Machine Learning Algorithms for Sourcing and Evaluating VC and PE investment deals

Yijie Sun, Prerna Khullar, Andrew Matangaidze

I. INTRODUCTION

The hedge fund community is well known for using quantitative models. However, in private investment markets, many decisions are being made on a gut feel. Traditional finance calls for simple ratio analysis and cash flow valuation based on financial metrics. Current VC and PE investors network to identify potential deals. With limited operating history and tight budgets, a whole lot of features could be incorporated in the analysis to pick those deals that will unlock value for investors.

Using Pitchbook data (courtesy of Professor of Finance Ilya A. Strebulaev at Stanford Graduate School Business), we applied machine learning techniques to predict the probability of a startup's success. Given the nature of the problem at hand and the structure of the dataset, we focused on binary classification algorithms including logistic regression, regularized logistic with L1 and L2 penalty, Support Vector Machine, Random Forest and Neural Network. Our research will enable VC and PE investors to proactively know which companies to talk to, help in sourcing deals and due diligence activities that form the core of the growth investment process. Moreover, it will also serve as a self-assessment tool for start-up founders so that they can make sound decisions during incubation.

II. RELATED WORK

The first classic success/failure study was conducted by Lussier and Pfeifer[1] in 2001 and was extended to markets in the United States, Chile and Croatia[2]. Since then, several similar studies using machine learning algorithms have been performed, but with unsatisfactory outcome. These include research by Wei et al.[3] (2009) on the use of ensemble classifiers to predict acquisition on 600 cases, which gave a global accuracy rate of 88% and a precision of 40%.

Only recently have larger datasets been considered for such studies: Da Silva Ribeiro Bento (2018) compared logistic regression, SVM and Random Forest for predicting M&A or IPO[4] of 80,000 startups. Due to a huge class imbalance, the author generated synthetic variables to deal with sparsity and oversampled minority class to get a precision close to 92%.

Another study by Arroyo et al. (2019) used a Crunchbase dataset with 120507 companies on a three-year window to predict whether a start-up will move to the next round of financing[5]. They found that the Gradient Tree Boosting had the best performance with an accuracy of 82.2%, followed by Random Forest, Extremely Randomized Tree and SVM.

So far, most related work used accuracy as an indicator of model performance. However, due to class imbalance, we concluded that the Area Under Receiver Operating Characteristic (AUROC) and precision are more meaningful metrics. Futhermore, none of these studies considered deep learning algorithms.

In our research, we started with algorithms that have been shown to perform well and moved on to train a neural network. Our work differs from past literature by focusing on a longer period since growth investors (both VC and PE firms) play long. We used a different dataset with a variety of qualitative and quantitative features, and defined labels differently.

III. DATASET AND FEATURES

A. Overview

Our dataset includes a high-level company summary, transactions made (debt, fund raised, M&A), as well as information on investors, location, industry and founder.

Invariably, we observed that if an entry had missing values for employee count and total funds raised to date, it usually had incomplete information for other features as well. To avoid making assumptions about the missing values, we filtered out the sparse entries. Further, in the interest of chronological relevance and to control for geographic discrepancies, we restricted the scope of analysis to the 21403 companies founded *after 1997 in the United States.*

The data was standardized before being split into a training set of size 17122 and a test set of size 4281 in accordance with the 80/20 rule. The examples were labelled by their business status spanning over

TABLE I

Numerical Features

Directly obtained from dataset

Feature	Representing
Employee count	Team size
Number of competitors	Awareness of competitive landscape
Number of board members	Team diversity
Deal number	Financing Stage
Number of tranches	Investor base
Number of sellers	Number of exiters

Derived features

Feature	Preprocessing	Representing
Total funds raised to date	Aggregated by company from transactions table	Availability of cash
Total debt	Aggregated by company from transactions table	Financial discipline on the company due to loan covenants

TABLE II

Categorical Features

Obtained from data

Feature	No. levels	Representing
Year founded	22	Regime dynamics
State	53	Availability of support services
Industry	7	Revenue bucket

Derived features

Feature	No. levels	Preprocessing	Representing
Gender of co-founders	2	Dummy variable with value 1 if the founding team includes a female founder and 0 otherwise	Team dynamics
Financing status	2	Originally contain over 20 levels, merged into 2 levels depending on whether the company was previously PE-backed or VC-backed	Current finance stage
Stock type	6	Originally contain over 40 levels, merged into 6 levels including: common stock, convertible preferred, participating preferred, options, preferred stock and combination stock	Governance dynamics and capital structure

17 levels. The status of stealth, bankruptcy and out of business were categorized as failure. All other levels, including product development, generating revenue and profitable were categorized as success.

Since the Pitchbook data has been updated over the last two years, we define the success or failure of a company, only as of today. Our analysis is not a time-series analysis of each company over the entire period from 1997 to 2019.

B. Feature Engineering

According to a comprehensive research conducted by CB Insights[6], the top five factors that affect the survival of startups are no market need, ran out of cash, not the right team, get out-competed and pricing or cost issues.

Using the above factors and our financial intuition, we came up with 14 features (excluding dummy levels) that would help classify the companies. The numerical features are summarized in Table 1 and the categorical features in Table 2.

IV. METHODS

Due to selection bias that successful startups are more likely to report their statistics to Pitchbook, our data is highly imbalanced, with a ratio of success/failure = 92/8. Without accounting for the class imbalance, the logistic regression will predict all startups as successful, achieving a high accuracy but a low AUROC. Therefore, for all models except for the Neural Net, we penalized the misclassification on the minority class by adjusting weights inversely proportional to the class frequency using the `class weight = balanced` argument in the `sklearn` library. The hyperparameters were tuned by grid searching over the parameter space and choosing the set of parameter values that yields the largest AUROC based on 5-fold cross-validation on the training set.

A. Logistic Regression

Logistic regression is a simple binary classification algorithm which served as our baseline. The probability that a startup will become successful is modeled by the sigmoid function:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

To find the optimal θ vector, we optimized the log loss through coordinate gradient descent:

$$l(\theta) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)}))$$

B. Logistic Regression with L1 regularization

Given that there is a total of 98 features (including dummy levels), the regular logistic behaved rather poorly possibly due to multicollinearity. To improve the test set performance, we controlled for the issue of overfitting through regularization. In particular, we imposed an L1 penalty on the θ vector.

By performing grid search on $\lambda \in [10^{-4}, 10^{5}]$ followed by a finer search on $\lambda \in [1, 100]$, we selected $\lambda_{opt} = 8.33$.

C. Logistic with L2 regularization

Similarly, we imposed a L2 penalty on the θ vector and minimizes the cost function:

$$l(\theta) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)})) + \lambda||\theta||_2^2$$

The regularization constant was tuned to be $\lambda_{opt} = 0.18$.

D. Support Vector Machine (SVM)

Despite the improvement in performance due to regularization, the logistic regression has a linear decision boundary which cannot adequately capture the distinction between classes.

The SVM is an extension of the maximal margin classifier which enlarges the feature space by means of kernels so that the separating hyperplane produced is nonlinear in the original feature space.

By performing a grid search over the choices of kernels, we determined that the Laplacian kernel outperforms the linear, the polynomial, the radial basis function (rbf) and the sigmoid kernels. The Laplacian kernel has the form:

$$K(x, y) = exp(-\gamma||x-y||_1)$$

We also grid searched the regularization constant and found that $\lambda_{opt} = 0.01$. The gamma parameter for the Laplacian kernel was tuned to be $\gamma_{opt} = 0.00055$. In order to speed up the computation, we relied on the Nystroem method to construct an approximate feature maps to the kernels.

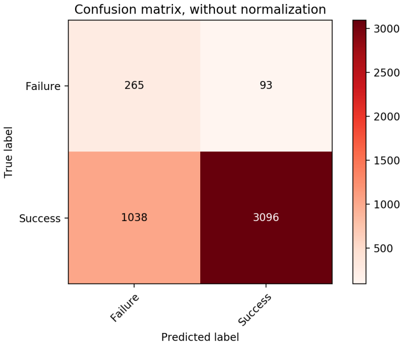

Fig. 1. Confusion Matrix for SVM

E. Random Forest

The Random Forest algorithm builds a large number of decision trees through bootstrap sampling. At each split, the tree is allowed to consider a random sample of 10 predictors. By forcing the algorithm to choose from the square root of the number of available features, we can de-correlate the trees grown from the bootstrap samples, thus reducing the variance of the classifier. The optimal number of trees in the forest was selected to be 600 and the maximal depth is 20 according to the cross-validated random search.

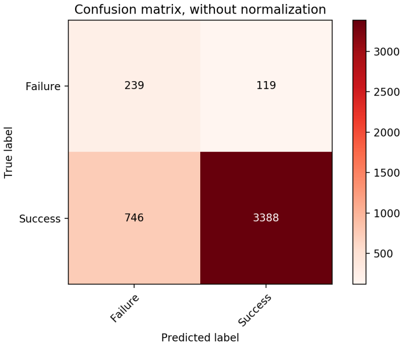

Fig. 2. Confusion Matrix for Random Forest

F. Neural Network (Multi Layer Perceptron)

We trained a fully-connected regularized network with logistic activation. To account for data imbalance, we used Synthetic Minority Oversampling Technique (SMOTE) which relied on a combination of over-sampling the minority class and under-sampling the majority class to improve classification performance.

The output of the layer i is calculated as:

$$a^{[i]} = g(W^{[i]}x + b^{[i]})$$

We used the regularized cross-entropy loss such that the L2 regularization constant $\lambda_{opt} = 0.05$, with the Adam stochastic gradient-based solver:

$$J = CE(y^{(i)}, \hat{y}^{(i)}) + \lambda \sum_{i} ||W^{[i]}||^2$$

By grid searching the hyperparameters, we found the Neural Network performed the best with 4 hidden layers (Fig. 3) of shape (100, 50, 50, 20) and a learning rate of $\alpha_{opt} = 0.0001$.

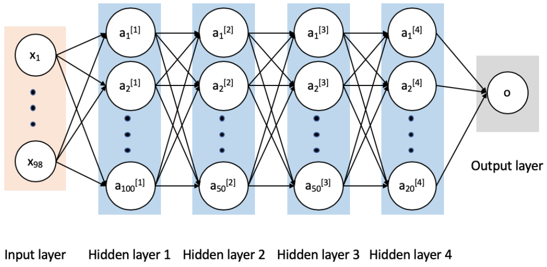

Fig. 3. Neural Network

G. Other methods tried

To deal with the data imbalance, we initially performed softmax classification with the original business status labels. However, as most companies in the dataset are generating revenue, running a multi-class classification does not help with the issue of class imbalance, so we kept our classification binary.

Besides, we also considered other types of sampling techniques including random under-sampling, random over-sampling, tomek links and cluster centroids. The SMOTE method gave the best results in terms of computation time and introduced a moderate bias and variance relative to other methods, so we decided to perform SMOTE for the Neural Network.

In addition to the logistic regression, another baseline model we tried is the LDA. Due to the large number of categorical features, we encountered the issue of multicollinearity among variables. To resolve this issue, we clustered the states and the years based on domain knowledge, but that did not seem to impact the performance of the models. Thus, we conclude the unsatisfactory performance of the LDA is probably due to the violation of multivariate normal assumption on the conditional distribution.

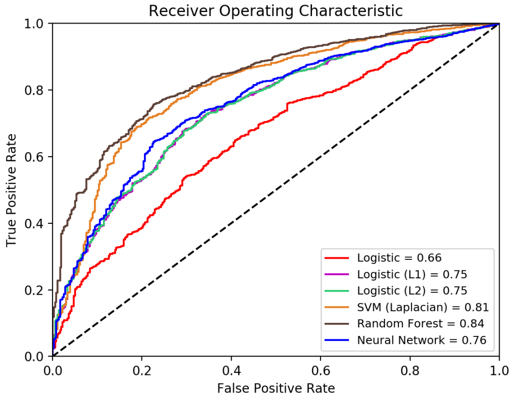

Fig. 4. AUROC comparison

V. Results

A. Evaluation metrics

As mentioned above, a high accuracy can be achieved by predicting all companies as successful. Therefore, when tuning the hyperparameters, we used the AUROC as our evaluation metric (Fig. 4) because it captures both the true positive rate and the false positive rate.

Additionally, we note that the VC and PE investors only target companies that will become good investments, so we considered the metric precision, which measures the fraction of true positive instances among all instances predicted as positive:

$$precision = \frac{TP}{TP + FN}$$

At the same time, growth investors do not want to miss any startups that can become next Facebook or Tesla. Therefore, we also computed the recall as the fraction of all positive instances that were actually predicted as positive:

$$recall = \frac{TP}{TP + FN}$$

The F_β score is a metric which summarizes the weighted combination of precision and recall. Depending on how risk-averse the investor is, we may want to vary the value of β. The larger the β, the more the investor care about not missing a potential deal. Here we have chosen $\beta = 0.5$ to reflect the decision-making of a conservative investor:

$$F_{0.5} = (1 + 0.5^2) \times \frac{precision \times recall}{(0.5^2 \times precision) + recall}$$

B. Discussion

Among the models considered, Random Forest had by far the best performance (Fig. 2, Tab. III and Fig. 4), probably because it effectively handled the

TABLE III

Results Summary

	Train	Test			
Models	AUROC	AUROC	Precision	Recall	$F_{0.5}$
Logistic	0.63	0.66	0.92	1	0.94
Logistic (L1)	0.78	0.75	0.96	0.66	0.88
Logistic (L2)	0.78	0.75	0.96	0.66	0.88
SVM (Laplacian)	0.85	0.81	0.97	0.75	0.92
Random Forest	0.95	0.84	0.97	0.82	0.93
Neural Network	0.81	0.76	0.97	0.69	0.89

large number of categorical variables. While in theory, Random Forest would control for the variance by averaging predictions from a large number of bootstrap samples, we still observed a generalization gap between the train AUROC and the test AUROC. We suspected this occurred due to the large search space for parameters. Therefore, we tried to mitigate the issue of overfitting by capping the max depth of trees at 20.

The first runner-up was SVM (Fig.1, Tab. III and Fig. 4) with the Laplacian kernel. The selection of the Laplacian kernel over other kernels could possibly be attributed its local behavior, in the sense that only nearby training observations have an effect on the predicted class labels.[4]

Neural Network did not perform as well as we expected. Oversampling the minority class in the training set may have led to a generalization gap in the test set. Moreover, our dataset only contains 21403 examples and is likely insufficient to train a good Neural Network model. The limitation of data size could be a reason why past literature did not consider deep learning models.

For the Random Forest classifier, we also evaluated the importance of features, measured as the average amount the Gini index decreases by splitting over a given predictor.

Most numerical features are ranked high on the feature importance chart (Fig. 5 and Tab. 1). In particular, the features funds raised to date, number of competitors, deal number, number of board members and employee count also have the largest coefficients in the regularized logistic. Categorical features such as being in California, having a female co-founder and launching in years 2014 to 2017 are also deemed critical to a startup's success.

We hypothesized that these features were shown to be the top because California is the Entrepreneurial hub, having a female co-founder reflects the diverse mindset of the company founders and a healthy work environment, and the years 2014 to 2017 were those following the recovery from the Great Recession and therefore, start ups were blooming.

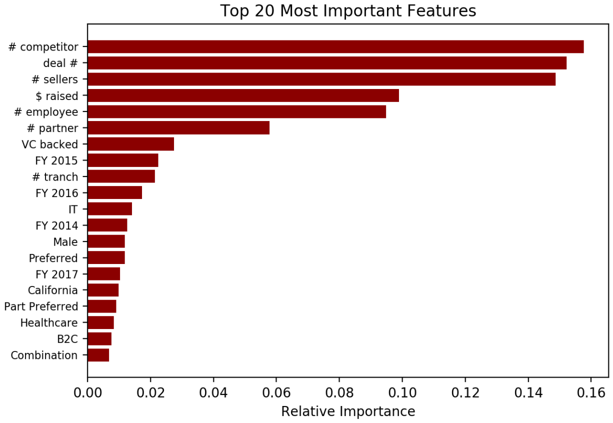

Fig. 5. Top 20 Features

VI. Future work

Based on our results, we plan to enhance our model by acquiring additional numerical features, including financial metrics such as revenue, net income, enterprise value and market cap, as well as statistics on social media presence, e.g. number of Twitter followers, number of web visitors, etc.

Since Random Forest yielded the best results, we are motivated to try additional tree-based models such as Extremely Randomized Trees and Gradient Tree Boosting.

Our raw data contains two textual features - the company description and the competitors, which we did not analyze in this paper. Performing NLP on these features could yield crucial intricate insights into the companies.

Furthermore, we plan to restore the detailed business status labels and run multi-class classification algorithms after balancing the classes.

A final topic we are interested in exploring is how to use generative models to produce feature values such that a company will almost certainly become successful, if we start with a base set of features.

VII. Contributions

Yijie Sun worked on modeling, data preprocessing, poster and report writing.

Prerna Khullar worked on coding, data cleaning, poster and report writing.

Andrew Matangaidze sourced data from Professor Ilya A. Strebulaev, worked on feature engineering and data cleaning, and led the motivation, literature survey and data component of the final report.

Please visit https://github.com/yijiesun97/CS229 for our code.

References

[1] Lussier, Robert N., and Sanja Pfeifer. "A crossnational prediction model for business success." Journal of Small Business Management 39, no. 3 (2001): 228-239.

[2] E. Halabí C, N. Lussier R. A model for predicting small firm performance: Increasing the probability of entrepreneurial success in Chile. Journal of Small Business and Enterprise Development. 2014 Feb 11;21(1):4-25.

[3] Wei CP, Jiang YS, Yang CS. Patent analysis for supporting merger and acquisition (ma) prediction: A data mining approach. InWorkshop on E-Business 2008 Dec 13 (pp. 187-200). Springer, Berlin, Heidelberg.

[4] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013 Feb 11.

[5] Bento FR. Predicting start-up success with machine learning (Doctoral dissertation).

[6] Arroyo J, Corea F, Jimenez-Diaz G, Recio-Garcia JA. Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. Ieee Access. 2019 Aug 30;7:124233-43.

[7] “The Top 20 Reasons Startups Fail”, CB Insights(November, 2019).