2016 IEEE 16th International Conference on Data Mining Workshops

Predicting the Outcome of Startups: Less Failure, More Success.

Amar Krishna, Ankit Agrawal, Alok Choudhary*

Abstract

On an average 9 out of 10 startups fail(industry standard). Several reasons are responsible for the failure of a startup including bad management, lack of funds, etc. This work aims to create a predictive model for startups based on many key things involved at various stages in the life of a startup. It is highly desirable to increase the success rate of startups and not much work have been done to address the same.

We propose a method to predict the outcome of a startups based on many key factors like seed funding amount, seed funding time, Series A funding, factors contributing to the success and failure of the company at every milestone. We can have created several models based on the data that we have carefully put together from various sources like Crunchbase, Tech Crunch, etc.

Several data mining classification techniques were used on the preprocessed data along with various data mining optimizations and validations. We provide our analysis using techniques such as Random Forest, ADTrees, Bayesian Networks, and so on. We evaluate the correctness of our models based on factors like area under the ROC curve, precision and recall. We show that a startup can use our models to decide which factors they need to focus more on, in order to hit the success mark.

Keywords. Startups, Weka, Precision, Accuracy, Prediction

1 Introduction

The first success of a startup begins with a great idea which later turns into a great hypothesis. A significant portion of entrepreneurs/innovators attempting to establish a business lead to failure. As per the statistics, 9 out of 10 startups fail. It has always been the need of entrepreneurs to know the key factors involved in creating a successful company. Each and every entrepreneur wants his/her hypothesis to work out which can further lead to a successful enterprise. They want to create a product which is liked by their customers and at the same time the company manages to get enough traction. Few factors which help creating a successful enterprise are traction, capital, management, skilled individuals, viable product, so on.

Few studies have been done trying to figure out the real reasons behind a startup failure. Many companies (tech/non-tech) have been working on the same issue for quite sometime now. One of the works done is deciding the success failure factors in the pre-startup phase [15]. Our work tries to create an accurate predictive model to predict if a startup will succeed or fail.

Our work involves, the data mining analysis of more than 11,000 companies, data (7,000 companies still in operation and 4,000 closed/acquired companies). We modeled our data into 9 model sets where each model represents a different milestone for the company. Each model was created based on the facts like seed funding, series A funding, etc. We analyze this data based on key factors like when the company was founded, how much seed funds it raised, how many months it took to raise the seed funds, factors which were affecting the growth of the company both positive and negative. Experiments with more than 30 classifiers were conducted to find that many meta classifiers used with decision trees can give impressive results, which can be further improved by combining the resulting prediction probabilities from several classifiers. Our results were represented in terms of parameters like AUC (Area Under the Curve), Recall Values and Precision Values. The figures for these parameters were pretty impressive which further led to developing a successful prediction model.

The rest of the paper is organized as follows: Section 2 summarizes the recent research relevant to the problem, followed by a description of the major classification schemes used in this study in Section 3. Key factors involved in our analysis is discussed in Section 4. The success/failure prediction system is presented in Section 5. Experiments and results are presented in Section 6 and the conclusion and future work is presented in Section 7.

2 Related Work

Failures of startups have drawn massive attention and most of the companies are working on designing various kinds of prediction/futuristic models to successfully predict the fate of a new company. Few researchers have done some interesting work trying to find the success/failure patterns of a startup. One of the works discusses the success and risk factors involved in a pre-

*EECS Department, Northwestern University.IL-60208.

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.103

startup phase [15]. The authors focus on estimating the relative importance of a variety of approaches and variables in explaining pre-startup success. They created a framework, which suggests that start-up efforts differ in terms of the characteristics of the individual(s) who start the venture, the organization that they create, the environment surrounding the new venture, and the process by which the new venture is started.

The work done in paper [4] closely addresses our problem. Research on personality characteristics relates dispositions such as risk-taking, locus of control, and need for achievement to the emergence and the success of entrepreneurship (for an overview, see [13]).

Greenewood et al. [8] have studied differences in motives as a success factor in nascent entrepreneurship. They find that women who start for internally oriented reasons, and men who start for externally oriented reasons (like perceiving a need in the market) have greater chances of successfully completing the pre-start-up phase. Another work was on crowd sourcing which gets a mention in [7]. In this paper the authors focus on how crowd sourcing can help creating a successful organization. Work of paper [14] focuses on developing a research program to investigate the major factors contributing to success in new technical ventures. Another way to construct networks is through the strategic alliances between firms. Another work on new venture failure is done in the paper [12]. In this paper, the authors demonstrate two ways to investigate new venture failure - testing for moderating effects of new venture failure on the relationship between startup experience and perceived startup expertise with a sample of 220 entrepreneurs; and second, by qualitatively exploring the nature of these relationships, drawing insights from interviews with these 220 entrepreneurs.

Different research has been done trying to figure out several aspects of entrepreneurship and how some of them can lead to a successful company. Work done in paper [2] addresses similar issues. Another famous work is by R. Dickinson in his article [5] where he discusses the critical success factors and small businesses. He also explains how these factors can be tweaked to create a successful enterprise. Article [6] discusses a lot of problems faced by innovators. This article focuses more on the hurdles faced by the innovators in terms of capitol, management etc. [10] is one article which address the market orientation for entrepreneurs. Research paper [11] discusses factors which can create successful companies.

3 Classification Techniques

Classification/regression techniques in machine learning are a supervised learning techniques that can used a labeled dataset to build predictive models for an outcome/target variable based on several independent input variables. Such techniques have been used with great success in various fields of science and engineering such as health-care, social media [17], materials science and engineering [1], and so on. We have used several classification techniques available through Weka [9]. Out of 30 techniques used, we have limited our results to 6 best techniques. We will define the top few techniques.

1. **Lazy lb1**: In most clustering algorithms, you fix the number of clusters to find, which means the number of instances inside the cluster depends on the size of the dataset. Nearest neighbor search algorithms (like IB1 and IBk), on the other hand, are fixing the number of instances that they consider as neighbors for each single instance; they generate as many neighborhoods as there are instances in the dataset. Defined in the paper named Lazy Associative Classification [16], the author overcomes this problem by focusing on the features of the given test instance, increasing the chance of generating more rules that are useful for classifying the test instance.

2. **Random Forest** : The Random Forest [3] classifier consists of multiple decision trees. The final class of an instance in a Random Forest is assigned by outputting the class that is the mode of the outputs of individual trees, which can produce robust and accurate classification, and ability to handle a very large number of input variables. It is relatively robust to overfitting and can handle datasets with highly imbalanced class distributions.

3. **NaiveBayes** : A Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood.

4. **ADTree**: ADTree is an AND/OR graph. Knowledge contained in the tree is distributed as multiple paths must be traversed to form predictions. Instances that satisfy multiple splitter nodes have the values of prediction nodes that they reach summed to form an overall prediction value. A positive sum represents one class and a negative sum the other in the two-class setting. The result is a single interpretable tree with predictive capabilities.

5. **Bayesian Network**: It is a dual natured Bayesian classification technique which contains two stages:- first learn a network structure, then learn the probability tables. A full Bayesian Network is used as the structure and a decision tree is learned for each CPT. The resulting model is called Full Bayesian network classifiers (FBCs). In learning an FBC, learning the decision trees for CPTs (conditional probability tables) captures essentially both variable independence and context-specific independence.

6. **SimpleLogistic**: Uses a stage-wise fitting process to construct the logistic regression models that can select relevant attributes in the data in a natural way, and can be used to build the logistic regression models at the leaves by incrementally refining those constructed at higher levels in the tree.

4 Key Factors Involved

The lifetime of a startup involves a lot of key factors, which are sometimes equally important for their success or failure. We took in consideration more than 20 key factors which led to the success or failure of the companies. Key factors like Seed funding, Series A funding, Series B funding, Severity scores are part of the 9 models we have created, mentioned in Table 1. Few of the key factors are defined as follows:-

(a) **Seed funding**: Seed funding acts as an initial fuel for any startup. Seed fund helps startups through their initial hurdles and at the same time act as an accelerator too.

(b) **Time to get seed funding**: Time and money are the two important parts of any startups. No start up wants to spend tones of time just to get a product in market which no one wants to buy. This parameter measures the number of months it took for the companies to get a initial seed/angel funding.

(c) **Rounds of Funding**: Refers to the number of rounds of funding a company raises/raised. This includes the seed/angel investment and venture capital.

(d) **Severity factors**: These are the most important factors contributing to the authenticity of our prediction models. These factors and their corresponding scores are used by many institutions like S&P, to evaluate companies. We have divided them into two segments. One is positive factors like : Plenty of traction, low burn rate, good management system, good use of funds and time, a vision to monetize from the very beginning, social skills-networking with the targeted audience, discipline, determination, ability to adapt to changes, fund raising skills, unwavering belief, the composition of capital structure, prospects of future earnings.

Table 1: Key Factors in a startup's timeline

Factors	Description
Start Date	When the company was founded
Seed Funding	Initial funds raised by a startup
Total Rounds of Funding	Includes seed and venture capital
Time for Seed(in months)	Months it took to raise seed funds
Severity Scores	Factors responsible for company's growth/fallout
Average Severity Score	Mean of the positive negative scores
Weighted Average	Weighted Average of the positive negative scores
Series A, B , C..., G funding	Venture Round funding
Valuation	Valuation of the company after each round of funding
Defunct Date	Date when the company dead-pooled(failed companies)
Months Active	Number of months the company is active in market
Market Value	Current Market Value of the company
Total Funds	Seed funds + Venture Funds
Burn Rate	Total fund/No of months the company is active

Table 2: Positive Factors Ranking(on a scale of 5, followed by major institutions)

Reason	Value
Determination	3
Fundraising	4
Execution	5
Social Skills	4
Discipline	3

Negative factors:- A small similar or non-scalable idea, No competitive research - wrong market positioning, no go-to-market strategy, no focus-lack of traction, no flexibility, no passion or persistence, wrong or incomplete leadership, unmotivated team, no mentors or adviser's, no revenue model, high burn rate, less capital than needed, no VC experience, no long term road-map for Return of Investment, bad luck or timing, market competition. Further we rate these factors on a scale of 1-5 (positive), few examples depicted in Table 2, with one being less severe and 5 being more severe and -1 to -5 (negative) with -1 being less severe and -5 more severe, few examples depicted in Table 3.

Table 3: Negative Factors Ranking(on a scale of -1 to -5)

Reason	Value
Badluck or timing	-5
Bad Revenue Model	-5
No Flexibility	-2
High Burn Rate	-5
No roadmap	-3

7. **Series A funding(Venture Round)**: This acts as a catalyst after the initial seed round. This also helps the company getting enough traction in market and from investors. At the same time helps the company to scale. All the venture rounds are termed as Series A, B .. G.

8. **Valuation after each round**: Valuation of a company is decided by the amount of funds it raises at each seed round or venture round. This is a key factor in deciding, if the company is going in a right direction or not. In our analysis, we calculate the valuation of the company after each round of funding. After seed round the valuation of the company is calculated using the formula (100*(seed amount)/15). In case of Series A round, the valuation is calculated using the formula (100 *(Series A amount)/8). For series B to G the formula for calculating the valuation is same as (100 *(Series A/B/D/E/F/G)/5). The valuation at Series B-G tend to stay of the same order because later stages of funding brings capital on less equity.

9. **Burn Rate**: This is the amount of time a company takes in order to burn all of its fund/cash. On general, higher the burn rate the higher the chances of failure, similarly lower the burn rate the company will sustain for a longer period of time.

5 Startup Success/Failure Prediction System

Data preprocessing is the one of the key steps in the whole data mining process. Preprocessing involves understanding and cleaning the data for further analysis. Therefore careful preprocessing of raw startup data collected from Crunchbase holds a great importance in the entire process. The proposed predictive systems consists of 4 stages. They are:

1. **CrunchBase data preprocessing**: This the first step of the entire research. The principal steps involved are :
   (a) Convert apparently numeric attributes to nominal e.g. Seed Funding (Yes or no).
   (b) Construct the time attributes (like seed funding date, Series A funding date etc), we take in consideration the number of months it took to reach every milestones.

2. **Predictive Modeling**: This is where data mining classifiers are employed to construct predictive models for startups success/failure, on the preprocessed data. The two straightforward steps of this stage are:
   (a) Split the preprocessed data in training and testing sets (or use cross validation).
   (b) Construct a model on the training data using data mining classifiers, e.g. Naive Bayes, logistic regression, decision trees, etc., including an ensemble of different classifiers.

3. **Evaluation**: This stage mostly comprises of evaluation of predictive model on testing data.
   (a) Compare the success/failure predictions from the predictive model on unseen data (testing set) against known successful/failed companies.
   (b) Calculate the performance metrics like accuracy (percentage of predictions that are correct), precision (percentage of positive predictions that are correct), recall/sensitivity (percentage of positive labeled records that were predicted as positive), specificity (percentage of negatively labeled records that were predicted as negative), area under the ROC curve (a measure of discriminative power of the model), etc.

6 Evaluation and Results

In our study we collected the data of 7000 successful companies and 4000 failed companies from Crunchbase (a wiki like database for all companies). This data consists of several factors as explained earlier (collected from other sources like TechCrunch and Forbes). In our experiments, we used the WEKA toolkit for classification, analysis and modeling. During our preprocessing step, we concluded that there are many key factors which significantly change the predictive models. Some of them are seed funding amount the company has raised or the rounds of funding it has gone through, if

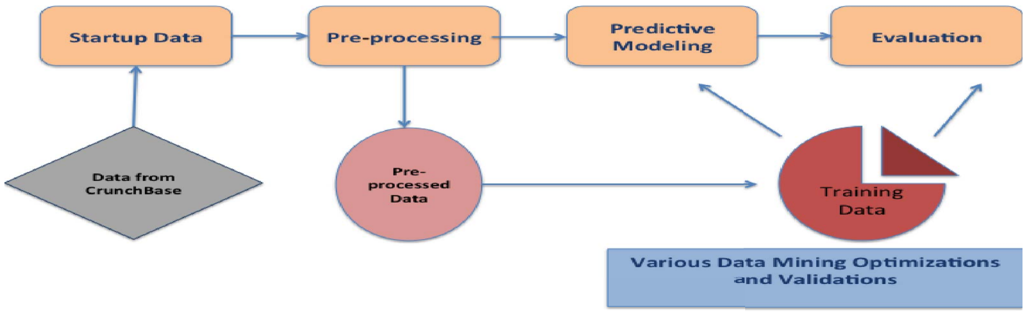

Figure 1: **Block Diagram of Success/Failure Prediction System**

at all the company was able to raise seed funds etc. Also two important key factors which also affected the outcomes were the Burn Rate of the company and few severity factors (mentioned in Section 4).

We collected data of the companies from 1999 (dot com bubble) to 2014. The prime reason for considering this time window is two recessions period and two dot com bubbles which can give us a clear picture of how companies can thrive even when the economy crashes and how some companies go bankrupt. We took into account almost 70 factors for each company. We did not consider the companies which were not able to raise any kind of funding. We did not consider companies with potential legal issues (rare cases of failure). We created several milestones in order to create different models e.g. M0 model refers to company with factors like start date, seed fund raised or not. M1 refers to M0 + seed funding amount. Similarly M2 refers to M1 + Series A funding amount and so on. Similarly, we were able to develop a total of 9 models.

For classification, we developed predictive models using more than 30 different classification schemes. The labels for these classes were if the company has failed or succeded based the data obtained from CrunchBase. We selected the following top 6 schemes:

1. NaiveBayes
2. ADTrees
3. BayesNet
4. Lazylb1
5. RandomForest
6. SimpleLogistics

Table 4: Values of AUC, Recall, Precision and Accuracy for the models

Model	Description	Best AUC	Best Precision	Best Recall	Accuracy
M0	Doesn't include seed fund amount	0.616	0.733	0.783	0.793
M1	Includes seed fund amount	0.865	0.863	0.864	0.883
M2	Includes Series A amount	0.89	0.881	0.888	0.895
M3	Includes Series B amount	0.87	0.875	0.888	0.895
M4	Includes Series C amount	0.872	0.867	0.861	0.887
M5	Includes Series D amount	0.914	0.884	0.888	0.895
M6	Includes Series E amount	0.92	0.864	0.861	0.887
M7	Includes Series F amount	0.92	0.879	0.861	0.889
M8	Includes Series G funding	0.91	0.87	0.88	0.888
M9	Includes all the factors(as mentioned in Table 1)	0.972	0.963	0.966	0.987

We conducted experiments with the above mentioned 6 classification schemes. We used Leave-One-Out Cross Validation (LOOCV) for evaluation, which in this case is equivalent to 11000-fold cross validation.

Since accuracy results can be often misleading due to imbalanced classes, the area under the ROC curve (AUC) is considered a better metric to measure the ability of the model to discriminate between the different class values. Fig. 2 presents the area under the ROC curve (AUC) for the same. We also present the recall values for our dataset.

Figure 4 shows the recall values for all the top 6 classifiers. High precision means that an algorithm returned substantially more relevant results than irrelevant, while high recall means that an algorithm returned most of

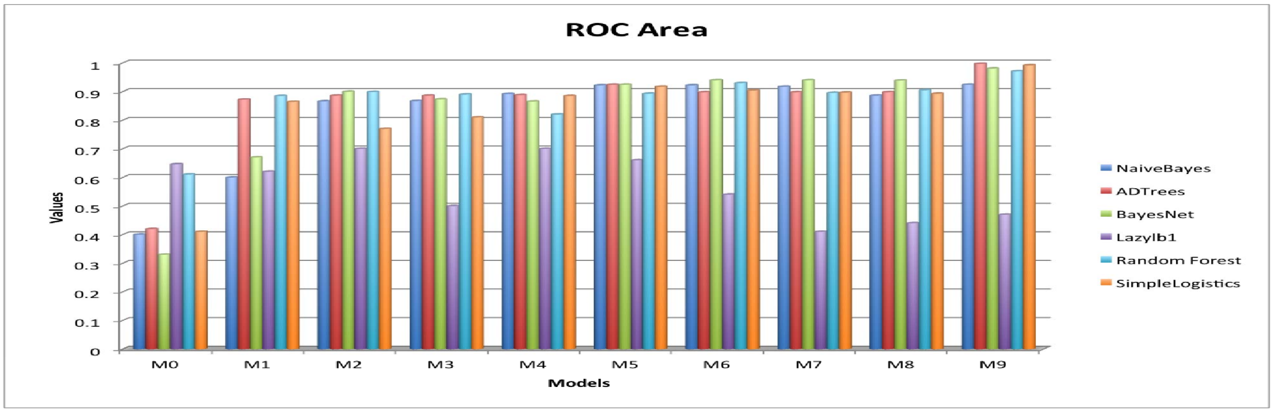

Figure 2: **Area under the curve (ROC) values for different classification schemes for Models M0-M9**

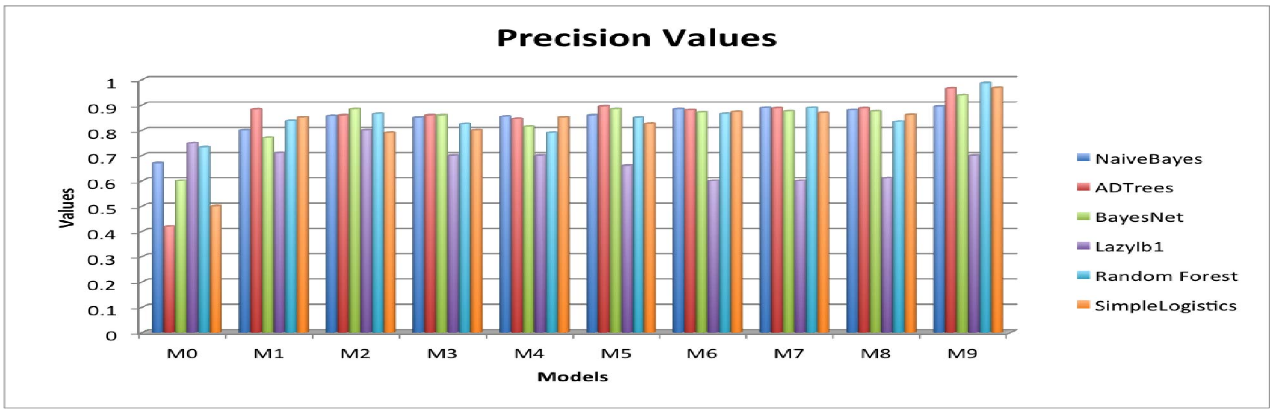

Figure 3: **Precision Values for different classification schemes for Models M0-M9**

the relevant results.

We will now analyze our results in terms of precision/accuracy. In Fig. 2, we can see that, AUC's(area under the curve) best results were for the Random Forest and SimpleLogistic classification schemes. The model performed badly for the lazy algorithms. In Fig. 3, we can see that the models for precision values performed magnificently under the ADTrees, RandomForest and SimpleLogistic classification schemes. Similarly in the Fig. 4, we can see that recall values performed best under RandomForest and ADTrees classification schemes at the same performed badly under lbk1. The area under the curve showed good results in the case of Model 1 to Model 9, which depicts the fact that the key factor in our prediction models is "amount of fund raised" like seed funding amount and so on.

Some of the best result precision results we got were 73.3% for Model 0 (No seed funding), 86.3% for Model 1 (M0 + including seed funding amount), 88.1 % for Model 2 (M1 + including Series A funding), 87.5% for Model 3 (M2+ including Series B funding), 86.7% for Model 4 (M3 + including Series C funding), 88.4% for Model 5 (M4 + including Series D funding) , 86.4% for Model 6 (M5 + including Series E funding), 87.9% for Model 7(M6 + including Series F funding), 87% for Model 8 (M7 + including series G funding), 96.3% for Model 9(M8 + including all the factors mentioned earlier in Section 4). Detailed values of the recall and precision are mentioned in Table 4.

We tested our predictive models on few startups (currently in market) which have both succeeded and failed. At first we tested our predictive models on a company named Spotify (founded in 2006). The company fulfills almost all the parameters of our models like the amounts of seed funding and venture funding (total amount raised 537.8mn dollars). The mean severity scores of the company at all the stages are close to 4.5 on a scale of 5. The burn rate of the company is moderate and under control. Factors like huge customer base, lot of traction, amazing services, well built product

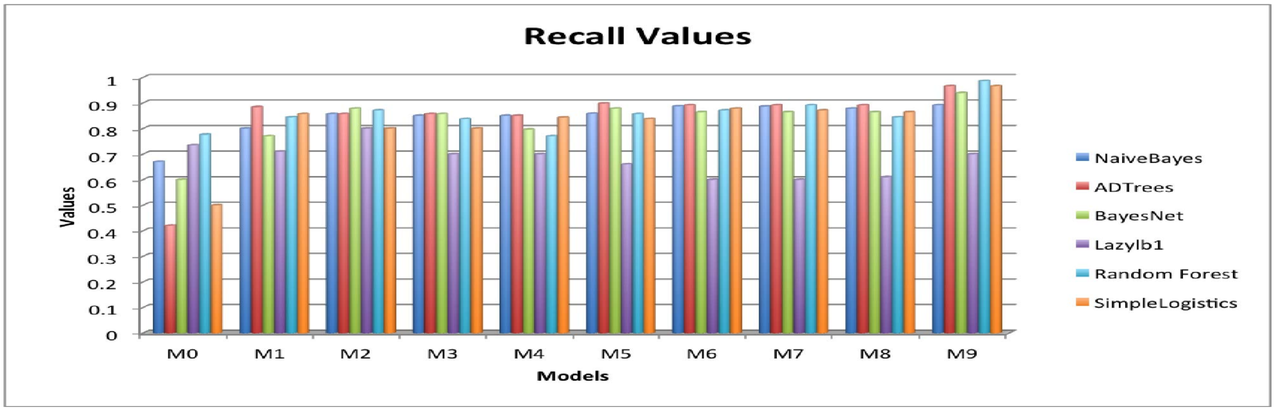

Figure 4: **Recall for different classification schemes for Models M0-M9**

etc helped the company to scale. Our model predicted a success probability of 88.9% which makes it a successful company. A example of failed company predicted by our model is the company named Everpix. The company was able to raise a total of 2.4mn dollars in seed and venture funding but the mean severity scores of the company were bad i.e. close to -1. Factors like bad management, less traction , extremely high burn rate led to the company's failure. Our model prediction for this company was close to 44.2% which puts it into the category of a failed company.

7 Conclusion

In this paper, we used several supervised learning classifiers to construct models for success/failure prediction of early stage startups. Precision accuracies of 73.3%, 86.3%, 88.1 %, 87.5%, 86.7%, 88.4%, 86.4%, 87.9%, 87%, 96.3% for Models 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 respectively. Further we also studied the values for the ROC area and recall. Given the prediction quality we can certainly say that any early stage startup can use our prediction models (at every milestone) to predict their outcome. This will give them an insight of how to be on correct path from the very beginning by taking correct measures and by not committing the same mistakes the failed startups did. Future work involves increasing our accuracy and precision values by incorporating more severity factors. We will also like to use few other classification techniques to derive better prediction models with much higher accuracies. We also want to develop a web tool based on our current approach. The web tool will be handy to use for the entrepreneurs and innovators.

8 Acknowledgment

This work is supported in part by the following grants: NSF awards CCF-1029166, IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330; AFOSR award FA9550-12-1-0458; NIST award 70NANB14H012; DARPA award N66001-15-C-4036.

References

[1] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integrating Materials and Manufacturing Innovation*, 3(8):1–19, 2014.

[2] T. M. Begley and W.-L. Tan. The socio-cultural environment for entrepreneurship: A comparison between east asian and anglo-saxon countries. *Journal of international business studies*, pages 537–553, 2001.

[3] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.

[4] J. Brüderl, P. Preisendörfer, and R. Ziegler. Survival chances of newly founded business organizations. *American sociological review*, pages 227–242, 1992.

[5] R. Dickinson. Business failure rate. *American Journal of Small Business*, 6(2):17–25, 1981.

[6] W. B. Gartner. Who is an entrepreneur? is the wrong question. *American journal of small business*, 12(4):11–32, 1988.

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

[8] P. G. Greene, M. M. Hart, E. J. Gatewood, C. G. Brush, and N. M. Carter. Women entrepreneurs: Moving front and center: An overview of research and theory. *Coleman White Paper Series*, 3:1–47, 2003.

[9] G. Holmes, A. Donkin, and I. H. Witten. Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361. IEEE, 1994.

[10] N. K. Malhotra. *Marketing Research: An Applied Orientation, 5/E.* Pearson Education India, 2008.

[11] D. C. McClelland. Characteristics of successful entrepreneurs*. *The journal of creative behavior*, 21(3):219–233, 1987.

[12] R. K. Mitchell, J. Mitchell, and J. B. Smith. Failing to succeed: new venture failure as a moderator of startup experience and startup expertise. *Frontiers of entrepreneurship research*, 2004.

[13] A. Rauch and M. Frese. Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners' personality traits, business creation, and success. *European Journal of Work and Organizational Psychology*, 16(4):353–385, 2007.

[14] R. Stuart and P. A. Abetti. Start-up ventures: Towards the prediction of initial success. *Journal of Business Venturing*, 2(3):215–230, 1987.

[15] M. Van Gelderen, R. Thurik, and N. Bosma. Success and risk factors in the pre-startup phase. *Small Business Economics*, 24(4):365–380, 2005.

[16] A. Veloso, W. Meira, and M. J. Zaki. Lazy associative classification. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 645–654. IEEE, 2006.

[17] Y. Xie, Z. Chen, K. Zhang, Y. Cheng, D. K. Honbo, A. Agrawal, and A. Choudhary. Muses: a multilingual sentiment elicitation system for social media data. *IEEE Intelligent Systems*, 99:1541–1672, 2013.