Recommender System for Clean Technology Investment to Maximize Impact on Climate Change

Param Somane
psomane@ucsd.edu
University of California, San Diego
La Jolla, CA, USA

Anushri Eswaran
aeswaran@ucsd.edu
University of California, San Diego
La Jolla, CA, USA

Yashraj Gavhane
ygavhane@ucsd.edu
University of California, San Diego
La Jolla, CA, USA

Abstract

Climate change poses significant challenges requiring substantial investment in clean technologies. This paper presents a comprehensive recommender system designed to advise investors on optimal clean technology sectors for investment, considering where, how much, and when to invest to maximize impact on climate change mitigation and societal well-being. Leveraging a rich dataset of global clean technology fundraisers and incorporating macroeconomic, social, and environmental indicators, we perform extensive exploratory data analysis, feature engineering, and develop predictive models for investment recommendations. Our system integrates text embeddings, time-series forecasting using Prophet and LightGBM, and evaluates model performance using metrics such as RMSE and MAE. We also analyze the alignment between funding allocation and greenhouse gas emissions by sector, identifying potentially underserved areas for impactful investment.

Keywords

Recommender System, Climate Change, Time-Series Forecasting, Clean Technology Investment, Feature Engineering, Environmental Indicators

1 Introduction

Climate change is one of the most pressing global challenges of our time, necessitating substantial investments in clean technology sectors to mitigate its effects and promote sustainable development. Investors are increasingly seeking data-driven insights to guide their investment decisions in clean technology, aiming to maximize both environmental impact and financial returns.

This paper presents a comprehensive recommender system designed to advise investors on optimal investment strategies in clean technology sectors. The system aims to answer critical questions such as which sectors to invest in, where geographically, how much to invest, and when to invest to maximize impact on climate change mitigation and societal well-being.

We leverage a rich dataset of global clean technology fundraisers, enriched with macroeconomic, social, and environmental indicators. Through extensive exploratory data analysis and feature engineering, we uncover key patterns and correlations that inform our predictive modeling. Our system integrates advanced machine learning techniques, including text embeddings using Sentence Transformers, time-series forecasting with Prophet and LightGBM, and interaction term creation for capturing complex relationships.

We evaluate our models using standard metrics and perform an analysis of funding allocation versus greenhouse gas emissions by sector to identify potentially underserved areas for impactful investment. Our findings provide valuable insights for investors aiming to make data-driven decisions in the clean technology space.

2 Related Work

Prior research has explored investment decision-making in clean technology using machine learning and recommender systems. Kraemer-Eis et al. [1] utilized supervised machine learning to classify European cleantech firms in the Orbis database, improving sector mapping and highlighting machine learning's potential in identifying investment opportunities.

Zhao et al. [21] proposed a risk-aware recommendation framework for venture capital investment, integrating risk management to balance expected returns and risks, addressing challenges like data sparsity and risk hedging.

Luef et al. [7] developed a recommender system for early-stage enterprise investment, matching investors with startups based on profiles and histories using collaborative filtering.

Lerner and Nanda [5] critiqued venture capital's narrow focus on certain sectors, underscoring the need to expand investment into areas like clean energy and new materials that offer broader societal benefits.

Zhou et al. [22] explored bipartite network projection for recommendation systems, principles applicable to investment recommendations in clean technology sectors.

Jang et al. [3] examined how venture capital investment affects startups' sustainable growth, emphasizing the importance of considering startup characteristics like absorptive capacity in investment decisions.

Our work differs by integrating macroeconomic, social, and environmental indicators into the investment recommendation process to maximize climate impact and societal well-being. We incorporate diverse features relevant to the clean technology domain and leverage advanced machine learning techniques such as text embeddings [17], time-series forecasting with Prophet [18] and LightGBM [4], and feature engineering.

Unlike the European-focused dataset in Kraemer-Eis et al. [1], our global dataset offers a more comprehensive investment landscape. While prior systems [7, 21] focus on matching investors to startups based on historical data, our approach emphasizes aligning investment with environmental impact, identifying underfunded sectors with high greenhouse gas emissions.

By integrating environmental indicators such as greenhouse gas emissions per capita [16], renewables share [11], and temperature anomalies [12], our recommender system prioritizes investments that significantly contribute to climate change mitigation. This addresses the gap highlighted by Lerner and Nanda [5] regarding

the need for venture capital to support a broader range of innovative sectors with societal benefits.

Our work builds upon prior research by incorporating a multidimensional perspective on clean technology investment, leveraging advanced machine learning models, and focusing on maximizing environmental impact, contributing to more effective investment strategies aligned with global sustainability goals.

3 Dataset Creation and Preparation

3.1 Data Collection Methodology

We developed a comprehensive dataset by scraping newsletters and blog posts from climate-focused websites, specifically *CTVC by Sightline Climate* [2] and *Keep Cool* [20], collected up to February 5, 2024. We utilized Selenium for web scraping, extracting fundraiser information such as entity names, amounts raised, dates, descriptions, and associated links.

To extract structured information from unstructured text, we employed GPT-4 for natural language processing. The model was prompted to parse the content and output a tabular representation of the fundraisers, including classification into clean technology sectors.

3.2 Data Cleaning and Processing

We performed data cleaning steps to ensure accuracy and consistency:

- **Normalization of Amounts**: Converted fundraising amounts into numerical USD values using currency conversion rates from the *CurrencyConverter* library.
- **Deduplication**: Applied fuzzy string matching using Levenshtein Distance to identify and merge duplicate fundraising entities.
- **Location Standardization**: Scraped addresses of fundraising businesses from online databases like *Crunchbase* and *CB Insights*. Used *GeoPy* to standardize locations and obtain GPS coordinates.
- **Feature Extraction**: Extracted telephone numbers and country codes from fundraiser descriptions.
- **Geolocation Processing**: Determined country and continent information for each fundraiser using GPS coordinates and the *pycountry* library.

Approximately 30 fundraiser records with amounts exceeding $100 billion tied to large governmental initiatives were omitted to focus on private sector investments. Records related to fines or companies going out of business were also excluded. Manual review ensured location accuracy and correct sector classification, leveraging zero-shot text classification with *facebook/bart-large-mnli* [6].

3.3 Additional Data Sources

We enriched our dataset with macroeconomic, social, and environmental indicators from reputable sources:

- **Macroeconomic Indicators**: GDP [19], GNI per Capita [8].
- **Social Indicators**: Happiness Score [9], Support for Climate Policies [14], Belief in Climate Threat [13], Support for Public Action [15].
- **Environmental Indicators**: Renewables Share [11], GHG Emissions per Capita [16], Primary Energy Consumption [10], Temperature Anomaly [12].

4 Exploratory Data Analysis

4.1 Dataset Statistics

To provide an overview of the fundraisers dataset, we present key statistical measures (Table 1).

Table 1: Summary Statistics of Fundraisers Dataset

Statistic	Amount Raised (USD)
Number of Records	3,348
Mean	$105,196,754
Median	$10,000,000
Standard Deviation	$1,038,774,190
Minimum	$20,000
Maximum	$60,000,000,000

4.2 Correlation Analysis

We computed the correlation matrix for key numerical variables to identify significant relationships (Figure 1).

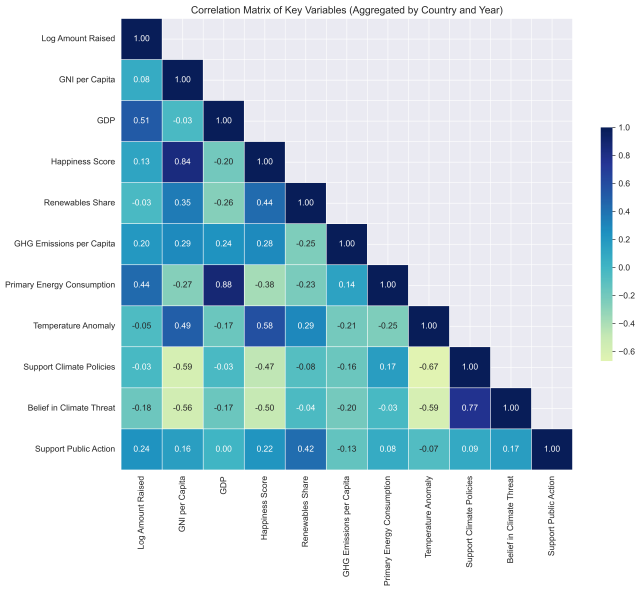


Figure 1: Correlation Matrix of Key Numerical Variables

Observations:

- **Log Amount Raised** is positively correlated with **GDP** and **Primary Energy Consumption**, suggesting that larger economies with higher energy consumption attract more investment in clean technology.
- **GNI per Capita** has a strong positive correlation with **Happiness Score**, indicating that wealthier countries tend to report higher happiness levels.

- **Support for Climate Policies** is negatively correlated with **GNI per Capita** and **Temperature Anomaly**, suggesting that higher-income countries and those experiencing greater temperature anomalies may show less public support for climate policies.

Interpretation: These correlations highlight the complex interplay between economic factors, public opinion, and investment in clean technology. The positive correlation between investment amounts and both **GDP** and **Primary Energy Consumption** suggests that larger economies with higher energy needs are attracting more investment, potentially due to greater resources and market opportunities. However, the negative correlation between **Support for Climate Policies** and both **GNI per Capita** and **Temperature Anomaly** indicates that wealthier countries and those experiencing greater climate impacts may exhibit less public support for climate initiatives. This counterintuitive finding may be influenced by factors such as a perceived ability to adapt, differing media narratives, or political influences, and warrants further investigation.

4.3 Funding Distribution by Clean Technology Sector

We analyzed total funds raised across clean technology sectors to identify areas with high investment and those with potential for growth (Figure 2).

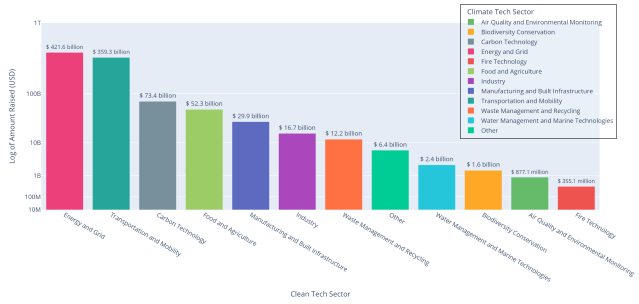


Figure 2: Funding Distribution by Clean Technology Sector

Key Insights:

- **Top-Funded Sectors**: Energy and Grid ($309.8B) and Transportation and Mobility ($338.6B) lead in funding, reflecting their high impact potential.
- **Mid-Funded Sectors**: Sectors such as Food and Agriculture ($31.1B) and Manufacturing and Built Infrastructure ($27.9B) have substantial but lower funding.
- **Lower-Funded Sectors**: Water Management and Marine Technologies ($2.39B), Biodiversity Conservation ($1.56B), and Air Quality and Environmental Monitoring ($0.87B) received limited funding, suggesting these areas may be overlooked despite their importance to sustainability.

Interpretation: The funding distribution highlights investor priorities in sectors with immediate, large-scale impact, while some critical areas—such as Water Management and Biodiversity Conservation—remain underfunded. These gaps suggest opportunities for targeted investment in less saturated but environmentally significant sectors.

The bar plot uses a **logarithmic scale** to clearly show disparities across sectors, emphasizing the concentration of funding in high-impact areas. This insight can inform the recommender system by identifying underfunded sectors with potential for impactful investment.

4.4 Distribution of Temperature Anomalies

We examined the distribution of temperature anomalies across countries (Figure 3).

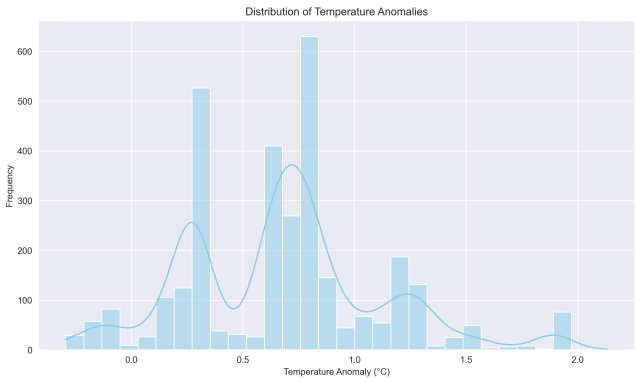


Figure 3: Distribution of Temperature Anomalies

Observations:

- The distribution peaks around 0.3°C and 0.7°C, indicating that most countries experience moderate temperature anomalies.
- The distribution is right-skewed with a long tail towards higher anomalies, signifying fewer countries facing extreme anomalies above 1.0°C.

Interpretation: The prevalence of moderate anomalies suggests that climate change impacts are widespread but vary in intensity. Regions with higher anomalies might be priority targets for investment to mitigate severe climate effects.

4.5 Investment vs. Temperature Anomalies

We analyzed the relationship between the amount raised and temperature anomalies (Figure 4).

Observations:

- No strong correlation between the amount raised and temperature anomalies.
- Investments are spread across regions with varying temperature anomalies.

Interpretation: The lack of correlation between investment amounts and temperature anomalies suggests that current investment patterns are not directly aligned with the severity of climate impacts in different regions. This presents an opportunity for investors and policymakers to realign funding strategies to target areas where investments can have the most significant environmental impact.

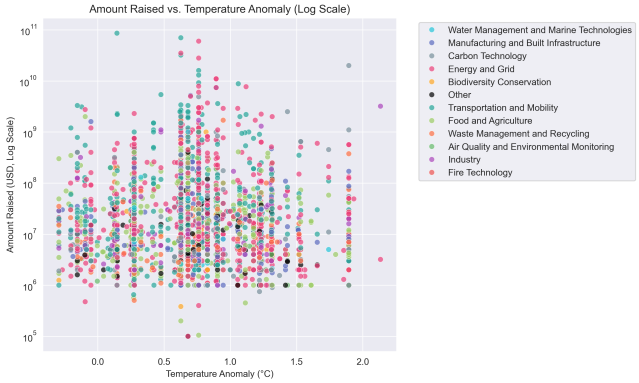


Figure 4: Amount Raised vs. Temperature Anomalies

5 Methodology

5.1 Predictive Task Definition

The primary predictive task is to develop a recommender system that advises investors on:

- Which clean technology sectors to invest in.
- Where (geographically) to invest.
- How much to invest.
- When to invest.

The goal is to maximize impact on climate change mitigation and societal well-being.

5.2 Implementation Details

5.2.1 Model Implementation. To address the predictive tasks, we implemented a combination of machine learning techniques tailored to our objectives.

Text Embedding of Fundraiser Descriptions. We utilized Sentence Transformers [17] to generate embeddings for fundraiser descriptions. Specifically, the *all-MiniLM-L6-v2* model was chosen for its balance between performance and computational efficiency. This allowed us to capture semantic similarities between fundraisers effectively.

Feature Engineering. Additional features were engineered based on the correlation analysis and domain knowledge:

- **Lag Features**: Created lagged variables (e.g., *lag_1_amount*, *lag_3_amount*) to capture temporal dependencies in fundraising amounts.
- **Moving Averages**: Computed moving averages (e.g., *ma_3_amount*) to smooth out short-term fluctuations.
- **Interaction Terms**: Generated interaction terms between macroeconomic indicators (e.g., *gni_renewables_interaction*) to model complex relationships.

5.2.2 Similarity Computation. To recommend similar fundraisers, we compute a combined similarity score that integrates textual, sectoral, and country-level similarities. The computation involves the following steps:

Description Similarity. We obtain description embeddings $\mathbf{e}_i$ for each fundraiser i using the Sentence Transformer model *all-MiniLM-L6-v2*. The description similarity between two fundraisers i and j is calculated using cosine similarity:

$$S_{\text{desc}}(i,j) = \cos(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}$$

Sector Similarity. We represent each clean technology sector with feature vectors derived from sector-level indicators such as emissions and funding data. The sector similarity is computed as:

$$S_{\text{sector}}(i,j) = \cos(\mathbf{s}_i, \mathbf{s}_j)$$

where $\mathbf{s}_i$ is the sector feature vector for fundraiser i.

Country Similarity. Country similarity is based on macroeconomic and environmental indicators aggregated at the country level. We compute the similarity between countries using:

$$S_{\text{country}}(i,j) = \cos(\mathbf{c}_i, \mathbf{c}_j)$$

where $\mathbf{c}_i$ is the country feature vector for fundraiser i.

Combined Similarity Score. The overall similarity score between fundraisers i and j is a weighted sum of the individual similarities:

$$S_{\text{combined}}(i,j) = \alpha \cdot S_{\text{desc}}(i,j) + \beta \cdot S_{\text{sector}}(i,j) + \gamma \cdot S_{\text{country}}(i,j)$$

where the weights are set to $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$, ensuring $\alpha + \beta + \gamma = 1$. These weights prioritize textual content while still considering sectoral and geographical factors.

Recommendation Generation. For a given fundraiser, we compute S_{combined} with all other fundraisers and rank them in descending order to generate recommendations. This method allows us to recommend fundraisers that are similar in content, operate in related sectors, and are situated in economically or environmentally comparable countries.

Regression Models. For predicting investment amounts and timing, we employed both Prophet [18] and LightGBM [4]. Prophet was utilized for time-series forecasting to model temporal trends and seasonality in investment amounts. However, to capture complex nonlinear relationships and interactions in the data, we integrated LightGBM, a gradient boosting framework known for its efficiency and accuracy with large datasets.

LightGBM Implementation. We prepared the data by sorting fundraisers by country, clean technology sector, and date. Lag features such as *lag_1_amount*, *lag_3_amount*, and moving averages like *ma_3_amount* were created to capture temporal dependencies. These features were included in the model along with macroeconomic and environmental indicators. The LightGBM regressor was configured with a learning rate of 0.01 and 1000 estimators to ensure gradual learning and sufficient model capacity, while setting the number of leaves to 31 and subsample ratio to 0.8 to balance complexity and prevent overfitting.

Feature Importance. After training the LightGBM model, we analyzed feature importances to understand the factors influencing investment amounts. The model provided insights into which features were most predictive, aiding in interpreting the results and refining the model.

5.3 Alternative Models and Justification

5.3.1 Alternative Models Considered and Challenges Faced. Initially, we explored collaborative filtering methods such as Bayesian Personalized Ranking (BPR) and latent factor models like Singular Value Decomposition (SVD). These models are effective in traditional recommender systems where user-item interaction matrices are dense and user preferences can be inferred from historical behavior.

However, our dataset presented unique challenges:

- **Data Sparsity**: The interaction matrix between investors and fundraisers was extremely sparse, as there were only a few investing entities that had multiple fundraiser records in the dataset.
- **Cold-Start Problem**: Many fundraisers and investors in the clean technology sector are new or have limited historical data, making it difficult for models like BPR and SVD to make accurate predictions.
- **Lack of Explicit Feedback**: The absence of explicit user ratings or preferences hindered the applicability of latent factor models that require such feedback for training.

We also attempted to apply matrix factorization using investment amounts as implicit feedback. However, the models failed to converge to meaningful representations due to the sparsity of the data and the absence of repeated interactions. Popularity-based recommendation methods were considered but did not account for investor preferences or alignment with climate impact goals, leading to suboptimal recommendations.

5.3.2 Justification for Chosen Model. Given these challenges, we opted for a content-based approach that leverages the rich textual information available in fundraiser descriptions and integrates macroeconomic and environmental indicators. The use of Sentence Transformers for text embeddings allowed us to capture semantic similarities between fundraisers effectively. By combining description, sector, and country similarities, we tailored recommendations to align with investors' interests and climate impact objectives.

5.3.3 Optimization Strategies and Challenges. Processing text descriptions and computing similarity matrices posed computational challenges. We addressed this by utilizing efficient pre-trained models (*all-MiniLM-L6-v2*) and leveraging GPU acceleration where available. To prevent overfitting in our regression models, we employed regularization techniques inherent in LightGBM, such as limiting the number of leaves to 31 and applying a subsampling coefficient of 0.8. The sparsity of time-series data for certain sectors and countries affected the forecasting models. We mitigated this by aggregating data at monthly intervals and using log transformations to stabilize variance.

5.3.4 Feature Representations.

Effective Features:

- **Lag Features and Moving Averages**: Captured temporal dependencies effectively.
- **Macroeconomic Indicators**: Provided context on economic conditions influencing investments.
- **Temperature Anomaly**: Included environmental impact considerations, aligning with climate investment goals.

Ineffective Features:

- **Interaction Terms**: Some interaction terms did not significantly improve model performance, possibly due to multicollinearity or lack of complex nonlinear relationships captured by the model.

5.3.5 Strengths and Weaknesses of the Models.

Strengths:

- **Content-Based Similarity**: Effectively handles new or unseen fundraisers by relying on their content, overcoming the cold-start problem.
- **Integration of Multiple Similarities**: Combining textual, sectoral, and country-level similarities provides a holistic recommendation approach.
- **LightGBM Regression**: Captures complex nonlinear relationships and interactions between features, improving forecasting accuracy.

Weaknesses:

- **Lack of User Interaction Data**: Without explicit investor preferences or interaction history, personalization is limited.
- **Computational Complexity**: Computing similarity matrices for large datasets can be resource-intensive.
- **Limited Interpretability**: While LightGBM provides feature importances, the overall model can be seen as a black box, making it harder to interpret specific parameter effects.

5.3.6 Investment Timing and Amount Prediction. In addition to Prophet and LightGBM, we evaluated simpler forecasting methods, such as using historical averages and naive forecasts based on the previous period's investment amounts.

Baseline Model. The baseline model predicts future investment amounts using the mean of past investments. We compared our LightGBM model's performance against this baseline using paired t-tests.

Evaluation Metrics and Statistical Significance. The paired t-test results on the original scale (Table 6) show a significant improvement of our model over the baseline with a p-value of 0.0393. However, on the log-transformed scale, the t-test yielded a t-statistic of 1.5314 and a p-value of 0.1352, indicating that the improvement was not statistically significant at the 5% level. The lack of statistical significance on the log scale suggests that our model's performance is not substantially better than the baseline when considering relative changes in investment amounts. This could be due to the high variability and unpredictability inherent in investment data on a logarithmic scale.

6 Results

6.1 Evaluation Results

6.1.1 Similar Fundraiser Recommendation. The recommender system was evaluated on the test set, yielding the following metrics:

Table 2: Evaluation Metrics for Similar Fundraiser Recommendations

Metric	Precision@5	Recall@5	NDCG@5	MRR@5
Value	0.0231	0.0019	0.0234	0.0498

Interpretation: The evaluation metrics indicate that while the precision and recall are low due to dataset sparsity and the novelty of fundraisers (cold-start problem), the NDCG and MRR values suggest that relevant fundraisers, when recommended, are ranked higher. This reflects the model's ability to prioritize more relevant recommendations effectively.

6.2 Case Study: Samsara Eco Recommendation

To demonstrate the effectiveness of our recommendation system, we present an example where we generate recommendations for the fundraising entity *Samsara Eco*, an Australia-based recycling company utilizing enzyme-based technology to break down plastic.

Table 3: Top 4 Recommendations for *Samsara Eco*

Rank	Fundraising Entity	Country	Similarity Score
1	Protein Evolution	USA	0.8549
2	Epoch Biodesign	UK	0.8355
3	Impact Recycling	UK	0.8316
4	Umincorp	Netherlands	0.8233

Interpretation: The recommendations primarily consist of companies involved in waste management and recycling, aligning with *Samsara Eco*'s sector. The top recommendation, *Protein Evolution*, is a US-based recycling company using enzyme-based technology, indicating high textual and sectoral similarity.

This example showcases the system's ability to identify fundraisers with similar technologies and objectives, even across different countries, highlighting the model's effectiveness in capturing nuanced similarities.

6.2.1 Investment Timing and Amount Prediction. For time-series forecasting, the models achieved the following:

Table 4: Forecast Evaluation Metrics

Metric	Log Scale	Original Scale
RMSE	2.4843	$11,867,685,005.33
MAE	2.0734	$4,276,705,464.96

Interpretation: The RMSE and MAE on the log scale indicate reasonable predictive accuracy considering the variability in the data. On the original scale, the large RMSE reflects the wide range of investment amounts, highlighting the challenges in predicting high-magnitude financial data.

Top Investment Timing and Amount Recommendations. Using the trained LightGBM model, we forecasted future investment amounts for the *Energy and Grid* sector in the *United States of America* over the next three years. The top investment timing recommendations are presented in Table 5.

Table 5: Top Investment Timing and Amount Recommendations

Date	Predicted Amount (USD)	Rank
December 31, 2026	$959,993,500	1
December 31, 2025	$800,031,400	2
December 31, 2024	$666,681,400	3
November 30, 2026	$606,418,000	4
September 30, 2026	$510,759,300	5

Interpretation: The model forecasts increasing investment amounts in the *Energy and Grid* sector, with the highest predicted amount in December 2026. This trend aligns with the growing emphasis on renewable energy and infrastructure development in the United States. Notably, the top investment recommendations are concentrated in the fourth quarter (Q4) of each year, suggesting a potential seasonal or cyclical pattern where investments peak toward the calendar year-end. This pattern may reflect budget cycles, strategic planning processes, or an annual increase in funding demand in anticipation of new fiscal initiatives.

Visualization. We visualized the forecasting results using Prophet and LightGBM (Figure 5).

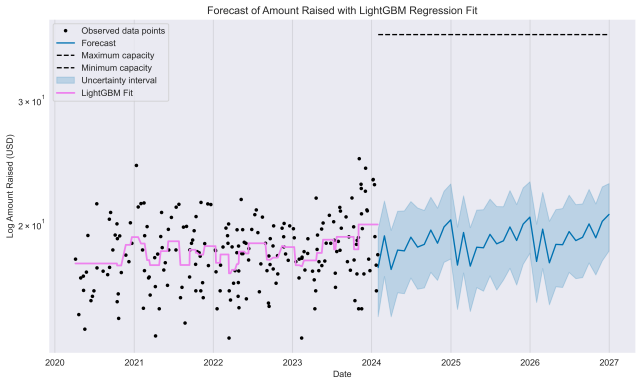


Figure 5: Energy and Grid Sector Investment Amount Forecasting Using Prophet and LightGBM

Interpretation: The plot illustrates the model's fit to historical data and its predictions for future investment amounts. The upward trend in the forecast aligns with expectations given the increasing focus on clean energy. The LightGBM regression fit closely follows the historical data points, demonstrating the model's effectiveness in capturing temporal patterns.

6.3 Statistical Significance Testing

We conducted a paired t-test to compare our model's performance against a baseline model using the errors on the original scale for fundraiser amounts. The baseline model predicts investment amounts using a simple historical average of past investments.

The null hypothesis (H_0) is that there is no significant difference between our model and the baseline model. The alternative hypothesis (H_1) is that our model performs significantly better than the baseline.

Table 6: Paired T-test Results

	T-statistic	P-value
Original Scale	2.1460	0.0393

Interpretation: With a t-statistic of 2.1460 and a p-value of 0.0393, we reject the null hypothesis at the 5% significance level. This indicates that our model performs significantly better than the baseline model, providing valuable predictive insights for investment timing and amounts in the clean technology sector.

6.4 Sector-Specific Time-Series Analysis

We analyzed investment trends across different clean technology sectors (Figures 6 and 7).

Observations:

- **Energy and Grid** and **Transportation and Mobility** sectors exhibit stable funding trends with high investment amounts.
- **Carbon Technology** shows a gradual upward trend, indicating growing interest.
- Sectors like **Water Management** and **Biodiversity Conservation** have lower funding levels but show potential for growth.

Interpretation: The stable funding in mature sectors suggests market saturation, whereas emerging sectors with upward trends may offer high-impact investment opportunities.

6.5 Funding Allocation vs. Emissions Impact

Our analysis revealed a misalignment between funding allocation and greenhouse gas emissions by sector (Figure 8).

Observations:

- **Energy and Grid** and **Transportation and Mobility** sectors receive substantial funding, aligning with their high emissions.
- Sectors like **Manufacturing and Built Infrastructure** (Emissions: 37.4 Gt CO_2e; Funding: \$27.9B) and **Food and Agriculture** (Emissions: 23.8 Gt CO_2e; Funding: \$31.1B) have high emissions but relatively lower funding.
- **Biodiversity Conservation** (Emissions: 5.4 Gt CO_2e; Funding: \$1.56B) is significantly underfunded despite its environmental importance.

Interpretation: There is potential for impactful investment in underfunded sectors with high emissions, such as **Manufacturing and Built Infrastructure** and **Food and Agriculture**. Redirecting funds to these areas could enhance overall climate change mitigation efforts. Investors may consider these sectors as opportunities for high environmental impact and potentially untapped financial returns.

6.6 Global Investment Distribution

We analyzed the global distribution of fundraisers and total amounts raised by continent (Figure 9).

Observations:

- **North America** leads in both the number of fundraisers and total funding.
- **Europe** follows, with significant but lower figures.
- **Asia** shows moderate activity, indicating potential growth opportunities.
- **Africa** and **South America** have minimal fundraising activity, highlighting underserved regions.

Interpretation: Investment is concentrated in developed regions, suggesting a need to encourage funding in emerging markets, which may offer high-impact opportunities due to less competition and significant environmental challenges.

7 Code and Data Availability

The source code for this research project is available on GitHub at https://github.com/ChestnutKurisu/CSE258_A2_FA24. The climate tech fundraiser dataset used in this study can be accessed on HuggingFace at https://huggingface.co/datasets/Xcissa/climate-codex.

8 Conclusion

This study presents a recommender system that integrates diverse data sources and advanced modeling techniques to advise investors on clean technology investments, aiming to maximize impact on climate change mitigation and societal well-being. Our analysis highlights misalignments between current investment patterns and areas of greatest environmental need, particularly in underfunded sectors with high emissions such as **Manufacturing and Built Infrastructure** and **Food and Agriculture**.

The evaluation results demonstrate the system's capability to provide meaningful recommendations despite challenges like data sparsity. The statistically significant improvement over baseline models in investment forecasting underscores the practical utility of our approach.

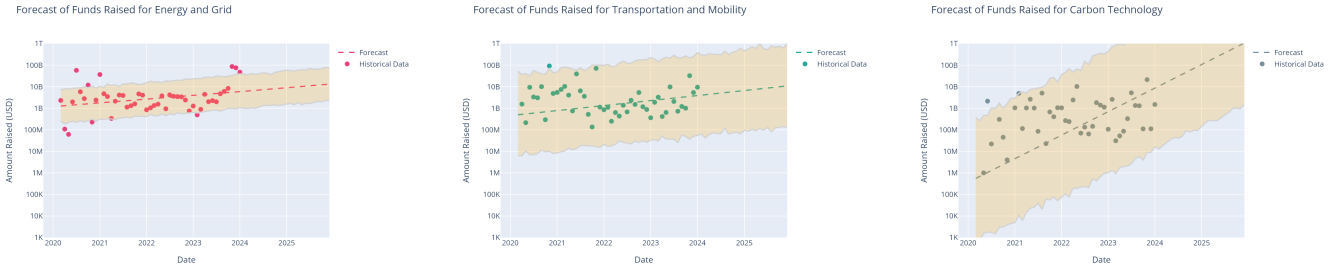
Future work includes enhancing the models with more granular data, incorporating real-time indicators, and exploring policy frameworks to better align investment strategies with global climate goals. Additionally, expanding the system to consider the socio-economic context of underrepresented regions could further optimize the impact of clean technology investments.

References

[1] Matteo Ambrois, Vincenzo Butticè, Federico Caviggioli, Giovanni Cerulli, Annalisa Croce, Antonio De Marco, Andrea Giordano, Giuliano Resce, Laura Toschi, Elisa Ughetto, and An Zinilli. 2023. *Using machine learning to map the European cleantech sector.* EIF Working Paper Series 2023/91. European Investment Fund (EIF). https://ideas.repec.org/p/zbw/eifwps/202391.html

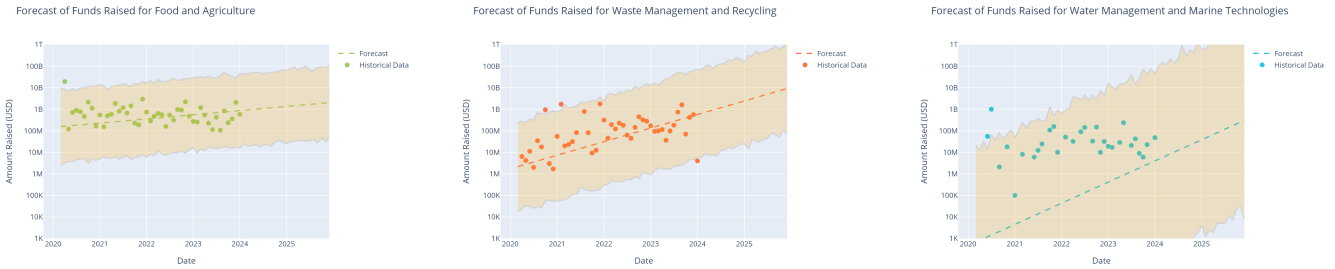
[2] Climate Tech VC. 2024. Climate Tech VC. https://www.ctvc.co/ Climate tech's leading perspective newsletter and market intelligence platform.

[3] Jihye Jeong, Juhee Kim, Hanei Son, and Dae-il Nam. 2020. The Role of Venture Capital Investment in Startups' Sustainable Growth and Performance: Focusing



(a) Energy and Grid **(b) Transportation and Mobility** **(c) Carbon Technology**

Figure 6: Time-Series of Investment Amounts by Sector (Part 1)



(a) Food and Agriculture **(b) Waste Management and Recycling** **(c) Water Management and Marine Technologies**

Figure 7: Time-Series of Investment Amounts by Sector (Part 2)

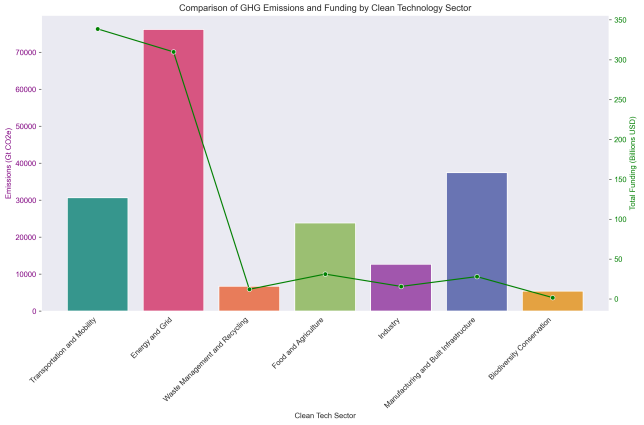


Figure 8: Comparison of GHG Emissions and Funding by Clean Technology Sector

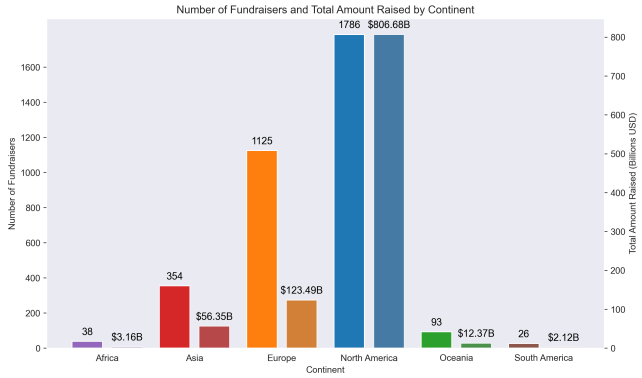


Figure 9: Number of Fundraisers and Total Amount Raised by Continent

on Absorptive Capacity and Venture Capitalists' Reputation. *Sustainability* 12, 8 (2020). https://doi.org/10.3390/su12083447

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

[5] J. Lerner and R. Nanda. 2020. Venture Capital's Role in Financing Innovation: What We Know and How Much We Still Need to Learn. *Journal of Economic Perspectives* 34, 3 (2020), 237–261. https://doi.org/10.1257/jep.34.3.237

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

[7] Johannes Luef, Christian Ohrfandl, Dimitris Sacharidis, and Hannes Werthner. 2020. A recommender system for investing in early-stage enterprises. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (Brno, Czech

Republic) *(SAC '20)*. Association for Computing Machinery, New York, NY, USA, 1453–1460. https://doi.org/10.1145/3341105.3375767
[8] Our World in Data. 2022. Gross National Income Per Capita. https://ourworldindata.org/grapher/gross-national-income-per-capita-undp UNDP, Human Development Report (2022) – with minor processing by Our World in Data.
[9] Our World in Data. 2023. Happiness Index (Cantril Ladder). https://ourworldindata.org/grapher/happiness-cantril-ladder World Happiness Report (2023) – processed by Our World in Data.
[10] Our World in Data. 2023. Primary Energy Consumption. https://ourworldindata.org/grapher/primary-energy-cons U.S. Energy Information Administration (2023); Energy Institute - Statistical Review of World Energy (2023) – with major processing by Our World in Data.
[11] Our World in Data. 2023. Renewable Share of Energy. https://ourworldindata.org/grapher/renewable-share-energy Energy Institute - Statistical Review of World Energy (2023) – with major processing by Our World in Data.
[12] Our World in Data. 2024. Average Temperature Anomaly, Global. https://ourworldindata.org/grapher/temperature-anomaly Met Office Hadley Centre (2024).
[13] Our World in Data. 2024. Beliefs about Climate Change. https://ourworldindata.org/grapher/share-believe-climate Vlasceanu et al. (2024). Addressing climate change with behavioral science: A global intervention tournament in 63 countries.
[14] Our World in Data. 2024. Support for Policies to Tackle Climate Change. https://ourworldindata.org/grapher/support-policies-climate Vlasceanu et al. (2024). Addressing climate change with behavioral science: A global intervention tournament in 63 countries.
[15] Our World in Data. 2024. Support for Public Action on Climate Change. https://ourworldindata.org/grapher/support-public-action-climate Andre et al. (2024). Globally representative evidence on the actual and perceived support for climate action.
[16] Our World in Data. 2024. Total Greenhouse Gas Emissions Per Capita. https://ourworldindata.org/grapher/per-capita-ghg-emissions Climate Watch (2023); Population based on various sources (2023) – with major processing by Our World in Data.
[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410
[18] Sean J. Taylor and Benjamin Letham. 2018. Forecasting at Scale. *The American Statistician* 72, 1 (2018), 37–45. https://doi.org/10.1080/00031305.2017.1380080 arXiv:https://doi.org/10.1080/00031305.2017.1380080
[19] The World Bank. 2024. Global GDP. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD World Bank national accounts data, and OECD National Accounts data files.
[20] Nick van Osdol. 2024. Keep Cool. https://www.keepcool.co/ Newsletter telling underpriced stories at the intersection of climate & business.
[21] Xiaoxue Zhao, Weinan Zhang, and Jun Wang. 2015. Risk-Hedged Venture Capital Investment Recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems* (Vienna, Austria) *(RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/2792838.2800181
[22] Tao Zhou, Jie Ren, Matú š Medo, and Yi-Cheng Zhang. 2007. Bipartite network projection and personal recommendation. *Phys. Rev. E* 76 (Oct 2007), 046115. Issue 4. https://doi.org/10.1103/PhysRevE.76.046115

A Additional Visualizations

A.1 Geographical Distribution of Fundraisers

Due to their size, the geographical plots are included in the appendix.

Observations:

- High concentration of fundraisers in **North America** and **Europe**, especially in **Energy and Grid** and **Transportation and Mobility**.
- Sparse activity in **Africa**, **South America**, and parts of **Asia**, indicating potential for growth.
- Sector diversity is greater in developed regions, suggesting mature investment ecosystems.

A.2 Geographical Distribution in the USA

Observations:

- Significant fundraising activity in **California** and the **Northeastern U.S.** across multiple sectors.
- **Texas** and the **Midwest** also show notable activity, primarily in **Energy** and **Manufacturing**.
- Low activity regions may represent untapped markets for investors.

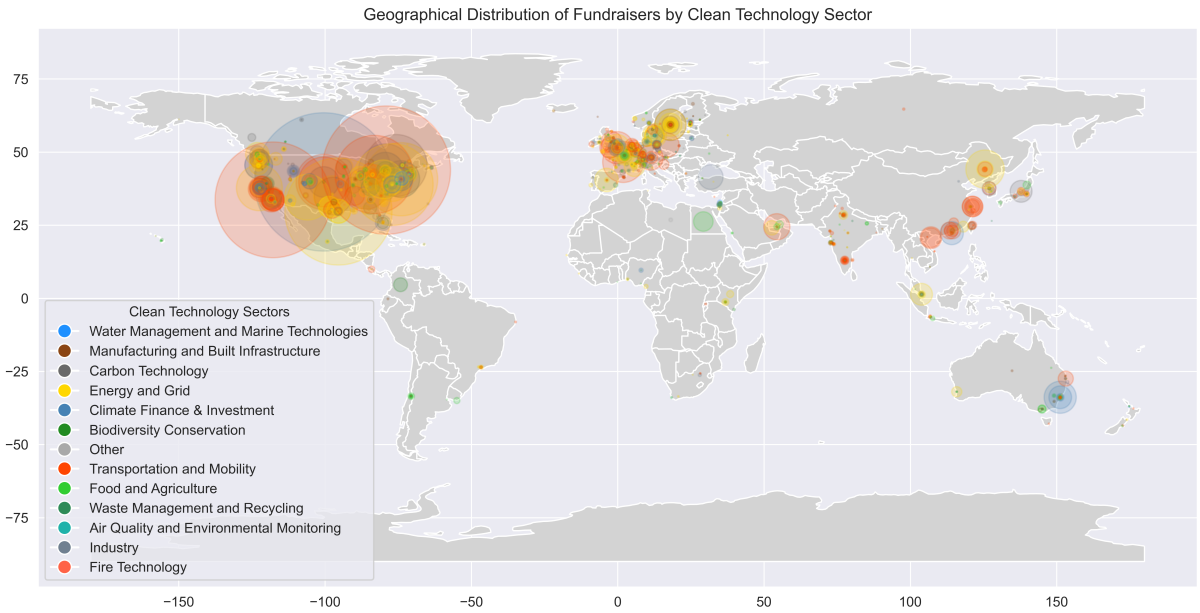


Figure 10: Global Distribution of Fundraisers by Clean Technology Sector

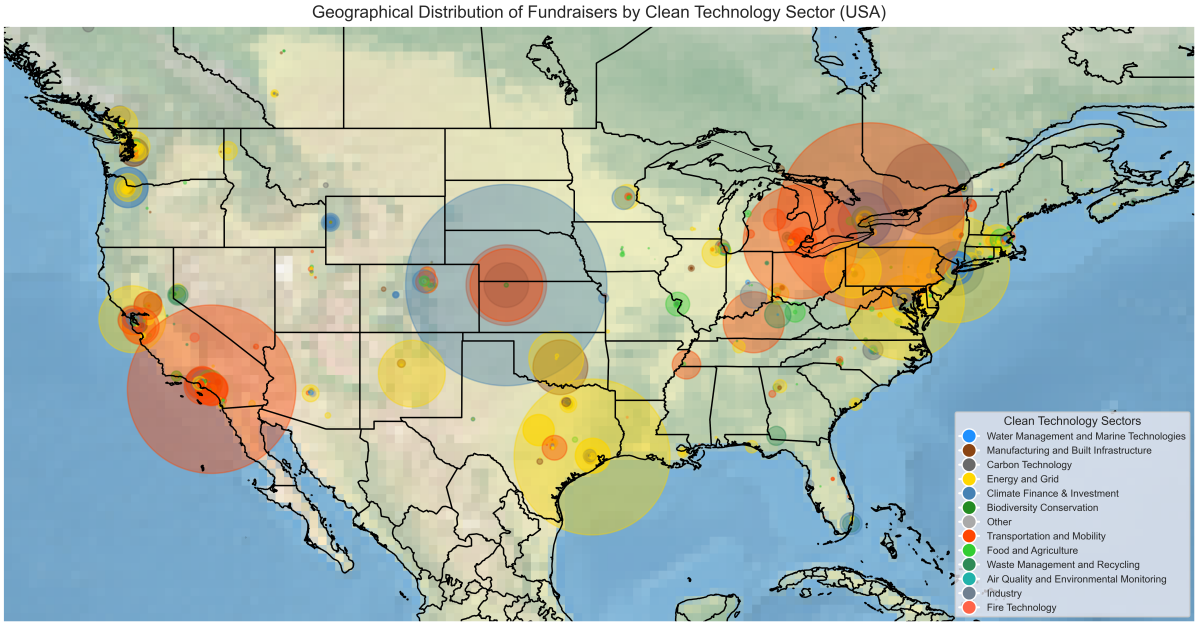


Figure 11: Distribution of Fundraisers in the USA by Clean Technology Sector