A Project Report On

Kickstarter.AI – A tool for Investors to assess startups

Submitted in partial fulfillment of the requirement for the 8^{th} semester

Bachelor of Engineering

in

Computer Science and Engineering

DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE & ISO 9001:2008 Certified)

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' grade

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560111

Submitted By

Mantej Singh Tuli 1DS19CS086

Sree Chand R 1DS19CS164

Shreyas G 1DS19CS202

Dheemanth A N 1DS19CS710

Under the guidance of

Prof. Anupama Girish Mam

Assist. Professor , CSE , DSCE

Prathul M

Co-guide - Industry

2022 - 2023

Department of Computer Science and Engineering

DAYANANDA SAGAR COLLEGE OF ENGINEERING

Bangalore - 560111

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Dayananda Sagar College of Engineering

(An Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE & ISO 9001:2008 Certified)

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' grade

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560111

Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled **Kickstarter AI – A tool for Investors to assess startups** is a bonafide work carried out by **Mantej Singh Tuli [1DS19CS086]**, **Shreyas G [1DS19CS202]**, **Sree Chand R [1DS19CS164]** and **Dheemanth A N [1DS19CS710]** in partial fulfillment of 8th semester, Bachelor of Engineering in Computer Science and Engineering under Visvesvaraya Technological University, Belgaum during the year 2022-23.

Prof. Anupama Girish	**Dr. Ramesh Babu D R**	**B G Prasad**
(Internal Guide)	Vice Principal & HOD	Principal
Asst Prof. CSE, DSCE	CSE, DSCE	DSCE
Signature:.....	Signature:.....	Signature:.....

Name of the Examiners:	Signature with date:
1.....	.....
2.....	.....

Acknowledgement

We are pleased to have successfully completed the project **Kickstarter AI – A tool for Investors to assess startups**. We thoroughly enjoyed the process of working on this project and gained a lot of knowledge doing so.

We would like to take this opportunity to express our gratitude to **B G Prasad**, Principal of DSCE, for permitting us to utilize all the necessary facilities of the institution.

We also thank our respected Vice Principal, HOD of Computer Science & Engineering, DSCE, Bangalore, **Dr. Ramesh Babu D R**, for his support and encouragement throughout the process.

We are immensely grateful to our respected and learned guide, **Prof. Anupama Girish**, Professor CSE, DSCE and our co-guide **Prathul M** for their valuable help and guidance. We are indebted to them for their invaluable guidance throughout the process and their useful inputs at all stages of the process.

We also thank all the faculty and support staff of Department of Computer Science, DSCE. Without their support over the years, this work would not have been possible.

Lastly, we would like to express our deep appreciation towards our classmates and our family for providing us with constant moral support and encouragement. They have stood by us in the most difficult of times.

Mantej Singh Tuli 1DS19CS086
Dheemanth A N 1DS19CS170
Shreyas G 1DS19CS202
Sree Chand R 1DS19CS164

Abstract

Artificial intelligence (AI) has emerged as a powerful technology with the potential to revolutionize many different industries and fields. With its applications ranging from healthcare, finance, education, production, manufacturing etc., we have decided to explore the potential of AI in the field of finance, particularly focusing on Investing in Early-Stage Startups and Venture Capital. We have decided to choose this niche domain of building an AI assistant to Venture Capitalists in order to simplify their decisions when it comes to financing a startup as there is a lot of human intuition involved in the present scenario and from our survey it became very clear that there are lots of inefficient processes involved in choosing whether investing in a startup is a fruitful decision or not. When there a lot of inefficiencies, we decided that a scientific, methodical and a through technical approach could aid the investors in making a decision and explaining the idea behind that particular decision rather than simply placing our decision on intuition and funding a startup that has no foreseeable good exit/outcome for the investor. With the use of AI, we have developed a classification model that decides based on our input of the company's fundamentals and financials, whether or not it's a good investment and appropriately outputs the decision. We have integrated the model with a web interface that allows investors to log in, create a new company, run analysis on that particular company, generate a report for the analyzed company and accordingly carry their decision. The interaction with the model takes place through API calls, the model hosted on the cloud is referenced in the backend through an API call, that posts all the required details of the company provided by the VC, fills in the keys of the API through a POST request, calculates the result and returns a response to the backend.

The aim of our project is to eliminate the high number of inefficiencies that lies in decision making when it comes to financing a startup, as the data tells that 7 out of 10 startups never see daylight. With such a high number of startups crashing, it becomes important to focus on backing the ones with through fundamentals, financials and a clearly defined exit route for VCs.

Table Of Contents

List of Figures

Chapter 1

Introduction

1.1 The Problem

Venture capital is a privately held equity usually run by a network of few high-net-worth individuals who pool in their money in order to finance upcoming project, companies and startups that shows great potential in order to multiply their investments. Most of the investors invest in such companies where they can see their investments becoming 100X of their initial invested amount.

Presently, most of the startups first start with an idea, build a small team and immediately gets in to the VC hunting zone, where in they start looking for VCs to finance and back their companies in exchange of equity or in some cases even raising debts from VCs or sometimes a convertible note.

The startups then proceed to build their products, expand their teams, decide a figure for the cash burn, marketing and advertising their products and finally edging towards profitability.

Presently, there are 3 exit routes for the investors after investing in a company. The first exit route is IPO, where the investors wait till a company does an IPO and sell their stakes in exchange for cash. The second route is acquisition by a bigger player, and thereby the VCs dilute their shares in exchange for cash and the third route is VCs exiting the company by selling their equity in further rounds of funding when the company subsequently receives a higher valuation.

With that said and done, there a lot of challenges that VCs face when analyzing whether or not to invest in a company, the following are some of the challenges faced by them.

1. Misjudging the market: One of the biggest mistakes that VCs can make

is misjudging the market for a startup's products or services as seen. This is a very tricky problem as investing in a sector that does not have a lot of competitors / players may mean that there might be no potential in that field or can even mean that the there is still time for product market fit . An example of this problem is a company called 'Second Life', that actually built the first Metaverse and an online social hangout platform in 3D, way back in 2003 .The company miserably failed to the likes of Facebook, Orkut but say a company like that had to come in today, then it would be a huge hit especially when it comes to receiving investments and funding from VCs.

2. Underestimating the risks: The rule of the VC game is very simple, 'High risk and High reward'. As we outlined earlier, VCs typically invest in such companies where they see their investments becoming 100X of their initial corpus. But stats also tell us that 7 out of 10 companies fail, even if a company succeeds, inching towards profitability and doing an IPO or making a significant dent in competitors' business and getting acquired by them is still a mammoth task and hence there lies a great deal of risks that are associated with any startup and decisions have to be made scientifically and efficiently.

3. Overvaluing the startup: This is one of the most common mistakes that is done intentionally both by the founders as well as VCs. The founders always want to negotiate the best possible deal for their company and usually quote a high multiple for the overall valuation of their company. Looking at some of the companies' data, the valuation numbers sometimes sound ridiculously absurd as there is no mathematical or logical reasoning behind this. Showing an inflated valuation is fairly common in VC and funding scenarios and the VCs with selection bias end up overvaluing a sector or a company and decide to invest in that company.

4. Failing to negotiate a fair deal: This is always a tricky problem to solve as both VCs and founders act with very little data on their hands, and are looking out for themselves rather than the opposite person. While VCs sometimes should finance companies that are doubtful to make it big but still worthy of giving a shot through debts with or without interest. The founders always find it favorable to seek investments involving distribution of equity as they are not liable to return the original corpus, even if the company crashes.

1.2 Organisation of Project Report

The project report is organized as follows:

In Chapter (2) we discuss the problem statement and the proposed solution. We also take a look at the systems that exist today and the drawbacks they face.

Chapter (3) takes a more in-depth look at various hardware and software based solutions that exist, with a survey on existing literature available.

Chapter (4) looks at the architecture of the proposed solution with an overview of the system design, utilizing system block diagrams and data flow diagrams.

Chapter (5) dives into the Implementation of the solution, by describing the hardware and software requirements, along with dataset descriptions and implementation details.

Chapter (6) describes our testing process, while Chapter (7) looks at our experimentation process and the obtained results.

Chapter (8) summarizes our findings and concludes the paper.

Chapter 2

Problem Statement and Proposed Solution

2.1 Problem Statement

To develop a solution that analyzes the fundamentals and financials of a startup and suggests whether or not it is a good investment for the investor.

2.2 Existing Systems

Multiple researches and literature surveys carried out on this topic, outlines problems in the existing system that is mostly based on intuition and a need for a scientific, methodical and a technical approach with little or no selection bias involved. Studying the existing systems, the previous approaches carried out can be split into 2 categories depending on the amount of data pertaining to financials and fundamentals of the company subjected to analysis, that is publicly available or available to an investor.

2.2.1 Early-stage startup with little or no data available

The problem with investing in early-stage startups is there is little or no data available, be it financial or fundamental data, this can lead to selection bias as well as a high chance of investors losing their corpus. Previous research on this has outlined approaches such as setting up an ESI [Early-Stage Startup Investment] Framework that gives investors a checklist and is mostly based on parameters such as the founder's track record, founder's alma mater, country in which the company is incorporated in, age of the founders, background of the founding team etc. Assigning weights

to parameters like these can help the investor arrive at a particular score which can determine if it is worth investing or not based on the median scores of other companies that started out in the same domain.

2.2.2 Mature companies with financial and fundamental data available

This is a different ball game altogether as there is a lot of data available for the investors to carefully take their decisions. The usage of Machine Learning models is highly recommended as the sample size is relatively large and there is a lot of data through which pattern recognition and analysis become an integral part of determining whether or not it's a sound decision to invest in a company at a fairly mature stage. While investors can certainly, hedge the risk of investing in a mature-stage startup as opposed to investing in an early-stage startup, the metrics such as valuation which can be overblown out of proportion, or give a blind eye to sectors that are relatively less known and a lack of subject matter expertise in a particular domain can potentially stop an investor from making a sound decision. Hence with the use of Machine learning models in these domains, accurate and timely results can be generated which can be beneficial to both investors as well as the companies.

2.3 Proposed Solution

Our research focuses on creating an advanced machine learning (ML) model that uses extensive data analysis from Crunchbase, a well-known start up information portal. The primary goal is to forecast start-up success or failure by analyzing numerous input criteria such as Category List, Total Funding Raised, Country of Operation, Number of Funding Rounds, and Funding Dates. After considerable research and comparison of several techniques, we found that the Random Forest ensemble methodology is the best strategy for our prediction model. Serialization, vectorization, and other methods are used.

The trained model will be pickled and effortlessly incorporated into a user-friendly website, allowing users to enter important start-up facts and obtain estimates on their chances of success, including the possibility of going public or being purchased by a larger business. Our ML model will combine crucial methods such as serialization, vectorization, and others to effectively handle the data and give reliable predictions.

2.3.1 Detailed Proposed Solution

Data Collection and Pre-processing

We will use Crunchbase's Figure 2.1 vast dataset to acquire relevant start-up information. This dataset will be used to extract useful information such as the Category List, Total Funding Awarded, Operating Country, Number of Funding Rounds, and Funding Dates. We will undertake substantial pre-processing of the data to prepare it for machine learning model training, including data cleansing, missing value management, and categorical variable encoding. These processes ensure that the data is properly structured for subsequent examination.

Figure 2.1: Crunchbase Database

Feature Engineering

We may investigate using additional feature engineering strategies to improve the accuracy of our prediction model for start-up success. This might include developing new features based on existing data or integrating additional data sources that provide useful insights into elements that contribute to company success.

ML Model Selection and Training

During the ML model selection and training phase, we will do experiments with several ML approaches such as Random Forest, Logistic Regression, Gradient Boosting, and Support Vector Machines. Our goal is to find the best effective technique for

our start-up success prediction model. Following a thorough examination and comparison, we found that the Random Forest Figure 2.2 ensemble approach is the best option. This conclusion is based on its capacity to handle complicated data linkages, manage high-dimensional datasets, and generate strong predictions.

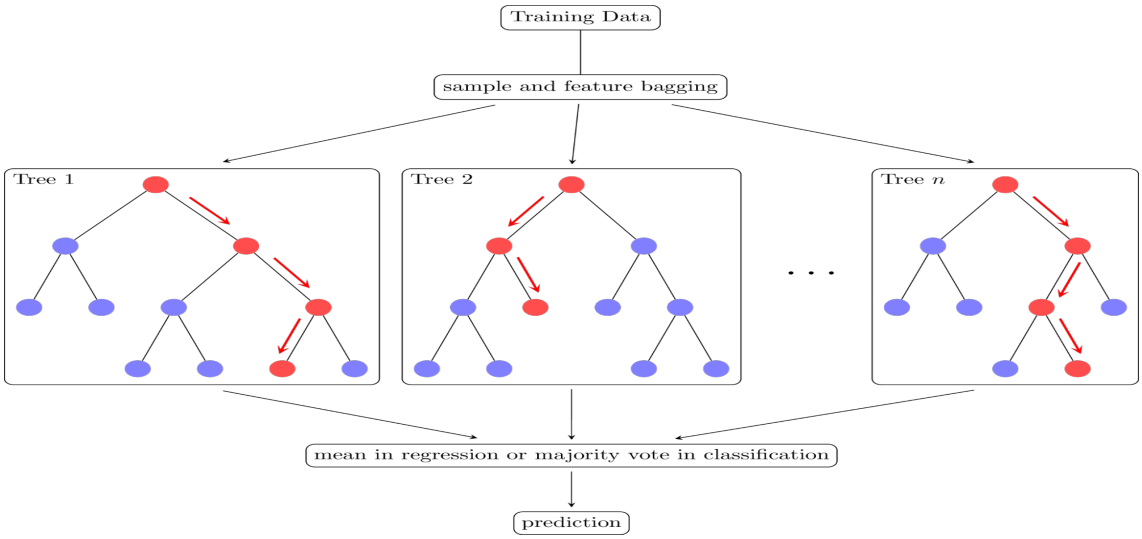


Figure 2.2: Standard working of a random-forest model

Pickling and Model Integration

After training and assessing the Random Forest model, we will serialize it for storage via pickling [Figure 2.3]. This serialized format makes it simple to integrate with the website, providing fast loading for real-time forecasts. The pickled model will be a critical component of our ML integration, allowing users to submit start-up data via the website's user interface. This connection will make it easier to use the model's predictive capabilities.

Website Integration and API Calls

The compiled model is hosted on the cloud, and a designated endpoint leads us to getting results from the model. A form that requires all the relevant details have to be first filled out by the investor (assuming due diligence has been thoroughly carried out and the data that is being submitted is true), which on submitting calls the API that hosts the model and the relevant data is sent to the API through a POST request, and the user is redirected to another page that houses the report of the particular analyzed company. With the submitted form details and getting back the result from the API, an appropriate report card is generated for the company that is analyzed and whether or not the investor should consider investing in that company. This is backed

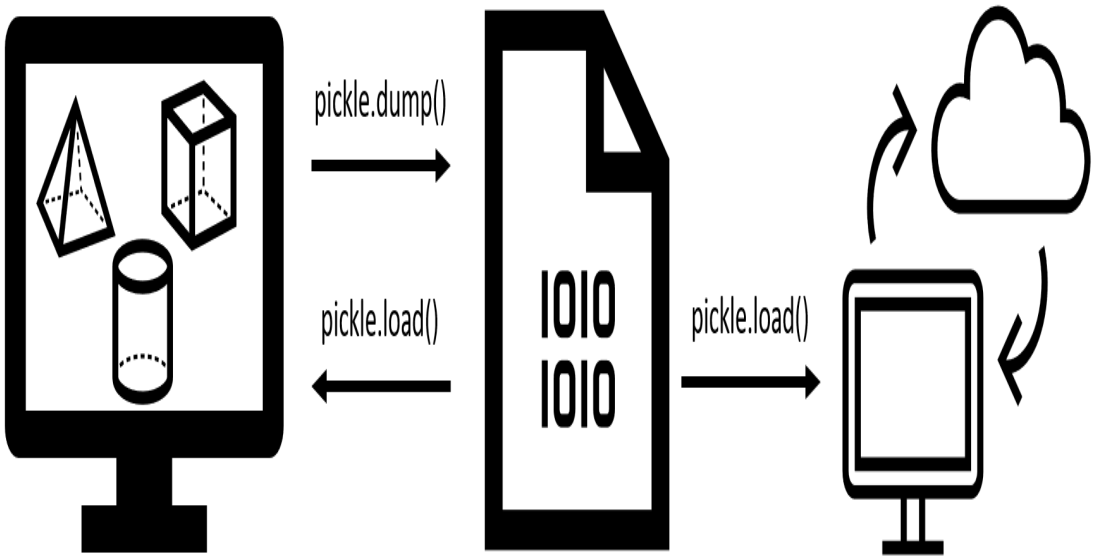


Figure 2.3: Working of pickle file in a machine-learning model

up by the national average of the funding received by all the previously successful companies in that sector to give an enhanced view of the overall report.

Prediction Generation and Result Display

After receiving user input, the website will make API calls to provide the required parameters to the pickled ML model. The ML model will analyse the information and create predictions about the start-up's chances of success using techniques like vectorization and serialization. These forecasts may contain information on the prospects of going public or being bought by a larger corporation. The forecast findings will be shown on the website, giving users useful insights into the future prospects of their companies.

In a nutshell, our project offers a comprehensive method for forecasting start-up success. We allow entrepreneurs and investors to make educated decisions based on data-driven forecasts by combining Crunchbase data and sophisticated ML methods, as well as integrating our proprietary ML model into a user-friendly website. This strategy increases the likelihood of business success and fosters development within the entrepreneurial ecosystem.

Upon receiving user input, the website will pass the relevant parameters to the pickled ML model via API calls. The ML model will process the input using techniques

such as vectorization and serialization, generating a prediction regarding the startup's potential for success, including the likelihood of IPO transition or acquisition by a larger company.

The prediction results will be displayed on the website, providing users with valuable insights into the future prospects of their startups. In conclusion, our project proposes a comprehensive solution for predicting startup success by harnessing Crunchbase data and employing advanced ML techniques. By developing a user-friendly website integrated with our pickled ML model, we empower entrepreneurs and investors to make informed decisions based on data-driven predictions, ultimately increasing the chances of startup success and facilitating growth within the entrepreneurial ecosystem.

Chapter 3

Literature survey

Nick Skillicorn et al.,[1] A study, published in the Journal of Business Research , found that the use of AI in venture capital can help VCs to make more accurate predictions about the future performance of companies and industries and to identify potential investment opportunities that might have been missed by human analysts. The study from **Tomer Dean et al.,**[3] also found that AI can help VCs to make more efficient and effective use of their time and resources, freeing them up to focus on more strategic and creative thinking. Human analysts can miss potential investment opportunities for a variety of reasons. Some common reasons include:

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

2. Confirmation bias: Human analysts may be prone to confirmation bias, which is the propensity to look for and analyses data in a manner that supports one's preexisting views or hypotheses. This can lead them to overlook or discount information that contradicts their preconceptions and may cause them to miss potential investment opportunities.

3. Sunk cost fallacy: Human analysts may be affected by the sunk cost fallacy (**Jared Council et al.,**[5]), which is the tendency to continue investing in a company or project even when it is not performing well, in order to avoid feeling like the time and resources already invested have been wasted. This can lead them to miss

potential investment opportunities that may be more profitable and sustainable

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

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

Hyoung Jun Kim et al.,[9] By assisting potential entrepreneurs, venture capital (VC) is a crucial engine fueling innovation and economic progress. Making educated VC investment decisions necessitates precise forecasts of future business performance as well as the identification of previously ignored prospects. While human analysts have traditionally played an essential part in investment evaluation, new research has looked into the revolutionary potential of artificial intelligence (AI) in VC decision-making. This comprehensive literature review seeks to dive into the existing body of research in order to identify the various benefits of incorporating AI into VC and resolving the limitations of human analysts.

AI's Advantages in VC Decision Making:

Unprecedented prediction Accuracy: New research, like that done by Jared Council (**Jared Council et al.,[5]**) and published in the prestigious Journal of Business Research(**Tin Kam Ho et al.,[6]**), shows that AI may significantly improve VC decision-making accuracy. AI algorithms enable VCs to make more exact projections about a company's future success by evaluating massive amounts of data and detecting hidden trends. AI algorithms find investment possibilities that human analysts may have missed by minimizing human biases and utilizing large databases.

Optimized Efficiency and Resource Allocation: The Council's report also shows AI's tremendous potential in expediting VC decision-making processes, optimizing efficiency, and allocating resources. When analyzing investment prospects, human ana-

lysts frequently face restrictions such as limited time and resources. Artificial intelligence (AI) technology may automate labor-intensive processes like data collecting, analysis, and market research, freeing up human analysts to focus on strategic and creative thinking. AI integration enables venture capitalists to make better use of their resources, spending valuable time to discovering and developing interesting investment ideas.

Overcoming the Limitations of Human Analysts:

Taming Overconfidence: One of the human analysts' intrinsic limitations is their proclivity for overconfidence, which causes them to ignore investing possibilities that depart from their preconceived assumptions. Analysts may accidentally dismiss unusual or disruptive companies with significant growth potential by relying on their experience and instincts. AI systems, on the other hand, are not impacted by such prejudices and may investigate a larger variety of investment opportunities.

Mitigating Confirmation Bias: Another cognitive hazard experienced by human analysts is confirmation bias, which occurs when people seek and interpret information that confirms their previous views or hypotheses while discounting contrary data. This prejudice might make it difficult to examine alternate perspectives and new information that may indicate good investment prospects beyond their previous opinions. AI-driven algorithms, powered by data and statistical patterns, give a new and impartial lens through which to analyze investment possibilities, lessening the impact of confirmation bias **(Arroyo et al.,[10])**.

Avoiding the Sunk Cost Fallacy: Human analysts may fall victim to the sunk cost fallacy, continuing to invest in underperforming firms or initiatives to avoid acknowledging that previous expenditures were a waste of time. This prejudice stops people from identifying potentially successful investment options with more long-term sustainability. AI systems, free of emotional attachments or previous investments, may objectively analyze an opportunity's viability and potential, allowing VCs to avoid the sunk cost fallacy and explore more fruitful chances.

Human analysts face limitations in their personal experiences and areas of expertise,

which can restrict their awareness of investment opportunities beyond their immediate fields of knowledge. AI, with its ability to analyze a wide range of data sources, offers a broader perspective, identifying investment possibilities that may have been overlooked by human analysts operating within narrower frameworks. This limitation impedes the exploration of ventures in diverse industries or emerging markets, resulting in missed opportunities for significant returns.

A significant challenge faced by human analysts is the constraint of limited time and resources, preventing them from thoroughly researching and evaluating investment opportunities. By leveraging AI's capabilities for data processing, pattern recognition, and automation, VCs can uncover hidden investment prospects that demand more comprehensive scrutiny, ensuring no stone goes unturned in the pursuit of fruitful investments **(Torben Antretter et al.,[12])**.

The integration of AI into venture capital decision making represents a transformative paradigm shift, offering a plethora of benefits to overcome the limitations of human analysts. By enhancing predictive accuracy, optimizing efficiency, and addressing cognitive biases, AI empowers VCs to identify overlooked investment opportunities, improve decision-making processes, and maximize returns. Striking a delicate balance between human judgment and AI-driven insights is crucial to ensure AI serves as a powerful decision support system, augmenting the expertise of human analysts rather than replacing it entirely. Further research in this domain will undoubtedly contribute to refining AI models, addressing ethical concerns, and advancing the overall efficacy of venture capital decision making.

Chapter 4

Architecture and System Design

The overview of the system is represented in Fig.4.1. It shows the modules involved in building the system i.e.,

- Web APP
- Frontend UI/UX
- Backend ML model

4.1 Software Overview

4.1.1 System Block Diagram

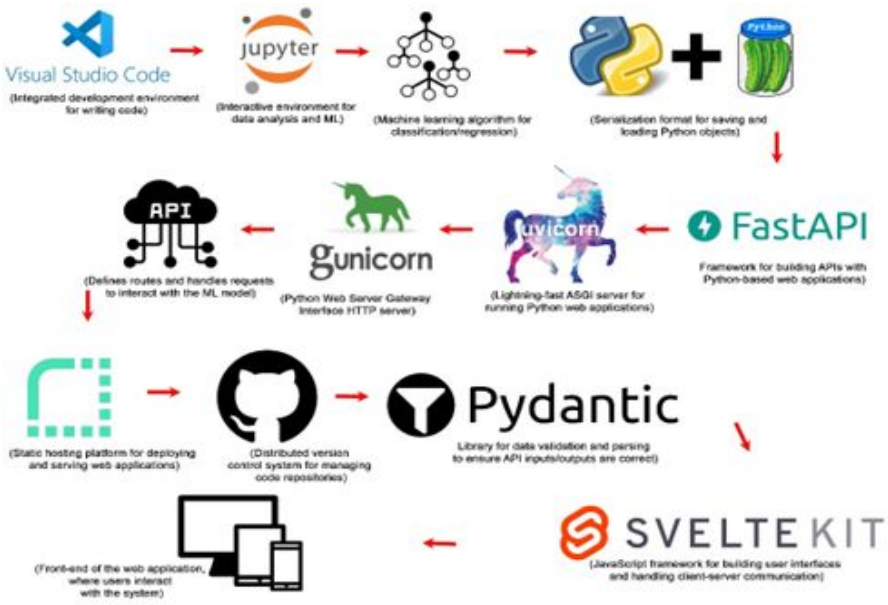


Figure 4.1: System Block Design

A client-facing application for investors to analyses, audit, and understand startups and their financials. Using FastAPI and machine learning and hosted with the help of Uvicorn and Gunicorn, the front end was developed with Svelte to make an

integrated web application. This creates an intuitive UI that is easy to handle on different platforms and produces accurate results using the ML models, which are deployed as an API over the network Figure 4.1

4.1.2 Data Flow Diagram

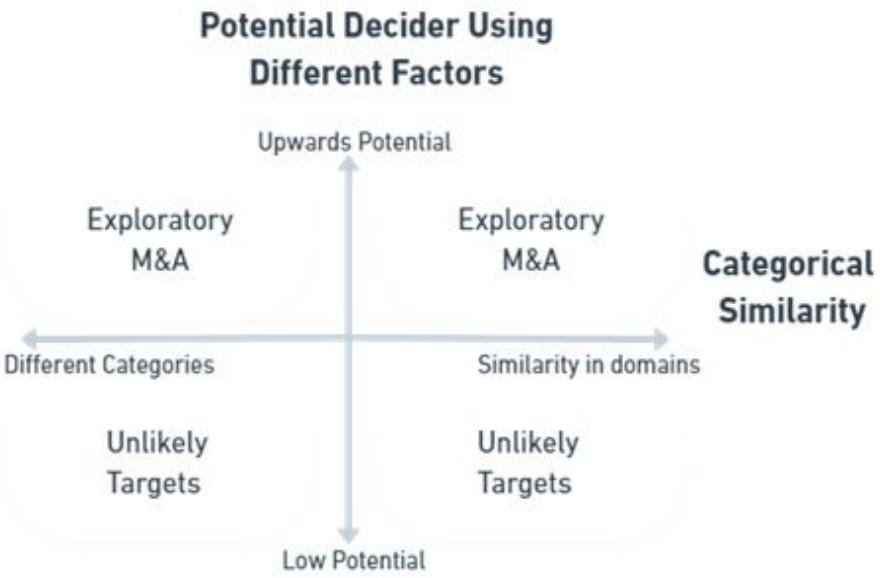


Figure 4.2: Dependencies of the Features in the Models

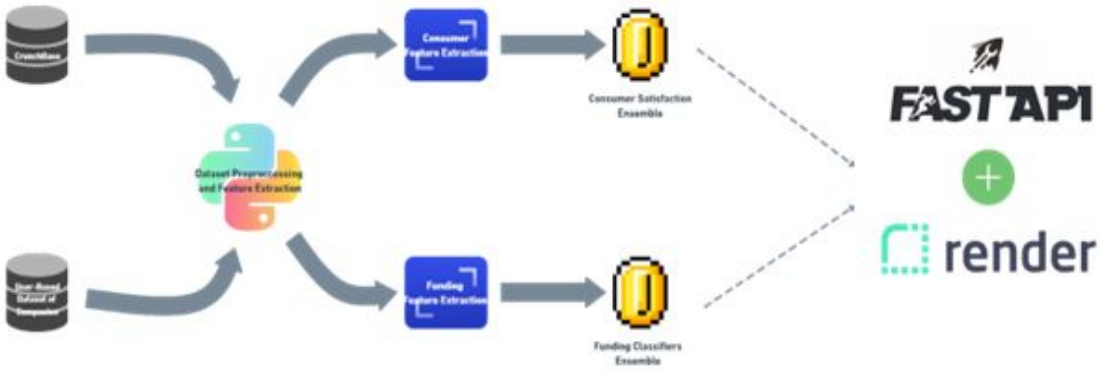


Figure 4.3: Data Flow Diagram

4.1.3 Sequential Diagram

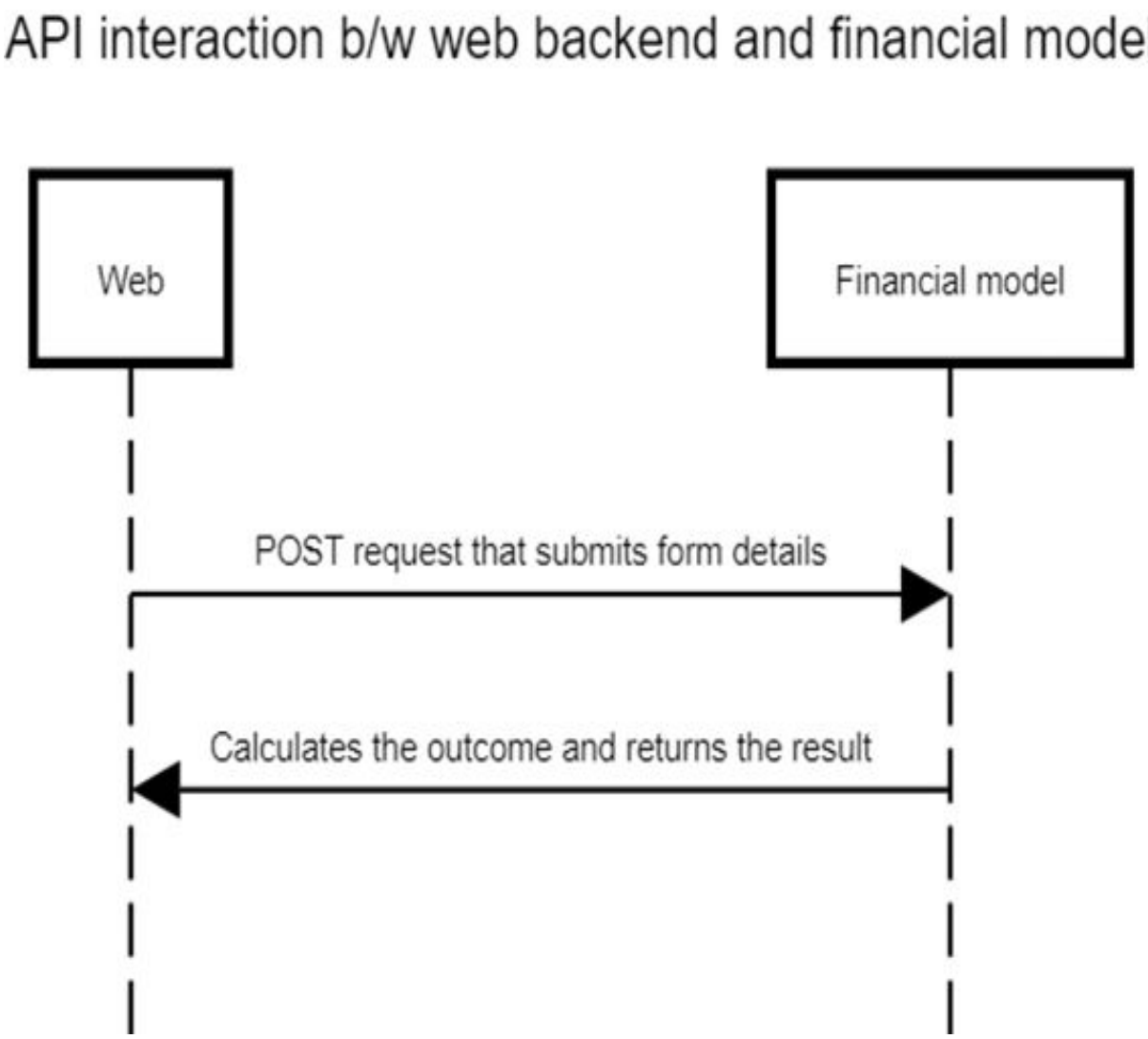


Figure 4.4: Sequential diagram that shows the interaction b/w financials model and the web backend

The Sequential diagram shown in Figure 4.4 depicts the interaction between the financial model and the web, first the details are submitted via a POST request to the model which in turn returns the outcome after the computation

The sequential diagram shown in Figure 4.5 gives an overview on the interaction between the website and the fundamentals models which deals with the fundamentals of an organization. A POST request is sent to the model which contains various details of the firm and the model after computation return the results back to the web.

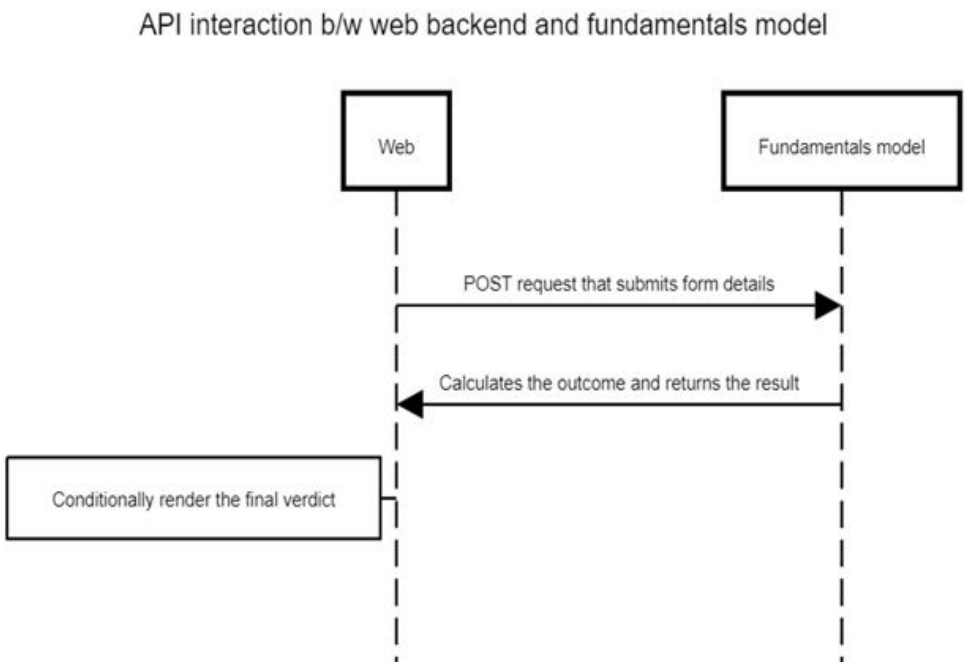


Figure 4.5: Sequential diagram that shows the interaction b/w fundamentals model and the web backend

4.1.4 Use Case Diagram

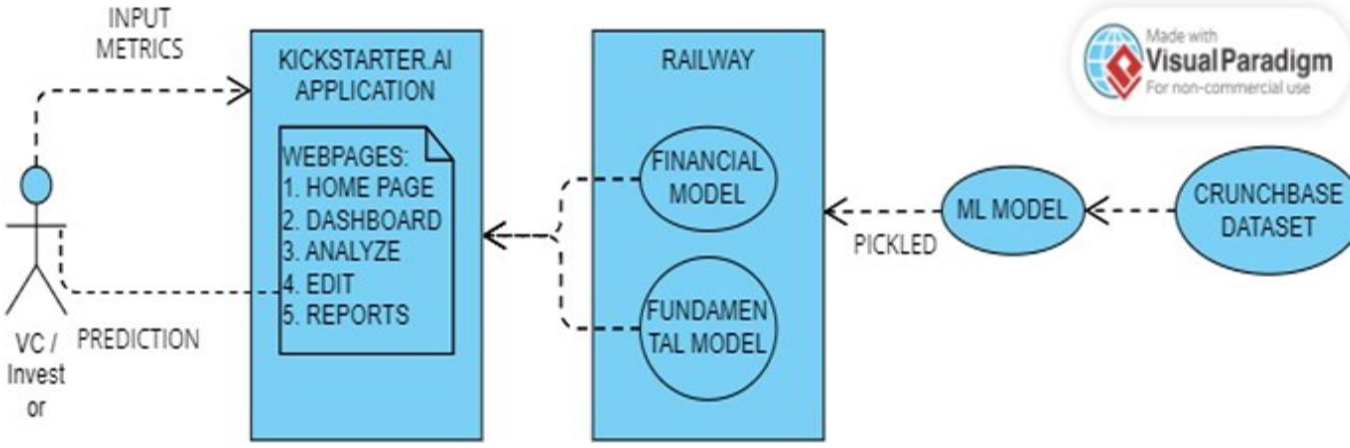


Figure 4.6: Use case diagram

The above Figure 4.6, showcases the use case of the project where in the two-model work hand in hand to predict if the investment will be favorable or not and sends the obtained result to the web which can be viewed by the user.

Chapter 5

Implementation

5.1 Implementation Platform

5.1.1 Hardware

GPU: NVIDIA GeForce GTX 1080 Ti or higher (or equivalent)

Processor: Intel Core i7 or AMD Ryzen 7 (or higher) RAM: 8 GB or higher

Storage: Solid State Drive (SSD) with at least 500 GB capacity

Internet Connection: High-speed and stable internet connection

5.1.2 Software

Operating System: Windows 10 or Linux (Ubuntu 20.04 or higher)

Python: Version 3.8 or higher

Development Environment: Anaconda or Miniconda

Integrated Development Environment (IDE): Visual Studio Code

ML Libraries: scikit-learn, pandas, NumPy, matplotlib, seaborn, TensorFlow, pydantic, uvloop, uvicorn

Full Stack Framework [Frontend + Backend]: Sveltekit [JavaScript framework]

Database:MongoDB, Mongoose ODM

Styling and CSS : Tailwind CSS

Authentication: AuthO

API tools : FastAPI,Postman

Deployment : Railway

5.2 Implementation Details

5.2.1 Organisation of files

Project Root:

README.md: Documentation providing an overview of the project, installation instructions, and usage guidelines.

requirements.txt: File listing all the required Python packages and their versions.

Data:

dataset.csv: The Crunchbase dataset or any other relevant data file used for training and testing the ML model.

Notebooks:

data preprocessing.ipynb: Jupyter Notebook for data cleaning, preprocessing, and feature engineering tasks.

model training.ipynb: Notebook containing code for training and evaluating the ML model.

model evaluation.ipynb: Notebook for evaluating the performance of the trained model.

Models:

trained model.pkl: Pickle file containing the trained Random Forest ensemble model (or any other chosen model).

Website:

src : This folder consists of all the routes and components used in the website

models: This folder consists of individual files that exports a model pertaining to a particular schema of a MongoDB collection.

routes: Sveltekit implements file based routing and accordingly we have established the following routes :

‘/’:root

‘/dashboard’:To view the dashboard page that allows a user to analyze,edit and view reports.

‘/analyze’:To analyze a particular company’s financials and fundamental.

‘/report’:To view the report of a particular company that is analyzed.

‘/edit’:To edit the details of an existing company ‘/api’: This folder houses all the APIs that bridges the model and the web interface along with creation of user in the database, authentication,saving company details in the database and editing the

details of an already analyzed company.

static: This folder consists of all the images and other static assets.

5.2.2 Implementation Workflow

Data Preprocessing

Load the Crunchbase dataset or any relevant data. Perform data cleaning, handle missing values, and remove outliers if necessary. Perform feature engineering and transformation on the dataset. Split the data into training and testing sets.

To begin, we load the Crunchbase dataset or other relevant data into our system. We diligently clean the dataset before analysis, fixing missing values and eliminating outliers as appropriate. This guarantees that our future modeling method is founded on a high-quality and trustworthy dataset. In addition, to extract relevant insights from the data, we use feature engineering and transformation approaches. Finally, we divide the dataset into training and testing sets to ensure that our models are assessed on separate data to appropriately assess their performance.

Model Selection and Training:

Choose appropriate ML techniques for the project (e.g., Random Forest Ensemble). Initialize the chosen ML model with desired parameters. Train the model using the preprocessed training dataset. Evaluate the model's performance using appropriate metrics.

We carefully choose the most appropriate machine learning approaches for this project, taking into account aspects such as the nature of the data and the project goals. In this scenario, we use an ensemble approach recognized for its ability to handle complicated interactions and deliver robust forecasts, such as Random Forest. We start the Random Forest model with carefully adjusted parameters to get the best balance of bias and volatility. Following that, we train the model with the preprocessed training dataset, allowing it to discover patterns and correlations from the data supplied. Finally, we assess the model's performance using relevant measures including accuracy, precision, and recall.

Model Evaluation and Tuning:

Analyze the model's performance metrics and identify areas for improvement. Perform model tuning by adjusting hyperparameters to enhance performance. Validate the tuned model using cross-validation techniques. Iterate the tuning process if necessary to achieve desired performance.

We analyze the acquired metrics after analyzing the model's performance to suggest

areas for improvement. We obtain insights into potential improvements that might improve the model's forecasting skills by attentively assessing its strengths and flaws. We go through a thorough hyperparameter tuning procedure, rigorously adjusting critical parameters and analyzing their influence on performance. We aim to attain the right balance of model complexity and generalizability through repeated modifications and experimentation. We evaluate the customized model using cross-validation techniques to guarantee resilience, offering a realistic estimate of its performance on unseen data.

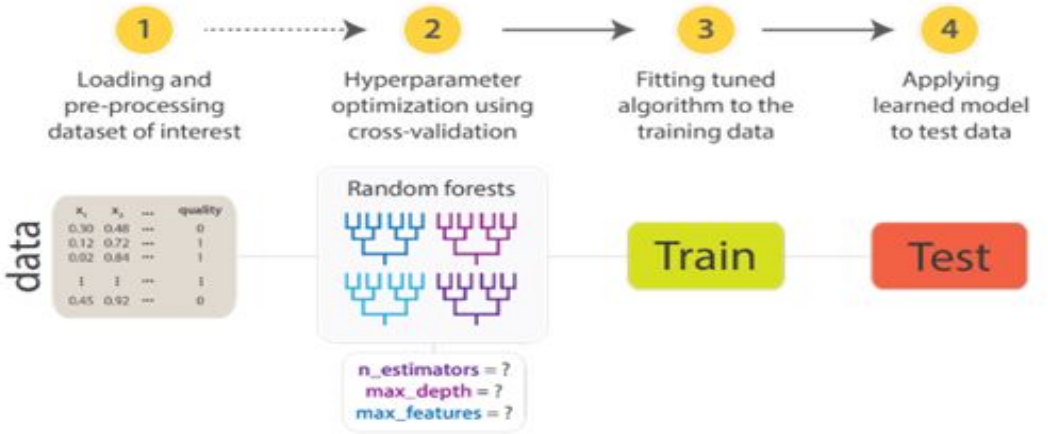


Figure 5.1: Model Evaluation

Model Persistence:

Serialize and save the trained model using pickle or similar techniques. Store the serialized model file in the designated "Models" directory.

After obtaining a well-performing model, we serialize and preserve it for future use. Serialization entails transforming the learned model into a format that can be efficiently stored and loaded. We use libraries like Pickle to serialize the model, resulting in a serialized model file. This file is then safely saved in the selected "Models" directory, allowing for simple deployment and prediction.

API Development:

The API was developed using the python programming languages major library for deployment which is FastAPI in conjunction with Uvicorn and Gunicorn. We created and used FastAPI in-built tools for API testing

on the local machine and also created the main framework around the API using this tool and the deployment and testing of these API was handled by the Uvicorn and Gunicorn libraries which works very well in conjunction with the FastAPI library. After the API's were tested locally we pushed all these files over to GitHub into a

public repo and used an open hosting platform for static websites called render to deploy our GitHub repo over on their open network. Due to this deployment now anyone in the work can use our API's and predict a company's worth by filling out the necessary parameters.

The API's deployed are linked below:
•https://kickstarter-fin-model.onrender.com
•https://kickstarter5.onrender.com

We put up the FastAPI framework to handle incoming HTTP requests and responses to ensure smooth interaction with our model. We may construct API endpoints for taking user input and making predictions using this lightweight and high-performance framework. We create the appropriate routes and request handlers to ensure that incoming requests are processed as quickly as possible. To ensure security, we employ Auth0 authentication procedures to ensure that only authorized users may access and use our API.

Frontend Development:

Create the frontend components using a framework like SvelteKit or similar. Design the user interface to capture input parameters for the prediction. Implement client-side validation for input data. Connect the frontend components with the API endpoints using HTTP requests.

We design straightforward and visually appealing front-end components that improve user experience. We provide a user interface that takes input factors needed for prediction using frameworks such as SvelteKit. To ensure data integrity, we employ client-side validation to ensure that user inputs follow the desired format and limitations. To link the frontend and back end, we use HTTP requests to communicate with the previously specified API endpoints.

Prediction and Result Display:

Accept user inputs from the frontend. Make a POST API request to '/reports/[company-name]' endpoint with the user inputs.

This request does a POST request to the API hosted on railway deployment platform that runs the Machine Learning Algorithm on the submitted data and delivers a result accordingly ,the result is a JSON that is returned from the Railway endpoint.

This is a sample request body that has to submitted to the endpoint :

{
“category”: ”3D—Adaptive Equipment—Advertising Networks”,
“total_funding”: 0.0,
“country_code”: ”BHR”,
“total_funding_rounds”: 0,
“first_funding_date”: ”1861-01-01”,
“last_funding_date”: ”1861-01-01”
}

The result can be one among the following :

{
“success”:1
}

If the model analyses the request parameters and returns the result that the company can be a successful investment on the basis of its financials and funding rounds , from the data that was submitted by the user, including fields like company's industries , total funding amount raised in USD, total number of funding rounds, first and the last funding round, difference between the first and the last funding round in days.

Similarly , the model returns the following result in case of a failure,

{
“success”:0
}

Before the final result is submitted to the model another API POST request is done on the backend, which now returns a result from the data analysis that was carried out on various countries and the metrics that matter to a company in order to turn itself into a successful investment for the interested investor.
Since, the user inputs the field country name in the frontend, a data analysis is done for that particular country and the alpha3 code of the country is sent to the ‘Railway’ deployment platform that hosts the data analysis results for all the countries.

Since the user submits the complete country name in the frontend , which is then carried to the backend, the npm package i18n-iso-countries is used in order to extract the alpha3 code of the submitted country name.

We use country.getAlpha3Code() function to extract the 3 letter alpha3 code and then successfully submit it as a POST request to the data analysis API hosted on Railway. The API returns the following data for a particular country,

- Total funding average
- Average gap b/w first and last funding rounds
- Average no of funding rounds

The following are the results of all the successful companies in a particular country and the data returned from the analysis of their data.

An example of this can be:
Request body:
{
"country_code": "IND"
}

Response:
{
"funding_total_average": 24696251.63829787,
"funding_duration_avg": 258.531914893617,
"funding_rounds_avg": 2
}

Finally, after the data analysis , the awaited code then moves to doing a final API request for the analysis of the fundamentals of a company .
This is done because analysing details regarding financials and funding might not provide the overall result with very good accuracy . There are a lot of companies that are essentially bootstrapped and analysing their details keeping in mind just the funding aspect can always deliver a negative response or a failure.
In order to mitigate the risks of presenting false information and decreasing the accuracy of the model, analysing fundamentals of a company such as total no of milestones achieved, no of years that took the company to achieve its first milestone, investment of a company in advert labels, answering if a company is listed amongst the top 500 in

its country or the world etc can provide a holistic answer to the question of whether investing in a company is worth or not.

Hence , all this data is POSTed to the endpoint deployed on 'Railway' deployment platform and appropriate result is returned.

The following is a sample request body that has to be submitted to the endpoint:

{
"advert": "Yes",
"age_fund": 0,
"age_mile": 0,
"relation_score": 100,
"signi_event": 4,
"second_round": "Yes",
"num_employ": 5,
"top500": "Yes"
}

The result can be one among the following:
If the model analyses the request parameters and returns the result that the company can be a successful investment on the basis of its fundamentals, from the data that was submitted by the user.

{
"success":1
}

Similarly , the model returns the following result in case of a failure, {
"success":0
}

Reporting and Documentation:

The overall result from the financials and fundamentals model that is returned is aggregated to provide a unified decision to the investor.

Financials model	Fundamentals model	Verdict
success = 1	success = 1	Great Investment
success = 1	success = 0	Good Investment
success = 0	success = 1	Risky Investment
success = 0	success = 0	Bad Investment

Figure 5.2: Decision making reference table

The following table Figure 5.2 displays the various decisions that can be given out by the model depending on the state of the data returned by the financials and fundamentals model:

Along with that graphs supporting the decisions and displaying clear distinction between various metrics are appropriately plotted to give clearer insights to the investor.

The various graphs that are plotted are (we have considered Cred's details over here) :

The '/reports' page consists of all the previously saved reports and pulls it from MongoDB. In order to view reports of individual companies, the user can click on the tile corresponding to the company name, which then pulls out the data of that company along with the decision of whether or not to invest in that particular company.

Along with this a thorough reasoning is given to the end user regarding the decision, followed by a data analysis of the funding threshold that has to be carried out by the company to surpass the national funding average of the companies belonging to that particular sector that are successful. A side-by-side overview of the current company's funding amount vs the national average of the companies belonging to that particular sector clearly communicates the reasoning behind the overall decision given by the model.

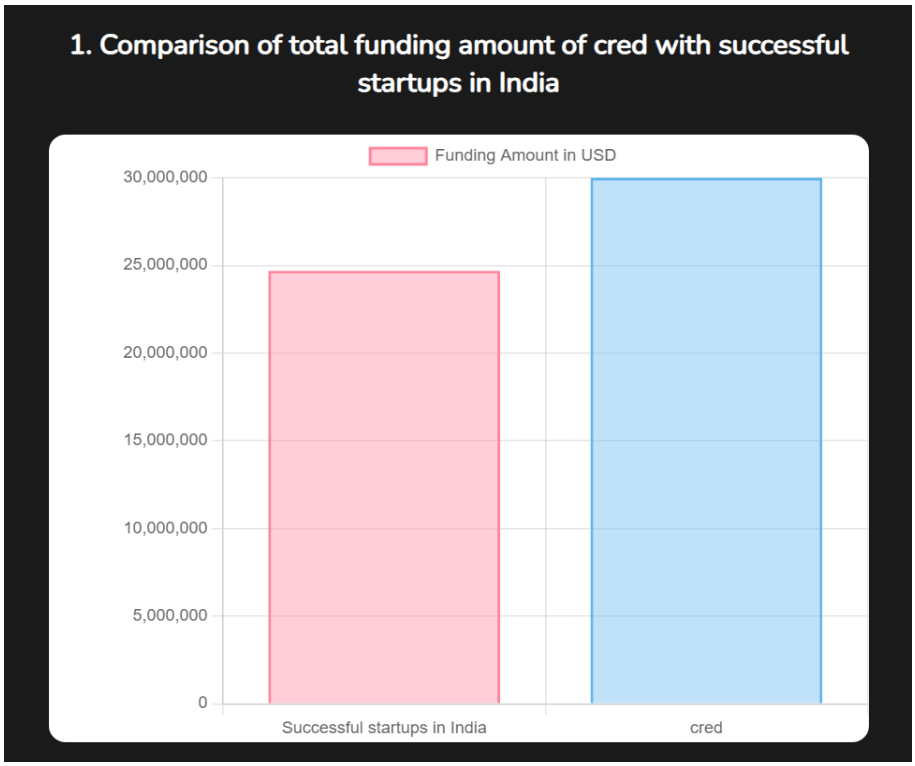


Figure 5.3: Comparison of total funding amount of Cred with successful startups in India

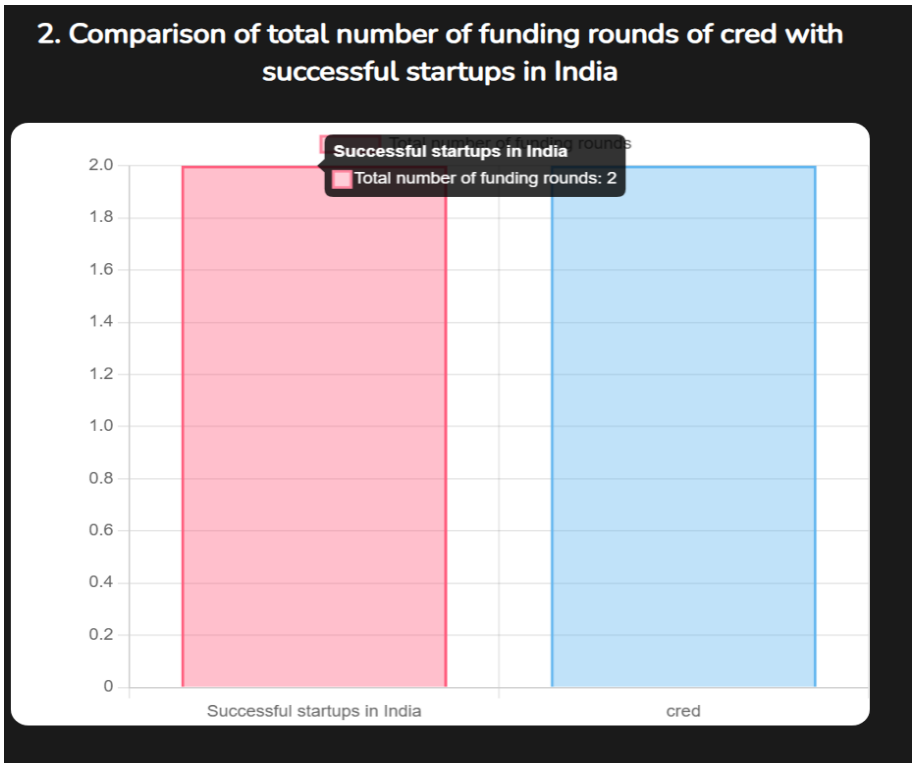


Figure 5.4: Comparison of total number of funding rounds of Cred with successful startups in India

The Figure 5.8 above is a screenshot of our product's homepage.

The Website Home page serves as the primary interface for users, offering a seamless

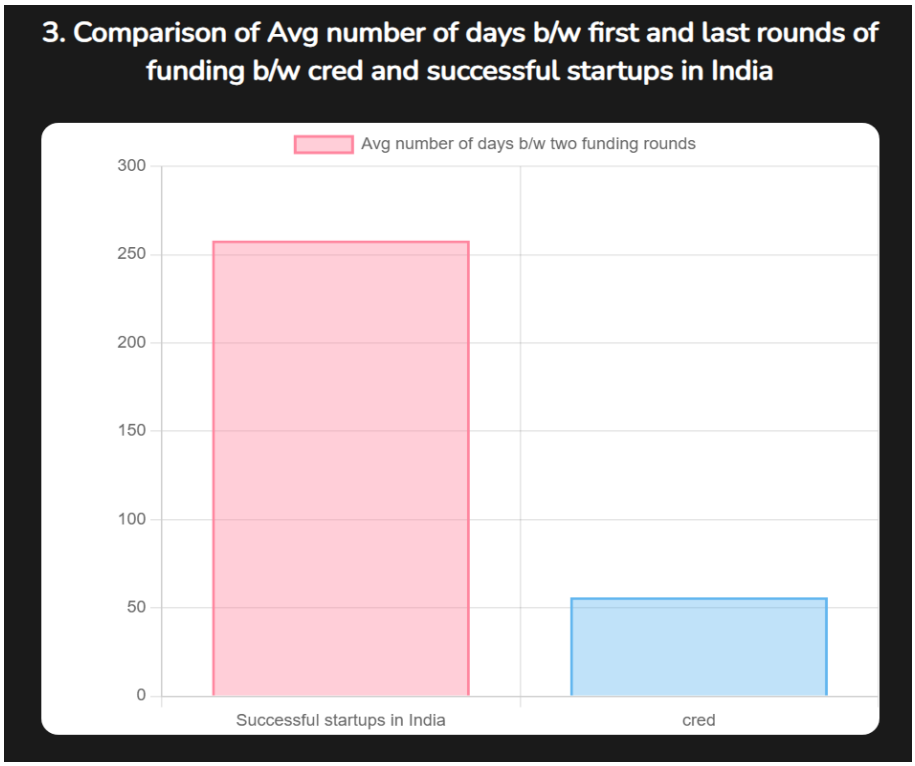


Figure 5.5: Comparison of Avg number of days b/w first and last rounds of funding b/w Cred and successful startups in India

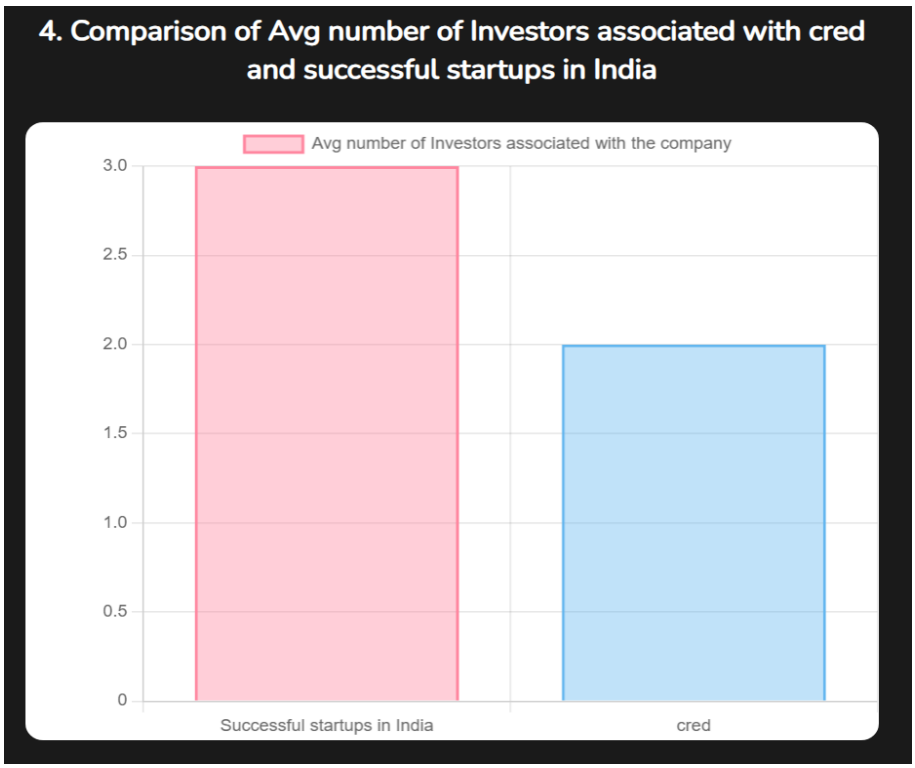


Figure 5.6: Comparison of Avg number of Investors associated with Cred and successful startups in India

gateway to navigate through our digital landscape. This report delves into our Home page's professional design and functionality, highlighting its crucial role in facilitating

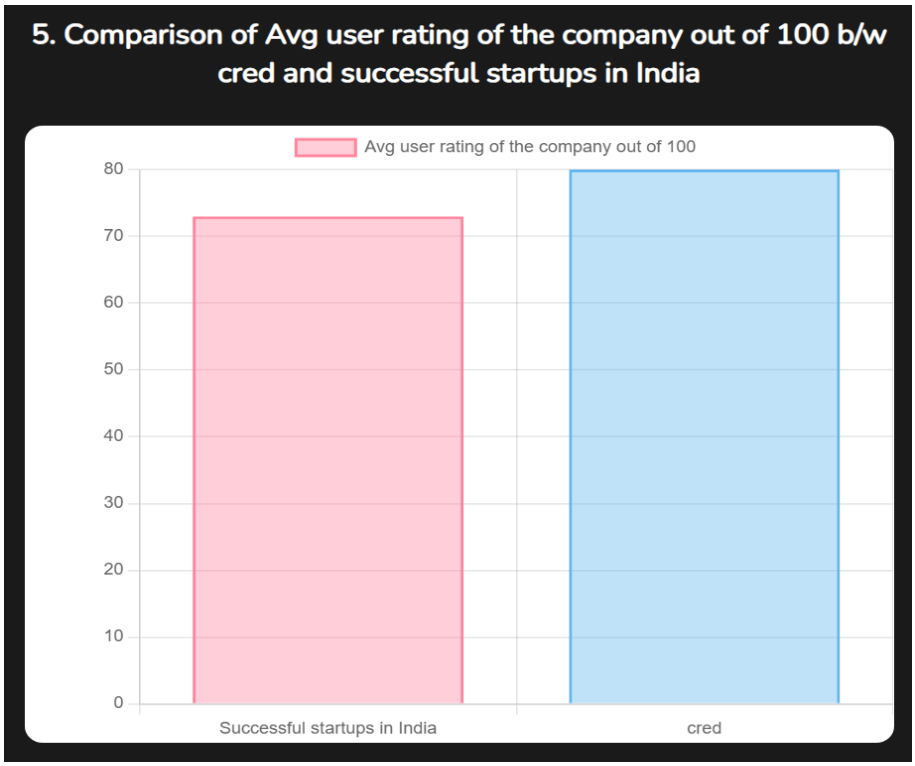


Figure 5.7: Comparison of Avg user rating of the company out of 100 b/w Cred and successful startups in India

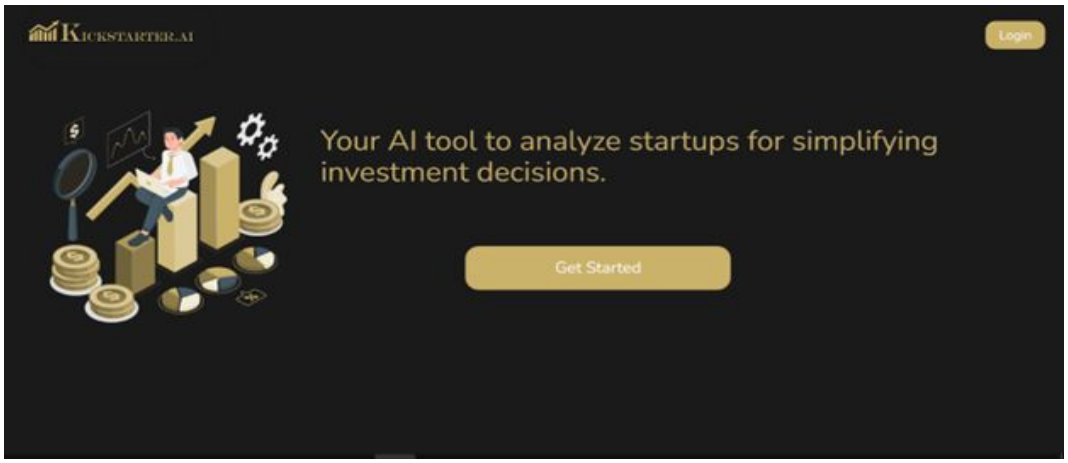


Figure 5.8: Home Page

user engagement and directing visitors to key pages of significance.

The Figure 5.9 represents Kickstarter. Ai's dashboard.

The Dashboard is our product's command centre, offering users a full set of options for analysing and customizing start-up data. This description digs into the Dashboard's multidimensional capabilities, allowing users to discover new companies, fine-tune prior assessments, and get detailed information on previously analysed initiatives.

The Figure 5.10 shows the Response Editing page allows users to easily customize and alter the metrics that were previously entered during startup analysis. It has an

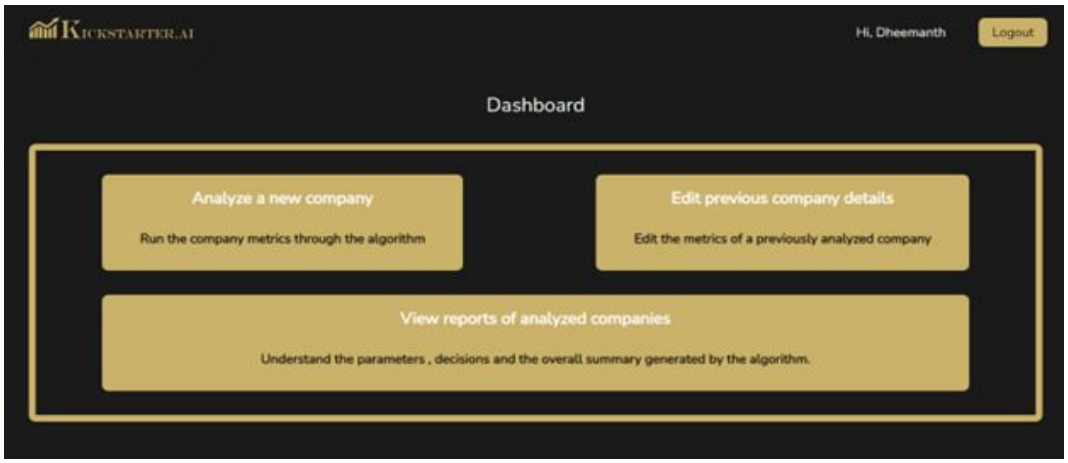


Figure 5.9: Dashboard Page

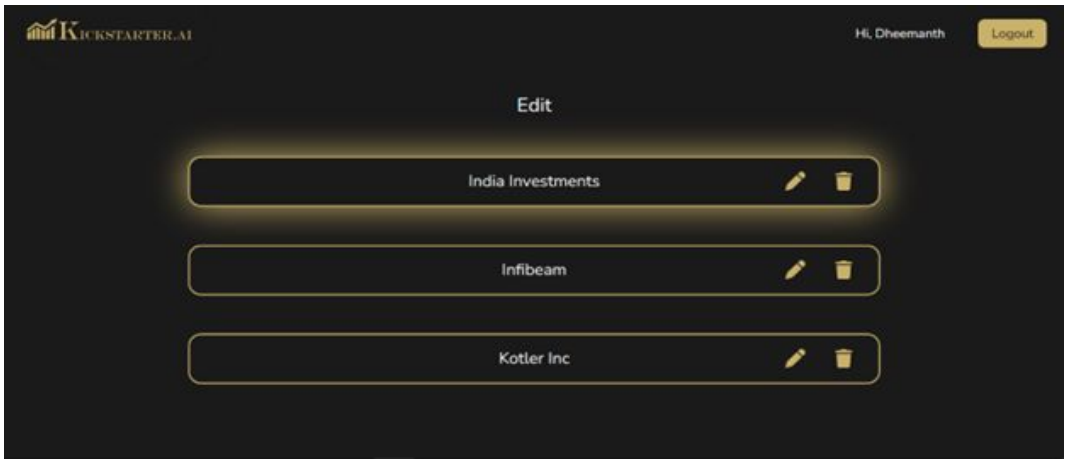


Figure 5.10: Page to edit Response

easy-to-use interface that allows users to simply change critical data points, ensuring that the analysis represents the most accurate and up-to-date information available.

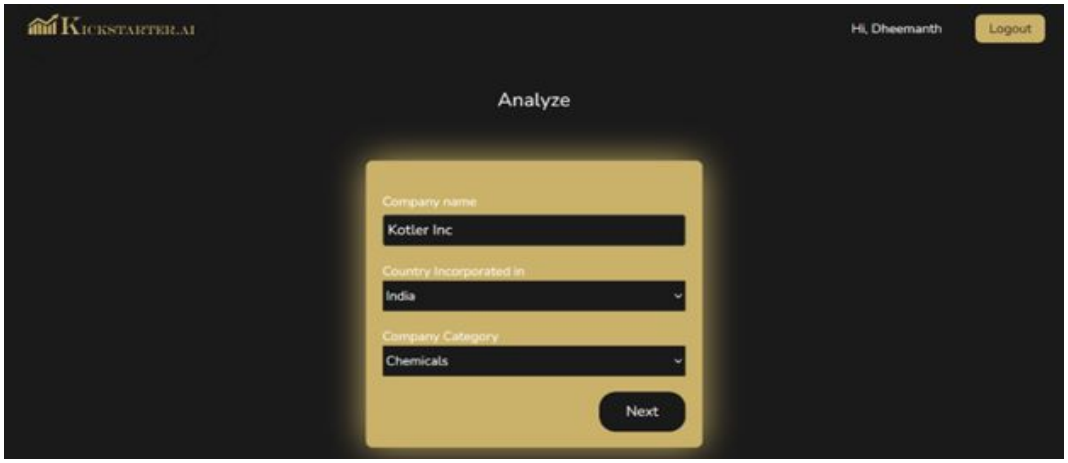


Figure 5.11: Analyse Page 1

The Analyze page Figure 5.11 provides a robust financial modeling feature that enables users to input precise details related to a startup's funding rounds. By cap-

turing the number of funding rounds, their corresponding dates, and the amounts raised in USD, users can create a comprehensive financial profile of the startup. This level of granularity allows for more accurate and insightful analysis, fostering informed decision-making.

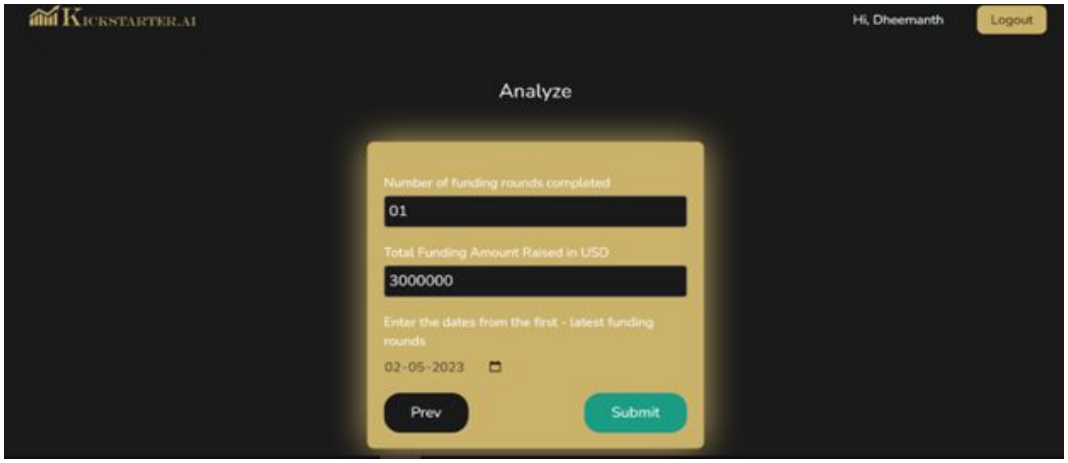


Figure 5.12: Analyse Page 2

Analyze page's Figure 5.12 Financial Model section plays a pivotal role in empowering users to conduct comprehensive startup assessments. By capturing key financial details, users can gain valuable insights into a startup's funding history and financial performance. The ability to conduct scenario planning and sensitivity analysis further enhances the analytical capabilities, enabling users to make informed decisions based on a deep understanding of the startup's financial landscape.

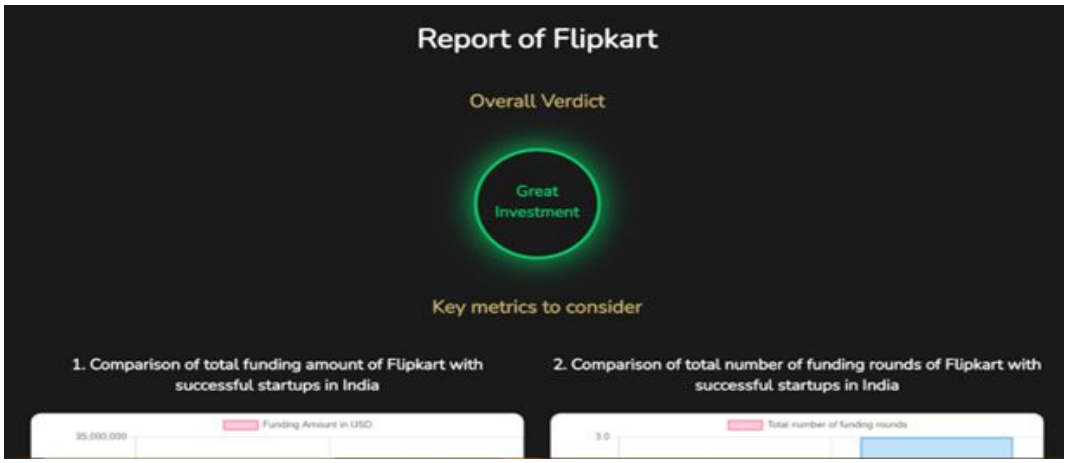


Figure 5.13: Report Classification

The image above illustrates the post-analysis report, which includes a categorization system that divides firms into four unique classifications depending on their investment potential:

Great Investment: Companies in this category have outstanding investment potential. They have good financial indications, significant growth potential, and a high probability of providing excellent profits. Investors are recommended to explore these firms as excellent investment prospects.

Companies classed as good investments have favourable qualities and provide strong investment potential. They have demonstrated strong financial performance, consistent growth patterns, and the capacity to deliver satisfying returns. These firms are regarded as trustworthy investment candidates.

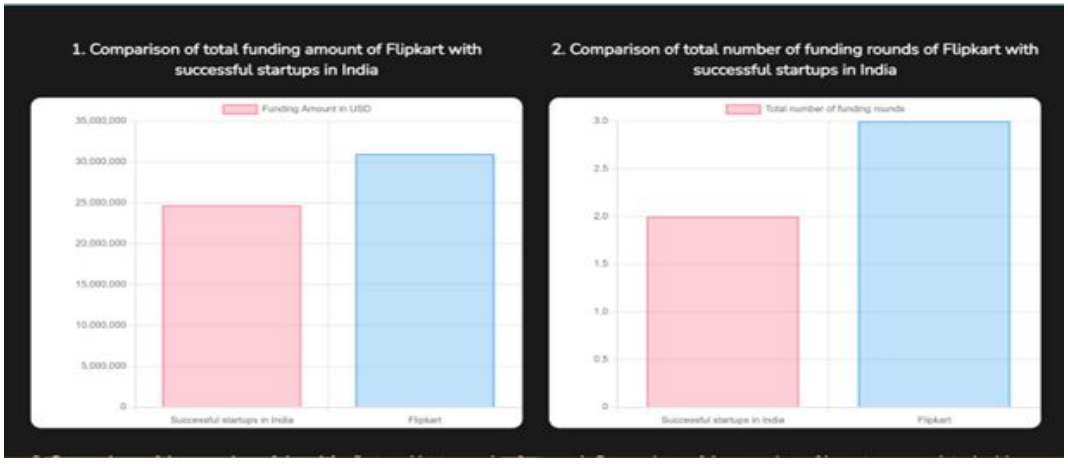


Figure 5.14: Graph Visualization

The visualization component as depicted in Figure 5.14 in the analysis of start-ups offers a comprehensive overview of each start-up and presents key statistics compared to the average statistics of companies within the same country. This visualization provides a more intuitive and informative representation of their performance and characteristics, using graphs, charts, and other visual elements to help users grasp the relative position of a start-up in terms of various metrics. These metrics serve as benchmarks for evaluating the performance of individual start-ups and provide insights into their relative strengths or weaknesses. By comparing a start-up's statistics to the country's average, users can gain a better understanding of its performance in relation to the broader market. This contextualization allows for a more nuanced assessment of a start-up's potential, considering the specific dynamics and competitive landscape of its operating country.

The visualization component aids in identifying start-ups that stand out positively or negatively from the average. Start-ups with metrics exceeding the country's average can be identified as potential high performers, attracting greater attention for invest-

ment consideration. Conversely, start-ups with metrics significantly below the average may raise concerns and warrant further scrutiny. The visualized statistics provide users with a clear visual representation, facilitating easier interpretation and comparison, enhancing the overall user experience, and enabling more informed decision-making when evaluating investment opportunities. Overall, integrating visualizations of start-ups and their statistics against the average values of their respective countries offers a comprehensive and intuitive representation of a start-up's performance within its operating environment, enabling users to gain valuable insights and make more informed investment decisions based on the relative positioning and performance of each start-up.

Figure 5.15: Donut chart and final verdict

Correlation analysis is a crucial tool in understanding the relationship between various metrics in the financial and fundamental model. By analysing the interdependencies and impact of each metric on the overall evaluation process, stakeholders can gain valuable insights. To present this information visually, a donut chart Figure 5.15 is employed to visually represent the weightage assigned to each metric within the model. This visualization serves as a valuable tool for decision-making, highlighting the key factors that contribute most significantly to the final evaluation. The AI-powered model, utilizing advanced algorithms and machine learning techniques, analyses the comprehensive set of metrics and factors considered during the evalua-

tion process. It takes into account the correlation analysis, impact weightage of each metric, and other relevant inputs to generate an unbiased and data-driven recommendation. This enhances the decision-making process, providing stakeholders with an objective and informed perspective on the investment potential of a start-up. However, stakeholders should also consider their own expertise, market knowledge, and risk tolerance when making investment decisions.

In summary, correlation analysis provides valuable insights into the relationship between metrics, and the impact weightage is visualized through a donut chart, offering a clear representation of the relative significance of each metric. The final verdict, generated by the AI-powered model, provides an unbiased and data-driven recommendation based on the comprehensive evaluation process, empowering stakeholders to make more informed investment decisions.

5.3 Dataset

The Crunchbase dataset plays a crucial role in our project as it provides valuable information about start-ups and their funding details. Crunchbase is a comprehensive database that collects and curates' data on companies, venture capital firms, investments, and other relevant information in the start-up ecosystem. It serves as a valuable resource for understanding the start-up landscape, identifying trends, and making data-driven decisions.

In our project, we leverage the Crunchbase dataset to train and test our ML models for predicting start-up success or failure. The dataset contains various parameters that are important for our prediction task. Some of the crucial parameters we consider include:

Category List: This parameter provides information about the industry or sector in which the start-up operates. It helps us understand the specific domain and market dynamics relevant to the start-up's success.

Total Funding Raised: This numerical value represents the total amount of funding the start-up has raised from various sources. It gives us insights into the financial resources available to the start-up, which is an important factor in determining its growth potential. Country: This parameter indicates the country in which the start-up is based. It allows us to analyse how the start-up's geographic location may impact its success, considering factors such as market size, regulatory environment, and avail-

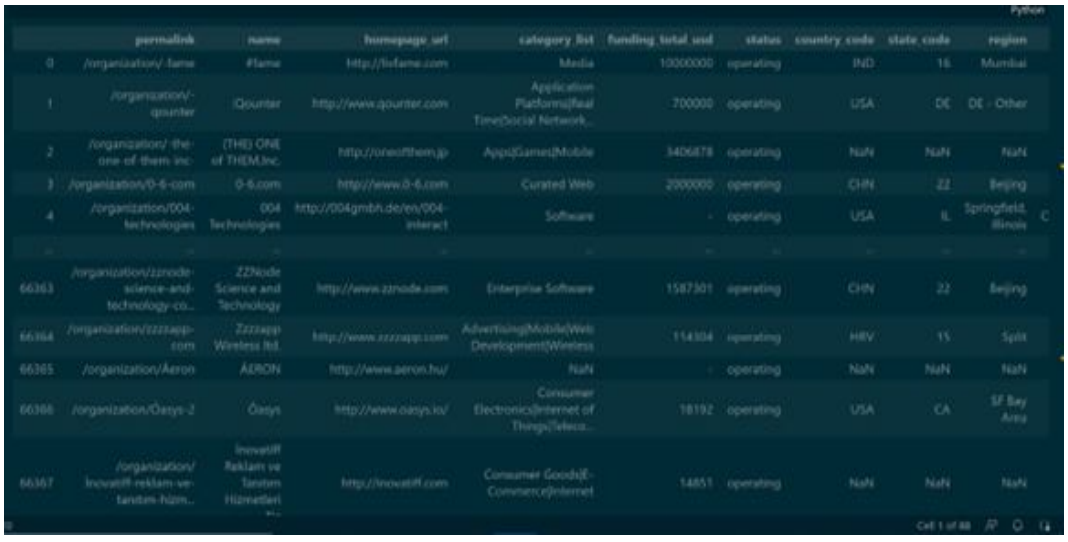
able resources.

Number of Funding Rounds: This numerical value represents the number of times the start-up has gone through funding rounds. It provides an indication of the start-up's ability to attract investment and its progress over time.

Funding Dates: The funding dates parameter includes the specific dates on which the start-up received funding. Analysing the timing of funding rounds can help identify patterns or trends that may influence the start-up's success, such as a concentrated period of funding or long intervals between rounds.

By incorporating these important parameters from the Crunchbase dataset, we can build ML models that learn patterns and relationships between the start-up characteristics and their eventual outcomes, such as IPO transition or acquisition by a larger company. The dataset allows us to capture the diverse factors that contribute to start-up success and failure, enabling us to develop accurate prediction models.

Additionally, the Crunchbase dataset provides a vast amount of additional information that can be explored to extract further insights and enhance the predictive power of our models. This includes details about company founders, team size, market trends, investor profiles, and more. By incorporating these features and leveraging the richness of the dataset, we aim to create robust ML models that provide valuable predictions and support decision-making in the start-up ecosystem.



	permalink	name	homepage_url	category_list	funding_total_usd	status	country_code	state_code	region
0	/organization/-fame	#fame	http://livefame.com	Media	10000000	operating	IND	16	Mumbai
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	700000	operating	USA	DE	DE - Other
2	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	3406878	operating	NaN	NaN	NaN
3	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	2000000	operating	CHN	22	Beijing
4	/organization/004-technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	-	operating	USA	IL	Springfield, Illinois
...	...	...	...	...	...	...	...	...	...
66363	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	Enterprise Software	1587301	operating	CHN	22	Beijing
66364	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	Advertising|Mobile|Web Development|Wireless	114304	operating	HRV	15	Split
66365	/organization/Aeron	AERON	http://www.aeron.hu/	NaN	-	operating	NaN	NaN	NaN
66366	/organization/Oasys-2	Oasys	http://www.oasys.io/	Consumer Electronics|Internet of Things|Teleco...	18192	operating	USA	CA	SF Bay Area
66367	/organization/Inovatiff-reklam-ve-tanitim-hizm...	Inovatiff Reklam ve Tanitim Hizmetleri	http://inovatiff.com	Consumer Goods|E-Commerce|Internet	14851	operating	NaN	NaN	NaN

Figure 5.16: Dataset prior to pre-processing

Before preprocessing steps Figure 5.16, it is crucial to understand the dataset, which includes financial details, funding rounds, and other start-up-related data. The dataset may exhibit characteristics such as missing values, outliers, inconsistencies, and potential data quality issues. Preprocessing steps should be tailored to address

these challenges and maximize the reliability and utility of the data for subsequent analyses. The dataset may consist of both numerical and categorical variables, such as funding amounts, funding rounds, and industry categories. Assessing the scale and distribution of the variables is crucial for determining appropriate preprocessing techniques.

Additionally, the dataset may contain temporal information, such as funding round dates, which can provide valuable insights into start-up dynamics and growth patterns over time. Overall, a comprehensive examination of the dataset is vital for making informed decisions regarding data cleaning, feature engineering, and modelling.

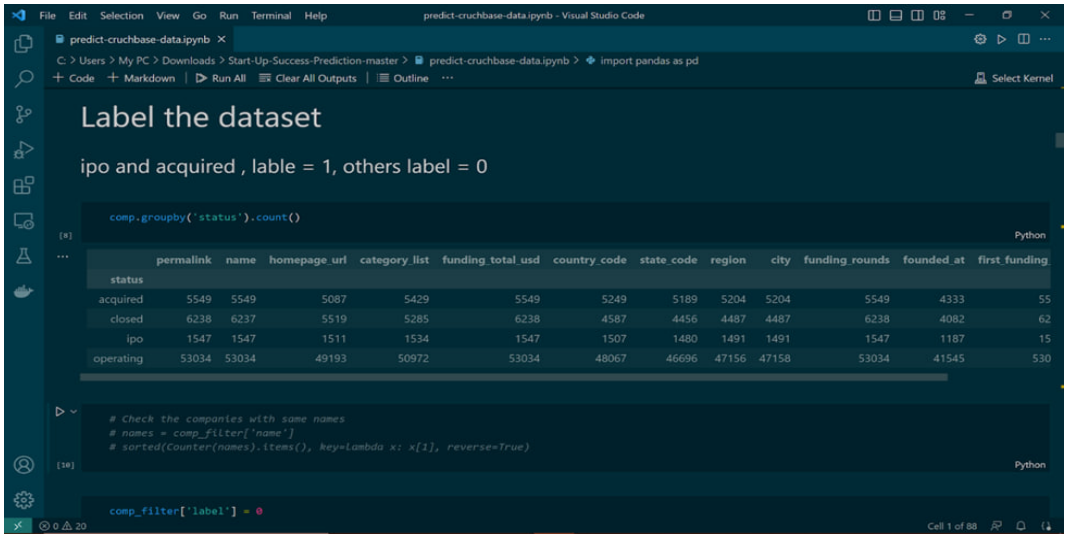


Figure 5.17: Labelling the dataset

The labelling process is a crucial step in supervised machine learning, as it involves assigning appropriate target labels or classes to each instance in the dataset. The choice of labels depends on the specific task and objectives of the analysis. To ensure accurate and reliable labelling, clear criteria, domain expertise, prior knowledge, or established guidelines are essential. An analyst or subject matter expert may review each instance and assign the most appropriate label based on the desired outcomes. Automated or semi-automated approaches can be employed to label a subset of instances, which can then be used to train a model that predicts labels for the remaining unlabelled instances.

Maintaining consistency and accuracy is crucial throughout the labelling process, as the quality and accuracy directly impact the performance and generalization capabilities of the machine learning models trained on it. Regular quality checks and inter-rater reliability assessments can verify the consistency and accuracy of the labelling. In conclusion, labelling the dataset is a crucial step in supervised machine

learning, laying the foundation for effective models for analysis and prediction Figure 5.17.

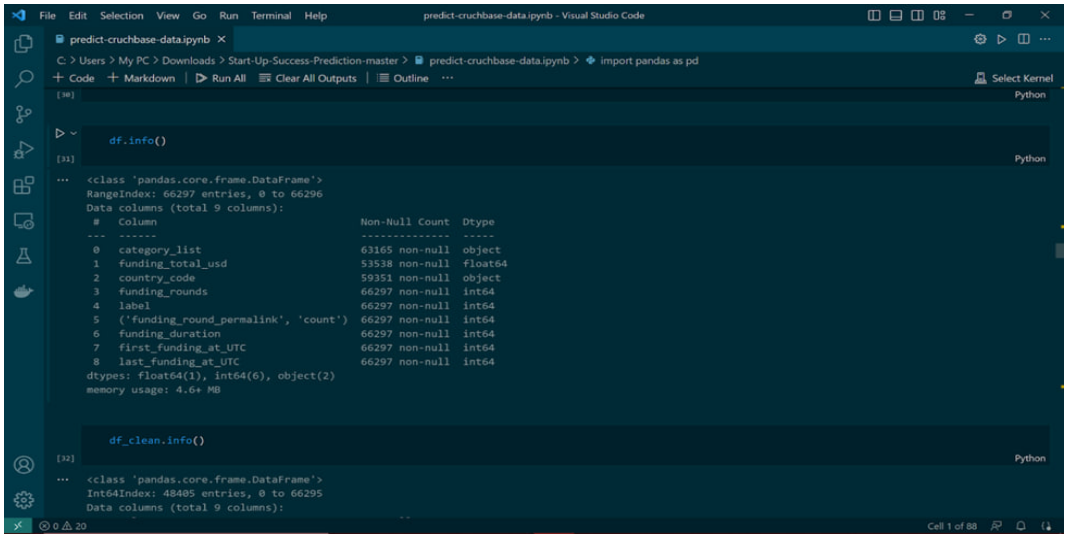


Figure 5.18: Collecting meaning attributes post pre-processing

Collecting meaningful attributes Figure 5.18 from a pre-processed dataset is crucial for subsequent analysis. This involves carefully examining the pre-processed dataset to identify the variables or features that are most relevant to the specific analysis or modelling task at hand. This process requires a deep understanding of the domain and objectives of the project. During preprocessing, variables and their transformations are reviewed, considering the range of features obtained from the initial dataset. Each attribute should be evaluated in terms of its potential contribution to the analysis and its ability to capture meaningful information. Domain knowledge and expert insights play a vital role in identifying the attributes that are most likely to have a direct impact on the desired outcomes.

Statistical analysis techniques, such as correlation analysis or feature importance ranking, can be employed to assess the relevance and predictive power of each attribute. Striking a balance between the number of attributes and the risk of overfitting or introducing unnecessary complexity is essential. Selecting too many attributes can lead to model instability and reduced generalization capabilities while omitting important attributes can result in the loss of valuable information.

The selection of meaningful attributes should also consider the computational feasibility of the subsequent analysis. Prioritizing or selecting a subset of attributes that strike the right balance between relevance and computational efficiency may be necessary. The goal of collecting meaningful attributes is to create a refined and focused dataset that contains the most informative variables for the analysis. By carefully

selecting relevant attributes, subsequent modelling and analysis processes are based on the most important and impactful features, leading to more accurate and insightful results.

Chapter 6

Testing

Testing our ML models is a meticulously orchestrated phase, characterized by a thorough evaluation process that yields accurate and reliable predictions with wide-ranging generalizability. We embark on this critical endeavour by carefully partitioning the Crunchbase dataset into distinct training and testing sets, forming the foundation for our evaluation journey. The training set becomes the breeding ground for honing our models, with notable contenders including the versatile K-Nearest Neighbours (KNN), the robust Random Forest, and the interpretable yet powerful Logistic Regression.

When subjecting our models to rigorous evaluation, an ensemble of metrics takes centre stage, allowing us to scrutinize their performance from various angles. Accuracy, precision, recall, and F1-score emerge as stalwart companions in our quest for pinpointing the models' ability to effectively classify start-ups as either destined for success or teetering on the brink of failure. These metrics serve as guiding lights, illuminating the path to understanding the models' predictive accuracy, their knack for uncovering intricate relationships, and their propensity for capturing nuanced patterns residing within the dataset's vast expanse.

The KNN model, with its ability to discern similarities and differences within the data, undergoes meticulous evaluation, leaving no stone unturned. As we unravel its true potential, we embrace evaluation metrics that shine a light on its classification prowess, empowering us to unravel the subtleties of start-up success prediction. The Random Forest, a force to be reckoned with, faces a rigorous assessment encompassing accuracy, precision, recall, F1-score, and the formidable area under the receiver operating characteristic curve (AUC-ROC). Through this intricate evaluation dance, we delve into the model's ability to navigate complex decision boundaries and capture

the essence of start-up triumph. Meanwhile, the Logistic Regression model, renowned for its interpretability, stands under the scrutiny of similar evaluation metrics, illuminating its capability to uncover critical insights into the factors that steer start-ups towards success or failure.

Our relentless pursuit of reliable and generalizable models leads us to embrace cross-validation techniques of the highest calibre. Among these techniques, the esteemed k-fold cross-validation takes centre stage, enabling us to traverse diverse folds and meticulously train and test the models on distinct subsets of the dataset. By doing so, we obtain a comprehensive understanding of their performance across a multitude of scenarios, ensuring their resilience in the face of varying data distributions and preserving their ability to provide insightful predictions.

A noteworthy aspect of our testing process lies in the art of ablation studies, where we systematically strip away input parameters to assess their individual impact on the models' performance. This surgical precision allows us to fine-tune our models, uncovering the most influential factors in the prediction process. The intricate dance of feature importance unravels before us, aiding in the refinement of our models and enhancing their predictive prowess.

Throughout the testing phase, we meticulously document the models' performance, capturing their strengths, weaknesses, and idiosyncrasies. These invaluable records fuel iterative improvements, laying the groundwork for enhanced accuracy and informed decision-making. Our unwavering commitment to transparency and explain ability culminates in the delivery of comprehensive reports. These reports not only unveil the models' predictions but also provide profound insights into the underlying factors that shape each prediction, empowering stakeholders with a profound understanding of the models' decision-making journey.

By meticulously adhering to this rigorous testing process, our triumvirate of KNN, Random Forest, and Logistic Regression models emerge as stalwart allies, armed with the precision and robustness needed to predict start-up success with unwavering accuracy. The amalgamation of comprehensive evaluation metrics, robust cross-validation techniques, and intricate ablation studies bestows upon us a holistic view of their capabilities and limitations. In this ever-evolving landscape of start-up investment, our tireless testing efforts culminate in accurate and reliable predictions, arming stakeholders with invaluable insights for strategic decision-making.

Chapter 7

Experimentation and Results

7.1 Experimentation phase

In our experimentation phase after surveying multiple papers, we got to the point where we narrowed down our search to 3 machine learning models that are: Random forest, linear regression, and K-nearest neighbours. Using these 3 models we find the best model for our purpose out of them. The first model for our implementation purpose was the logistic regression model which is a very basic model but helped us create a baseline of what a good model should look like in the comparison to the base model. The base parameters like accuracy, TPR(True Positive rate), FPR(False Positive Rate) were found out for this model using the previous distinction of train, dev and test datasets, which were named as X_train_con, y_dev and y_test respectively.

For a more pictorial representation a confusion matrix heatmap was also generated for the results of the same model. The next model for testing was the Random forest where we used the RandomForestClassifier() function from the tensor library. Similar to the previous implementation, the random forest model was also tested on its accuracy of the prediction that it was making using the predict() function with addition of a confusion matrix heatmap generated for the model. The next mode in our consideration was KNN which uses the KNeighborsClassifier() function again from the tensor library. Similar to our previous approaches the accuracy score from the prediction and a confusion matrix heatmap was made for the same.

The accuracy score, TPR, FPR and the confusion matrix generated from these models were used to better understand and find the best model for our purpose of predicting the success of an organization using the financial data, and the funding amount and time period of that organization. This obviously didn't take into account

whether the company fairs on good terms with the customers if they had any.

The second model that we experimented and developed used such kind of satisfaction data from the customers and provided an outward analysis of the company rather than giving an inward analysis which was done by our previous model.

The same approach of choosing multiple models namely logistic regression, KNN, random forest and SVC was used in the development of this model however it's focus and data was centred around different params like the number of employees in a company and rating of a company from 1 to 100 by a customer, so the random forest again cannot be used for the same.

Thus, the same testing process for the models started and we ended up with the conclusion that the SVC(support vector classification) model was used for this prediction base by analysing the prediction accuracy of the model.

7.2 Results

The results of our extensive experimentation and evaluation of the Logistic Regression, K-Nearest Neighbours (KNN), and Random Forest models have provided us with valuable insights into their performance in predicting start-up success. The Logistic Regression model achieved an accuracy of 73.39%, indicating that it correctly classified start-ups as succeeding or failing in 73.39% of cases Figure7.1 . Furthermore, the True Positive Rate (TPR) of 67.21% showcases the model's ability to accurately identify start-ups that would succeed, while the False Positive Rate (FPR) of 25.71% indicates the rate of incorrect predictions for succeeding start-ups. Although the Logistic Regression model demonstrates a reasonably high accuracy, it exhibits room for improvement in terms of its ability to correctly classify both succeeding and failing start-ups.

The KNN model yielded a similar performance, with an accuracy of 73.06%. The TPR of 70.13% indicates a relatively high capability of correctly identifying succeeding start-ups, while the FPR of 26.52% suggests a slightly higher rate of incorrect predictions for succeeding start-ups Figure 7.3. The KNN model shows promise in accurately predicting start-up success, but also displays room for refinement to reduce the false positive predictions.

On the other hand, the Random Forest model exhibited outstanding performance, achieving an impressive accuracy of 85.91%. Although the TPR of 28.25% indicates a relatively lower ability to accurately identify succeeding start-ups, the FPR of 5.68%

showcases the model's superior capability in minimizing false positive predictions Figure 7.2. The Random Forest model's high accuracy and low false positive rate make it a valuable asset in predicting start-up success, albeit with a focus on refining its ability to identify succeeding start-ups more accurately.

These results highlight the importance of selecting the appropriate model for the task at hand. While the Logistic Regression and KNN models demonstrate respectable performance, the Random Forest model outshines them by delivering superior accuracy and minimizing false positive predictions. This suggests that the Random Forest model's ensemble learning approach, leveraging multiple decision trees, enables it to capture complex relationships and patterns within the Crunchbase dataset more effectively.

Furthermore, these results emphasize the significance of understanding the trade-offs associated with different models. The Logistic Regression and KNN models may be more suitable in scenarios where minimizing false negatives is crucial, while the Random Forest model excels in scenarios where reducing false positives is paramount. The choice of model depends on the specific needs and requirements of the application, allowing stakeholders to make informed decisions based on the desired balance between precision and recall.

In conclusion, the evaluation results shed light on the strengths and limitations of the Logistic Regression, KNN, and Random Forest models in predicting start-up success. The Random Forest model emerged as the most promising candidate, exhibiting a remarkable balance between accuracy and false positive rate. These findings guide us in selecting the most appropriate model for our project, ensuring accurate predictions and empowering stakeholders to make informed decisions in the dynamic landscape of start-up success prediction.

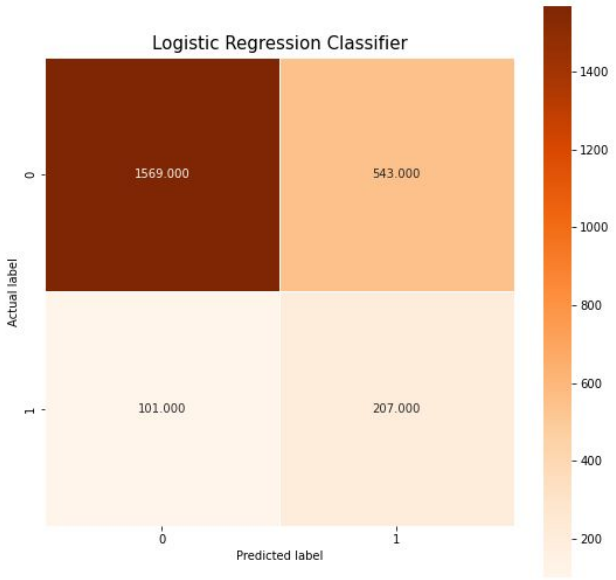


Figure 7.1: LRC confusion Matrix

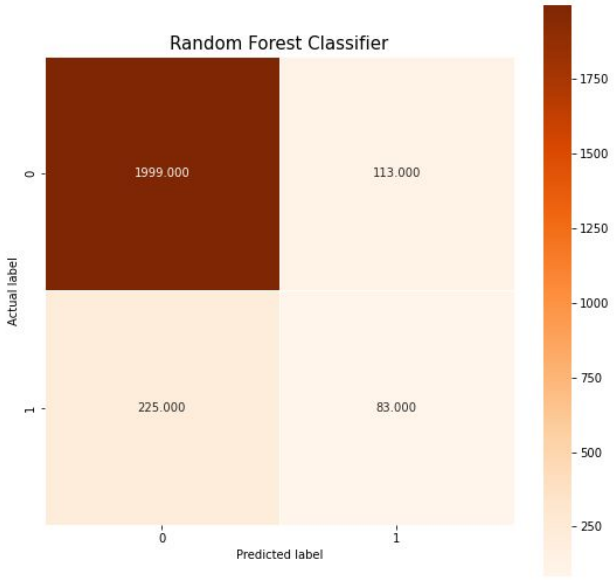


Figure 7.2: Random forest confusion Matrix

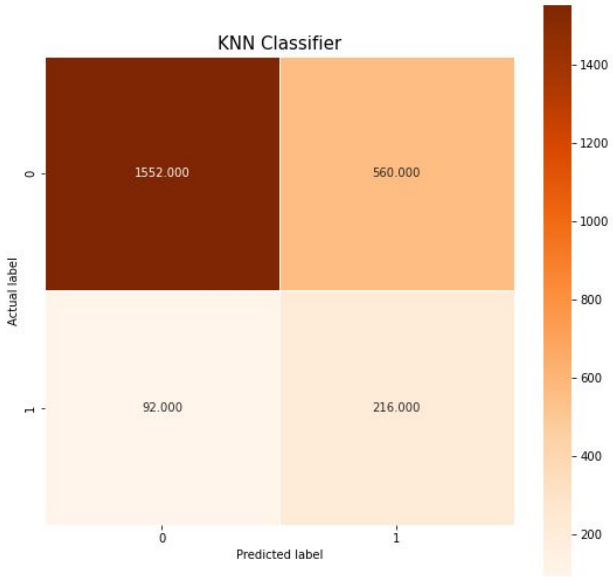


Figure 7.3: KNN Classifier confusion Matrix

Conclusion

In conclusion, our project on predicting start-up success using the Crunchbase dataset has yielded significant insights and outcomes that contribute to the field of entrepreneurial research and decision-making. Through a meticulous and rigorous approach, we have successfully developed a robust solution that empowers stakeholders to make informed decisions and allocate resources effectively.

By leveraging the extensive Crunchbase dataset, we have gained valuable insights into the key factors influencing start-up success. The dataset's rich information on categories, funding, geographical location, and funding rounds has allowed us to uncover patterns, trends, and relationships that contribute to a start-up's trajectory. This comprehensive understanding of the start-up ecosystem provides valuable knowledge for investors, accelerators, and entrepreneurs alike.

Our data pre-processing phase was crucial in ensuring the quality and relevance of the dataset. By applying techniques such as cleaning, normalization, and feature engineering, we were able to handle missing data, eliminate noise, and transform the raw data into a suitable format for our machine learning models. This pre-processing step enhanced the accuracy and reliability of our predictions, laying a solid foundation for the subsequent stages of our project.

The selection and evaluation of machine learning models played a vital role in the project's success. Through careful experimentation and comparison, we identified the Random Forest model as the most effective in predicting start-up success. Its ability to handle complex relationships, mitigate overfitting, and provide interpretable results made it an ideal choice for our project. Additionally, we tested and evaluated other models such as Logistic Regression and K-Nearest Neighbours, providing comprehensive insights into their performance and capabilities.

The testing and evaluation phase provided valuable validation for our models. By using various evaluation metrics, including accuracy, true positive rate, and false positive rate, we were able to measure the models' effectiveness in predicting start-up

success. The comprehensive evaluation process, including train-test splitting, cross-validation, and ablation studies, ensured the reliability, generalizability, and interpretability of our models' predictions.

Furthermore, we have taken our project beyond the realm of machine learning by integrating our models into a web application. By utilizing technologies such as FAST API, Svelte Kit, and Auth0 authentication, we have provided a user-friendly interface for stakeholders to input parameters and obtain predictions. This integration enhances the accessibility and usability of our models, allowing users to make data-driven decisions conveniently and securely.

Ultimately, our project has achieved its objectives by developing a comprehensive solution for predicting start-up success. The insights, models, and web application we have created empower stakeholders, including investors and venture capitalists, to make informed decisions, allocate resources efficiently, and identify promising start-ups for investment or collaboration. Our project contributes to the advancement of entrepreneurial research, providing valuable tools and methodologies for understanding and predicting start-up success in a rapidly evolving business landscape. By harnessing the power of data and machine learning, our project equips stakeholders with the necessary tools to navigate the complex start-up ecosystem and maximize their potential for success.

References

1. Nick Skillicorn, 65% of Venture Capital-backed Deals Fail to Return Investment, and Only 4% Make Substantial Returns, IDEA TO VALUE (Oc.t 18, 2018)

2. Nicolás Cerdeira Kyril Kotashev, Startup Failure Rate: Ultimate Report + Infographic [2021], FAILORY (Mar. 25, 2021)

3. Tomer Dean, The Meeting That Showed Me the Truth About VCs, TECHCRUNCH (Jun. 1, 2017)

4. Sam Reynolds, VCs Lose a Lot of Money Hunting for the 10x Return, WCCF TECH (Sept. 17, 2019)

5. Jared Council, VC Firms Have Long Backed AI, Now, They Are Using It, THE WALL STREET J. (Mar. 25, 2021)

6. Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

7. Amit, R., Brander, J., Zott, C. Why do venture capital firms exist? Theory and Canadian evidence. Journal of business Venturing, 13(6), 441-466(1998)

8. Arvidsson, V., Holmström, J., & Lyytinen, K. Information systems use as strategy practice: A multi-dimensional view of strategic information system implementation and use. The Journal of Strategic Information Systems, 23(1), 45-61(2014)

9. Hyoung Jun Kim, Tae San Kim, So Young Sohn, Recommendation of startups as technology cooperation candidates from the perspectives of similarity and potential: A deep learning approach, Decision Support Systems, Volume 130, 2020, 113229, ISSN 0167-9236

10. Arroyo, Javier et al. "Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments." IEEE Access 7 (2019): 124233-124243

11. Jan K. Woike, Ulrich Hoffrage, Jeffrey S. Petty, Picking profitable investments: The successor equal weighting in simulated venture capitalist decision making, Elsevier (2015)(JBR-08367)

12. Torben Antretter, Ivo Blohm, Diemtar Grichnik, Joakim Wincent, Predicting new venture survival: A Twitter-based Machine Learning approach to measure online legitimacy, Journal of Business Venturing Insights 11 (2018) e00109

13. Cockburn IM, Macgarvie MJ. Patents, thickets and the financing of early-stage firms: evidence from the software industry. J Econ. 2009;18(3):729e773.

14. Mann RJ, Sager TW. Patents, venture capital, and software start-ups. Res Pol. 2007, March;36(2):193e208.

15. Greg Ross, Sanjiv Das, Daniel Sciro, Hussain Raza, CapitalVX: A machine learning model for startup selection and exit prediction, The Journal of Finance and Data Science, Volume 7, 2021, Pages 94-114, ISSN 2405-9188

16. Puri M, Zarutskie R. On the lifecycle dynamics of venture-capital- and non-venture-capital-financed firms. J Finance.2012;67(6):2247e2293.

17. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016, August:785e794.

18. Santos RS, Qin L. Risk Capital and Emerging Technologies: Innovation and Investment Patterns Based on Artificial Intelligence Patent Data Analysis. Journal of Risk and Financial Management. 2019; 12(4):189.

19. Ajai Mishra, Dharm Singh Jat, and Durgesh Kumar Mishra. 2022. Machine Intelligence for Predicting New Start-ups Success: A Survey. In Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence (DSMLAI '21'). Association for Computing Machinery, New York, NY, USA, 99–105. https://doi.org/10.1145/3484824.3484919

20. J. Arroyo, F. Corea, G. Jimenez-Diaz and J. A. Recio-Garcia, "Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments," in IEEE Access, vol. 7, pp. 124233-124243, 2019, doi: 10.1109/ACCESS.2019.2938659.

21. Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2021). A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. Proceedings of the International AAAI Conference on Web and Social Media, 6(1), 607-610

22. Y. Zhao, Y. Shen, Y. Zhu, et al., "Forecasting wavelet transformed time series with attentive neural networks," in Proceedings of the International Conference on Data Mining (ICDM), Singapore, November 2018. View at: Publisher Site — Google Scholar

23. AlexSherstinsky, "Fundamentals of Recurrent Neural Network and Long short term memory", 2019.

Report

ORIGINALITY REPORT

7%	5%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	v1.overleaf.com Internet Source	3%
2	Submitted to Visvesvaraya Technological University Student Paper	1%
3	Submitted to Middle East College of Information Technology Student Paper	<1%
4	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1%
5	www.science.gov Internet Source	<1%
6	studymoose.com Internet Source	<1%
7	Submitted to CSU, Dominguez Hills Student Paper	<1%
8	Submitted to Arab Open University Student Paper	<1%

9	ir.uitm.edu.my Internet Source	<1%
10	www.coursehero.com Internet Source	<1%
11	github.com Internet Source	<1%
12	Submitted to BMS College of Engineering Student Paper	<1%
13	Submitted to University of Wales Institute, Cardiff Student Paper	<1%
14	qdosd.squiz.cloud Internet Source	<1%
15	repozitorij.uni-lj.si Internet Source	<1%
16	"Message from the chairman", 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017 Publication	<1%
17	Mohamed Aly Bouke, Azizol Abdullah. "An Empirical Study of Pattern Leakage Impact During Data Preprocessing On Machine Learning-Based Intrusion Detection Models Reliability", Expert Systems with Applications, 2023	<1%

	Publication	
18	elibrary.tucl.edu.np Internet Source	<1 %
19	hdl.handle.net Internet Source	<1 %
20	innovative-technology.com Internet Source	<1 %
21	mp.jvolsu.com Internet Source	<1 %
22	web.wpi.edu Internet Source	<1 %
23	www.biorxiv.org Internet Source	<1 %
24	www.kscst.iisc.ernet.in Internet Source	<1 %
25	www.repository.cam.ac.uk Internet Source	<1 %
26	Greg Ross, Sanjiv Das, Daniel Sciro, Hussain Raza. "CapitalVX: A machine learning model for startup selection and exit prediction", The Journal of Finance and Data Science, 2021 Publication	<1 %

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

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

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

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

1. Abstract

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

2. Introduction

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

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

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

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

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

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

3. Previous research

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

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

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

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

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

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

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

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

4. Data overview

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

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

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

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

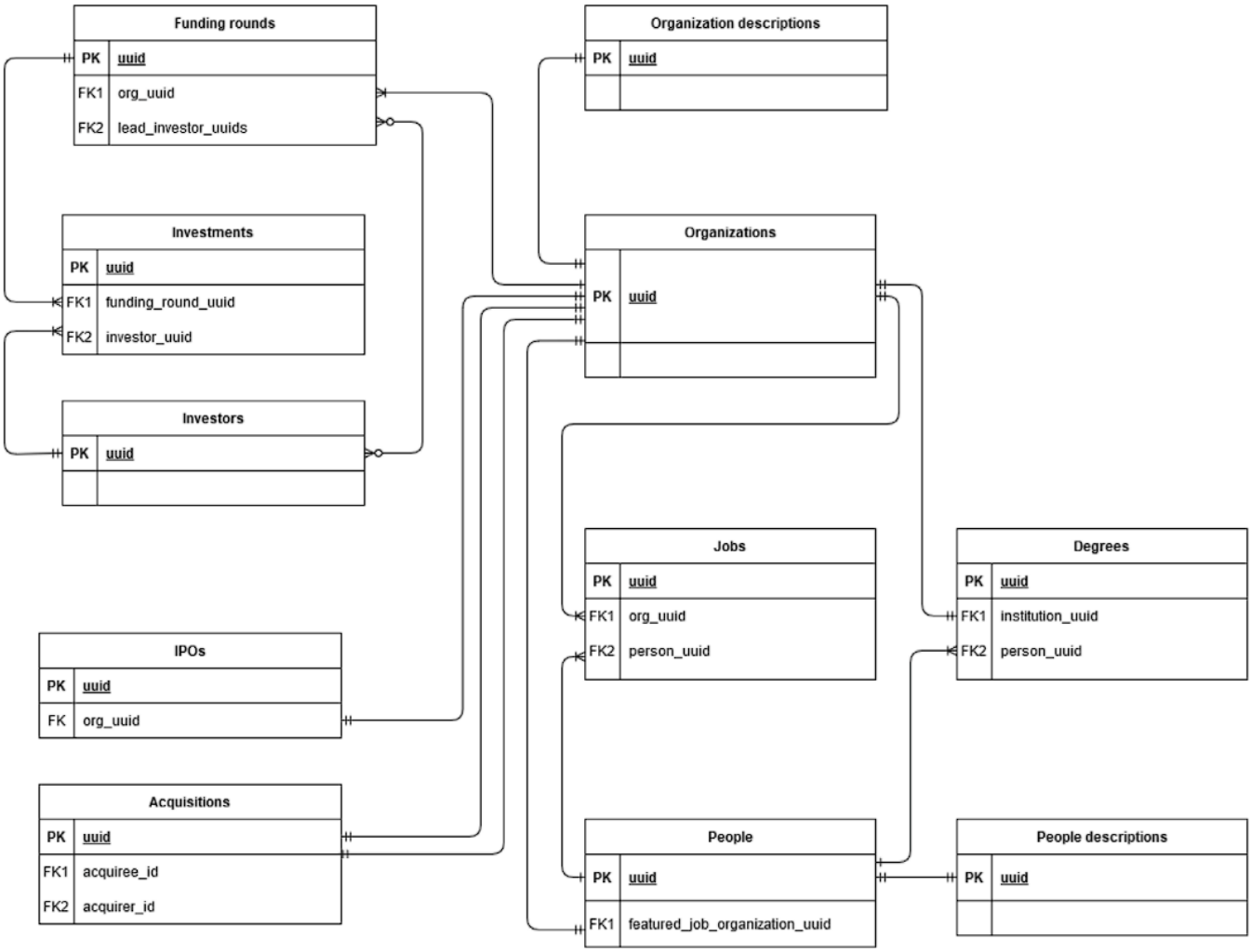


Fig1. Crunchbase data simple ERD diagram

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

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

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

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

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

Table 2: Summary of Additional External Data

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

4.1 Data processing

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

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

4.2 SQL-based feature selection and engineering

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

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

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

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

Table 3:
Dataset parameters to be used in the implementation

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

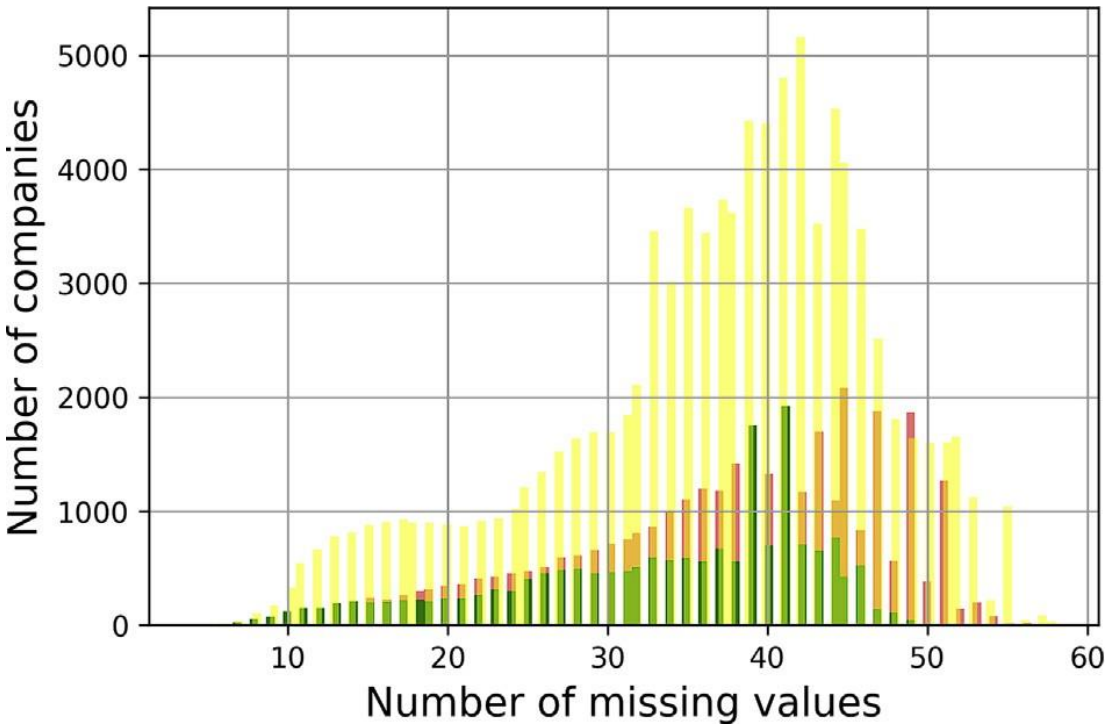


Fig2. The graph represents missing values from companies in all domain public acquired or failed companies. Red represents failures, Yellow represent acquired companies, while Green shows public companies

4.3 How patent data from USPTO is useful

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

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

5. Question to be answered by analysis

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

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

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

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

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

6. Model implementation

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

6.1 Splitting the data for training and maintaining the class balance

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

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

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

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

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

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

6.2 Concept of multiple-layered perceptrons

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

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

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

6.3 Logistic Regression

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

6.4 Random Forest

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

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

6.5 Xtreme gradient boosting algorithm (XGB)

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

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

6.6 Two-step classification

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

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

6.7 Robustness

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

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

7. Expandability

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

8. Conclusion

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

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

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

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

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

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

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

References

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

DOI: 10.55041/IJSREM17996

ISSN: 2582-3930

Impact Factor: 7.185

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

Mantej Singh Tuli

in recognizaton to the publication of paper titled

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

published in IJSREM Journal on Volume 07 Issue 03 March, 2023

Editor-in-Chief
IJSREM Journal

DOI: 10.55041/IJSREM17996

ISSN: 2582-3930

Impact Factor: 7.185

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

Shreyas G

in recognizaton to the publication of paper titled

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

published in IJSREM Journal on Volume 07 Issue 03 March, 2023

Editor-in-Chief
IJSREM Journal

DOI: 10.55041/IJSREM17996

ISSN: 2582-3930

Impact Factor: 7.185

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

Dheemanth A N

in recognizaton to the publication of paper titled

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

published in IJSREM Journal on Volume 07 Issue 03 March, 2023

Editor-in-Chief
IJSREM Journal

DOI: 10.55041/IJSREM17996

ISSN: 2582-3930

Impact Factor: 7.185

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

Sree Chand R

in recognizaton to the publication of paper titled

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

published in IJSREM Journal on Volume 07 Issue 03 March, 2023

Editor-in-Chief
IJSREM Journal

DOI: 10.55041/IJSREM17996

ISSN: 2582-3930

Impact Factor: 7.185

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

Anupama Y K

in recognizaton to the publication of paper titled

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

published in IJSREM Journal on Volume 07 Issue 03 March, 2023

Editor-in-Chief
IJSREM Journal