International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Smart Start-Up Analyzer: Prediction Model, Analysistool for Venture Capitals using Machine Learning

Mantej Singh Tuli[a,*], Shreyas G[a], Dheemanth A N[a], Sree Chand R[a] , Anupama Y K[a]

[a]*Department of Computer Science, Dayananda Sagar College of Engineering, Bangalore, Karnataka*

1. Abstract

Artificial intelligence (AI) has emerged as a powerful technology with the potential to revolutionize many different industries and fields. In recent years, AI has been applied in a variety of areas, comprising healthcare, finance, and even venture capital. Venture capital is a form of private equity that involves investing in startups and other early-stage companies with high growth potential. Venture capitalists often provide funding, expertise, and other resources to help these companies succeed and generate returns for investors. Our research suggests that AI has the potential to play an important role in the venture capital industry, helping venture capitalists make more informed and profitable investments. In the following sections of the paper, we will provide a more detailed background on the history and evolution of AI, and discuss the current state of the art in AI technology. We will then present a literature review of previous research on the use of AI in venture capital, and outline our research approach. Ultimately, we will bestow our findings and review their implications for the future of venture capital and AI.

2. Introduction

Venture capital is a form of private equity that involves investing in startups and other early-stage companies with promising growth capability. Venture capitalists provide funding, expertise, and other resources to help these companies succeed and generate returns for investors. However, venture capitalists face many challenges and obstacles during the decision-making process, and there are many ways in which they can go wrong. Some of the key ways in which VCs can go wrong during their decision to invest in a startup include:

1. Misjudging the market: One of the biggest mistakes that VCs can make is misjudging the market for a startup's products or services as seen in [1]. This can happen if a VC overestimates the size of the market, fails to anticipate changes in consumer preferences or trends, or ignores the competition. As a result, the VC may invest in a startup that is unable to generate sufficient demand for its products or services, and ultimately fail to generate a profit.

2. Underestimating the risks: Venture capital is a high-risk investment, and VCs must be able to manage and mitigate the risks involved. However, VCs can sometimes underestimate the risks associated with a startup as can be seen in [1][2], and fail to properly assess the potential pitfalls and challenges that the startup may face. For example, a VC may invest in a startup that relies on a single product or customer and fail to consider the risks of losing that product or customer.

3. Overvaluing the startup: Another common mistake that VCs can make is overvaluing the startup they are investing in. This can happen if a VC is overly optimistic about the startup's potential. As a result, the VC may be willing to invest more money in the startup than it is worth, and ultimately end up paying too much for an investment that does not generate a profit.

4. Failing to negotiate a fair deal: VCs must be able to effectively negotiate with founders and other stakeholders in order to secure a favorable investment deal or losses incur like [4]. However, VCs can sometimes fail to negotiate a fair deal and end up with terms that are unfavorable to the investors [3]. For example, a VC may agree to invest in a startup without securing the right to appoint a board member or observer, and thus lose influence and control over the startup's direction.

Overall, there are many ways in which VCs can go wrong during their decision to invest in a startup. These mistakes can result in investments that fail to generate a profit, and can ultimately harm both the investors and the startups themselves. It is important for VCs to carefully consider these risks and avoid making common mistakes, in order to maximize the chances of success for both the investors and the startups.

3. Previous research

A study, published in the Journal of Business Research [6], found that the use of AI in venture capital can help VCs to make more accurate predictions about the future performance of companies and industries and to identify potential investment opportunities that might have been missed by human analysts. The study from Jared Council [5] also found that AI can help VCs to make more efficient and effective use of their time and resources, freeing them up to focus on more strategic and creative thinking.

Human analysts can miss potential investment opportunities for a variety of reasons. Some common reasons include:

1. Overconfidence: Human analysts may be overconfident in their ability to accurately predict the future performance of companies and industries, and may overlook potential investment opportunities [1] that do not fit with their expectations or beliefs.

2. Confirmation bias: Human analysts may be prone to confirmation bias, which is the propensity to look for and analyse data in a manner that supports one's preexisting views or hypotheses. This can lead them to overlook

or discount information that contradicts their preconceptions and may cause them to miss potential investment opportunities.

3. Sunk cost fallacy: Human analysts may be affected by the sunk cost fallacy [7], which is the tendency to continue investing in a company or project even when it is not performing well, in order to avoid feeling like the time and resources already invested have been wasted. This can lead them to miss potential investment opportunities that may be more profitable and sustainable.

4. Limited perspective: Human analysts may be limited by their own personal experiences and perspectives, and may not be aware of potential investment opportunities that are outside of their immediate field of expertise or knowledge.

5. Limited resources: Human analysts may be limited by the amount of time and resources they have available to research and evaluate potential investment opportunities, and may miss opportunities that are not immediately obvious or that require more in-depth analysis.

4. Data overview

We gathered information from 17 sets of data, including various attributes of companies, by utilizing academic access to Crunchbase data. The raw data included past events and a snapshot of all firms at the time of the data extraction. After conducting an examination of the features, we selected 7 sets of data as shown in Table 4.1 to create a historical view of start-ups. Additionally, we incorporated external data, as shown ibelow, in order to amplify the models' aptness to make predictions [23].

Dataset	Details of the data utilized
Acquisition	Information about corporate purchase activities, including the acquisition date and the unique ID of the target business.
Degrees	Information on each student's educational background at the individual level, including name, degree received, graduation date, and institution.
Funding Rounds	Information on financial support received by a company, including identification of the company, details of the funding round, and date of the financing round.
IPOs	Information on initial public offerings by a company, title of the company and time of the IPO.

Jobs	Information on an individual's professional experience, including identification of the person, identification of the company they worked for, job title, start and end date of employment, and indication of whether it is still their current job.
Organizations	Information on an individual's professional experience, including identification of the person, identification of the company they worked for, job title, start and end date of employment, and indication of whether it is still their current job.
People	Personal information, including: identification of the person, gender, and country of origin.

Table 1: All the .csv files present in Crunchbase dataset which was exported

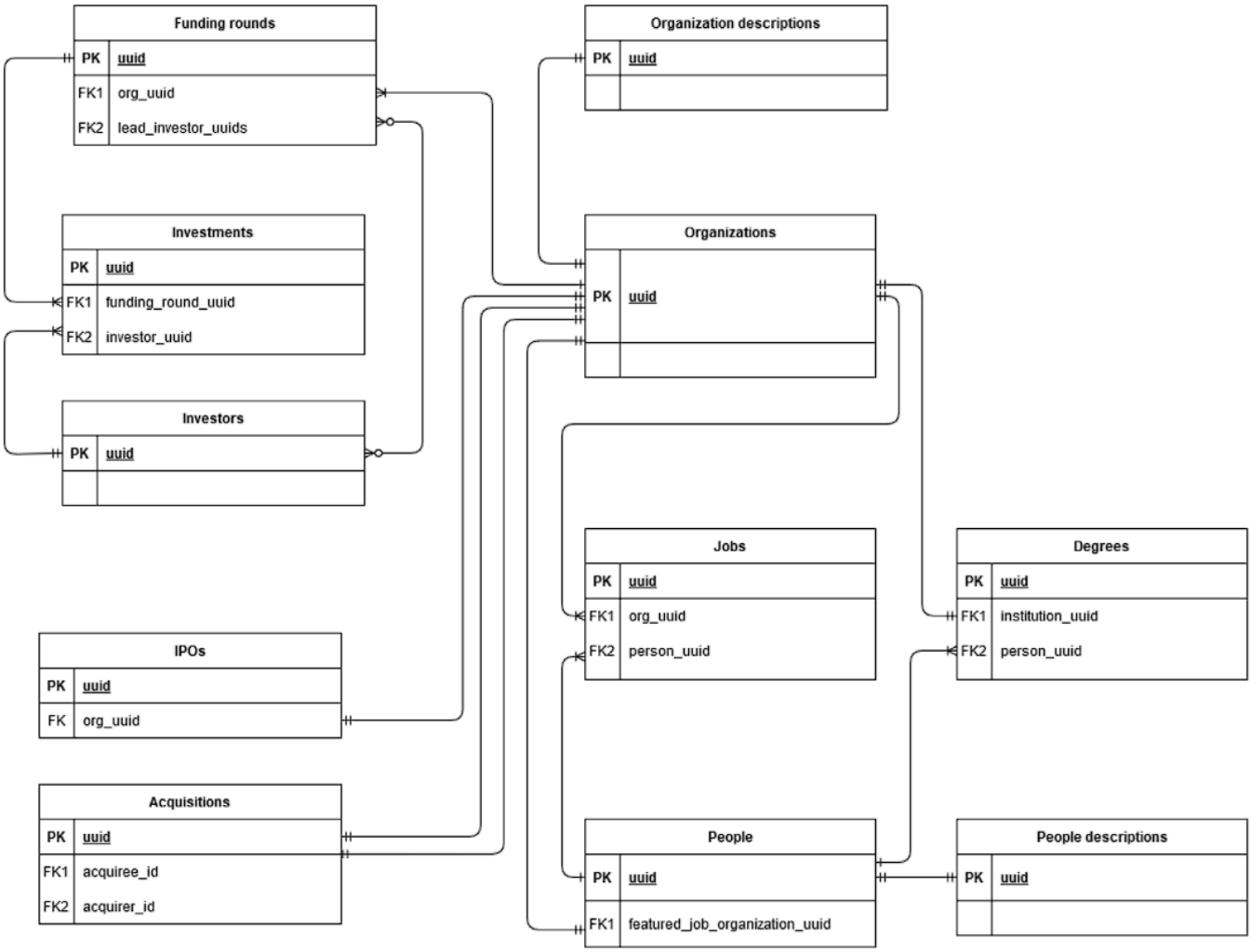

Fig1. Crunchbase data simple ERD diagram

The above information and Fig1. points were all sourced from the Crunchbase dataset, a open-sourced database, with additional information being sourced from United States Patent and Trademark Office (USPTO). This

database will be the best choice, with it being continuously updated to ensure that all dimensions, distributions are filtered out with additional entries which have haywire data points for our study.

Though the USPTO dataset was used, all the features to be used for model creation will be mostly derived from the Crunchbase data. The set that will be used was retrieved in January 2023. The fields extracted from the dataset that are of utmost importance include the number of funding rounds a startup has participated in, information on those funding rounds, the total amount invested, and the job positions held by the individuals working in the startup.

The representation of the progress of the companies will be done differently. Rather than using time as a measure of progression, we use the cumulative records of the company as a base. The reasoning behind this is to make it easier to create models around the companies as a base, rather than having their funding rounds determine the final outputs. This said, the time between funding rounds a startup is participating in is a very important measure of a startup's growth. For this reason, we will combine the funding information into time between funding rounds and total funding parameters. For this study, the company categorization idea was obtained from a data distribution of Crunchbase [24].

Data Point	Description
Economic Indicator Data	Economic indicators including consumer price index (CPI), the 10 year government bond with, GDP growth rates, M1 and M3 money supply, unemployment rates, government asset purchase levels (quantitative easing), and private household consumption and expenditure were added to the data on a 3-month lagged basis for each country. These macroeconomic indicators were included to indicate economic developments that could impact a startup.

Table 2: Summary of Additional External Data

The data used to train models for predicting the success of a company's fundraising efforts includes a variety of fields such as the number of rounds, contact information, funding amount, and employee and founder job titles. This data may come in different formats, including unstructured text and numerical values, and must go through a process of encoding, transforming, and conditioning before it can be used. To simplify the analysis, the authors chose to represent each company with a single record; it has data about funding rounds into parameters like the time between fundings, total funding information, and the most recent funds taken by the startup. Additionally, the information is kept cross-sectional when building train-test splits for model development, so that earlier rounds of the company are free of any effects from future rounds [23].

4.1 Data processing

Prior to the creation of the data set from the 68,400 American and UK companies' statistics, the figures should be cleaned to rule out any abnormalities. The scope of this study encompasses various types of startups, regardless of their funding structure. This includes startups that have obtained funding through loans or other forms of debt, rather than selling ownership shares in the company. The study also includes startups that have undergone funding rounds after already going public (in the case of an IPO) or issuing tokens through an ICO. Additionally, the study will consider startups that have experienced periods of time during which they sought out investment from external sources. Furthermore, startups that have been acquired by larger companies will also be considered. The study will also take into account startups that have gone public by issuing shares to the public through an IPO, as well as those that have shut down operations permanently. Finally, the study will not exclude any startups without a name from consideration. 64,197 startups are left after deleting these anomalies with 9541 being from the UK and 54656 being from the US. The Crunchbase data sets for these remaining businesses will then be integrated to produce a single, coherent time series with the purpose of capturing the altering evolution of companies and connecting these to the likelihood that a venture capital investment will be successful or unsuccessful.

Investors in venture capital have enough time to assess and spend in startups that these forecasting algorithms were able to point out because forecasts are made at the conclusion of each month for a goal that exhibits activities that take place over the following 3 or several months.

4.2 SQL-based feature selection and engineering

The information release used in our endeavors contains many CSV files, which each denotes a tabular database that MySQL is capable of importing. The analysis is then restricted to businesses that potentially serve as targets for such ventures by leaving out VC-affiliated and other institutions that make investments. There are currently 942,605 firms operating in the United States, of which 18,419 are reportedly publicly traded, 94,225 have been bought, 33,298 have been shut down, and the remaining 796,663 are reportedly privately held and active. The next phase will involve converting date information to numerical format using string formatting. In addition, new features will be extracted, such as the duration between financing rounds and the educational backgrounds of the startup's founders, including any degrees earned from top universities. According to d31, the first stage of feature engineering is finished by combining the original and new features into a single features table when investors and entrepreneurs have similar educational backgrounds. a subset of the 370 traits discovered in this way. The data from the characteristics table can be further evaluated to produce more intriguing factors lying dormant in this database.

One noteworthy aspect of our dataset is the number of missing fields for each organization. This information is used as a feature during model training and also provides an indication to help balance the classes. Additionally, keeping track of these sums adds a unique element to our dataset and highlights the value of this form of open data. To address the issue of missing data, we use logical defaults to fill in some of the blank fields. For example, if the overall financing amount is left blank, we assume a default value of zero. Despite the challenges of working with missing values in a public dataset, our analysis indicates that we can still use the Crunchbase data to develop a highly accurate classification model. Figure 1 displays the distribution of missing features across the three departure categories, with an average of 30.89 missing fields for publicly traded companies, 36.82 for acquired companies, and 37.57 for failing businesses.

Additionally, Puri and Zarutskie [17] discovered that businesses with venture capital support expand more quickly and had a lower failure rate within the first five years of operation.

Table 3 shows the parameters which are to be used from the overall assimilation of the Crunchbase dataset. With that there is a graph in Fig2. representing the missing value from the dataset for different organizations according to their current status.

Table 3:
Dataset parameters to be used in the implementation

Feature	Description
period_between_funding	time between funding rounds
no_female_found	number of founder who are females
no_male_found	number of founder who are male
no_patents	patent count
state	state code
country_id	country code
category	company category
descrip_leng	length of company description
no_degree	number of employee degree
top_degree	number of people with top school degrees
domain_info	whether company has a web domain
investor_preference	Angel,VC investors
no_events	number of company events
acquisitions	number of acquisitions

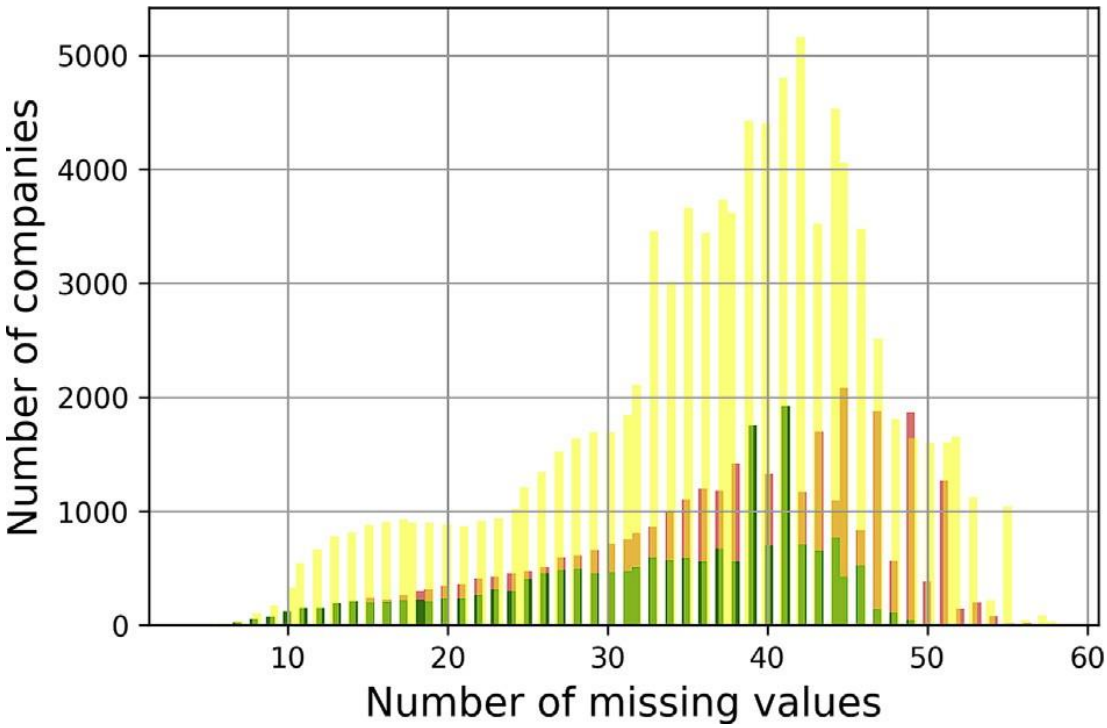

Fig2. The graph represents missing values from companies in all domain public acquired or failed companies. Red represents failures, Yellow represent acquired companies, while Green shows public companies

4.3 How patent data from USPTO is useful

Businesses that grant patents in a market with significant levels of competition, according to Cockburn and Macgarvie [13], are going public or attracting additional investment is more likely for patent activity. Patents can work as an organization value indicator to the investors. It was in [14] investigated the correlation between the advancement of the cycle followed by VC's and patents which were filed by the software related startups. They found a strong correlation between the filing of patents and the financial and long-term profitability of businesses. The 7,426,601 USPTO patents taken in account from way back to 1980 were collected as an extra data collection as a result of these discoveries.

The goal was to determine how many patents each company owned based on Crunchbase data. This can be the tricky part as the patents are filed in a variety of ways either under the original name or some pseudo name for some reasons or other.

5. Question to be answered by analysis

The goal or vision of this project is to determine how well ML can be used to predict whether a company will succeed or fail in the long run. Many other previous studies have taken this same idea and run with it in a simple

and head-on fashion, but our target is to make this vision a multi-faceted one, with all minute details contributing towards decision-making. Multiple categories inclusive machine learning models have a question on their minds: whether they can predict a company's future endeavors from the perspective of an investor. VCs who already use the Crunchbase dataset as a base for making their investments in a company will have an additional tool for assimilating this data, making their decision-making process more robust, successful and streamlined.

The next question is whether the company can or will receive a follow-on funding round, which is directly beneficial for the early investors as the valuation of the company goes up and their investment grows in size. This task, if performed by a machine learning model with a desirable rate of success, can become a boon to the investors, who now can back their bets according to the stats presented by the machine learning model. This will not only increase their investments success rates but can also increase the number of calculated investments they make.

The final issue then remains regarding the model selection process for these tasks. The datasets used have some perplexing issues like not reporting information on time, large volumes of missing data, and sometimes changing recording formats altogether. As a result, the data, some of which has already been mentioned, must be interpreted carefully. Additionally, we assess numerous models, including model ensembles, Random Forest, XGBoost, and feed-forward neural networks.

We will study how the model performance on the mentioned question is different across different parameters. Each of these models has a variety of hyperparameters variables. We also want to assess how well the models perform at different stages of development. We then go over these various trials with these questions as a point of reference.

6. Model implementation

When necessary, the class imbalance in the target variable was taken into account when implementing each model in Python using the appropriate Scikit-learn1 packages. The penalties that will be utilized to account for the categorization of an imbalance in the class system in the random forest, regression, support vector machines (SVM), and extreme gradient boosting (XGB) models are weighted because this could otherwise skew the model predictions. While there are other ways to balance the classes, which are the SMOTE (synthetic minority oversampling technique (SMOTE) and the ADASYN (adaptive synthetic sampling approach), researched and demonstrates that weighted penalties, a cost-sensitive learning approach, outperforms over- and undersampling techniques almost always.

6.1 Splitting the data for training and maintaining the class balance

The figures describe 942,889 businesses, of them 18,789 are shown as public offerings, 94,567 as having been acquired, 33,598 as having been shut down, and the remainder 795,935 as being private entities and still in operation. It is recognized to train end models on publicly traded, acquired, and unsuccessful businesses as well as follow-on funding modeled business in order to estimate the likely course of action for private businesses. Due to the crowd-funded nature of the Crunchbase data, many companies have meager information. This is especially evident with expansive, publicly traded corporations. This may be expected as a perception that as these

businesses are no longer start-ups or just scraping by, the founders would have less incentive to update the information. Nevertheless, many of those sparse items are eliminated to prevent biasing the algorithms toward associating less data with IPO or acquisition. This aids in maintaining a balance among the classes. Equation (1) presents the cutoff for excluding organizations with missing values.

$$Threshold = \mu_m - \frac{\sigma \quad m}{\alpha} \tag{1}$$

Where μ_m and σ_m indicates minute specific constants which are dependent on class, and the mean deviation and the standard deviation of the count of not available values for businesses within a class. To describe a class, we take it is as A, in the funding models we have to calibrate for the IPO class and acquired class to define the parameter which describes the exit of a startup.

As the larger numeration of missing fields is deemed to constitute an important indicator for failure, therefore this constraint is not used for exit of an organization class. There are still 6986 public corporations and 15,527 acquired organizations still remaining after the constraint of public and acquired firms is applied. So that the number of failed businesses matched the number of acquired businesses, they were chosen at random. Keeping in mind that fewer than one third of the companies are publicly traded.

To balance the numbers, oversampling (using SMOTE) was used, however it was discovered that this can have a negative impact on later modeling, hence the existing balance of the class was retained for the exit data of the corporations. The data from the other models of importance underwent the same process. Following filtering, 6157 businesses made it from one investment round to the next. That leaves 180,666 businesses that received investment in the beginning but did not receive more funding when snapshots of Crunchbase database were taken in the period (July 2018 to April 2020). A random value for companies with no-growth, equalized to the number of positive-growth can used to have a class balance. The sample is different every time the training is performed over the models if the random number generator is not given a seed. Standard 80 for training and 20 for validation can be used for all evaluations. The results of the model on the validation data will be used for arriving at the conclusion from the models.

6.2 Concept of multiple-layered perceptrons

When a multilayer perceptron model has numerous hidden layers, it is recognised as a deep neural network. Multilayer perceptron models are in concept categorized as feed-forward neural networks. The network's multiple binary classifiers (perceptrons) work together to ensure that both supervised and unsupervised learning are possible, which gives the network's classifiers a wide range of application possibilities and subsequent classification capabilities. This considerably accelerates the minimization of loss when combined with backpropagation of mistakes, making them functionable for parameters which have dimensionality and cardinality set to high.

We will test numerous presets, varying the regularization, quantities of hidden layers, and quantities of neurons in each layer. The most effective network, which consisted of 5 hidden layers with 32 neurons each and alternating dropout layers with a dropout rate of 0.2, produced the best results according to CaptitalVX [24]. Every layer, with the exception of the final one, employed the rectified linear activation function (ReLU) to maximize efficiency in their research. This is clearly a multi-varied problem because, as in the exit case, the objective is to assign all private companies to one of 3 classes with a certain probability. As a consequence, the softmax function

as the activation layer can be chosen for the final layer and sparse categorical cross-entropy as the preferred loss function.

6.3 Logistic Regression

A categorical response variable's chance of occurring can be modeled using logistic regression (i.e. target or dependent variable). It is a nonparametric technique, as opposed to linear regression, and as such makes no assumptions on the data distribution or the residuals of the model. The odds of the occurrence Y = 1 are given by: For example, Y be a binary variable and let p be the probability of Y given X.

6.4 Random Forest

Decision trees, which can be either regression or classification trees, serve as the foundation for the random forest (RF) technique. Although the same technique can handle both categorical and numerical data as inputs and outputs, we will show the categorical classification tree algorithms given the purpose of this study and the structure of the dependent variable as with boosting system study [15].

Before analyzing the classification tree algorithm, we will first provide a brief overview of tree-based methods. After discussing the benefits and drawbacks of this strategy, we'll look into the random-forest algorithm which addresses these issues, because the random forest mixes several trees to anticipate the class of the dataset, some decision trees may predict the proper output while others may not. But when all the trees are taken into account, they accurately predict the outcome.

6.5 Xtreme gradient boosting algorithm (XGB)

The extremely complex algorithm known as extreme gradient boosting (XGB) by Chen T, Guestrin [16] expands upon ideas found in many other algorithms. The Gradient Boosting (GB) algorithm, which works on the basis of classification trees, is already implemented in the RF algorithm, which could be a modified version of XGB. When compared to the GB algorithm, the extreme part of XGB is caused by the lengths the algorithm takes to optimize and accelerate calculations.

Extreme gradient boosting (XGB) is a significantly altered version of the Gradient Boost algorithm that includes multiple modifications and extraneous steps to increase speed and predictability. In XGBoost, gradient boosting is used to correct previous tree model errors by fitting new tree models.

6.6 Two-step classification

In order to classify the firm between IPO, acquisition and failure, the complexity of the model would increase and the model would have to generalize a lot of features. But if we go with a simple binary classification of success vs fail, the model would prove to be in-efficient. Hence it will be necessary to come up with a two-step classification. In the first step, only failure or successful exit is classified. In the second step the likelihood of IPO or acquisition is classified for the successful firms. In order to avoid any target leaks in the second step, the model

was only trained with organizations that became public or acquired by bigger firms. As the parameters in both the steps are different, ensembles of these would result in higher accuracy.

6.7 Robustness

It's important for the model to maintain its accuracy across different subsets of the data. One major factor to partition data is based on the funding rounding it has achieved. As some VCs would want to invest in early-stage start-up's and would restrict themselves in later stages of funding rounds. As seen in the data, the majority of the data partition occurs at the seed funding stage and the count of companies decline as the subsequent funding rounds are raised. The model shows a good accuracy level across all stages of funding except round F, as very less data was available.

Out of the 18 companies that went public after the last round, the model could identify six of them and no false positives. Due to the association between the most recent fundraising round and the probability of a successful exit, the IPO recall is low for early-stage start-ups. This is when false negatives predominate.

7. Expandability

The machine learning model is often viewed as a black-box. Hence it becomes important to explain why a model arrived at a particular result for a particular input. This can be done in two phases, in the first phase, the importance of each feature across the data (global explainability) and in the second phase, importance of specific features for that very input(instance-level explainability). In case of a regression model, it is easier to explain in terms of the coefficient associated with a particular feature. In case of Decision-trees the instance-level explainability doesn't come into picture as the model takes into account the global explainability during the training phase itself.

8. Conclusion

A comparative study of different ML and DL algorithms was carried out to find out the best among these. The results obtained were as follows:

1.) In the first model SVM with XGBoost and Random Forest were used, the data from Kaggle included around 40 different characteristics of almost 22,000 companies. The features were categorized and the important features were analyzed. A heat map was used to study the correlation between different parameters in the data. While XGBoost used boosting and random forest using bagging technique, it was found that XGBoost after fine tuning the parameters was better than random forest, as it reduced the overfitting of data. Therefore, it will enhance the accuracy [18].

2.) The second study was based on exploring the impact of tech news along with the other features. The CrunchBase data was used for the analysis. This approach resulted in TP between 60% and 79.8% with a considerable FP of 0% to 8.3% with very few missing attributes in the CrunchBase [19].

3.) In the third study, different classifiers like Decision tree, Logistic Regression, MLP were used on the Kaggle dataset, which is a real time data, it was found that the Decision tree resulted in the best accuracy with 98% for this kind of data [20].

4.) The fourth study proposed a multi-class classification technique unlike the previous binary classification techniques, rather than predicting if a start-up is merely good or bad, it also predicts how many rounds of funding the start-up can take and also predicts if a start-up would be acquired or not by a bigger firm. There are some classification errors which aren't harmful while making the decision [21].

5.) The fifth study focuses the data on twitter posts, which the firm posts and along with its financial data. The model yields an accuracy of up-to 76%, showcasing that the firms survived for five or more years in the industry [22].

6.) The sixth study shows the impact of AI patent or any emerging technology patent that can influence the decision making of VC's, another interesting insight drawn from this study is that the patent with high degree of knowledge coupling attracts VC [23].

References

[1] Nick Skillicorn, 65% of Venture Capital-backed Deals Fail to Return Investment, and Only 4% Make Substantial Returns, IDEA TO VALUE (Oc.t 18, 2018)
[2] Nicolás Cerdeira & Kyril Kotashev, Startup Failure Rate: Ultimate Report + Infographic [2021], FAILORY (Mar. 25, 2021)
[3] Tomer Dean, The Meeting That Showed Me The Truth About VCs, TECHCRUNCH (Jun. 1, 2017)
[4] Sam Reynolds, VCs Lose a Lot of Money Hunting for the 10x Return, WCCF TECH (Sept. 17, 2019)
[5] Jared Council, VC Firms Have Long Backed AI, Now, They Are Using It, THE WALL STREET J. (Mar. 25, 2021)
[6] Amit, R., Brander, J., & Zott, C. (1998). Why do venture capital firms exist? Theory and Canadian evidence. Journal of business Venturing, 13(6), 441-466
[7] Arvidsson, V., Holmström, J., & Lyytinen, K. (2014). Information systems use as strategy practice: A multi-dimensional view of strategic information system implementation and use. The Journal of Strategic Information Systems, 23(1), 45-61
[8] Hyoung Jun Kim, Tae San Kim, So Young Sohn, Recommendation of startups as technology cooperation candidates from the perspective of similarity and potential: a deep learning approach, Elsevier
[9] Javier Arroyo, Francesco Corea, Guillermo Jimenz-Diaz, and Juan A. Recio-Gracia, Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments, IEEE(Sept. 13, 2019)
[10] Jan K. Woike, Ulrich Hoffrage, Jeffrey S. Petty, Picking profitable investments: The successor equal weighting in simulated venture capitalist decision making, Elsevier (2015)(JBR-08367)
[11] Torben Antretter, Ivo Blohm, Diemtar Grichnik, Joakim Wincent, Predicting new venture survival: A Twitter-based Machine Learning approach to measure online legitimacy, Journal of Business Venturing Insights 11 (2018) e00109
[12] Cockburn IM, Macgarvie MJ. Patents, thickets and the financing of early-stage firms: evidence from the software industry. J Econ. 2009;18(3):729e773.
[13] Mann RJ, Sager TW. Patents, venture capital, and software start-ups. Res Pol. 2007, March;36(2):193e208.
[14] Ho TK. Random decision Forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC. 1995, August:278e282.
[15] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016, August:785e794.
[16] Puri M, Zarutskie R. On the lifecycle dynamics of venture-capital- and non-venture-capital-financed firms. J Finance. 2012;67(6):2247e2293.
[17] Jinze Li, Prediction of the success of start-up companies based on SVM and random forest
[18] Guang Xiang, A supervised approach to predict company acquisition with factual and news articles on TechCrunch
[19] Fardin Rahman Akash, Start-up success prediction using classification algorithms
[20] Javier Arroyo, Assessment of ML performance for decision support in venture capital investments
[21] Torben antretter, predicting new venture survival: A twitter-based ML approach
[22] Roberto S Santos, Risk capital and emerging technologies: Innovation and investment patterns based on AI patent Data analysis
[23] Thomas Hengstberger, Increasing Venture Capital Investment Success Rates Through Machine Learning
[24] Greg Ross, Sanjiv Das, Hussain Raza, and Daniel Sciro, CapitalVX: A machine learning model for startup selection and exit prediction