Application of Startup Success Prediction Models and Business Document Extraction Using Large Language Models to Enhance Due Diligence Efficiency

Vito Christian Samudra and Dicky Prima Satya

September 6, 2024

Application of Startup Success Prediction Models and Business Document Extraction Using Large Language Models to Enhance Due Diligence Efficiency

Vito Christian Samudra
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
18220080@std.stei.itb.ac.id

Dicky Prima Satya
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
dicky@itb.ac.id

Abstract—**Startups face extreme uncertainty and high failure rates, posing challenges for investors in identifying promising ventures. This research, based on a case study and interviews at a prominent Indonesian corporate venture capital firm, explores the due diligence process, typically taking 4-6 weeks depending on data completeness. Using Large Language Model (LLM) and Machine Learning (ML) technologies developed with the Team Data Science Process (TDSP) methodology, the research aims to enhance due diligence efficiency. Key development steps include data integration, ML model creation for startup success classification, and the integration of OpenAI's GPT-4 and Google Search APIs for comprehensive business analysis. The system's dashboard offers features such as pitch deck, financial, market trends, competitor, and founding team analyses, along with startup success prediction using the XGBoost model. This model, deployed via Flask, demonstrated consistent results through cross-validation. Customer acceptance testing, conducted with eight experienced startup investors, yielded a high satisfaction rate of 4.50 out of 5.00, indicating strong approval of the system's effectiveness.**

Keywords—Venture Capital, Large Language Model (LLM), Machine Learning, GPT-4, Google Search API

I. Introduction

A startup can be defined as an institution founded to develop new products and/or services under conditions of extreme uncertainty [1]. The potential for failure in a startup is very high, with more than two-thirds of startups failing to provide positive revenue to investors [2] and 65% of new startups failing within the first ten years [3]. Despite the high failure rate, startups are renowned for their exponential business growth, making them an attractive investment opportunity for investors seeking high potential returns despite the inherent risks.

The case study and interview process were conducted at a prominent corporate venture capital firm in Indonesia. This firm focuses on investing in companies that innovate and add value to its ecosystem. An Investment Manager at the firm provides insight into the due diligence process. Typically, the due diligence process for an investment plan takes approximately 4-6 weeks, depending on the completeness of the data provided by the target company. Due diligence at a venture capital firm featured in the case study aims to conduct comprehensive business analysis to identify critical aspects before committing to investments. This process involves examining the target company's management background, financial health, business model, value proposition, and associated risks, ultimately increasing investor confidence and minimizing potential risks. However, there are significant challenges during due diligence. Incomplete or inaccurate information from the target company can hinder risk evaluation and decision-making. Additionally, assessing risks can be difficult due to insufficient information about the target company's industry or comparable companies.

The development of Artificial Intelligence (AI) can assist investors in enhancing due diligence efficiency. Hone Capital claims that by combining Machine Learning with human recommendations, they can increase investment deal success by up to 3.5 times the industry average [4]. By using AI as a basis for decision-making, investors can significantly enhance the accuracy of their decisions, potentially doubling it [5]. The case study and interview underscore the need for AI to enhance the accuracy and efficiency of the due diligence process.

II. Methodology

The methodology used in this study is the Team Data Science Process (TDSP). TDSP is a data science methodology that employs an agile and iterative approach to efficiently deliver predictive analytics solutions and intelligent applications [6]. One of the key reasons for using this methodology is its user-centric approach, which is not typically found in other methodologies, particularly through its customer acceptance phase. TDSP consists of five main stages: business understanding, data acquisition and understanding, modeling, deployment, and customer acceptance.

1. Business Understanding: This stage involves understanding user needs. It is conducted through interviews and observations at a venture capital firm featured in the case study to gather the necessary requirements.

2. Data Acquisition and Understanding: The objective of this stage is to produce clean, complete, and high-quality data. This step is crucial to ensure that the data is ready for modeling.

3. Modeling: This stage involves determining the appropriate data features for the machine learning model, creating a machine learning model with the highest prediction accuracy, and selecting the suitable machine learning model for deployment.

4. Deployment: This stage focuses on integrating the developed system to ensure it is ready for the customer acceptance phase.

5. Customer Acceptance: The final stage ensures that the system meets the customer's objectives.

III. Solution Design

This section outlines the overall solution design following the Team Data Science Process (TDSP) methodology.

A. *Business Understanding*

Based on the results of the interviews conducted, the gap analysis is outlined in Table 1 to understand the ideal condition to be achieved.

TABLE I. GAP ANALYSIS

Current State	Ideal State	Gap
Data and information collection is manual.	Data and information collection is automated.	Manual and time-consuming.
Business and financial analysis is manual.	Analysis with the assistance of AI to improve efficiency and accuracy.	Less efficient and prone to human error.
Due diligence process takes 4-6 weeks.	Due diligence process is completed faster.	Long duration and labor-intensive.
Challenges with incomplete/inaccurate information.	Information collected is accurate and complete for more precise risk assessment.	Difficulty in obtaining accurate and complete information.

After identifying the gaps in the current state, the Table II outlines the system requirements needed to address these gaps. These requirements aim to optimize the due diligence process.

TABLE II. FUNCTIONAL REQUIREMENTS

Code	Description
F-01	The system can identify and analyze risks based on historical data and business documents.
F-02	The system should use predictive models to assess the potential success and failure of startups.
F-03	The system displays a dashboard of startup assessment metrics.
F-04	The system provides external information for automated competitor and industry trend analysis.
F-05	The system can automate the due diligence process through the analysis of business and financial documents.

B. *Data Acquisition and Understanding*

The data utilized in this research includes Pitch Decks, financial reports, and the Startup Success Prediction dataset sourced from Kaggle, as detailed in Table III. The Pitch Decks and financial reports provide crucial insights into the startup's business model, financial health, and overall potential. Additionally, the Startup Success Prediction dataset from Kaggle is employed to train machine learning models designed to predict startup success.

TABLE III. DATA SOURCES

Data	Description	Data Type
Pitch Deck	Presentation used to provide an overview of the startup to potential investors	Portable Document Format (PDF)
Financial	Report typically containing structured and systematic financial information of the startup	Excel (Xlsx)
Startup Success Prediction	Information regarding startup investment data sourced from Crunchbase	Object, int, float

The dataset for startup success prediction comprises 923 rows and 49 columns, including attributes like funding rounds, funding total, industry, location, and others. The primary goal is to predict whether a startup will succeed or fail. The distribution of startup success statuses is depicted in the Figure 1.

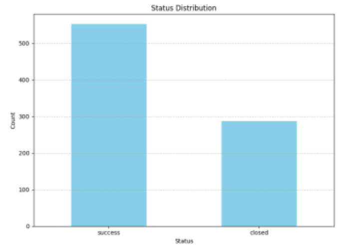

Fig. 1. Startup Status Distribution

C. *Modeling*

In the startup success prediction dataset, the data has been cleaned and prepared. Next, the data is modeled to predict startup status. Tree-based models were selected to address the binary classification problem in this research due to their capability to handle numerous variables with superior interpretability, robustness against overfitting, and efficient management of large and complex datasets [7]. The methodological process encompassed several stages: feature engineering, data splitting (80% training, 20% testing), and model selection and training using five models, such as Random Forest, XGBoost, LightGBM, AdaBoost, and Decision Tree. Subsequent steps included K-Fold Cross Validation and evaluation based on metrics such as accuracy, precision, recall, and F1 score. The final model was chosen based on its performance in these evaluation metrics.

To enhance the efficiency of due diligence and risk analysis processes, leveraging Large Language Models (LLMs) in conjunction with human expertise is crucial. This approach, which integrates reliable data sources and advanced analytical techniques, offers a comprehensive assessment of a company's financial health, reputation, and other critical factors. By doing so, it enables more accurate and informed investment decisions [8]. Figure 2 illustrates the workflow of the LLM used to extract answers based on the provided prompts.

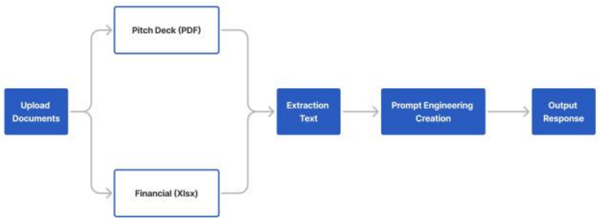

Fig. 2. The Process of Using the GPT-4 Model for Analyzing Pitch Deck and Financial Data

D. Deployment

During the deployment process, integration of the developed modules will be carried out, including startup success prediction and pitch deck and financial analysis, into a single dashboard. This dashboard helps optimize the due diligence process. Based on the functional requirements analysis, a use case diagram has been developed to illustrate the interactions between actors and the developed system. The resulting use case diagram is shown in the Figure 3.

Fig. 3. Use Case Diagram of the System

The developed technology encompasses both a front end and a back end. The front end, built with React and Tailwind CSS, provides an interactive and user-friendly dashboard for uploading documents, viewing analysis results, and downloading analysis results. The back end utilizes Express.js and Flask, with Flask handling machine learning model predictions and Express.js facilitating interactions with the GPT-4 model through the OpenAI API and gathering additional industry data via the Google Search API.

E. Customer Acceptance

The goal of this process is to ensure that the developed startup analysis system meets user expectations in terms of ease of use, accuracy, and relevance of the analysis results provided. This process aims to evaluate how well the system assists users, particularly startup investors, in performing due diligence more efficiently and accurately. Additionally, the testing focuses on measuring user satisfaction with the system's features and the impact of the system on user efficiency and speed. The objectives and metrics for the customer acceptance are shown in Table IV.

TABLE IV. COMPARISON OF EVALUATION METRICS ACROSS THE MODELS USED

Objective	Target	Metric
Evaluate Ease of Use	Users can operate the system easily without requiring special training.	Average score on a usability questionnaire (scale 1-5).
Assess Accuracy of Analysis Results	System analysis results align with uploaded data and user expectations.	User satisfaction level with accuracy (scale 1-5).
Measure Feature Relevance	Users find the features provided useful and aligned with their expectations.	Average score on a feature relevance questionnaire (scale 1-5).
Evaluate System Appearance and Interface	The system's appearance is intuitive and visually appealing.	Average score on a system appearance questionnaire (scale 1-5).
Measure User Satisfaction	Users find the system beneficial and supportive of their work.	Average score on a user satisfaction questionnaire (scale 1-5).

The number one represents "Strongly Disagree" and the number five represents "Strongly Agree". The target respondents for this process are investment analysts working at a venture capital firm featured in the case study and other respondents with similar experience and positions.

IV. SOLUTION IMPLEMENTATION

This section covers the implementation of each developed feature. Next, each developed feature will be integrated into a single system ready for use. There are three things explained in this section: implementation of pitch deck and financial analysis, implementation of startup success prediction, integration, and dashboard results.

A. Implementation of Pitch Deck and Financial Analysis

The technologies used in the implementation of this system are Large Language Model (LLM) and Google Search API. The process starts when a user uploads files, such as pitch decks and financial reports. First, the user selects the files, which are saved locally on their device, and then the "Upload" button sends these files to the server. While the files are being uploaded, a loading indicator shows the progress. Once the upload is complete, the files are processed to make them readable by GPT-4. For PDFs, the system extracts text, and for Excel files, it converts the data into a readable format. Any problems during this process will show an error message. The steps of this process are shown in the flowcharts provided.

Well-designed prompts can significantly enhance the quality and relevance of responses from LLM models, whereas poorly designed prompts may lead to unsatisfactory or incorrect responses [9]. The prompt engineering process aims to create specific commands to extract targeted information from pitch decks and startup financial documents, enabling the generation of accurate automatic answers. Once the prompts are prepared, the system is connected to the OpenAI API to provide responses based on these prompts. The entire process, from prompt creation to generating output, is illustrated in the Figure 4.

Fig. 4. The Process flow for Prompt Engineering to Generate Output

For data scraping, Google Search API is used to gather market/industry trends and competitor analysis data. The process flow for data scraping involves several key steps. It begins with defining the search link, followed by sending a query to the specified source. Once a response is received, the extracted results are then processed. These processed results are subsequently connected with the GPT-4 model to analyze and generate a comprehensive answer. The details of this process are outlined in Figure 5.

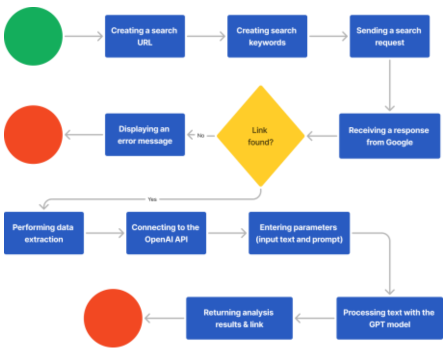

Fig. 5. Process Flow for Data Scraping with Google Search API and Analysis with GPT-4

B. Implementation of Startup Success Prediction

The selected models are tree-based models, including Random Forest, XGBoost, LightGBM, AdaBoost, and Decision Tree. These models are evaluated using cross-validation, a method that allocates data from the training sample to the validation sample to select the most effective model. Specifically, K-fold cross-validation is employed, which involves randomly dividing the dataset into K subsets to ensure robust model performance and reliable evaluation [10]. Afterwards, the models are evaluated based on the following evaluation metrics:

1. Accuracy: The proportion of correctly predicted instances out of the total instances [11].
2. Precision: The proportion of true positive predictions out of the total positive predictions [11].
3. Recall: The proportion of true positive predictions out of the total actual positives [11].
4. F1 Score: The harmonic mean of precision and recall, balancing the two metrics [11].

With 5 rounds of cross-validation, Table V shows the average evaluation results obtained. The chosen model is XGBoost because it has the best performance.

TABLE V. COMPARISON OF EVALUATION METRIC RESULTS

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	82.36%	82.92%	92,49%	87.33%
XGBoost	84.00%	85.30%	91.58%	88.23%
LGBM	82.06%	84.75%	88.84%	86.65%
AdaBoost	82.36%	82.85%	92.26%	87.28%
Decision Tree	75.93%	82.06%	81.33%	81.59%

The process is illustrated in Figure 6, showing how the XGBoost model is integrated into the system using Flask for real-time startup success predictions. Input parameters include metrics like initial funding age, funding amount, and startup milestones, which are processed to deliver accurate investor recommendations.

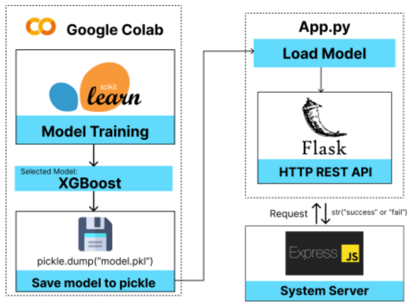

Fig. 6. Saving the Model in Pickle Format and Creating the Flask Server

C. Integration Results

Integration and implementation of a system consisting of front-end and back-end components. Additionally, the integrated system is deployed locally. The Figure 7 outlines the user journey for using the developed system. The system features multiple pages guiding users through the analysis process.

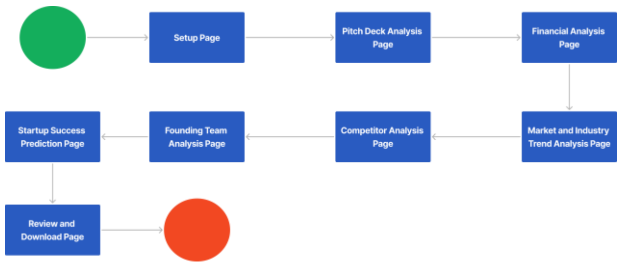

Fig. 7. User Journey Flow of System

As illustrated in Figure 8, the process utilizes Flask to manage the startup success prediction feature, which employs an XGBoost model stored in a pickle file. The application initializes a Flask object and enables Cross-Origin Resource Sharing (CORS) to handle requests from various domains. After loading the XGBoost model, user inputs are processed, and the "/predict" endpoint manages POST requests by sending the data to the model for prediction. The result, indicating 'success' or 'fail,' is returned in JSON format.

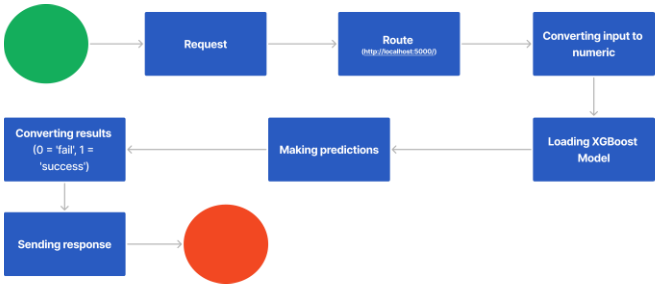

Fig. 8. Request Configuration for Startup Success Prediction

The routing implementation is developed using Node.js with the Express framework. The main server configuration, named "server.js," is set up with Express and includes CORS (Cross-Origin Resource Sharing) middleware. Incoming requests are directed to a router file, which defines various routes for handling requests related to pitch decks, financials, team, market size, and competitors. The request handling process is shown in Figure 9.

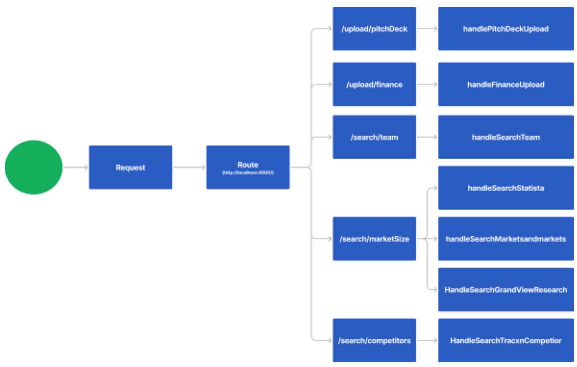

Fig. 9. Routing Handling Flow Diagram

Table VI displays the outcomes from the prompts and responses generated by the GPT-4 model. The system demonstrates its capability to extract text efficiently from uploaded pitch decks.

TABLE VI. SAMPLE PROMPT AND GPT-4 ANSWER

Prompt	Minimum 10 sentences of description of the Product/Service, Target Market, Customer Pain Points, Proposed Solutions, Key Features and Benefits, Unique Selling Proposition (USP), and Scalability if possible.
Answer	Startup X offers a comprehensive solution for shrimp farmers by addressing several key pain points. Farmers struggle with the maintenance and reliability of traditional paddlewheels, which are prone to breaking and can lead to significant losses. Startup X's system intelligently monitors the condition of paddlewheels, issuing emergency alarms when issues arise and notifying the community through messaging groups. The product is easy to use, solar-powered, and offers free WiFi coverage, ensuring device longevity of over 5 years without requiring on-grid electricity or internet quota fees. The target market includes shrimp farmers who can avoid potential losses of up to 6,700 dollars by investing in Jentera's system, which costs only up to 214 dollars. The system's unique selling proposition lies in its ability to predict and mitigate risks through three-layered risk mitigation, ensuring high profitability and minimizing operational downtime. Additionally, the solution is scalable to other aquaculture and manufacturing industries, potentially expanding its market reach significantly.

D. Dashboard Results

The results of the system implementation, as shown in Figure 10, are a dashboard that users can utilize to obtain analysis results by uploading pitch decks and financial documents and providing some required inputs.

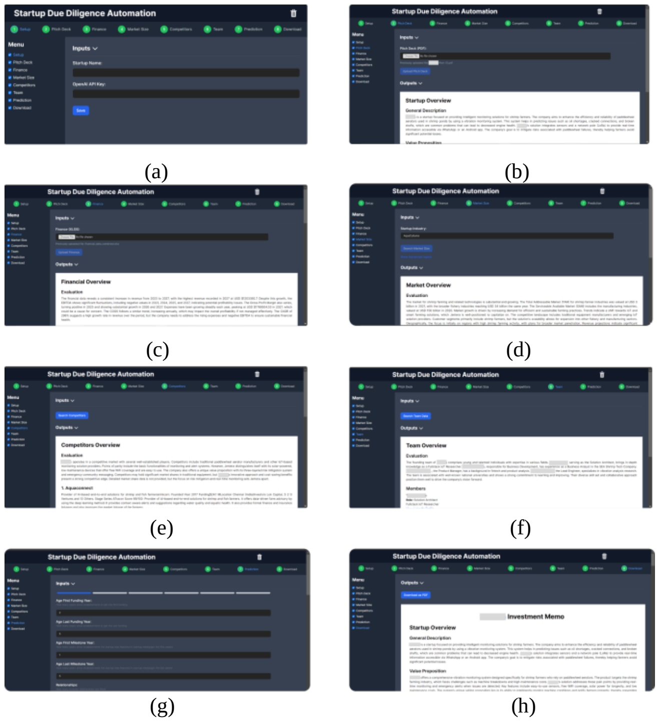

Fig. 10. Dashboard System Interface: (a) Setup Page, (b) Pitch Deck Page, (c) Financial Page, (d) Market Size Page, (e) Competitors Page, (f) Team Page, (g) Prediction Page, (h) Download Page

V. CUSTOMER ACCEPTANCE

The system is then tested and feedback is collected. The target users are professionals with experience as investment analysts or in similar roles at venture capital firms, particularly those from a venture capital firm featured in the case study, who are expected to have expertise in investment and startup analysis.

TABLE VII. PARTICIPANT RATINGS RESULTS

Objective	Average Score (1-5)
Assess system ease of use	4.56
Evaluate system appearance and interface	4.63
Assess accuracy of analysis results	4.38
Measure feature relevance to user expectations	4.40
Measure user satisfaction with the system	4.56
Acceptance Score (1-5)	**4.50**

Based on the five objective parameters listed in Table VII, no average rating for any parameter fell below four out of five. This suggests that the system performed well and met the participants' expectations. Overall, the system achieved an acceptance score of 4.50 out of 5.00, which is derived from the average of the five objective parameters.

VI. Impact and Risk Assessment

This research utilizes the OpenAI API platform, which offers enterprise-level data security and privacy. According to OpenAI's enterprise privacy policy, data from ChatGPT Team, ChatGPT Enterprise, or the API Platform is not used to train OpenAI models, and users have control over the duration of data storage, including rights to their input and output as per applicable laws [12].

However, data security remains a concern when using LLMs, especially regarding confidential information such as corporate data. The following is an analysis of the impacts and risks associated with using LLM and machine learning technology in this research.

1. The analysis results provided need further processing and cannot be used as the sole basis for final investment decisions.

   Currently, GPT-4 cannot replace an expert in business decision-making [13]. However, it can assist experts by enhancing productivity, improving time efficiency, and providing deeper analysis. In this study, the analysis results aid in streamlining the due diligence process but require further review before making investment decisions.

2. GPT-4 technology is not yet optimal for solving complex mathematical problems and reasoning tasks.

   The financial analysis feature faces several challenges that require system optimization. Financial report formats vary by company, and the large number of columns and rows can lead to errors in data extraction. Currently, the system only extracts data from the first page of financial reports, limiting its accuracy for multi-page documents. Additionally, GPT-4 technology remains suboptimal for solving complex mathematical problems and reasoning tasks. Research shows that complex math problems continue to challenge leading LLM models like GPT-4, even with external tools, due to frequent execution errors [14]. The model also struggles with reasoning, often exhibiting inconsistencies due to its inability to apply basic reasoning techniques and understand fundamental concepts, leading to potentially erroneous results [15].

3. Data and system security remain significant issues.

   Applications interacting with third-party LLM services are susceptible to various external attacks that can compromise security [10]. Sensitive corporate data is at risk from threats such as prompt injection attacks, where malicious inputs deceive the model into generating harmful outputs, SQL injection attacks, where crafted prompts induce the LLM to execute harmful SQL code, and data poisoning attacks, which manipulate training data to alter the model’s accuracy and results [16].

VII. Conclusion

This research demonstrates the application of Large Language Models (LLMs) and Machine Learning (ML) technologies to enhance the efficiency of due diligence processes for startup investments. By leveraging the Team Data Science Process (TDSP) methodology, we developed a system that integrates multiple data sources and analytical tools, including OpenAI's GPT-4 model and the Google Search API. The system's capabilities include analyzing pitch decks, financial reports, market trends, and competitor data, culminating in a startup success prediction using the XGBoost model.

The developed system has shown promising results, with a high customer acceptance score of 4.50 out of 5.00, indicating that it meets user expectations in terms of ease of use, accuracy, and relevance. Despite its success, there are notable challenges and limitations. Data security remains a significant concern, particularly regarding confidential corporate information. Additionally, while the system provides valuable insights, it cannot replace expert judgment and should be used as a supplementary tool in the decision-making process.

References

[1] E. Ries, Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. New York: Crown Business, 2011.

[2] T. Eisenmann, "Why Start-Ups Fail," Harvard Business Review, 2021. [Online]. Available: https://hbr.org/2021/05/why-start-ups-fail.

[3] J. Howarth, "Startup Failure Rate Statistics," Exploding Topics, Mar. 16, 2023. [Online]. Available: https://explodingtopics.com/blog/startup-failure-stats.

[4] McKinsey Quarterly, "A machine-learning approach to venture capital," 2017. [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/a-machine-learning-approach-to-venture-capital.

[5] C. M. Schmidt, "The impact of artificial intelligence on decision-making in Venture Capital Firms," PhD dissertation, 2019.

[6] N. Hotz, "What is TDSP? Data Science Project Management," Mar. 31, 2024. [Online]. Available: https://www.datascience-pm.com/tdsp/

[7] W. Loh, "Classification and Regression Trees," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 14-23, 2011. [Online]. Available: https://doi.org/10.1002/widm.8.

[8] D. Krause, "ChatGPT and Other AI Models as a Due Diligence Tool: Benefits and Limitations for Private Firm Investment Analysis," SSRN, 2023. [Online]. Available: https://ssrn.com/abstract=4416159.

[9] A. Setiawan, "Prompt Engineering untuk ChatGPT (Bab 12)," 2023.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of Machine Learning, 2nd ed. Cambridge, MA: MIT Press, 2018.

[11] H. Dalianis, "Evaluation Metrics and Evaluation," in Clinical Text Mining, Cham: Springer, 2018, pp. 101-124. [Online]. Available: https://doi.org/10.1007/978-3-319-78503-5_6.

[12] OpenAI, "Enterprise Privacy at OpenAI," [Online]. Available: https://openai.com/enterprise-privacy. [Accessed: 30-Jul-2024].

[13] E. L. Chuma and G. G. de Oliveira, "Generative AI for Business Decision-Making: A Case of ChatGPT," Management Science and Business Decisions, vol. 3, no. 1, pp. 5-11, 2023. [Online]. Available: https://doi.org/10.52812/msbd.63

[14] Y. Wu, F. Jia, S. Zhang, Q. Wu, H. Li, E. Zhu, Y. Wang, Y. T. Lee, R. Peng, and C. Wang, "An Empirical Study on Challenging Math Problem Solving with GPT-4," 2023.

[15] K. Arkoudas, "GPT-4 Can't Reason," Preprints, 2023080148, 2023. [Online]. Available: https://doi.org/10.20944/preprints202308.0148.v2.

[16] S. Tete, "Threat Modelling and Risk Analysis for Large Language Model (LLM)-Powered Applications," 2024.