University College London

ICSD 2023

Leveraging Open Data and AI to Measure SDG Progress in Resource-Limited Settings: A Case Study in the Venture Capital Sector.

Simon Levy

Supervisor: Dr Waqas Rafique

July 2023

Keywords: SDGs, Multi-Layer Perceptron, Transformers, Human-Centred, Venture Capital, Ecological Economics, Self Determination Theory

VentuESGator repository: https://colab.research.google.com/drive/1EEbGo-gO5vaPoFj5huXy6GFT4zAf0hC7?usp=sharing

1. Introduction

1.1 Background Information on the UN Sustainable Development Goals (SDGs) and Venture Capital Investment

The United Nations Sustainable Development Goals (SDGs) require considerable financial and strategic input, especially from the private sector. Yet, evaluating early-stage ventures' impact on the SDGs is difficult due to insufficient environmental, social, and governance (ESG) reporting (United Nations, 2015) (ISSB, 2021) (Simon Sharpe, 2021).

Our research addresses this by using Artificial Intelligence (AI), specifically Natural Language Processing (NLP) and Machine Learning (ML), to enhance SDGs' support in venture capital investments. We're developing an AI-based scoring system integrating ESG sentiment analysis and investor inputs for more informed, SDG-aligned decisions (HBS, 2021) (Ross, 2019) (Vinci, 2022).

Our methodology caters to the rising demand for impact investing, accelerates investment identification, and fosters venture capital investment transparency. After assessing our approach's effectiveness against conventional ML algorithms, we aim to establish a data-driven, human-centered strategy for SDG advancement in early-stage venture capital investments (Imperial, 2021; W. Arthur, 2012).

1.2 Limitations of Current Research and Practice in ESG Impact Assessment

ESG assessment, while progressing, still faces challenges like varying interpretations of ESG principles, making consistent application and comparison difficult. Especially in resource-limited settings, there's a lack of reliable ESG performance data (Ruppert, 2020).

Current machine learning and data science techniques, though successful in predicting startup success, are not fully utilized in ESG impact assessment due to the unstructured, qualitative nature of ESG data. The lack of methods for analyzing such data hampers ESG impact assessment and venture capital benchmarking.

There's a significant gap in research on integrating sentiment analysis for qualitative ESG impact assessment in venture capital investments, limiting the models' ability to capture ventures' nuanced SDG impacts. Given these limitations, innovative approaches harnessing sentiment analysis for ESG impact assessment are urgently needed, particularly in resource-limited settings.

2. Literature Review

2.1 The Use of Data Science and Machine Learning Techniques for Company Success Prediction

Data science and machine learning are being increasingly utilized in the venture capital (VC) industry for predicting startup success, thereby improving investment returns. For instance, the CapitalVX model from Santa Clara researchers uses public data to predict success, suggesting a need for hybrid intelligence approaches combining machine learning and human expertise (Ross et al., 2021).

Retterath's research similarly explores using machine learning for VC investment decisions, with XGBoost algorithm outperforming median VC by 25% and average VC by 29%. The study suggests focusing on promising opportunities through a scalable pre-selection process and identifies essential features for ML algorithms. However, it notes the importance of complete data and the need for hybrid systems (Retterath, 2020a; 2020b).

Retterath emphasizes the necessity for standardized data and highlights Crunchbase as a comprehensive source, though human judgment remains crucial due to limited disclosure requirements from private companies (Retterath, 2020b).

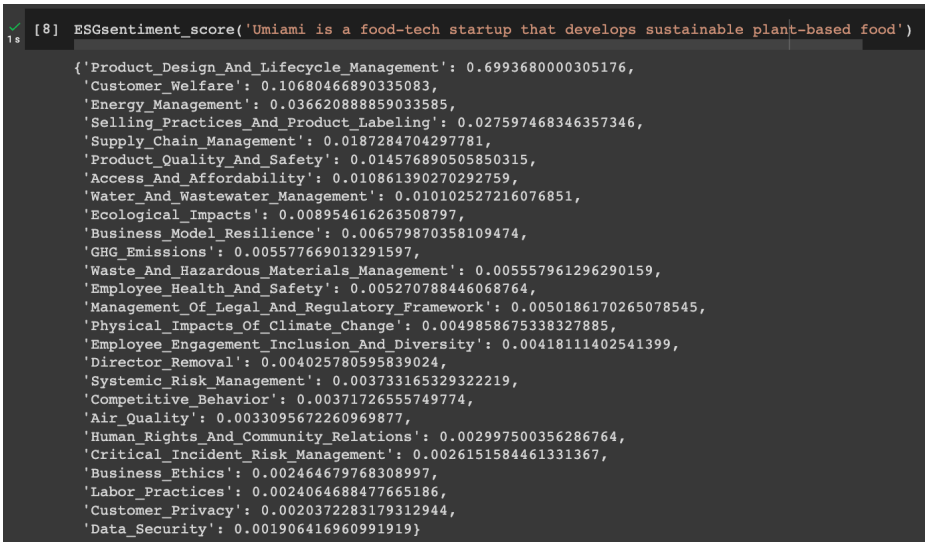
Other research from Stanford, Eindhoven, and Harvard Universities have also used similar approaches for startup success prediction, with accuracy levels around 70-80% (Ang, Chia, and Saghafian, 2021; Pan, Gao and Luo, 2018; Powell, 2021).

2.2 Investment for Qualitative Impact Assessment

The use of techniques to analyze qualitative data for decision-making about potential high-performing companies aligning with KPIs and UN's SDGs is growing among investors (Bice and Fischer, 2020). Automated textual analysis using natural language processing (NLP) and sentiment analysis, like Google's BERT, has proven successful but can struggle with domain-specific vocabulary (Ganesh, 2019).

Berkeley researchers have found NLP and sentiment analysis helpful for identifying and categorizing ESG-related information from unstructured textual data, aiding investors in sustainability-focused decision making (Mehra, Louka, and Zhang, 2022).

ESG-BERT, a model trained on unstructured text data to understand sustainable investing vocabulary, showed promising results with high accuracies and F-1 score in text classification tasks. Its use could revolutionize impact assessment in venture capital, creating a more efficient way to analyze ESG data and inform investment decisions.



```
[8] ESGsentiment_score('Umiami is a food-tech startup that develops sustainable plant-based food')

{'Product_Design_And_Lifecycle_Management': 0.6993680000305176,
 'Customer_Welfare': 0.10680466890335083,
 'Energy_Management': 0.036620888859033585,
 'Selling_Practices_And_Product_Labeling': 0.027597468346357346,
 'Supply_Chain_Management': 0.0187284704297781,
 'Product_Quality_And_Safety': 0.014576890505850315,
 'Access_And_Affordability': 0.010861390270292759,
 'Water_And_Wastewater_Management': 0.010102527216076851,
 'Ecological_Impacts': 0.008954616263508797,
 'Business_Model_Resilience': 0.006579870358109474,
 'GHG_Emissions': 0.005577669013291597,
 'Waste_And_Hazardous_Materials_Management': 0.005557961296290159,
 'Employee_Health_And_Safety': 0.005270788446068764,
 'Management_Of_Legal_And_Regulatory_Framework': 0.0050186170265078545,
 'Physical_Impacts_Of_Climate_Change': 0.0049858675338327885,
 'Employee_Engagement_Inclusion_And_Diversity': 0.00418111402541399,
 'Director_Removal': 0.004025780595839024,
 'Systemic_Risk_Management': 0.003733165329322219,
 'Competitive_Behavior': 0.0037172655574977 4,
 'Air_Quality': 0.0033095672260969877,
 'Human_Rights_And_Community_Relations': 0.002997500356286764,
 'Critical_Incident_Risk_Management': 0.0026151584461331367,
 'Business_Ethics': 0.002464679768308997,
 'Labor_Practices': 0.0024064688477665186,
 'Customer_Privacy': 0.0020372283179312944,
 'Data_Security': 0.00190641696099191 9}
```

Figure 1: Example of an ESG BERT output

2.3 Application of Data Science Techniques in Private Investment for Quantitative Impact Evaluation

Quantitatively assessing a company's environmental and societal impact is complex but vital for understanding their environmental footprint and guiding reduction efforts. However, data science is seldom used to evaluate ESG data for private sector companies (HBS, 2021; ISSB, 2021).

Harvard Business School addresses issues of company valuation and environmental pollution by automatically calculating company externalities into monetary values. This approach uses emissions and water pollution data to quantify environmental costs, promoting their inclusion in valuation standards (Eccles and Mirchandani, 2022; HBS, 2021).

In this approach, data on energy resources and water consumption is used to compute monetary emissions. Environmental costs are calculated using metrics like tons of CO2, NOx, SOx, VOC emitted, and PM 2.5 emitted, along with water withdrawal and discharge data. The total environmental cost, representing the company's overall environmental impact, is determined by adding monetary emissions and water costs (HBS, 2021; Gladys Velez Caicedo, 2021).

2.4 Limitations of Existing Research

Data science research on venture capital risk assessment lacks a hybrid intelligence system for more granular predictions and a unified method for analyzing both quantitative and qualitative data of private firms. Current impact evaluation is fund-specific, needing standardization for easier comparison. Though qualitative ESG data assessment shows promise, techniques to evaluate quantitative data remain underdeveloped. Harvard Business School's scalable methodology for computing environmental costs offers hope. We suggest a program integrating ESG sentiment analysis for assessing private firms' environmental impact as a vital step towards improving sustainability risk evaluation.

3. Methodology

We utilized open data and AI to measure progress towards the Sustainable Development Goals (SDGs) in resource-limited settings, specifically in the venture capital sector, using a dataset from Crunchbase (2015-2019) involving 55,665 companies. Our four-pronged approach involved:

1. Using ESG BERT AI model to analyze qualitative company information, successfully categorizing company descriptions under the SDGs.

2. Creating a start-up success prediction model using machine learning, with the multi-layer perceptron neural network model achieving 84% accuracy.

3. Designing a weighted scoring system requiring human input after AI-generated investment recommendations, using 30 venture capital funds data to provide an overall weighted company score.

4. Applying Harvard Business School's methodology to compute ESG externalities of private companies in monetary terms, converting gas emissions and water consumption into a single environmental cost using a Python framework.

Through this integrated approach, we developed a comprehensive system for sourcing, screening, and assessing potential investments. We applied ESG BERT to our dataset, focusing on company descriptions, and created a filtering mechanism to identify potential impact investments.

Our ESG sentiment analysis on the 'short description' attribute of companies measured the ESG BERT model's accuracy. However, model performance reduced when certain confidence thresholds were unmet, which led us to propose using sentence embeddings for semantic similarity measurements, needing further refinement for underperforming categories.

3.2 Startup Success Prediction Model

For the predictive model, we utilized Python's Scikit-Learn library. Our preliminary task was data cleaning and preprocessing, which entailed handling missing values, employing one-hot encoding to transform categorical variables, and normalizing numerical variables. A Random Forest classifier was used to pinpoint the most influential variables affecting a startup's success. Subsequently, we trained various machine learning models, including logistic regression, gradient-boosted trees, and a multi-layer perceptron (a type of neural network). We applied Principal Component Analysis (PCA) to diminish our data's dimensionality. The models' performance was appraised using metrics such as accuracy, recall, and precision. The multi-layer perceptron neural network model, in combination with PCA, outperformed the other models.

In the second segment of our project, we implemented a methodology to identify the factors contributing to successful startup acquisitions and train a definitive prediction model. Post data cleaning, we visualized the data to gain a better comprehension of the overall distribution of companies across different markets, their founding years, and the number of funding rounds they received. Having acquired a clearer understanding of the data's overall distribution and features' importance, and executed PCA to reduce data dimensionality, we compared the performance of different machine learning algorithms—with and without PCA—in predicting start-up success.

The process of model training commenced with splitting the data into features (X) and the target variable (y), segregating the dataset into input features and the target variable indicating a company's success. The categorical features ('name' and 'short_description') were then one-hot encoded to convert them into a numerical format compatible with machine learning algorithms. Subsequently, we combined numerical and categorical features by merging the one-hot encoded categorical features with the remaining numerical features to construct the complete feature set (X). To tackle the issue of class imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class, 'success' in this instance. The dataset was split into training (70%) and testing (30%) sets, enabling us to evaluate the model's performance on unseen data. Following this, we fitted different models on the training data using the selected features. Each trained model was saved using Python's pickle library for future use. The regression models were then used to predict on the test set. Finally, the model's performance was evaluated using a confusion matrix and classification report, offering metrics such as precision, recall, and F1-score. This methodology allowed us to preprocess the data, balance the dataset, train regression models, and evaluate their performance on the test set to predict companies' success.

To curb the running time, we limited the dataset to 600 companies. The algorithms used include:

Logistic Regression (LR): A statistical technique for predicting binary outcomes based on a set of predictor variables. The formula is as follows:

$$P(Y = 1|X) = 1/(1 + e^{-}(b0 + b1X1 + b2X2 + ... + bnXn))$$

Support Vector Machine (SVM): A machine learning model that finds the optimal decision boundary (hyperplane) separating different classes within the feature space. The decision function is as follows:

$$y(x) = w^T\varphi(x) + b$$

Gradient Boosted Trees (GBT): An ensemble technique that combines multiple weak decision trees to create a robust model. It doesn't have a specific formula as it iteratively adds new trees to the ensemble, each one trained to correct the errors made by the preceding trees.

Random Forest (RF): Another ensemble technique that combines multiple decision trees, each trained on a random subset of the data and features. It doesn't have a specific formula as its final prediction is based on averaging (for regression) or taking a majority vote (for classification) from all individual tree predictions.

Multi-Layer Perceptron classifier (NN - MLP): An artificial neural network made up of multiple layers of interconnected nodes or neurons. It's defined as follows:

$$y(x) = fL(WLfL - 1(WL - 1...f1(W1x + b1)... + bL - 1) + bL)$$

All assumptions for these models were validated, including binary outcome variables, no high intercorrelations or multicollinearity among predictors, no extreme outliers, linear relationships between explanatory variables and the logit of response variables, and a sufficiently large sample size.

3.3 Weighted Scoring System

In the second evaluation phase, we adopted a human-centric approach to empower investor decision-making, drawing from Self-Determination Theory. We established a scoring system incorporating investor assessments in four areas: market, product, team, and funding, applicable to companies flagged as successful by our algorithm with high ESG potential. This risk-averse strategy factors human evaluation of 26 variables.

Based on Retterath's research, our system demands input on five pillars: general company information, market, product, team, and funding, incorporating various factors. The assessment generates an overall company score through weighted inputs, grounded in literature reviews and survey validation from thirty venture capital funds. Scores are normalized to a 100-point scale for consistency.

For ESG Externalities Calculation, we calculate total environmental cost using available activity data, offering a monetary measure of the environmental footprint, following Harvard Business School's methodology. We compute overall monetary impact using quantifiable ESG metrics (e.g., CO2, NOx, SOx, VOC, PM 2.5 emissions) and water usage data, paired with social costs and water scarcity risk coefficients. By multiplying emissions and water usage totals by environmental costs, we establish the total annual environmental cost, providing an economic translation of a company's environmental impact.

The formula to calculate the total environmental cost is:

$$TotalEnvironmentalCost = \sum(EachTypeofEmission * ItsCorrespondingSocialCost) + (WaterUsed * GlobalWaterPrice * WaterScarcityRiskCoefficient)$$

4. Results

Our AI models, trained on open data, underwent a thorough analysis of data distribution. This extensive review, especially for key features such as the geographic location of company headquarters and funding stages, enabled the transformation of qualitative data into quantifiable metrics, a vital step for training our predictive models.

During our initial model training, we found that funding-related features, like 'last_funding_at' and 'days_since_funding', had a significant influence on predicting success. To delve deeper into our dataset and enhance its analysis, we utilized Principal Component Analysis (PCA). After applying PCA, we discovered that the first 30 principal components accounted for roughly 74.25% of the dataset's variance, which was satisfactory for further model training.

The model that excelled in its performance was the neural network MLP classifier, when combined with PCA. This model achieved an accuracy of 84%, along with an impressive balance between precision and recall. These metrics are crucial in the venture capital environment, where incorrect predictions can lead to substantial costs.

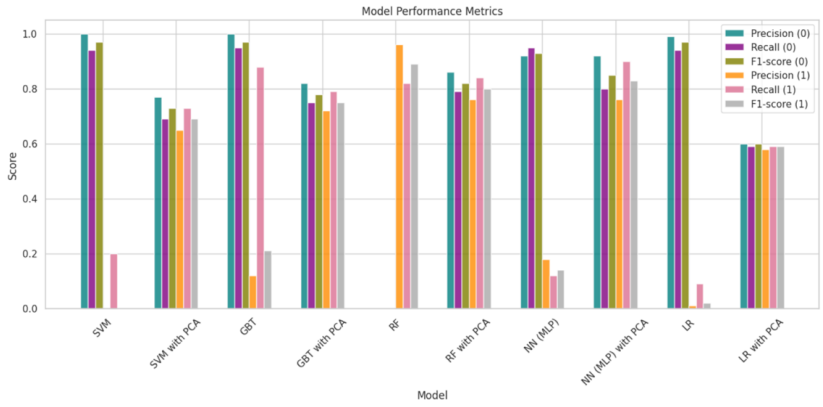


Figure 2: Visualisation - Machine Learning Model Performance Benchmark for Company Success Prediction

Our process aims to eliminate time-consuming tasks such as sourcing and screening, thereby enabling analysts to focus more on collecting accurate data and enhancing their contributions. The practical application of this approach was shown when our system accurately predicted the success of a company called Fyde.

Metric	Score
Overall company score	75.00/100
Overall market score	94.12/100
Overall product score	79.37/100
Overall team score	80.00/100
Overall funding score	61.22/100
Total Score	**79.58/100**

Figure 3: Fyde - Opportunity Scores

Nevertheless, our current system does have certain limitations, including the need for additional validations and interface enhancements. Despite these constraints, our system proves to be an effective tool for comparing different investment opportunities using readily available information.

Our system includes a human-centric mechanism – a comprehensive scoring system – which analysts can use to effectively select companies from the ESG BERT recommended list that align with their investment philosophies. According to the Self-Determination Theory (SDT), the inclusion of human intelligence in this process is crucial for fostering a sense of ownership and purpose, ultimately leading to improved performance.

ESG BERT is a critical tool in our approach for measuring SDG (Sustainable Development Goals) progress, especially in resource-limited environments. This model was designed to classify companies based on the highest ESG (Environmental, Social, and Governance) impact. In practical terms, it demonstrated its efficiency by associating the company Fyde, with the highest ESG BERT label, Data_Security, with a confidence level of 69%.

Despite the reliance on open data and potential human error during manual data entry, our system shows considerable potential. With further research and improvements, we expect an increased capacity to measure SDG progress and make more accurate predictions in the venture capital industry, even in settings with limited resources.

5. Conclusion

Our study underscores the importance of ESG factors in reaching the UN's Sustainable Development Goals (SDGs), particularly in resource-limited venture capital contexts, leveraging open data and AI. Our methodology integrates ESG sentiment analysis into investment evaluations, contributing to all 17 SDGs. The Neural Network MLP classifier improves performance benchmarks, achieving over 80% precision, recall, and accuracy, opening up a broad spectrum of sustainable investment opportunities.

We advocate for ethical AI, emphasizing transparency and investor engagement in decision-making processes, promoting AI-driven sustainable investment practices.

Recognized limitations include needing a more enriched dataset for diverse investment stages, the ESG sentiment analysis's token limit, potential enhancement of the survey-based scoring system with automated data collection, and the ESG cost model's refinement for accurate negative impact assessments. Further exploration of intuition's role in early-stage investments is also necessary.

Despite these, our research highlights ESG factors, open data, and AI's critical role in progressing towards SDGs in resource-constrained environments. It is a significant stride towards AI-supported sustainable venture capital investments and achieving the UN SDGs, laying the groundwork for a more sustainable future.

Ressources

1. United Nations (2015). The 17 Sustainable Development Goals. [online] Available at: https://sdgs.un.org/goals.

2. ISSB (2021). IFRS - International Sustainability Standards Board. [online] Available at: https://www.ifrs.org/groups/international-sustainability-standards-board/#about.

3. Sharpe, S. (2021). Freeing Sisyphus: new rules of thumb for policymaking on decarbonisation - Groupe d'études géopolitiques. [online] Available at: https://geopolitique.eu/en/articles/freeing-sisyphus-new-rules-of-thumb-for-policymaking-on-decarbonisation/#.

4. HBS (2021). Explore Our Data - Impact-Weighted Accounts - Harvard Business School. [online] Available at: https://www.hbs.edu/impact-weighted-accounts/Pages/explore-our-data.aspx.

5. Ross, S. (2019). What kind of financial reporting requirements does GAAP set out? [online] Investopedia. Available at: https://www.investopedia.com/ask/answers/011915/what-kind-financial-reporting-requirements-does-gaap-set-out.asp.

6. VinciWorks (2022). Is ESG Reporting Mandatory in the UK, EU & US? | VinciWorks. [online] Available at: https://vinciworks.com/blog/is-esg-reporting-mandatory-in-the-uk-the-eu-and-the-us/#:~:text=Today%2C [Accessed 7 Apr. 2023].

7. Imperial (2021). Private investor-backed finance is key to funding food security | Imperial News | Imperial College London. [online] Available at: https://www.imperial.ac.uk/news/240292/private-investor-backed-finance-funding-food-security/ [Accessed 7 Apr. 2023].

8. W. Arthur (2012). Details of source not provided.

9. Ruppert (2020). Details of source not provided.

10. Ross, G., Das, S., Sciro, D. and Raza, H. (2021). CapitalVX: A machine learning model for startup selection and exit prediction. The Journal of Finance and Data Science, 7, pp.94–114. doi:https://doi.org/10.1016/j.jfds.2021.04.001.

11. Retterath, A. (2020b). Human Versus Computer: Benchmarking Venture Capitalists and Machine Learning Algorithms for Investment Screening. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3706119.

12. Retterath, A. (2020a). Essays on Machine Learning and the Value of Data in Venture Capital. [online] Universitätsbibliothek der TU München. Available at: https://books.google.co.uk/books/about/Essays_on_Machine_Learning_and_the_Value.html?id=rLtgzgEACAAJ&redir_esc=y [Accessed 8 Apr. 2023].

13. Ang, Y.Q., Chia, A. and Saghafian, S. (2021). Using Machine Learning to Demystify Startups' Funding, Post-Money Valuation, and Success. Innovative Technology at the Interface of Finance and Operations, pp.271–296. doi:https://doi.org/10.1007/978-3-030-75729-8_10.

14. Pan, C., Gao, Y. and Luo, Y. (2018). Machine Learning Prediction of Companies' Business Success. [online] Available at: https://cs229.stanford.edu/proj2018/report/88.pdf.

15. Powell, W. (2021). Machine Learning & Startups: Predicting The Next Unicorn? Thesis committee. [online] Available at: http://arno.uvt.nl/show.cgi?fid=157833 [Accessed 17 Apr. 2023].

16. Bice, S. and Fischer, T.B. (2020). Impact assessment for the 21st century – what future? Impact Assessment and Project Appraisal, 38(2), pp.89–93. doi:https://doi.org/10.1080/14615517.2020.1731202.

17. huggingface.co. (n.d.). nbroad/ESG-BERT · Hugging Face. [online] Available at: https://huggingface.co/nbroad/ESG-BERT [Accessed 8 Apr. 2023].

18. Ganesh, P. (2019). Pre-trained Language Models : Simplified. [online] Medium. Available at: https://towardsdatascience.com/pre-trained-language-models-simplified-b8ec80c62217#:~:text=