is failing to become successful and this number can be minimized utilizing machine learning. This paper describes a method of predicting the future state of a start-up company by applying four prediction algorithms on a data set which holds data about companies. The applied algorithms are Decision Tree, Naive Bayes, Artificial Neural Network and Support Vector Machine. After doing pre-processing of the data set, we split the data set into training and testing parts. We create classifier model by using training set data and then predict our output class using test set data. There are three outcome variable of our classification class which are Successful, cancelled and failed. By applying C5.0 algorithm and Ripper algorithm, we have extricated a few if then rules which will assist organizations with making all the decisions more effectively. The practical usage of this model is rather applying thoughts indiscriminately to enhance the condition of the organization; organizations can precisely improve the choice of decisions by putting their attribute values on this model or by simply checking on the rules.

Acknowledgments

First of all, We would like to express our deepest gratitude to The Almighty for His blessings on us. Next, our special thanks go to thank our thesis advisor Dr. Shamim H. Ripon of Department of Computer Science and Engineering at East West University. The door to his office was always open whenever we ran into a trouble spot or had a question about our thesis. He consistently allowed this thesis to be our own work, but steered us in the right the direction whenever he thought we needed it.

We would also like to acknowledge all of the faculty members of Department of Computer Science and Engineering for their unconditional help and support whenever we needed it.

Finally, we must express our very profound gratitude to our parents for providing unfailing support and continuous encouragement throughout our years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you

Md Ashraf Uz Zaman Shahria
August, 2018

Rulia Islam
August, 2018

Sk. Sabit Faisal
August, 2018

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

There is the absolute confidence that in recent era machine learning is at the highest of the hype curve. Previously it was a buzzword but currently its a leading topic of all over the world. It is these much important not only for a closed sector but also its wide used in day to day life. Machine Learning is the method of obtaining computers to be told and act like humans do, and improve their learning over time in an independent way, by feeding them information and data within the sort of surveillance and real-world interactions. It is an associate application of artificial intelligence that gives systems the flexibility to mechanically learn and improve from expertise while not being expressly programmed. Machine learning focuses on the event of computer programs which will access information and use it learn for themselves. The procedure of learning begins with observations or knowledge, like examples, direct expertise, or instruction, so as to appear for patterns in knowledge and build higher choices within the future supported the examples that we offer. The first aim is to permit the computers learn mechanically while not human intervention or help and change actions consequently.

Nowadays Machine learning is becoming the most important part of a successful business. For some of its better features like- making user-generated content valuable, finding products faster, engaging with customers, understanding customer behavior, uses algorithms to help website rank better and giving birth to the world of search engine optimization, it is tremendously improving and fruitful for business sector [1] .

The use of ML is successful because of its godlike performance in an exceedingly wide

selection of activities. Like as, police investigation fraud and diagnosis illness. In the sphere of business, ML has a huge impact on the dimensions of earlier all-purpose technologies. Though it's already in use in thousands of firms round the world, most huge opportunities haven't nevertheless been tapped. The results of ML is exaggerated within the returning decade, as producing, retailing, transportation, finance, health care, law, advertising, insurance, recreation, education, and just about each alternative trade rework their core processes and business models to require advantage of it. The bottleneck now could be in management, implementation, and business imagination. [2]
Thousands of new emerging companies are trying to be successful each year and failing rate is too high. The efficient usage of resources can turn a startup company small to a very big thing. Obviously the main important issue here is the decisions they make in every step of constructing the organization. But unfortunately most of the time inexperienced companies make wrong choices and turn the organization into failed organization. Here our method comes; rather taking decisions blindly, it predicts a companys future more precisely which can lead to success. It delivers a solid arrangement of rules to generalize the idea of successful investment and simplifies the way of success. Our aim is to running on a selected business hassle domain . that's are expecting the fame of a web based begin up enterprise. We narrow down our attention on Kickstarter website , this website is a platform for beginning new enterprise . wherein humans can publish their undertaking idea . If venture idea is ideal than human beings can donate cash on those particular assignment . we're going to expect the commercial enterprise fame based on how lots money is wanted for finishing the mission , how a whole lot money this mission benefit and what amount of human beings donate cash for this particular mission .

1.1.1 Objectives

So the objectives of our thesis are-

- Train the models to find the most effective model that can classify future state of a company/project.

- Generating rules that help the users transforming a failed or cancelled project to become successful.
- Developing user between user and model interaction.

1.1.2 Contribution

Our contribution on fulfilling those objectives was satisfactory since we successfully deployed all the objectives.

- We trained sixteen models for different instances to get the best model. Four different algorithms were used to train the models and they are Support Vector Machine,Naive Bayes,Artificial Neural Network and Decision Tree. Each algorithm is applied for four different amount of data. The best model is chosen from all sixteen models for evaluating the state of new data.
- By using two algorithms we generated the rules. The algorithms are Decision Tree and RIPPER. These rules are if-then rules which are easier to understand. These rules not only gives a solution for a failed project but also describe the causes of success or being cancelled of a project.
- We developed an interface for the users. The interface maintains the interaction between the model and the user.

Incorporating Human Computer Interaction could be one of them which can be applied with the interface.

1.1.3 Outline

Chapter 1 introduction: This chapter includes the motivation behind the thesis. Describes what aspect motivated us to research on this particular topic. The objectives are clarified and contributions are attached.
Chapter 2 Background: This chapter describes the background researches on this area. It also includes the description of different sub areas in data mining particularly

while using for business oriented environment.

Chapter 3 Proposed Model: The model of work flow is described with a figure.

Chapter 4 Domain description This chapter includes the description of a start up and kickstarter. It includes the ideas of crowd funding and attributes of the dataset. Statistical analysis is attached.

Chapter 5 Mining Technique: How all four algorithms (ANN, SVM, DT, and NB) which are applied on dataset is described in this section. The packages of R and the scripts are also precisely described.

Chapter 6 Result and Analysis: The efficiency of the final model and how it was chosen among other candidate models is described. The process of adding an user interface and associating the interface with rules are described.

Chapter 7 Conclusion Summary of the thesis work and the future works are included.

Chapter 2

Background

Data mining is broadly viewed as essential in numerous business applications for compelling basic leadership. The significance of business information mining is reflected by the presence of various studies in the writing concentrating on the examination of related works utilizing information digging systems for taking care of particular business issues. There are 33 unique information mining systems utilized in eight distinctive application zones[13]. A large portion of them are regulated learning procedures and the application region where such methods are regularly observed is liquidation expectation, trailed by the zones of client relationship administration, extortion discovery, interruption recognition and recommender frameworks. Moreover, the broadly utilized ten information digging procedures for business applications are the decision tree (C4.5 decision tree and characterization and relapse tree), hereditary calculation, k-nearest neighbor, multilayer perceptron neural system, Naive Bayes and support vector machine as the managed learning methods and affiliation control, desire amplification and k-implies as the unsupervised learning strategies. Utilizing innovation to pick up an edge in business is definitely not another thought. At whatever point there is something new, business visionaries will rush to endeavor to discover an application for it in the business world to profit. Data mining (DM) and business intelligence (BI) are among the data innovation applications that have business esteem [14]. Data mining is the way toward looking through information utilizing different calculations to find examples and connections inside a database of data. Business knowledge, on the other hand, concentrates more on information joining and association. It will join information break down to encourage administrators make operational, strategic, or key business choices. Data mining can be utilized to help the goals of business insight framework[15]. Business Intelligence could be a thought of applying a gathering of innovations to change over data into which means information. Bismuth ways encapsulate information recovery, information handling, connected math investigation yet as data visual picture. Mammoth measures of knowledge of data beginning totally unique in a few various configurations furthermore, from various sources might be solidified and recover to key business learning. Presents a general read on anyway data square measure renovated to business insight. The technique includes every business advisors and specialized experts[16]. It changes over an outsized size of data to significance results consequently on offer basic leadership support to complete clients. Organizations can apply information mining with a specific end goal to enhance their business and pick up points of interest over the contenders. Have you ever consider the proposals you get when you shop on the web. In the event that you buy for instance a TV on the web, the site prescribes you another items that you truly need to get[17]. Likewise have ever consider the cautions you get from your bank when you complete a sudden utilization of your credit card in an alternate city. In reality these are cases of data mining which is the way toward finding helpful examples in a tremendous informational index. This gigantic information is made by coordinating current and verifiable information from various sources and store them halfway in an exceptional archive called Data Warehousing(DW)[18]. A few investigations utilized information digging for removing rules and foreseeing certain practices in a few regions of science, data innovation, HR, instruction, science what's more, drug. For instance, Beikzadeh and Delavari (2004) utilized information digging procedures for recommending upgrades on higher instructive frameworks. Al-Radaideh et al. (2006) likewise utilized data mining procedures to anticipate college understudies' execution. Numerous restorative scientists, then again, utilized information mining strategies for clinical extraction units utilizing the colossal patients information documents and chronicles, Lavrac (1999) was one of such specialists. Mullins et al. (2006) additionally chipped away at patients' information to separate malady affiliation rules utilizing unsupervised strategies. Karatepe et al. (2006) characterized the execution of a cutting edge worker, as his/her efficiency contrasting and his/her companions. Schwab (1991), then again, depicted the execution of college instructors incorporated into his examination, as the number of looks into refered to or distributed. By and large, execution is normally estimated by the units created by the representative in his/her activity inside the given timeframe[19]. Specialists like Chein and Chen (2006) have dealt with the change of representative choice, by building a model, utilizing information mining methods, to anticipate the execution of recently candidates. Contingent upon characteristics chose from their CVs, work applications and meetings. Their execution could be anticipated to be a base for leaders to take their choices about either utilizing these candidates or not. Past investigations determined a few characteristics influencing the worker execution[20]. A portion of these characteristics are close to home attributes, others are instructive lastly proficient properties were likewise considered. Chein and Chen (2006) utilized a few ascribes to foresee the worker execution. They indicated age, sexual orientation, conjugal status, encounter, instruction, significant subjects and school tires as potential factors that may influence the execution. At that point they avoided age, sexual orientation and conjugal status, with the goal that no separation would exist in the procedure of individual determination. Thus for their examination, they discovered that worker execution is profoundly influenced by instruction degree, the school tire, and the activity encounter. So it is anything but an old thought rather utilizing for quite a while now to join information mining in business condition. By and large, this proposal is a fundamental endeavor to utilize information mining ideas, especially order, to help business executives and leaders. assessing workers' information to ponder the fundamental properties that may influence the representatives' execution[21].

Chapter 3

Proposed Model

Our proposed model for Crowding fund based business domain is given in . The methodology starts with the data collected from kaggle data repository []. After the collection of data we perform data prepossessing to identify unnecessary feature. Here we identify several feature which is not important to create our model. Data cleaning is also applied to remove noisy data. Data pre-processing is a data mining technique that transform a data into an understandable format. After data pre-processing our working data just created. Now our goal is finding rules from data and find best model for classification. For finding rules we apply two data mining technique to generate rule . Then we evaluate our rule on basis of few parameter. For finding best classification model at first we split our data into train set and test set . We apply several data mining technique on training data then build few candidate model . Next we find our best model by comparing with predictions with known target value which is test data. Next we evaluate our model performance.

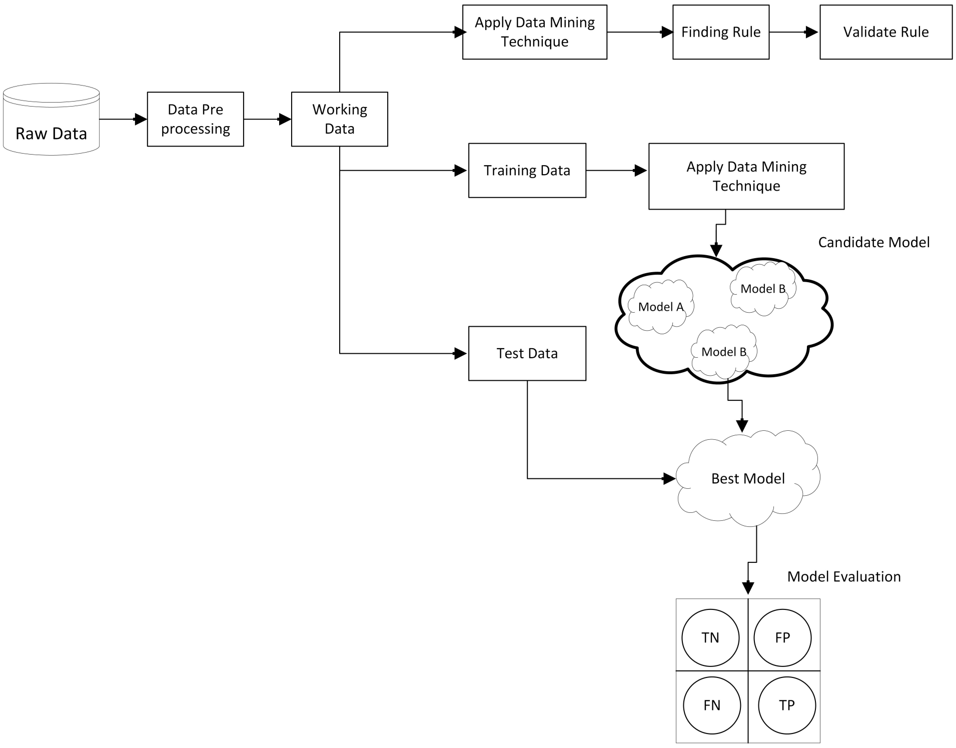


Figure 3.1: Proposed Model

Chapter 4

Domain Description

4.1 Overview of Start-Up

A startup is a new organization or in other words an ambitious initiative that is simply trying to get in market. New businesses are generally little and at first financed and worked by a bunch of organizers or one person. These organizations offer an item or administration that isn't right now being offered somewhere else in the market. In the beginning times, startups costs have a tendency to surpass their incomes as they need more money on promoting and creating the company. For running a startup smoothly, they require regular financing. New companies might be financed by customary private company advances from banks or credit associations, by government-supported Small Business Administration advances from neighborhood banks, or by awards from philanthropic associations and state governments. Some startups trade some control with the people who finances the business and its definitely fine because most of the investors and banks are not brave enough to put their valuable money on something which may fail in future. So it needs huge courage to back up a startup company that sometimes the investors expect a little more from the company owners except gratitude. The term startup has gained a lot of popularity in recent years because of the revolution around the Silicon Valley and other young tech companies around the world who are making huge amount of money with enhancing technology and bringing new opportunities to millions of people. Moreover economists believe that this the best time in history to invest in technology so investors from all around the world is breaking geographical barrier by doing investment in different continents with the help of the mighty internet.

Here we introduce a new fancy term which is Online Startup. The internet is enormous; there is literally a sea of need about anything. Its like a big town where people can sell anything since a large market always have someone to buy something. The boom of the internet in 21st century is so massive that startups which are opened and distributed in online have a bigger chance of success than local startups. Even people who sell products locally are also joining or creating websites so that they can be a part of this massive market. The banking system or paying method is becoming less complex everyday so people are more attracted to buy from online. As a result the ambitious startup owners from any corner of the world holding an idea to deploy it in this enormous market. Now companies are making targeted advertisements in online which will go to exact cluster of people which will hit them on precise. Strategies like email marketing literally turning visitors into buyers leading an online company to become more successful.

4.2 Overview of kickstarter

The internet also reduced the gap between investors and startup idea holders in many ways. One of them is Kickstarter. This website based on a bold idea which is known as Crowdfunding. The idea is pretty simple, when someone thinks he/she got a bold business idea to start with but does not have financial support then they can sign in to crowd funding websites to raise money. A lot of crowd funding websites are available there but Kickstarter is regarded as the best one since they deployed a lot of successful exotic project from previous years. After signing in the initiator needs to set up a campaign and usually people spend a lot of time in campaigning. They make videos to promote their project, they give a visual architectural overview on their overall thinking and mindset about the project. Some likes to share a basic prototype which is usually done by tech companies. Each project has a funding goal which may set by the initiator and a deadline to raise that goal inside the timeframe. The project owner offers a set of rewards for the pledgers. The rewards are often advised to rationalized with human choices; for example no one wants to spend thousand dollar to buy a AI robot rather they might spend some money to help the project owner to persue his/her dream. So the reward list needs to be attractive and market oriented. Owners need to understand what the better choices are for the people who will help the project to move on. Funding is everything in Kickstarter. The project is funded when it reaches the funding goal within the time period the project owner has given. Sometimes some projects raise more money than they actually asked from the market since they might be able to successfully attracted lot of people towards the project. One of the main focus of Kickstarter is making creative projects. The project can be arts oriented, can be movie or music oriented or can be a technological or innovative idea. So at the end of everything, the project owner must provide a solid product to the people who have given money. The primary danger of Kickstarter is it's win big or bust financing rules. In the event that you don't totally finance your venture, you don't get a dime of cash. Regardless of whether you're simply $10 shy of your objective, your entire undertaking will be denied. Its also important to understand that Kickstarter is not charity. The site does not help those projects who want to run a charity. Finally the most important factor is Kickstarter does not ensure equity which SharkTank does. People who give money to projects do not become business partners with the project owners. The investors are paying a flat fee in exchange of goods and services. Kickstarter is not a place where people can raise money for groceries like soft drinks or other beverages, items that encourage hate speech; it is also not the place for fundraiser who wants to sell drugs and other stuffs. Its a place for new startups companies which will bring a project that can eventually become a breakthrough. Some biggest companies which are initially emerged from Kickstarter are smart watch company Pebble, card game company Exploding Kittens and Coolest Cooler.

4.3 Attribute information

Dataset collected from kaggle machine learning repository . Kaggle is a platform for predictive modelling and analytic competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowd-sourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be most effective.

Our data set contain 15 column and 378661 instances. But for our investigation we include 7 column in our dataset. We choose those column by our intuitive sense . We discard name column because this column contain project description which is inappropriate for building our model. We discard currency column because currency and country both column contain same information so its better to discard currency column. We discard launched and deadline column because our model is not related with project completed timing. Pledged and goal column contain information about what amount of money already funded for a particular project and what amount of money is needed for completing the particular project. But these two column information is based on donor currency . This two column currency is converted into dollar currency on two individual column . So we discard pledege and goal column from our working dataset. Data set attribute information is giving below .

Data set attribute		
Attribute Name	Description	Data Type
Category	Sub category of project	String
Main category	Main Category of the project	String
Goal	Fund-raising goal The funding goal is the amount of money that a creator needs to complete their project.	Numeric
Pledged	Pledged amount in the project currency	Numeric
State	State of project	String
Backers	Number of backers	Integer
Country	Country	String

4.4 Statistical Analysis

Our classification class is State . That means we are going to classify the state of the particular project based on left over attribute . As we have mentioned earlier, our models main goal is to classify the state of a business project . We include three state for our research investigation these are failed, cancelled and successful . Among the whole data set we filtered out that three particular state . Lets examine the state class , In figure 4.1 shows an abstract view of states. A statistical view of our data set shows that that 53% projects were failed , 10% projects were cancelled and 36% projects were successful.

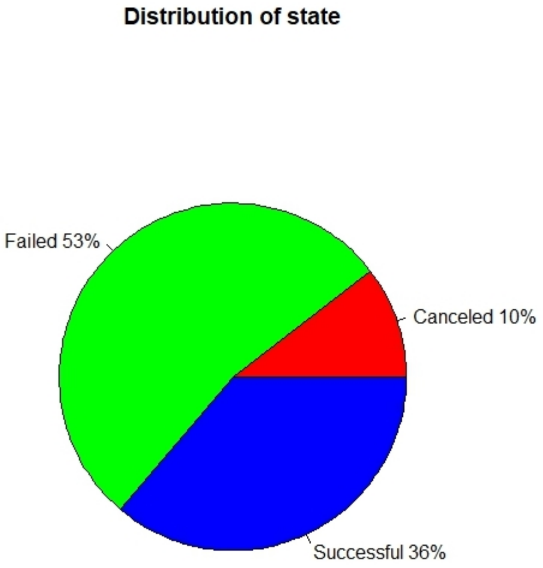


Figure 4.1: Distribution of State

For statistical analyzing purpose we are going to reveal some important fact that is hidden on the dataset. First statistical question arise that What types of projects are most popular ? . From the figure 4.2 we plot frequency of most popular project based on category . We can see that Film Video appears to be the most popular project category.

Now well do the same thing for subcategories. There are 159 subcategory levels, which is far too many to plot individually, so well just plot the ten subcategories with the greatest

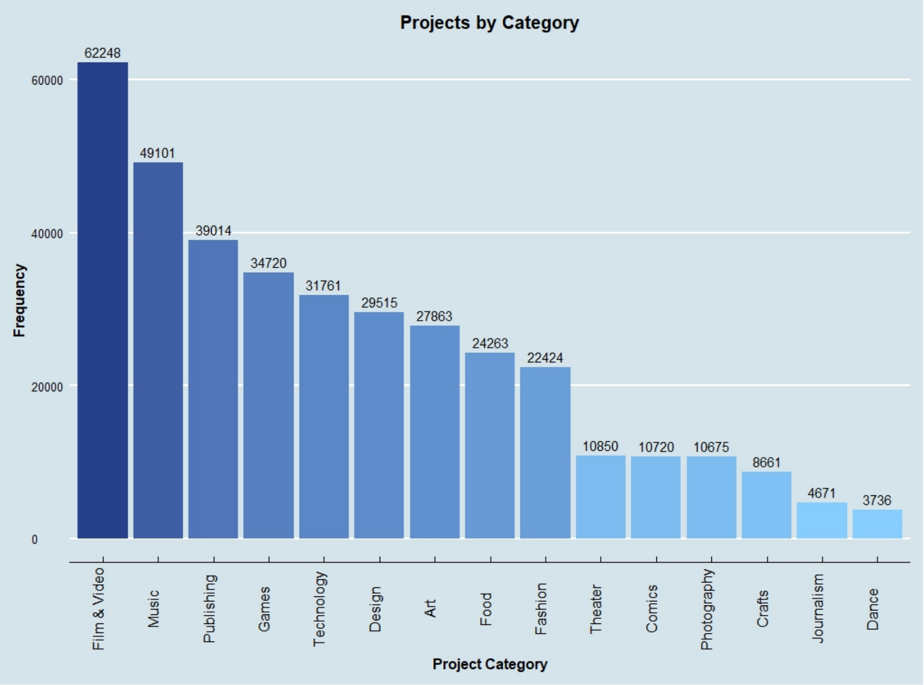


Figure 4.2: Distribution of project by category

number of projects. That shows in figure 4.3 , we find here that Product Design is the most popular subcategory, stemming from the category of Design.

Now every other query arrives that What forms of initiatives are being funded ? This query is akin to the primary question however phrased from the attitude of the backers. In other words, the maximum funded projects are the most famous initiatives within the angle of the backers. First, permits determine what sorts of initiatives funding goes closer to. Well try this through aggregating the quantity of finances pledged for every category, presenting us with the whole amount pledged for each class. In figure 4.6 , shows that Games, Design, and Technology are the highest grossing categories by far. Lets take a closer look at Games and see what subcategories are most popular within it .

Lets jump back to our first result in figure 4.2 and examine the category of Films Video further. Lets break down the number of projects into its various subcategories in figure 4.5. Documentaries seem to be very popular. Perhaps theyre under appreciated

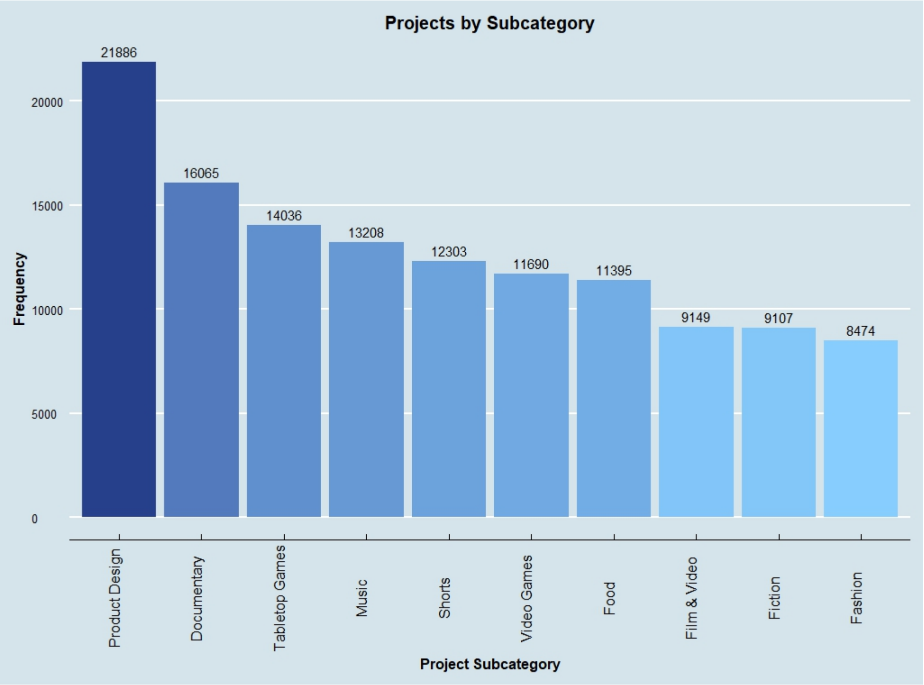


Figure 4.3: Distribution of top project by Subcategory

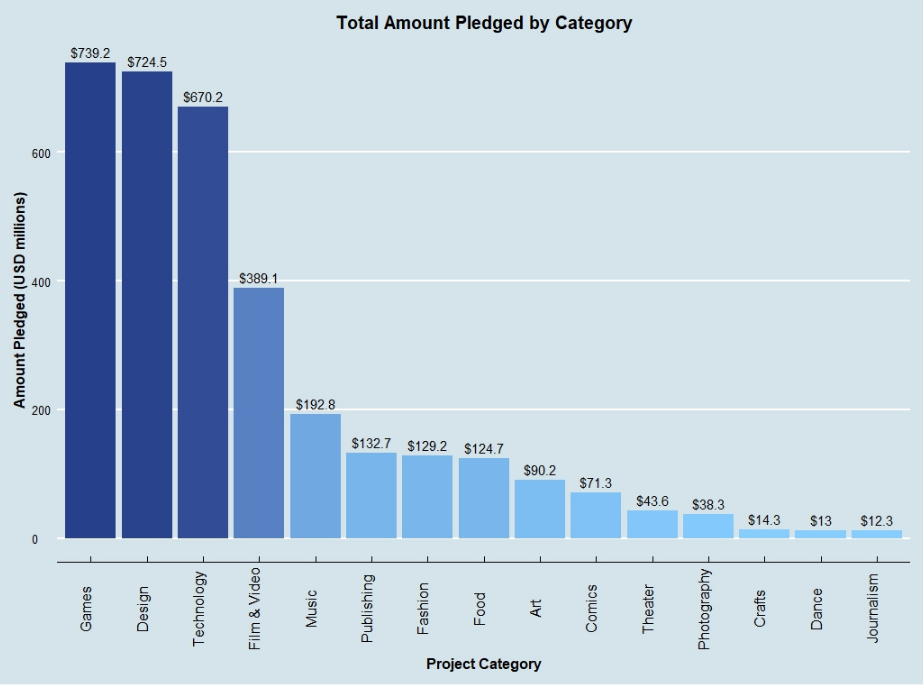


Figure 4.4: Total amount pledged by category

in mainstream media and require sources for alternative funding such as Kickstarter.

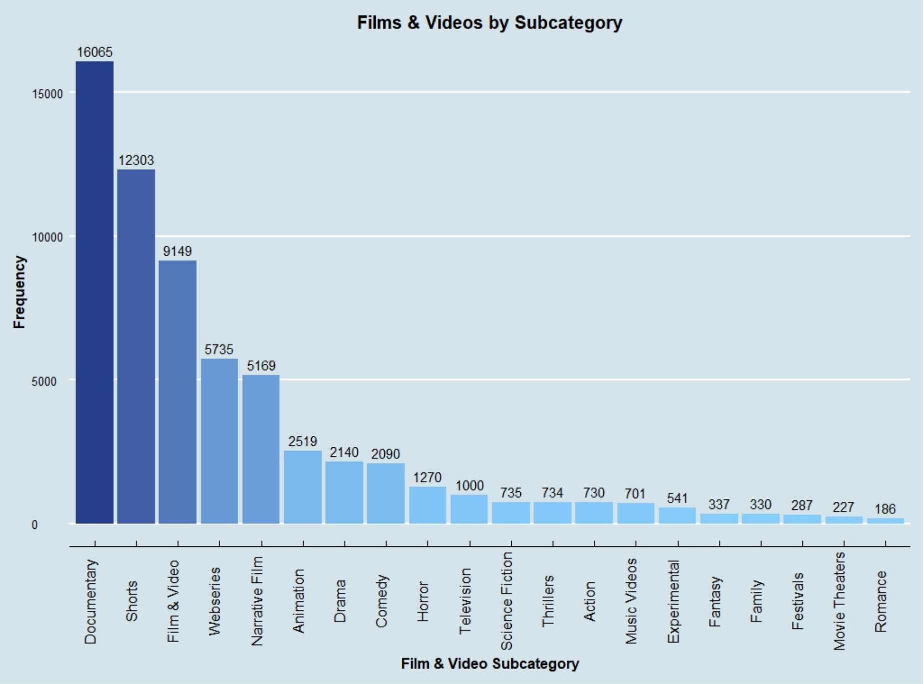


Figure 4.5: Film and Video Subcategory

Now we are going to investigate another statistical question What types of projects are being funded? . This question is akin to the first question but phrased from the perspective of the backers. In other words, the most funded projects are the most popular projects in the perspective of the backers. First, lets determine what types of projects funding is going towards. Well do this by aggregating the amount of funds pledged for each category, providing us with the total amount pledged for each category in figure 4.6. Games, Design, and Technology are the highest grossing categories by far. Lets take a closer look at Games and see what subcategories are most popular within it.

Now again back to previous result , from the previous result we found that games is the most funded project by backers . Now examine the games subcategory in figure 4.7. Surprisingly Tabletop Games are highly funded, which is a bit surprising to me. We seem to have underestimated their popularity as a hobby.

Now we examine the specifically at successful projects and examine the number of

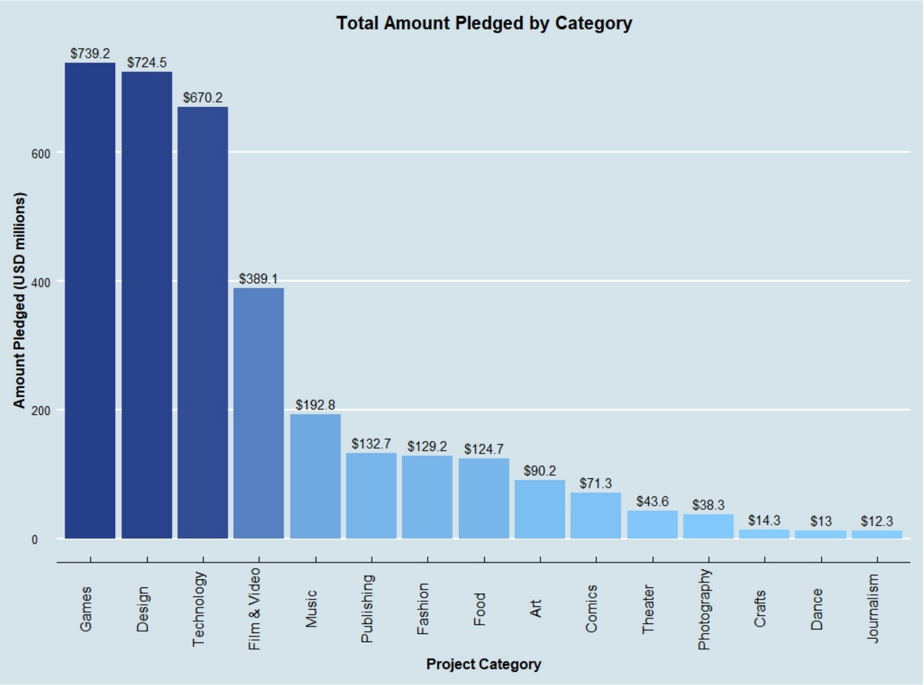


Figure 4.6: Total amount pledged by category

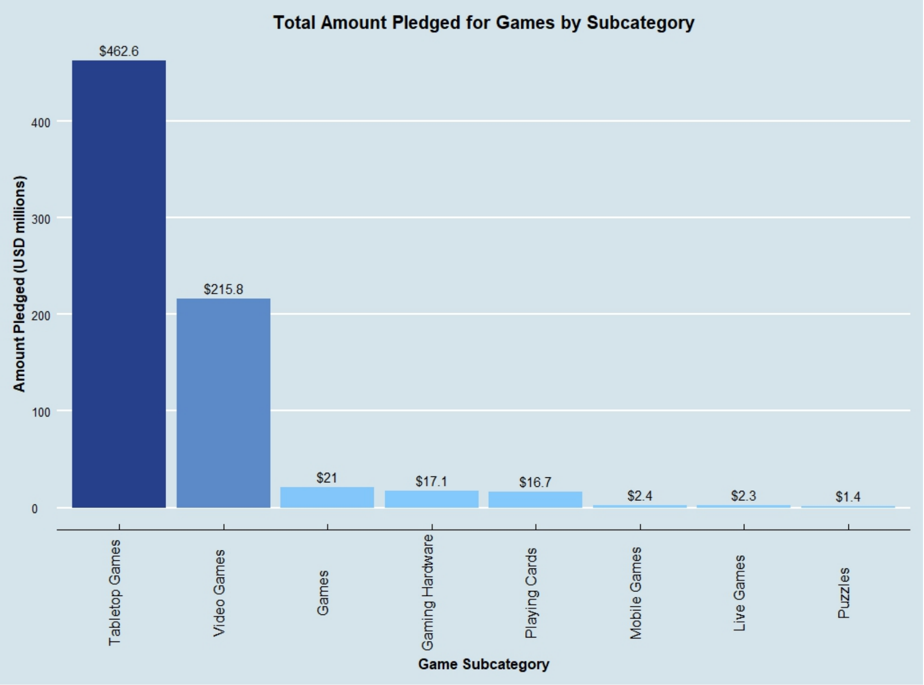


Figure 4.7: Total Amount Pledged for Games by Subcategory

projects by category in figure 4.8. From the result we can see that music is the top funded successful project.

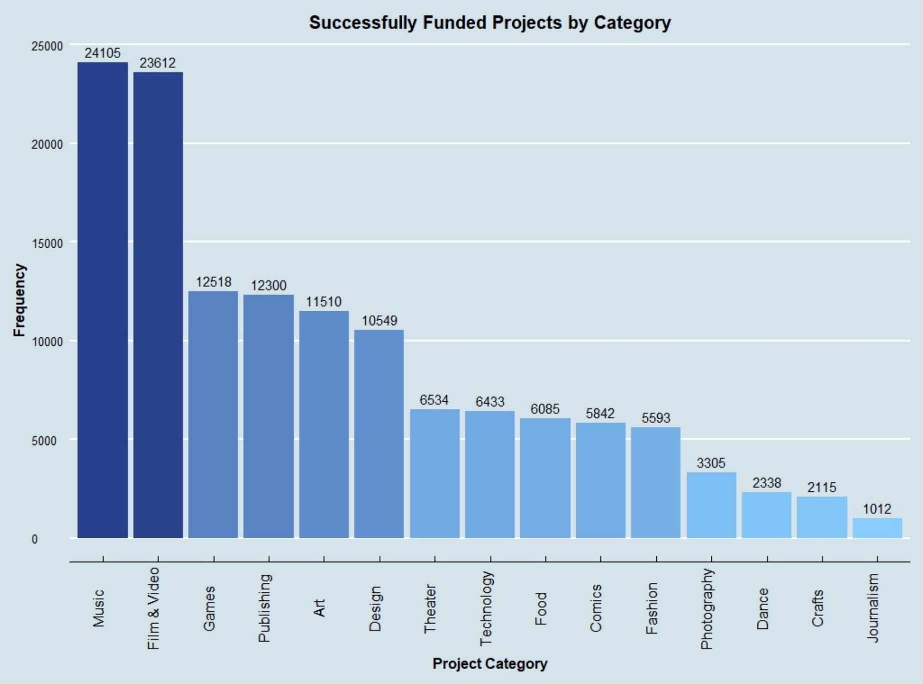


Figure 4.8: Successfully Funded Projects by Category

Perhaps it would be more helpful to know the proportions of success vs. failure for each category rather than just the frequency. Note that only using projects with the status of success or failure to calculate these success/failure proportions would represent all completed projects (i.e. projects that are not live and have not been cancelled) in figure 4.9. Dance, Theater, and Comics have the highest success rates and Technology, Journalism, and Crafts have the lowest. This is an interesting result because Dance, Theater, and Comics were on the lower end in terms of the aggregate and average amounts pledged, whereas Technology was on the higher end. This does however, agree with the results from the box plots above, which illustrated the distribution of the amounts pledged for individual projects. In other words, categories that received little funding on the individual project level had a higher category failure rate.

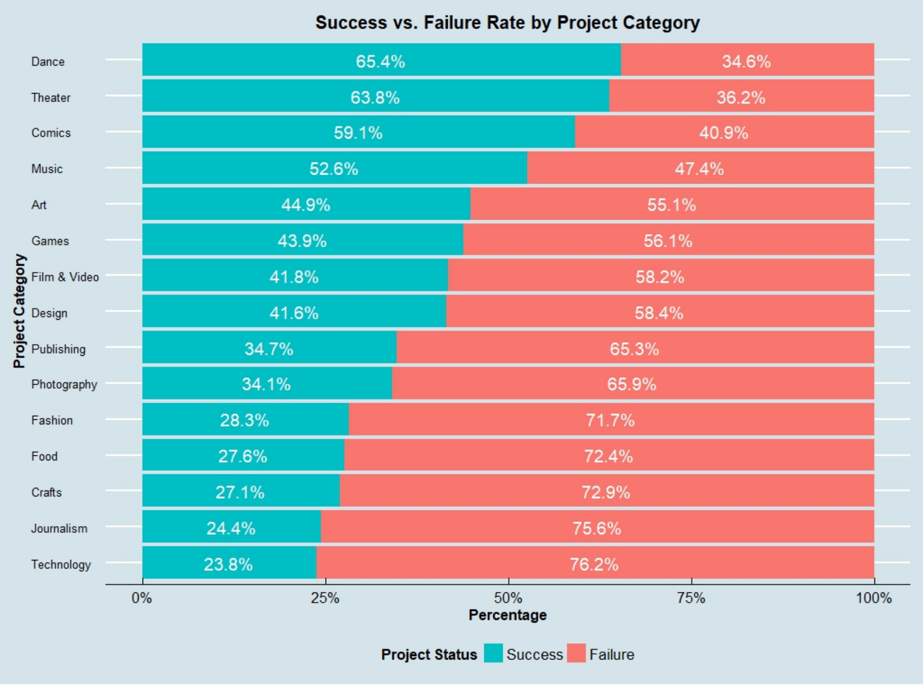


Figure 4.9: Success vs. Failure Rate by Project Category

Chapter 5

Mining Technique

5.1 Introduction

In this area we will expand which machine learning algorithms are utilized as a part of our proposed demonstration. How we have worked together with the dataset will be portrayed in different areas. This segment fundamentally centers on a short concise. We have applied five algorithms and they are: Artificial Neural Network (ANN), Nave Bayes Classifier, Support Vector Machine, Ripper and Decision Tree.

5.1.1 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a data processing paradigm that is enlivened by the way of the central nervous system or the human brain. The key component of this paradigm is the novel structure of the data processing framework. It is made out of an enormous number of profoundly interconnected handling components (neurons) working as one to take care of particular issues. ANNs, similar to individuals, learn by handling cases. An ANN is designed for a particular application, for example, pattern recognition or data processing, through a learning procedure. Learning in biological frameworks includes changes in accordance with the synaptic associations that exist between the neurons. This is valid for ANNs also.ANN is an awesome device for discovering designs which are unreasonably mind boggling or various for a human software engineer to concentrate and encourage the machine to perceive. Neural systems adopt an alternate strategy to critical thinking than that of regular PCs. Traditional machines utilizes an algorithmic approach. The machine takes after an arrangement of directions so as to take care of an issue. Except if the particular advances that the machine needs to take after are known, the machine can't take care of the issue. That limits the critical thinking ability of traditional PCs to issues that we as of now comprehend and know how to tackle. In any case, computers would be a great deal more helpful on the off chance that they could do things that we don't precisely know how to do. While neural systems (likewise called "perceptrons"), it is just over the most recent quite a few years where they have turned into a noteworthy piece of man-made brainpower. This is because of the entry of a method called "backpropagation," which enables systems to change their shrouded layers of neurons in Circumstances where the result doesn't coordinate what the maker is seeking after. Here is an example see in 5.1:

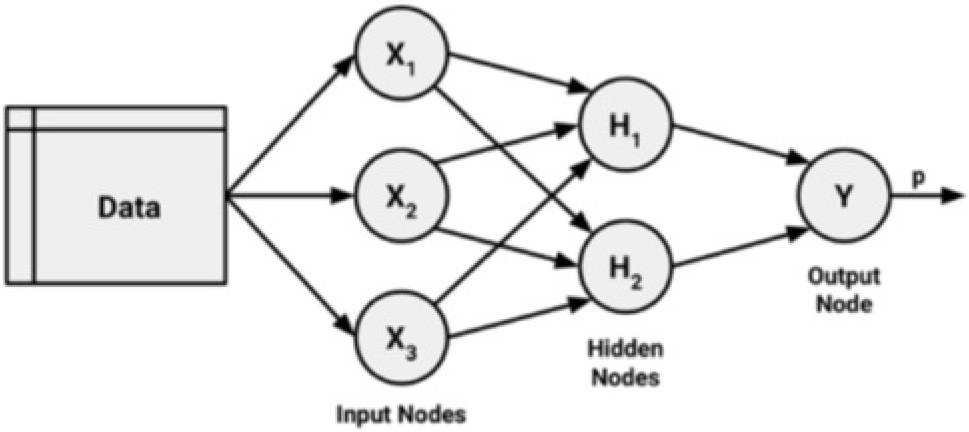
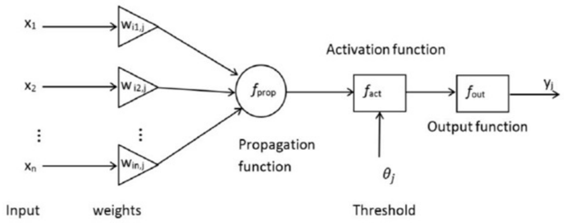


Figure 5.1: Artificial Neural Network

This is a general ANNs example with two hidden layer [3]. A simple function can define ANN which is $f : X- > Y$. The model of a single artificial neuron can be comprehended in wording fundamentally the same as to the natural model or a biological model. As portrayed in the accompanying figure, a coordinated system graph characterizes a connection between the info signals got by the dendrites (x factors), and the yield flag (y variable). Similarly likewise with the organic neuron, every dendrite's flag is weighted (w esteems) as indicated by its significanceoverlook, for presently, how these weights are resolved. The information signals are summed by the cell body and the flag is passed on as indicated by an activation function signified by f:



An ordinary artificial neuron with n input dendrites can be represented by this formula. The w weights permit every one of the n inputs (meant by xi) to contribute a more noteworthy or lesser add up to the aggregate of input signals. The net aggregate is utilized by the activation function f(x), and the subsequent flag, y(x), is the output axon:

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right)$$

5.1.2 Nave Bayes Classifier

Naive Bayes classifier is a clear and ground-breaking algorithm for the classification problems. Regardless of whether we are chipping away at an informational index with a huge number of records with a few characteristics, it is recommended to attempt Naive Bayes approach. Naive Bayes classifier gives extraordinary outcomes when we utilize it for textual information analysis. For example, Natural Language Processing. This kind of modeling turns out to be more helpful when conditional probabilities are utilized. These are values we work out by looking at the probability of seeing one value given we see another. Conditional probabilities are noted utilizing the bar '—' to isolate the conditioned from the conditioning value [3]. The formula of calculating Conditional Probability is:

$$P(H|E) = (P(E|H) * P(H))/P(E)$$

- P(H) = Probability of H (hypothesis) being true. Often called The prior probability.
- P(E) = Probability of the evidence (hypothesiss effect is not considered).
- P(E | H) = Probability of evidence where hypothesis is considered as true.
- P(H | E) = Probability of hypothesis which indicates evidence is there.

A Naive Bayes Classifier is a program which predicts a class esteem given an arrangement of set of characteristics. For each known class value these steps are taken,

- Calculate probabilities for each attribute, conditional on the class value.
- Utilize the product rule to acquire a joint conditional probability for the attributes.
- Use bayes rule to determine conditional probabilities for the class variable.

A niggling issue with the Naive bayes is the place the dataset doesn't give at least one of the probabilities we require. The NULL values create problems. We at that point get a probability of zero figured in with the general mish-mash. This may make divide by zero, or essentially make the last value itself zero. The least demanding arrangement is to overlook zero-valued probabilities or simply ignore them during pre processing. Analysts are fairly aggravated by utilization of the NBC in light of the fact that the gullible suspicion of freedom is quite often invalid in reality. Nonetheless, the technique has been appeared to perform shockingly well in a wide assortment of settings.

5.1.3 Support Vector Machine (SVM)

A support vector machine builds a hyperplane or set of hyperplanes in a high-or unending dimensional space, which can be utilized for classification, regression, or different undertakings like anomalies detection. Instinctively, a great partition is accomplished by the hyperplane that has the biggest separation to the closest preparing information purpose of any class (functional margin), since when all is said in done the bigger the edge the lower the generalization mistake of the classifier [3].

The benefits of support vector machines are:

- Compelling in high dimensional spaces.
- Still compelling in situations where number of measurements is more prominent than the quantity of tests (samples).
- Utilizations a subset of training points in the decision function, so it is likewise memory productive or in other words memory efficiency is very good.
- Flexible: diverse Kernel capacities can be indicated for the decision function. Basic kernels are given, yet it is likewise conceivable to indicate custom kernels.

It is most straightforward to see how to locate the greatest edge under the suspicion that the classes are straightly divisible. For this situation, the MMH(maximum edge hyperplane) is as far away as conceivable from the external limits of the two gatherings of data points. These external limits are known as the curved body or convex hull. The MMH is then the opposite bisector of the most brief line between the two curved structures. Refined machine calculations that utilization a system known as quadratic advancement are prepared to do finding the greatest edge along these lines. An option (however comparable) approach includes a hunt through the space of each conceivable hyperplane keeping in mind the end goal to locate an arrangement of two parallel planes that partition the focuses into homogeneous gatherings yet themselves are as far separated as could be expected under the circumstances. we'll have to characterize precisely what we mean by a hyperplane. In n-dimensional space, the accompanying condition is utilized

$$\vec{w} \cdot \vec{x} + b = 0$$

Utilizing this formula, the objective of the procedure is to locate an arrangement of weights that indicate two hyperplanes, as takes after

$$\vec{w} \cdot \vec{x} + b \geq +1$$
$$\vec{w} \cdot \vec{x} + b \leq -1$$

$$\frac{2}{||\vec{w}||}$$

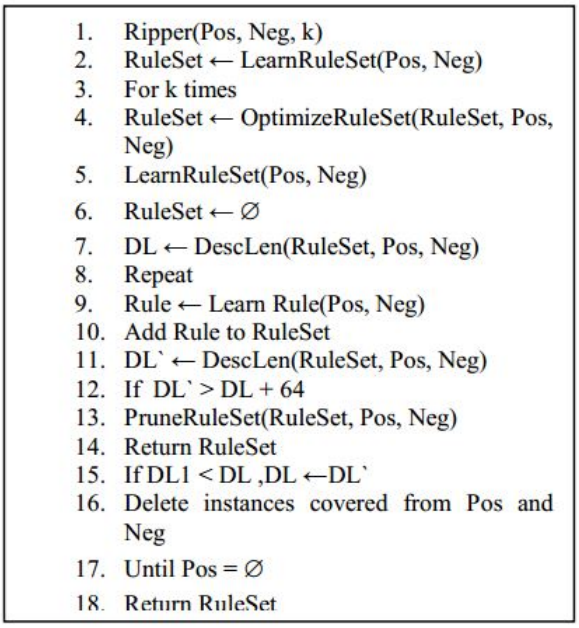
Vector geometry characterizes the separation between these two planes as:
Here, $||w||$ shows the Euclidean standard (the separation from the starting point to vector w). Since $||w||$ is in the denominator, to amplify separate, we have to limit $||w||$. The assignment is regularly re expressed as an arrangement of requirements, as takes after:

$$\min \frac{1}{2} \|\vec{w}\|^2$$
$$s.t. \;\; y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \forall \vec{x}_i$$

In spite of the fact that this looks untidy, it's extremely not very confused to get it thoughtfully. Fundamentally, the main line suggests that we have to limit the Euclidean standard (squared and isolated by two to make the figuring less demanding). The second line takes note of this is liable to, the condition that every one of the yi information focuses is effectively arranged. Note that y shows the class values (changed to either +1 or - 1) .

5.1.4 Ripper Algorithm

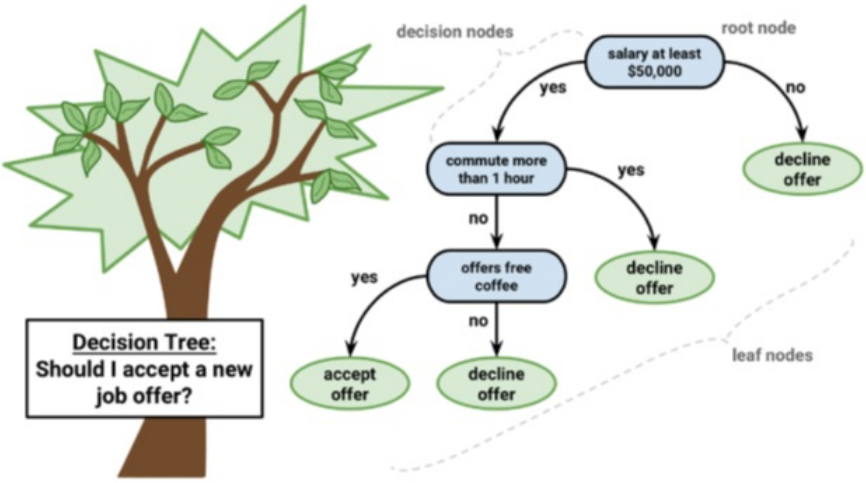
RIPPER is condensed as Repeated Incremental Pruning to Create Error Reduction. RIPPER is particularly more effective on extensive loud datasets .There are two sorts of circle in Ripper calculation. This calculation was planned by Cohen in 1995 in particular, Outer circle and Inner circle, often called inner loop and outer loop. There is a one rule base. Outer loops main activity is to perform an addition of a new rule in rule base. At a time only one rule can be added. The inner rules main activity is adding a new condition to the current rule. These loops continue until a negative example is created. The following figure shows the psedocode of Ripper [22].



```
Ripper(Pos, Neg, k)
RuleSet ← LearnRuleSet(Pos, Neg)
For k times
RuleSet ← OptimizeRuleSet(RuleSet, Pos,
Neg)
LearnRuleSet(Pos, Neg)
RuleSet ← ∅
DL ← DescLen(RuleSet, Pos, Neg)
Repeat
Rule ← Learn Rule(Pos, Neg)
Add Rule to RuleSet
DL` ← DescLen(RuleSet, Pos, Neg)
If DL` > DL + 64
PruneRuleSet(RuleSet, Pos, Neg)
Return RuleSet
If DL1 < DL ,DL ←DL`
Delete instances covered from Pos and
Neg
Until Pos = ∅
Return RuleSet
```

5.1.5 Decision Tree C5.0

Decision tree constructs regression or classification models as a tree structure. It separates a dataset into littler and littler subsets while in the meantime a related decision tree is incrementally created. The last outcome is a tree with decision nodes and leaf nodes. A decision node has at least two branches . Leaf node speaks to an order or choice. The highest decision node in a tree which compares to the best indicator called root node. Decision trees can deal with both unmitigated and numerical information. The basic calculation for constructing decision trees called ID3 by J. R. Quinlan which utilizes a top down, insatiable search (greedy search) through the space of conceivable branches with no backtracking. ID3 utilizes Entropy and Information Gain to develop a decision tree. In ZeroR model there is no indicator, in OneR model we attempt to locate the absolute best indicator, naive Bayesian incorporates all indicators utilizing Bayes' rules and the autonomy suspicions between predictors however choice tree incorporates all predictors with the reliance presumptions between predictors [3]. To better see how this functions practically speaking, we should think about the accompanying tree, which predicts whether an occupation offer ought to be acknowledged. An occupation offer to be considered starts at the root node, where it is then gone through choice nodes that require decisions to be made in view of the qualities of the activity. These decisions split the information crosswise over branches that show potential results of a choice, delineated here as yes or on the other hand no results, however sometimes there might be in excess of two potential outcomes. For the situation a ultimate conclusion can be made, the tree is ended by leaf nodes (too known as terminal nodes) that indicate the move to be made as the aftereffect of the arrangement of choices. On account of a prescient model, the leaf nodes give the normal result given the arrangement of occasions in the tree.



C5.0 utilizes entropy, an idea acquired from data hypothesis that measures the arbitrariness, or confusion, inside an arrangement of class values. Sets with high entropy are extremely different and give little data about different things that may likewise have a place in the set, as there is no clear shared characteristic. The choice tree would like to discover parts that lessen entropy, at last expanding homogeneity inside the gatherings. Ordinarily, entropy is estimated in bits. On the off chance that there are just two conceivable classes, entropy qualities can go from 0 to 1. For n classes, entropy ranges from 0 to log2(n). In each case, the base esteem shows that the example is totally homogenous, while the greatest esteem shows that the information are as different as would be prudent, and no gathering has even a little majority.

Equation for entropy calculation :

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \; log_2(p_i)$$

To utilize entropy to decide the ideal component to part upon, the calculation ascertains the adjustment in homogeneity that would come about because of a split on every conceivable include, which is a measure known as information gain. information gain for a feature F is computed as the contrast between the entropy in the fragment previously the split one and the parcels coming about because of the split two.

$$\text{InfoGain}(F) = \text{Entropy}(S_1) - \text{Entropy}(S_2)$$

One complexity is that after a split, the information is isolated into in excess of one segment. Along these lines, the capacity to figure Entropy(S2) needs to think about the aggregate entropy over the majority of the segments. It does this by measuring each segment's entropy by the extent of records falling into the segment. This can be expressed in an equation as:

$$\text{Entropy}(S) = \sum_{i=1}^{n} w_i \; \text{Entropy}(P_i)$$

Rules of classification can likewise be gotten specifically from decision trees. Starting at a leaf node and following the branches back to the root, you will have gotten an arrangement of choices. In our thesis we also figured the rules using decision tree itself.

5.2 Implemented Package in R

R packages are collections of functions and data sets developed by the community. They increase the power of R by improving existing base R functionalists, or by adding new ones.

5.2.1 Package For Support Vector Machine

We use e1071 package for implement support vector machine algorithm. In this package we have a function called *svm* . **svm** is used to train a support vector machine. It can be used to carry out general regression and classification (of nu and epsilon-type), as well as density-estimation. A formula interface is provided.

```
classifierModelSVM = svm(formula = state ~ .,
                         data = training_set,
                         type = 'C-classification',
                         kernel = 'linear')
```

Arguments

- formula a symbolic description of the model to be fit.
- data is a matrix, a vector, or a sparse matrix (object of class Matrix provided by the Matrix package, or of class matrix.csr provided by the SparseM package, or of class simple_triplet_matrix provided by the slam package).
- type of svm can be used as a classification machine, as a regression machine, or for novelty detection. Depending of whether y is a factor or not, the default setting for type is C-classification or eps-regression, respectively, but may be overwritten by setting an explicit value. Valid options are:
  - c-classification
  - n-classification
  - eps-regression
- the kernel used in training and predicting. You might consider changing some of the following parameters, depending on the kernel type. We use linear kernel because our data set show more promising result with this kernel.

5.2.2 Package For Artificial Neural Network

This is a package for running H2O via its REST API from within R. To communicate with a H2O instance, the version of the R package must match the version of H2O. When connecting to a new H2O cluster, it is necessary to re-run the initializer.

This package allows the user to run basic H2O commands using R commands. In order to use it, you must first have H2O running. To run H2O on your local machine, call h2o.init without any arguments, and H2O will be automatically launched at localhost:54321, where the IP is "127.0.0.1" and the port is 54321. If H2O is running on a cluster, you must provide the IP and port of the remote machine as arguments to the h2o.init() call.

H2O supports a number of standard statistical models, such as GLM, K-means, and Random Forest. For example, to run GLM, call h2o.glm with the H2O parsed data and parameters (response variable, error distribution, etc) as arguments. (The operation will be done on the server associated with the data object where H2O is running, not within the R environment).

We use this package for Building a feed-forward multilayer artificial neural network on an H2OFrame. How we implemented this package in our project is showing below .

```
classifierAnn = h2o.deeplearning(y = 'state',
                                 training_frame = as.h2o(training_set_ANN),
                                 activation = 'Rectifier',
                                 hidden = c(8,4),
                                 epochs = 10,
                                 train_samples_per_iteration = -2)
```

Arguments

- y is a name or column index of the response variable in the data. The response must be either a numeric or a categorical/factor variable. If the response is numeric, then a regression model will be trained, otherwise it will train a classification model.
- training_frame is Id of the training data frame.

- Activation function. Must be one of: Tanh, TanhWithDropout, Rectifier. Defaults is Rectifier.
- hidden parameter denote hidden layer sizes (e.g. [100, 100]). Defaults to [200, 200].
- epoch is How many times the dataset should be iterated (streamed), can be fractional. Defaults to 10.
- train_samples_per_iteration Number of training samples (globally) per MapReduce iteration. Special values are 0: one epoch, -1: all available data (e.g., replicated training data), -2: automatic. Defaults to -2.

5.2.3 Package For Naive Bayes

For implementing Naive Bayes we also use e1071 package . We use naiveBayes function for implement this algorithm.This function computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule. Our implementation is showing below

```
classifierNaive = naiveBayes(x = training_set_NV[-3],
                             y = training_set_NV$state)
```

Arguments

- x is a numeric matrix, or a data frame of categorical and/or numeric variables.
- y is a class vector.

5.2.4 Package For Decision Tree

For implementing decision tree we use c5.0 package . From this package we use c5.0 function for creating our model.This function fit classification tree models or rule-based models using Quinlans C5.0 algorithm .
Our implementation is showing below

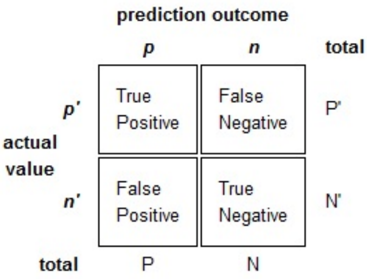
```
classifierDC = C5.0(formula = state ~ .,
                    data = training_set_DC)
```

Arguments

- a formula, with a response and at least one predictor.
- data is an optional data frame in which to interpret the variables named in the formula.

5.3 Model Evaluation Matrices

Confusion matrix
A confusion matrix is a table that arranges predictions as indicated by whether they coordinate the real value. One of the table's measurements demonstrates the possible categories of predicted values, while the other dimension demonstrates the same for actual values. The connection between the positive class and negative class forecasts can be portrayed as a 2 x 2 matrix that organizes whether expectations fall into one of the four categories (TP, TN, FP and FN) see figure 5.2.



		prediction outcome		
		p	*n*	total
actual value	*p'*	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
	total	P	N	

Figure 5.2: Confusion matrix

True positives (TP) Correctly classified as the class of interest. These are cases in which we predicted the state is successful, and the company is successful.

True negatives (TN) Correctly classified as not the class of interest. We predicted the company is canceled or failed, and they are failed or canceled.
False positives (FP) Incorrectly classified as the class of interest. We predicted the company as failed, but they don't actually fail- they are successful. (Also known as a Type I error)
False negatives (FN) Incorrectly classified as not the class of interest. We predicted successful, but they actually failed or canceled. (Also known as a "Type II error.")

Sensitivity
Sensitivity also called the true positive rate, the recall, or probability of detection in some fields; measures the proportion of actual positives that are correctly identified as such. In other words, Ratio of items predicted as Positives which are actually positive versus total number of Positive items. Increase in Sensitivity leads to increase in TPs and decrease in FNs.

$$Sensitivity = TP/(TP + FN)$$

Specificity Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such. In other words, Ratio of items predicted as Negative which are actually negative versus total number of Negative items. Increase in Specificity means increase in TNs and decrease in FPs.

$$Specificity = TN/(TN + FP)$$

Precision
Precision is defined as the fraction of the examples which are actually positive among all the examples which we predicted positive. In pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.

$$Precision = TP/(TP + FP)$$

Recall

We define recall as, among all the examples that actually positive, what fraction did we detect as positive. Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

$$Precision = TP/(TP + FN)$$

F-Measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure.

$$F - Measure = (2 * Precision * Recall)/(Precision + Recall)$$

Chapter 6

Result and Analysis

6.1 Classification model evaluation

In this chapter we are going to evaluate our model performance . These models are created with three different technique . First one is **Black Box Method** , second one is **Probabilistic method** and last one is **Divide and Conquer Method**

6.1.1 Evaluation with Black Box Method

Our model classify the business status of a start-up project . We choose a machine learning method which is sometimes called the Black Box Method [3]. We choose ANN and SVM from Black Box method .

We choose SVM because this algorithm not overly influenced by noisy data and not very prone to over fitting . It gives high accuracy .We choose ANN because this algorithm capable of modeling more complex patterns than nearly any algorithm and also it makes few assumptions about the data underlying relationship .

We have 370000 instances , as mentioned before we split our data set for training and testing set. Training set use for creating model and testing set use for predict the outcome class . In the svm we use "linear" kernel because solving the optimization problem for a linear kernel is much faster [8]. Hence in our data set has fewer feature and we don't need to map data to a higher dimensional space. For this particular situation choosing linear kernel has been the best option for the project. It is also faster to train the model . While applying ANN, we have used the Rectifier Activation Function which is considered as the best to train a neural network model [11] . The first hidden layer has 8 neurons and the second hidden layer has 4 neuron . Accuracy and Kappa statistic results is given below of two algorithms See in table 6.1 .

Algorithm Name	Accuracy	Kappa statistics
SVM	0.815	0.6468
ANN	0.8711	0.7572

Table 6.1: Accuracy and kappa value for SVM and ANN

Here we can see that , accuracy values of both algorithms are 0.815 and 0.8711 respectively . Kappa statistics values are 0.6468 and 0.7572 respectively which denote a good agreement between model prediction and true values.Now we are going to evaluate our model based on each individual class . We took sensitivity and specificity as a model evaluation parameter. In table 6.2 we show our sensitivity and specificity result for both algorithm.

Algorithm Name	Sensitivity	Specificity	Class
SVM	0.0049455	0.9981711	Canceled
SVM	0.9623	0.6565	Failed
SVM	0.8363	0.9640	Successful
ANN	0.0030434	0.9993652	Canceled
ANN	0.9740	0.7603	Failed
ANN	0.9753	0.9744	Successful

Table 6.2: Sensitivity and Specificity for SVM and ANN

From the following table we can say that in svm algorithm correctly identify canceled class .49% and but 99.50% times the class is undetected. 99% specificity shows that 99% times this model correctly identify Non-Cancelled class. For failed class 96% sensitivity measure that 96% times the model correctly identify failed class. 65% specificity measure that 65% times the model correctly identify non-failed class . For successful class 83% sensitivity means that 83% times the model correctly identify successful class. Now we going to describe our finding on ANN algorithm . For canceled class .3% times this model identify correctly. Specificity 99% measures that 99% times this model correctly identify non-cancelled class. Sensitivity 97% for failed class shows that 97% times this algorithm correctly identify failed class. 76% specificity means that 76% times this model correctly identify non-failed class. And last for successful class 97% sensitivity means that 97% times this model correctly identify successful class and specificity is also 97% here which means that our model correctly identify non-successful class by 97% times .

Now we are going to show precision , recall and f-measure value for both algorithm in table 6.3. With this comparison we can show that ANN is better model than SVM model. We got more accuracy on ANN . So based on f-measure we can say that ANN perform much better than SVM. For avoiding biasness of precision and recall we include f-measure value for our observation.If we consider successful class as our main consideration then model build with ANN give us best performance. If we consider failed class as a consideration then also ANN perform better.

Algorithm Name	Precision	Recall	F-measure	Class
SVM	0.004	0.24	0.007	cancelled
SVM	0.96	0.76	0.84	failed
SVM	0.83	0.92	0.87	successful
ANN	0.002	0.48	0.003	cancelled
ANN	0.97	0.82	0.88	failed
ANN	0.97	0.96	0.96	successful

Table 6.3: Precision, Recall and F-measure for SVM and ANN

Now we are interested in a particular domain that whether model performance improves if we increase our data .To evaluate this statement , we create three separate classification models .First model is trained with 100 thousand data , second model is trained with 200 thousand data and third model is trained with 300 thousand data . See in table 6.4 it proves that our model performance increases when we enrich our data from 100 thousand to 300 thousand for both algorithm. In SVM, accuracy increases from 79% to 81% and kappa values increases from 0.60 to 0.64. But ANN accuracy is not increase while its kappa value increases slightly from 0.73 to .74

Algorithm Name	Accuracy	Kappa statistics	Data Size
SVM	0.7941	0.6024	100k
SVM	0.803	0.6221	200k
SVM	0.815	0.6468	300k
ANN	0.8606	0.7376	100k
ANN	0.8669	0.7496	200k
ANN	0.8657	0.7489	300k

Table 6.4: Evaluate model performance by amount of data

6.1.2 Evaluation Based On Probability method

Then we have applied the Nave Bayes algorithm which is widely uses as a classification algorithm. Classifiers based on Bayesian methods utilize training data to calculate an observed probability of each outcome based on the evidence provided by feature values. When the classifier is later applied to unlabeled data, it uses the observed probabilities to predict the most likely class for the new features. It's a simple idea, but it results in a method that often has results on par with more sophisticated algorithms [4]. We choose Naive Bayes algorithm because it is simple , fast and very effective . It works well on noisy and missing data . Requires relatively few example for training , also work well with very large number of example. Accuracy and kappa statistics for Naive Bayes see in table 6.5

.

Algorithm Name	Accuracy	Kappa statistics
Naive Bayes	0.6421	0.2763

Table 6.5: Accuracy and kappa value for Naive Bayes

This result shows the accuracy value and kappa statistics value is comparatively low than the previous two models. Kappa statistics value denotes that there is a fare agreement between model prediction and true values . Now we are going evaluate this model based on sensitivity and specificity in table 6.6.

Algorithm Name	Sensitivity	Specificity	Class
Naive Bayes	0.020796	0.988573	Canceled
Naive Bayes	0.9759	0.2767	Failed
Naive Bayes	0.3319	0.9844	Successful

Table 6.6: Sensitivity and Specificity for Naive Bayes

From this table we can illustrate that 20% sensitivity for cancelled class means that 20% time our model correctly identify the cancel class and 98% specificity measures that our model correctly identify non-cancelled class by 98% times. Now for failed class 97% sensitivity means that 97% times our model correctly predict the failed class and specificity 27% means that our model correctly identify non-failed class by 27%. For successful class our model correctly identify successful class by 33% and correctly identify non-successful class by 98%. Now we show the model performance based on precision , recall and f-measure in table 6.7.

From the table we clearly show that our model performance based on f-measure NV perform lower than ANN and SVM. f1 score for our three main class respectively is .03 , .74 and .48. SVM perform better on classify failed class.

Like previously, now our goal is to answer the question Is the model performance

Algorithm Name	Precision	Recall	F-measure	Class
Naive Bayes	0.02	0.17	0.03	cancelled
Naive Bayes	0.97	0.6	0.74	failed
Naive Bayes	0.33	0.92	0.48	successful

Table 6.7: Precision, Recall and F-measure for Naive Bayes

is improved if we increase the amount of data.So we have created three different models which are trained with 100,200,300 thousand of data respectively. see in table 6.8.

Algorithm Name	Accuracy	Kappa statistics	Data Size
Naive Bayes	0.6674	0.3399	100k
Naive Bayes	0.6581	0.3172	200k
Naive Bayes	0.6381	0.2715	300k

Table 6.8: Evaluate model performance by amount of data

from this table 6.7 it is shown that Naive Bayes accuracy is quite low than previous two algorithm . Even if we increase our data accuracy remains the same. . Where SVM achieved 79% accuracy for 100 thousand data and ANN achieved 86% accuracy . Naive Bayes gain accuracy 66% for 100 thousand data and 63% for 300 thousand data .Here is an interesting fact , when we are increasing the amount of data, our data accuracy level decreases which means that this algorithm doesnt perform well on this specific problem domain .

6.1.3 Evaluation Based On Divide and Conquer Method

Next we implemented our last machine learning technique called **Divide and Conquer** . We choose C5.0 decision tree algorithm for creating our model . Decision trees are built using a heuristic called recursive partitioning. This approach is

also commonly known as divide and conquer because it splits the data into subsets, which are then split repeatedly into even smaller subsets, and so on and so forth until the process stops when the algorithm determines the data within the subsets are sufficiently homogeneous, or another stopping criterion has been met [5].
We choose decision tree for this particular business domain because this is an all-purpose classifier that perform well on most problems . Highly automatic learning process , which can handle both categorical and numerical data . Most beautiful feature of this algorithm is it can exclude unimportant feature [9]. Here is the accuracy and kappa statistics value for decision tree table 6.9

.

Algorithm Name	Accuracy	Kappa statistics
Decision Tree	0.8923	0.7969

Table 6.9: Accuracy and kappa value for Decision Tree

Now we are going to evaluate this model by class . We choose sensitivity and specificity like before for evaluating this model see in 6.10.

Algorithm Name	Sensitivity	Specificity	Class
Decision Tree	0.0055	0.9992	Canceled
Decision Tree	0.9976	0.7748	Failed
Decision Tree	0.9991	0.9942	Successful

Table 6.10: Sensitivity and Specificity for Decision Tree

From the table 6.10 we can show that for cancelled class sensitivity is .5% which is very low and its specificity is 99% which denote this model can correctly identify non-cancelled class. Now failed class sensitivity has 99% sensitivity that shows our model correctly identify failed class by 97%. At last for successful class our model correctly identify successful class by 99% . Which shows we had higher sensitivity for failed and successful class. Now we are going to evaluate the model based on precision , recall and f-measure see table 6.11 from the f-measure value it clearly show that our model can classify failed and successful class by 99%.

Algorithm Name	Precision	Recall	F-measure	Class
Decision Tree	0.005	0.57	0.009	cancelled
Decision Tree	0.99	0.83	0.9	failed
Decision Tree	0.99	0.99	0.99	successful

Table 6.11: Precision, Recall and F-measure for Decision Tree

From the table 6.9 we can see that decision tree obtain highest accuracy than any other algorithm . Its accuracy is 89% and kappa value is approximately .8 . Which mean this which denotes a very good agreement between model prediction and true values . Now like previous we also divide our data into 100 thousand , 200 thousand and 3 thousand and create three individual models to evaluate the model performance based on increase level of the data . See in table 6.12

Algorithm Name	Accuracy	Kappa statistics	Data Size
Decision Tree	0.895	0.8014	100k
Decision Tree	0.8934	0.7988	200k
Decision Tree	0.8916	0.795	300k

Table 6.12: Evaluate model performance by amount of data

6.1.4 Model evaluation in a Nutshell

In this section we are going to show our over all model performance with the help of visible graph so that we can visualize our model performance so easily. First graph is about our model performance based on accuracy and kappa value see 6.1. From the graph we clearly show that ANN and DC performance is higher than any other algorithm model. Because both algorithm has high accuracy and kappa value.

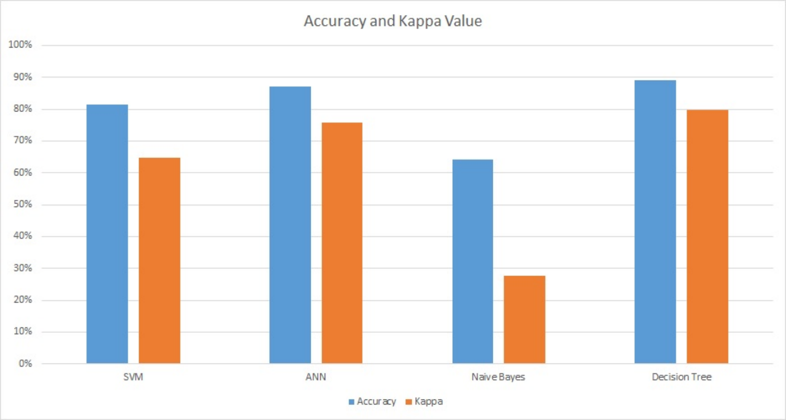


Figure 6.1: Accuracy and Kappa value for all model

Now as we told before that we divide our data into 100 thousand , 200 thousand and 300 thousand instances. Our goal is to check whether our model performance increase if we increase data in our model training.Now see figure 6.2, 6.3 and 6.4 . From those figure it clearly show that if we increase our data in our train model then model performance increases . But for the Naive Bayes algorithm model if we increase our data then model performance decrease .

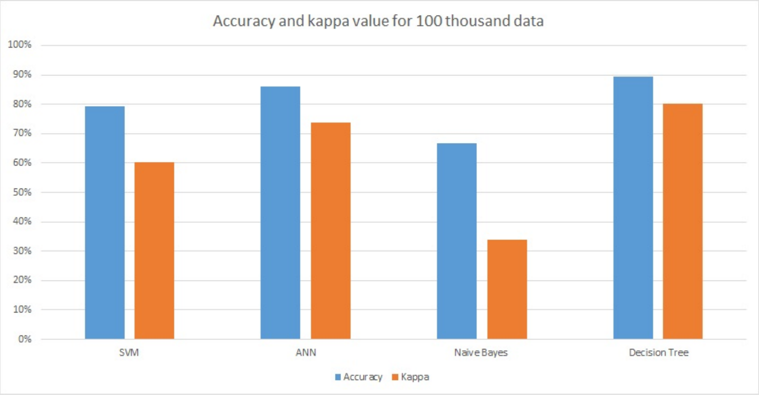


Figure 6.2: Accuracy and kappa value for 100 thousand

We evaluated our model based on precision , recall and f-score for individual outcome class see in figure 6.5 , 6.6 and 6.7.With this comparison we can show

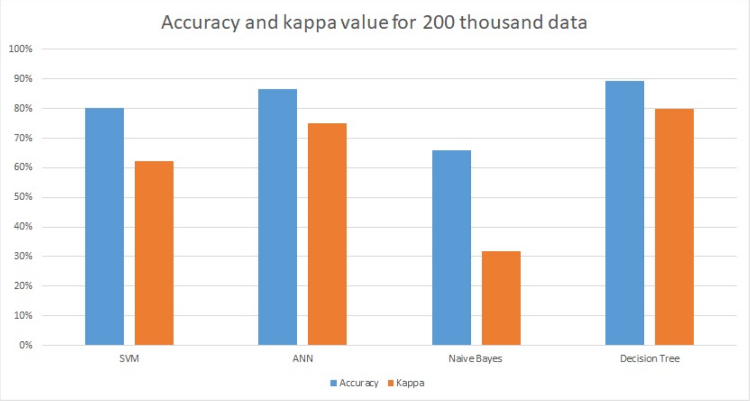


Figure 6.3: Accuracy and kappa value for 200 thousand

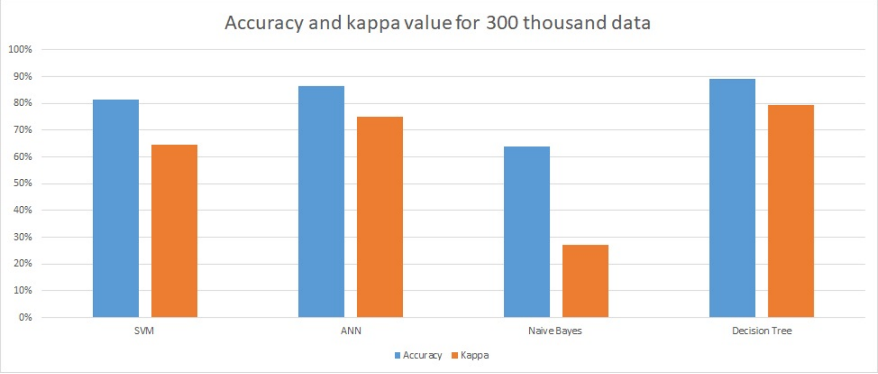
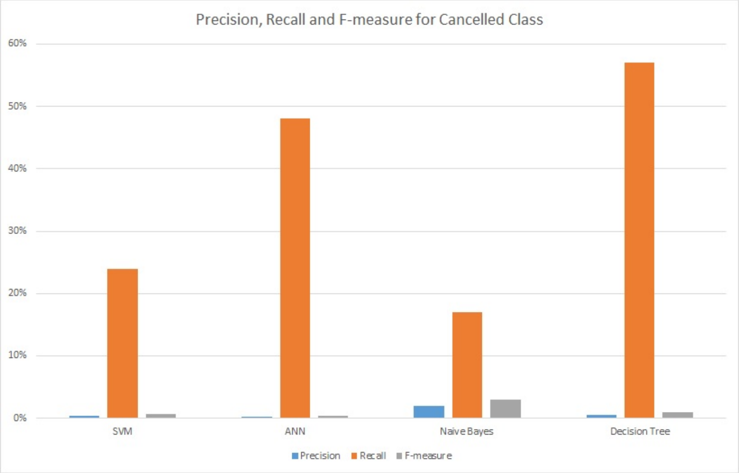


Figure 6.4: Accuracy and kappa value for 300 thousand

that ANN and Decision tree is better model than any other model. We got more f score value in Decision tree and ANN model. So based on f-measure we can say that ANN and Decision tree perform much better than SVM and Naive Bayes. For avoiding bias of precision and recall value we choose to evaluate our model performance based on f score value.If we consider successful class as our main consideration then model build with Decision tree give us best performance. If we consider failed class as a consideration then ANN perform better.



.

Figure 6.5: Precision, Recall and F-measure for Cancelled Class

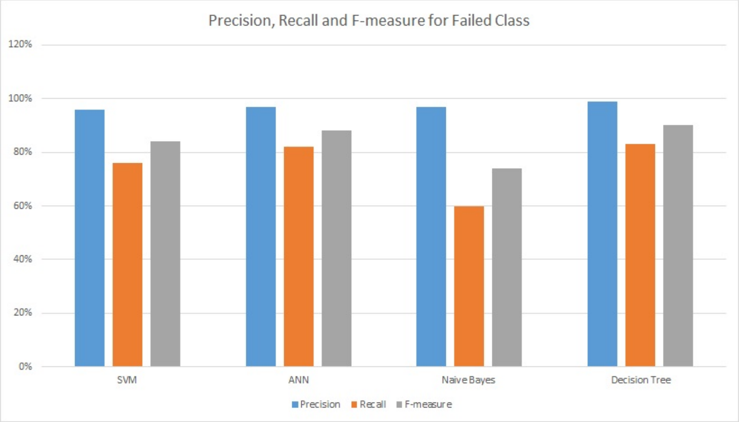


Figure 6.6: Precision, Recall and F-measure for Failed Class

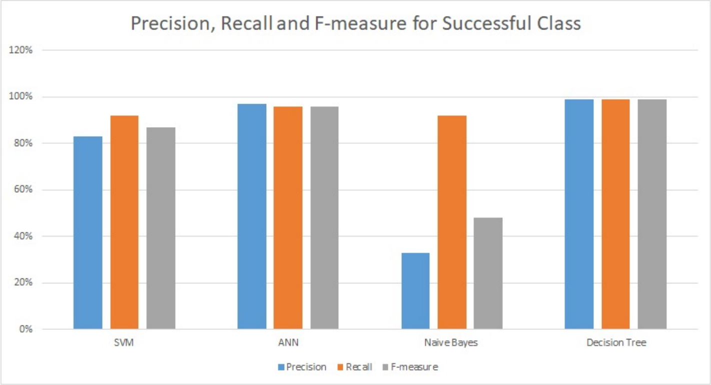


Figure 6.7: Precision, Recall and F-measure for Successful Class

6.2 Rule Extraction

Rule is extracted by two popular algorithm one is ripper algorithm and another is c5.0 algorithm. Both algorithm shows highest accuracy and lowest miss classification rate rule . This rule can be used as business purpose .

6.2.1 Finding Rule By Ripper Algorithm

We find 94 business rules by by ripper algorithm. We choose those rules which cover at least 2000 instances from our data set.

Rule 1 if(backers >= 21 && pledged >= 4926.8 && pledged <= 6894.16 && real >= 2515 && real <= 5397.49 =>) then class = successful ***Rule 2*** if(backers >= 18 && real <= 2515.99 && pledged >= 2026.68 && main_category = Music) then class = successful

Rule 3 if(backers >= 18 && real <= 4571.43 && pledged >= 3740.69) then class = successful

Rule 4 if(backers >= 18 && pledged >= 2496.64 && real <= 2844.14) then class = successful.

Rule 5 if(backers >= 18 && pledged >= 6894.16 && real <= 8407 &&

main_category = Music) then class = successful.

Rule 6 if(backers >= 16 && pledged >= 10000.01 && real <= 10280) then class = successful.

Rule 7 if(backers >= 18 && pledged >= 1487.12 && real <= 1889.29) then class = successful.

Rule 8 if(backers >= 14 && pledged >= 15145 && real <= 17541.6) then class = successful.

Rule 9 if(backers >= 13 && real <= 1307.53 && pledged >= 996.32) then class = successful.

Rule 10 if(backers >= 14 && pledged >= 5998.95 && real <= 6395.66) then class = successful.

Rule 11 if(backers >= 13 && pledged >= 25013.9 && real <= 25575.45) then class = successful.

Rule 12 if(pledged >= 3000 && real <= 3340.5 && backers >= 36) then class = successful.

Rule 13 if(backers >= 11 && real <= 665.76 && pledged >= 590.12) then class = successful.

Rule 14 if(backers >= 14 && real <= 2250 && pledged >= 1901) then class = successful.

Rule 15 if(pledged >= 30045.5 && real <= 36108.9) then class = successful.

Rule 16 if(pledged >= 49691.54 && real <= 59640.07) then class = successful.

Rule 17 if(backers >= 10 && real <= 549.99 && pledged >= 493.26) then class = successful.

Rule 18 if(real <= 23984.48 && pledged >= 20016.01) then class = successful.

Rule 19 if(real <= 339.24 && pledged >= 278.89) then class = successful

6.2.2 Finding Rule By C5.0 Algorithm

Next we extract rule by using C5.0 algorithm. We find total 67 rule here. We choose those rules which cover at least 10000 instances from our data set.

Rule 1 if(pledged <= 246 && goal > 244.99 && goal <= 1425) then class = failed

Rule 2 if(pledged <= 1469.09 && goal > 1425) then class = failed

Rule 3 if(backers > 14 && pledged <= 9986.01 && goal > 9939.6 && goal <= 20038.32) then class = failed

Rule 4 if(pledged <= 14970.06 && goal > 14950 && goal <= 20038.32) then class = failed

Rule 5 if(pledged <= 13682.15 && goal > 13279.18 && goal <= 20038.32) then class = failed

) ***Rule 6*** if(pledged <= 6473 && goal > 6097.75) then class = failed

Rule 7 if(pledged <= 24987.14 && goal > 24927.3 && goal <= 51625) then class = failed

Rule 8 if(pledged <= 17990 && goal > 17994.03) then class = failed

Rule 9 if(pledged <= 21437.71 && goal > 20038.32) then class = failed

Rule 10 if(backers <= 14) then class = failed

Rule 11 if(pledged <= 38157.52 && goal > 37868.33) then class = failed

Rule 12 if(pledged <= 49972.2 && goal > 49755) then class = failed

Rule 13 if(pledged <= 54169 && goal > 51625) then class = failed

Rule 14 if(pledged <= 100250 && goal > 99999) then class = failed

Rule 15 if(main_category in Art, Comics, Dance, Film Video, Food, Music,Photography, Publishing, Theater && pledged > 6473 && goal <= 9939.6) then class = successful

Rule 16 if(pledged > 2487.07 && goal > 1425 && goal <= 3197.16) then class = successful

Rule 17 if(pledged > 11347 && goal > 6097.75 && goal <= 13279.18) then class = successful

Rule 18 if(pledged > 1469.09 && goal <= 2032.52) then class = successful

Rule 19 if(pledged > 29885 && goal <= 37868.33) then class = successful

Rule 20 if(pledged > 897.4 && goal <= 1425) then class = successful

Rule 21 if(pledged > 587.12 && goal <= 895.68) then class = successful

Rule 22 if(pledged > 3192.31 && goal <= 5047.62) then class = successful

Chapter 7

Conclusion

Data mining is an important process to discover knowledge about your customer behavior towards your business offerings. It explores the unknown credible patterns those are significant for business success.According to Doug Alexander of the University of Texas it is actually defined as the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of data. With data mining, Business Organizations are able to make more accurate business decisions and incur more profits. From business, marketing advertising and introduction of new products or services, and everything in between. Here in our research we developed a model that can help to take decision in crowd funding based business problem. It will decrease the young entrepreneur risk for starting a new business. We also generate some if then rules that can also help to gain business success .

Bibliography

[1] *How real businesses are using machine learning* https://techcrunch.com/2016/03/19/how-real-businesses-are-using-machine-learning/

[2] *THE BUSINESS OF ARTIFICIAL INTELLIGENC* https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence/

[3] Brett Lantz *Machine Learning With R*, Page 219

[4] Min-Ling Zhang , Jos M. Pea ,Victor Robles *Feature selection for multi-label naive Bayes classification*,

[5] Brett Lantz *Machine Learning With R*, Page 126

[6] M. Vandromme , J. Jacques ,J. Taillard, A. Hansske, L. Jourdan, C. Dhaenens *Extraction and optimization of classification rules for temporal sequences: Application to hospital data*,

[7] S.Vijayarani, M.Divya *An Efficient Algorithm for Classification Rule Hiding*

[8] E. Deepak, G. Sai Pooja , R. N S Jyothi *SVM kernel based predictive analytics on faculty performance evaluation*

[9] Mu-Yen Chen *Predicting corporate financial distress based on integration of decision tree classification and logistic regression*

[10] https://www.kaggle.com/kemical/kickstarter-projects

[11] Bing Xu , Naiyan Wang , Tianqi Chen , Mu Li *Empirical Evaluation of Rectied Activations in Convolution Network*

[12] Pang-Ning Tan , Michael Steinbach, Anuj Karpatne,Vipin Kumar,*Introduction to Data Mining*

[13] Ruxandra PETRE, *Data Mining Solutions for the Business Environment*

[14] Business Intelligence Using Data Mining Techniques on Very Large Datasets Arti J. Ugale1 , P. S. Mohod

[15] Wei Chao Lin, Top 10 data mining techniques in business applications: a brief survey

[16] Ali Radhi Al Essa, Data Mining and are Housing

[17] Qasem A. Al-Radaideh, Eman Al Nagi, Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance

[18] Allen, M.W., Armstrong, D.J., Reid, M.F., Riemenschneider, C.K. (2009). IT Employee Retention: Employee Expectations and Workplace Environments, SIGMIS-CPR09 , May 2009, Limerick, Ireland.

[19] Al-Radaideh, Q. A., Al-Shawakfa, E.M., Al-Najjar, M.I. (2006). Mining Student Data Using Decision Trees, International Arab Conference on Information Technology (ACIT 2006), Dec 2006, Jordan.

[20] Chein, C., Chen, L. (2006) ”Data mining to improve personnel selection and enhance human capital: A case study in high technology industry”, Expert Systems with Applications, In Press.

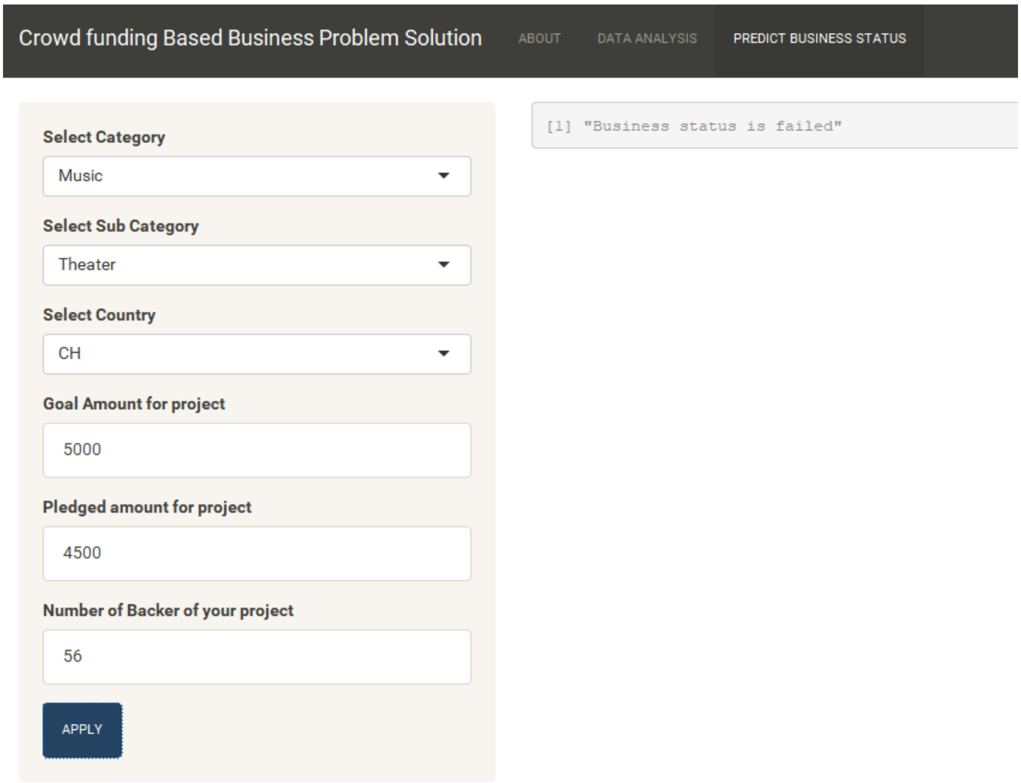
[21]] Cho, S., Johanson, M.M., Guchait, P. (2009). ”Employees intent to leave: A comparison of determinants of intent to leave versus intent to stay”, International Journal of Hospitality Management, 28, pp374-381.

[22] 1S.Vijayarani, 2M.Divya *An Efficient Algorithm for Generating Classification Rules*

Appendix A

Appendix

We develop a web application that can predict business status based on user input. Our young entrepreneur can input their business information here and after submitting save button then get predicted result whether their business should be successful or failed. Here is a screen shot about our developed web application see figure A.1



Select Category
Music

Select Sub Category
Theater

Select Country
CH

Goal Amount for project
5000

Pledged amount for project
4500

Number of Backer of your project
56

```
[1] "Business status is failed"
```

Figure A.1: Web application for predicting status