National College of Ireland

Optimizing Startup Funding Predictions: Genetic Programming and Machine Learning Synergy

MSc Research Project
Data Analytics

Chandrashekar Gettam Rajgopal
Student ID: x21226075

School of Computing
National College of Ireland

Supervisor: Mayank Jain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Chandrashekar Gettam Rajgopal
Student ID:	x21226075
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Mayank Jain
Submission Due Date:	14/12/2023
Project Title:	Optimizing Startup Funding Predictions: Genetic Programming and Machine Learning Synergy
Word Count:	6980
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Chandrashekar Gettam Rajgopal
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Startup Funding Predictions: Genetic Programming and Machine Learning Synergy

Chandrashekar Gettam Rajgopal
x21226075

Abstract

Startup finance is a vital and significant area that makes a considerable contribution to innovation and economic progress in the startup finance industry. Predicting startup funding accurately is important not just for investors, entrepreneurs and policymakers but also has a significant impact on the shaping economic landscape. This project delves into the application of genetic programming to optimize various machine learning models for the prediction of total funding in startup investments and addresses the difficulty of estimating total funding which is a task made more difficult by the complexity of startup ecosystems by analyzing a comprehensive dataset on startup investments which contains various financial indicators like seed, equity, venture and funding rounds. In this study, different machine learning models like Symbolic Regression, Random Forest, Linear Regression, XGBoost and K-Nearest Neighbors (KNN) are used using Genetic programming, which yields more precise prediction of startup funding forecasts and deeper insights through distinct viewpoints and gaps that have been found. Random Forest has performed better in predicting total funding when compared to other models. This research aims not only to contribute to the academic field of financial predictive analytics but also to provide reasonable tools for stakeholders in the startup and venture capital sectors and eventually aid in more informed decision-making processes.

1 Introduction

The financing environment for startups is a vital part of the expansion of the world economy since new businesses stimulate innovation and employment creation. Industry evaluations indicate that the startup ecosystem makes a substantial contribution to both the national and global Gross Domestic Product (GDP) (Singh and Ashraf; 2020) in addition to encouraging entrepreneurship. For investors and entrepreneurs, predicting total funding becomes critical due to the erratic and high-risk nature of startup investments. Accurate financing forecasting can result in more strategic investments and well-informed decision-making, which can eventually raise company success rates.

1.1 Motivation and Background

The driving force behind this study is the increasing demand for more trustworthy and precise tools that are used in the financial industry, especially in high-stakes startup financing environments. Many conventional prediction models often do not produce accurate results because of the complexity of the startup funding ecosystem. Due to this

advanced machine learning algorithms are used to fill this gap that can handle non-linear, complex and large datasets.

The use of Genetic programming (GP) (Muni et al.; 2006) in machine learning models has become a viable method in the field of predictive analytics, providing sophisticated instruments for the analysis of intricate and dynamic financial data. Also for optimizing the models, which are inspired by the ideas of natural selection and principles of evolution that offer robust mechanisms to improve the predictive capabilities of machine learning models. The goal of this research project is to maximize the forecast of startup total investments by utilizing these advanced computational methodologies to find meaningful insights and provide factors influencing startup funding beyond traditional prediction analytics. Such advancements in predictive modeling are significantly important and also address the practical implications for entrepreneurs, investors and policymakers.

1.2 Research Question and Objectives

1.2.1 Research Question

RQ: "In the developing landscape of startup financing, what are the challenges and limitations in accurately predicting total investment funding using genetic programming with diverse machine learning models to enhance prediction accuracy and model robustness?"

1.2.2 Research Objectives

The following objectives are outlined as a solution to the above-mentioned research question regarding the prediction of total funding details using GP and machine learning models:

Objective A: Implement and use GP techniques for the optimization of machine learning models through hyperparameter tuning and model selection tailored to the features of startup investment funding data.

Objective B: Evaluate and compare the predictive performance of different machine learning models using genetic programming, including Symbolic Regression, Random Forest, Linear Regression, XGBoost and KNN in the context of optimized total funding prediction.

Objective C: Analyze and interpret the results of the optimized models, identifying key factors and features that significantly influence total funding predictions and assessing the robustness and accuracy of each model.

1.2.3 Research Project Contribution

The fundamental contribution of this research project lies in the pioneering use of GP for predicting startup funding by optimizing traditional machine learning models. The unique quality of the GP to dynamically optimize model structures and automatic feature selection is used for developing efficient predictive models. This approach navigates through the complexities of startup financing where traditional models might struggle.

1.3 Document Structure

The document flows into the Literature Review which delves into existing research and provides a foundational understanding of both traditional machine learning applications

in startup funding and the role of Genetic Programming in various domains. Following the Literature Review, the Methodology section outlines the specific approaches and techniques used in the research, including a detailed description of Genetic Programming and its application in optimizing machine learning models.

After Methodology, the Design specification section provides information on the environment, packages, libraries and tools used. Next is the Implementation section which details model setup and configuration and is followed by the Evaluation section which discusses experiments and results and the last section is the Conclusion and Future work.

2 Related Work

2.1 Introduction

In the Literature review, discussion on how the evolution of approaches from traditional to advanced Machine Learning (ML) algorithms such as Symbolic Regression, Random Forest, Linear Regression, XGBoost, and K-Nearest Neighbors in predicting startup funding has historically been challenging because of the impact of various factors in the startup ecosystem, such as market trends, investor sentiment, and current economic conditions. Also, the review explores the use of Genetic programming (GP) (Gandomi and Roke; 2015) in the financial domain for funding prediction for handling high-dimensional data for improving the funding prediction accuracy. Furthermore, explores the gaps in the existing research in this domain and discusses novel approaches by integrating diverse ML models with GP to address these gaps.

2.2 Role of Machine Learning in Predicting Startup Funding and Success

Artificial Neural Network (ANN) and modified K-Means clustering algorithms were utilized in paper Misra et al. (2023) to predict the startup's success or failure. These models showcased the future status of the startup by classifying them into categorising like Acquired, Closed, Operating and IPO. Configuration used in the ANN model includes different layers with a combination of sigmoid activations, ReLu and softmax (Marcu and Grava; 2021) in the output layer also model uses Adam optimizer to maximize the model accuracy. When a traditional and modified k-means is used on the original dataset for clustering, the accuracy was 73% and 68% was with traditional K-means. The performance of both clustering improved better after preprocessing the data. Traditional K-means achieved 74% accuracy and K-means modified model achieved approximately 77%. The dataset with modified data and ANN algorithm achieved 85% accuracy and the same dataset with traditional K-means and ANN achieved around 80% accuracy. When the dataset was preprocessed the traditional k-means and ANN achieved 84% accuracy and modified K-means and ANN accuracy stood at 89%. However, in this study there were only a few attributes were used from the Crunchbase dataset and several temporal features like state code, region, etc were removed even though these features provide valuable temporal aspects in prediction.

(Żbikowski and Antosiuk; 2021) used different approaches to the prediction of startup success by employing companies founded only during the years 1995 and 2015 and using the category under which those companies fall. Also, the author has used features like

the number of employees in an organization, the educational background of those employees, gender, etc. In this study, Logistic regression, SVM, and XGBoost algorithms are employed and the accuracy stood at 90 percent for all these ML models. Selecting only features related to employees, and educational backgrounds and ignoring the funding details for prediction of the company's success is biased as it lacks the important features that represent the company's true status.

The research paperArroyo et al. (2019) discusses the use of machine learning to help venture capitalists evaluate startup companies for investment. In this study 120,000 The author proposes a multi-class approach to reduce risk and uncertainty when investing. The authors analyze the feature importance of the multi-class approach and suggest ways to refine the approach further. The dataset used in the study consists of over 120,000 startup companies retrieved from Crunchbase which are representative setting that tries to predict company progress in the 3-year time window i.e August 2015, August 2018 and a Warmup window of 4 years August 2011. These windows in this research represent that at the time of the investment these startup companies will not be older than 4 years and expect to raise new investments in not more than 3 years after the prediction. These windows are considered adequate for a startup given the high failure rate in the early years; for example, at four years was about 44 percent in the US [1]. Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Extremely Randomized Trees (ERT) and Gradient Tree Boosting (GTB) were used in this study and DT provided 74.6 percent accuracy and the rest of the models have performed less than DT. However, the main drawback is the bias in the target variable as the status column has more companies under operating and very few are in IPO or closed. So the author has merged the funding round, acquired, and IPO as one target variable which ignores other important features. Also, despite employing a variety of models, the study does not go into great detail about how these models might be used in conjunction or how to compare them when predicting startup funding. The study did not investigate how several models work better together or even outperform one another, particularly when combined with advanced ML techniques.

2.3 Challenges in Feature Selection and Skewed Data

This study by Yin et al. (2013) discusses the feature selection techniques for the classification of imbalanced data. The approach used initially in the paper is Bayesian learning (Hautsch and Hess; 2007) on data by showing the feasibility in controlled scenarios. Later the novel approach was applied by partitioning large classes into small subclasses and generating class labels for that subclass. And Hellinger Distance-Based (Kumari and Thakar; 2017) method was used for feature selection which is a metric for measuring distribution divergence. This method is insensitive to skewness as it does not involve prior class information. But still, there is a need for automated selection of features by considering the skewness in the complex dataset which was not discussed in this study.

Similarly, this research paper (Maldonado et al.; 2014) aims to tackle class imbalance and feature selection Backward Elimination (Mao; 2004) approach, which involves progressively eliminating less significant features based on the specific contribution measure. And Balanced Loss (Wu et al.; 2022) function is used to calculate an independent subset of data, ensuring the feature selection is standardized to the needs of imbalanced datasets. The approach is tested on highly imbalanced six microarray datasets which are well

[1] `https://smallbiztrends.com/2023/07/startup-statistics.html`

suited for evaluating the effectiveness of feature selection techniques in an imbalanced classed context. However, the method did not fully capture and utilize complex interactions between features, especially in datasets where non-linear relationships exist. Some methodologies are more skillful at uncovering and leveraging these intricate relationships to improve predictive accuracy.

2.4 The Rise of Genetic Programming Over Traditional Models

A significant change has been observed with the introduction of Genetic Programming (GP) Lambora et al. (2019) which is an extension of Genetic Algorithms (GA) in the domain of finance. One of the study by Etemadi et al. (2009) have showcased the effectiveness of Genetic programming in prediction especially when traditional models are compared in the financial sector. The strength of Genetic Programming lies in the evolutionary computation approach, which allows to efficient resolution of complex classification problems through the natural selection process (Lobo and Goldberg; 1997). The GP's adaptability makes it a powerful tool for various diverse applications including bankruptcy prediction of firms listed on exchanges like Tehran Stock Exchange. Compared to Multiple Discriminant Analysis (MDA) Yap et al. (2010), the GP model's accuracy has higher accuracy rates which shows GP superiority in terms of performance. MDA has achieved 77% and 73% in training and holdout samples respectively when compared to GP which achieved higher accuracy rates of 94% and 90%. These results are strengthened by a rigorous methodology, which includes a multiple-stage variable selection process, and the number of hits in the fitness function which demonstrates GP's capability in handling large datasets and complex financial variables Sette and Boullart (2001).

2.5 Innovating Startup Funding Predictions: Integrating Genetic Programming with Advanced Feature Selection

Etemadi et al. (2009) have already showcased the efficiency of GP in areas like bankruptcy prediction while highlighting its pertinence in complex financial predictions. However, the challenge in startup funding prediction presents in dealing with high and complex dimensional symbolic regression Chen et al. (2017) which often struggles with generalization. Hence feature selection becomes critical. The paper by Viegas et al. (2018) provides an innovative method for feature selection using Genetic programming in skewed and high-dimensional datasets. The challenges of "Curse of dimensionality" and data imbalance in which traditional feature selection methods struggle are approached in this paper. The authors propose an innovative GP-based strategy that harnesses different feature selection metrics to create an effective list of features and this method showcases the resilience to data skewness and improves performance.

The approach described in the paper is relevant to the startup funding prediction. In the fast-paced and data-intensive landscape of starts, accurate and efficient feature selection is crucial for making informed decisions, especially when high dimensional data is involved. The GP-based feature selection offers a robust solution to the startup funding prediction to reduce the complexity of the data while improving the accuracy of the predictive models.

In the study by Sandin et al. (2012) offers an enhancement to the GP approach for feature selection in high-dimensional and skewed datasets. This paper highlights the use of GP for making a composite feature selection metric which combines many basic

metrics to select a more efficient set of features in the view of data skewness. The authors highlight the automatic classification in skewed data when the majority class dominates, leading to more biased feature selection. GP strategy used in the paper aggressively reduces the dimensionality while efficiently managing the skewed data, exploring common feature selection metrics like Chi-Square, Odds ratio and Information Gain by combining the results to obtain a good unbiased estimate of each feature's discriminative power. This development is particularly important for startups as it offers a more advanced approach to handling complex and high-dimensional data by efficiently selecting the most informative features and overcoming the inherent biases in skewed datasets.

2.6 Limitatons

The literature emphasizes several limitations in the field of startup funding prediction. While addressing high-dimensional and skewed data the current feature selection strategies frequently fail to address complex feature interactions. In particular, non-linear datasets indicate the necessity for more advanced and nuanced techniques. Furthermore, even while strategies like balanced loss functions and backward elimination address class imbalance were used they could not function as well in highly skewed settings crucial events occur indicating the need for more resilient approaches in these extreme circumstances. Furthermore, there is still a need for the study to fully utilize the synergies between GP and a wider variety of sophisticated ML models. This is because the integration of this synergy is yet relatively unexplored. Also, there is a need to develop GP-based feature selection specifically tailored for financial data, as current feature selection methods do not fully address the unique challenges posed by financial variables in startup datasets. This would improve the depth and relevance of predictive analyses in this industry.

3 Research Methodology

This research project adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) (Schröer et al.; 2021) framework which is a widely recognized methodology in data mining and machine learning. CRISP-DM has been successfully employed in various data mining applications Hayat Suhendar and Widyani (2023) and its structured method serves as the backbone for this study. The subsequent sections detail each phase of this methodology as applied to the project.

3.1 Business Understanding

As shown in the figure 1 business understanding is the first step of CRISP-DM. Predicting startup funding and making lucrative investment decisions are difficult and time-consuming processes. Traditional manual procedures, while important, frequently ignore small but significant elements which are time-consuming and prone to mistakes. The advent of data mining and machine learning provides a chance to include a broader range of data more systematically and accurately. The initial goal of this project is to create an ML model that will assist entrepreneurs, policymakers, venture capitalists and angel investors in making better selections. The emphasis is focused on identifying the important elements that influence startup funding and applying machine learning to provide a more clear view of these factors.

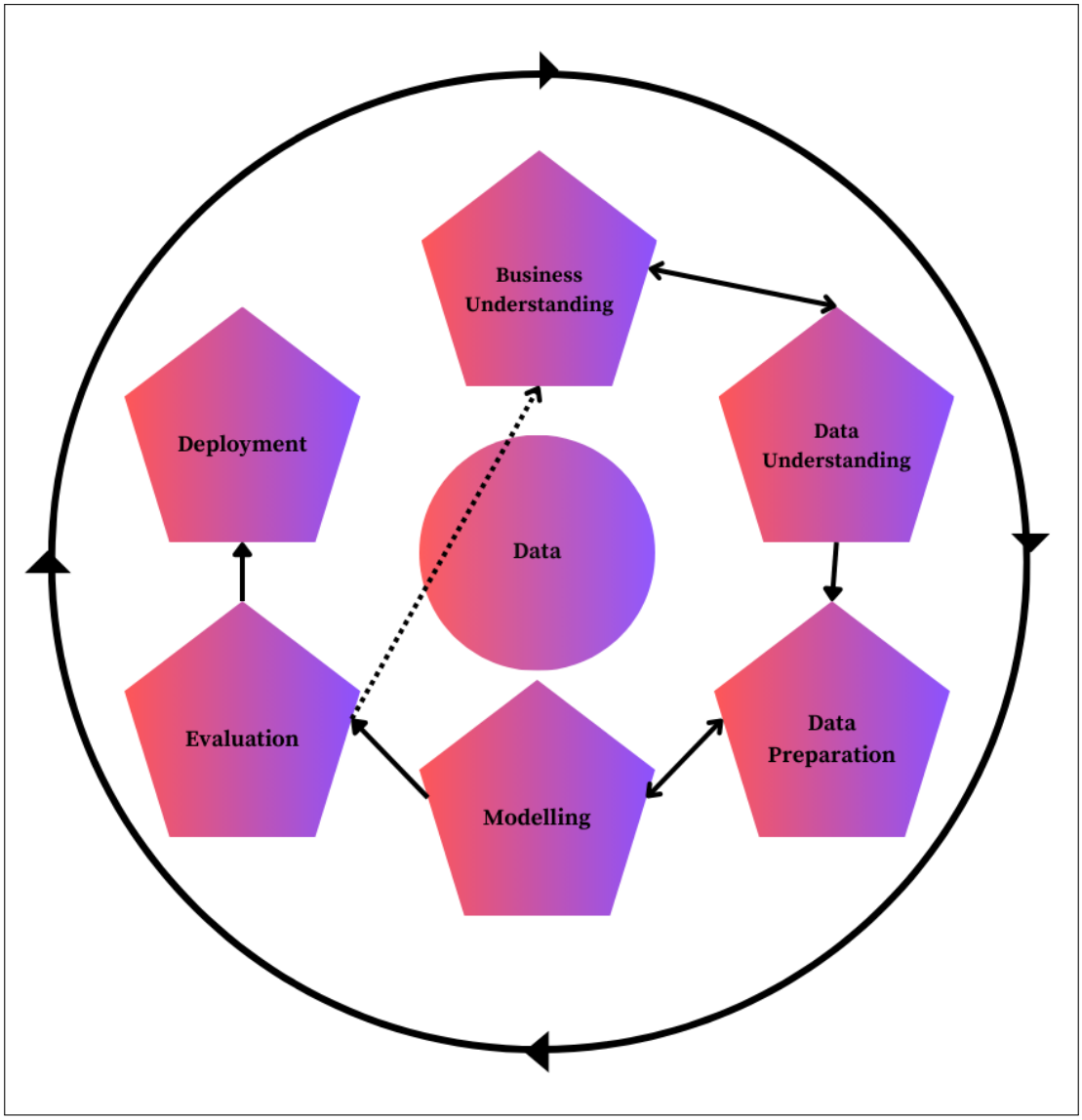

Figure 1: CRISSP-DM Hotz (2023)

3.2 Data Understanding

3.2.1 Data collection

The dataset used in this research was obtained from a comprehensive database of venture capital investments from Kaggle [2]. This dataset encompasses a wide range of information on investment patterns including company characteristics, funding details and geographical locations. The timeframe of the dataset spans from 1902 to 2014, providing a rich historical context for analyzing investment trends. The dataset includes unique identifiers, such as "permalink" and company-specific information, such as "name" "homepage_url" and "category_list". Key financial attributes, such as "funding_total_usd" and details on funding rounds ranging from initial seed to round h, enable us to examine the financial trajectories of these entities. Additional company characteristics, including "status", "country_code", "state_code" and "city" offer insights into the geographic distribution

[2] https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase

and operational status of these startups. Temporal aspects of their development are captured through "founded_at", "founded_month", "founded_quarter" and "founded_year". Moreover, "first_funding_at" and "last_funding_at" provide temporal context for funding events. This dataset serves as a valuable resource for conducting in-depth analyses and deriving meaningful insights into the dynamics of entrepreneurship and venture funding.

3.2.2 Data Exploration

Figure 2 shows the company names based on the total funding it has received through various funding rounds. From this, it is visible that Verizon Communications, Clearwire, and Charter Communications are some of the most funded companies.

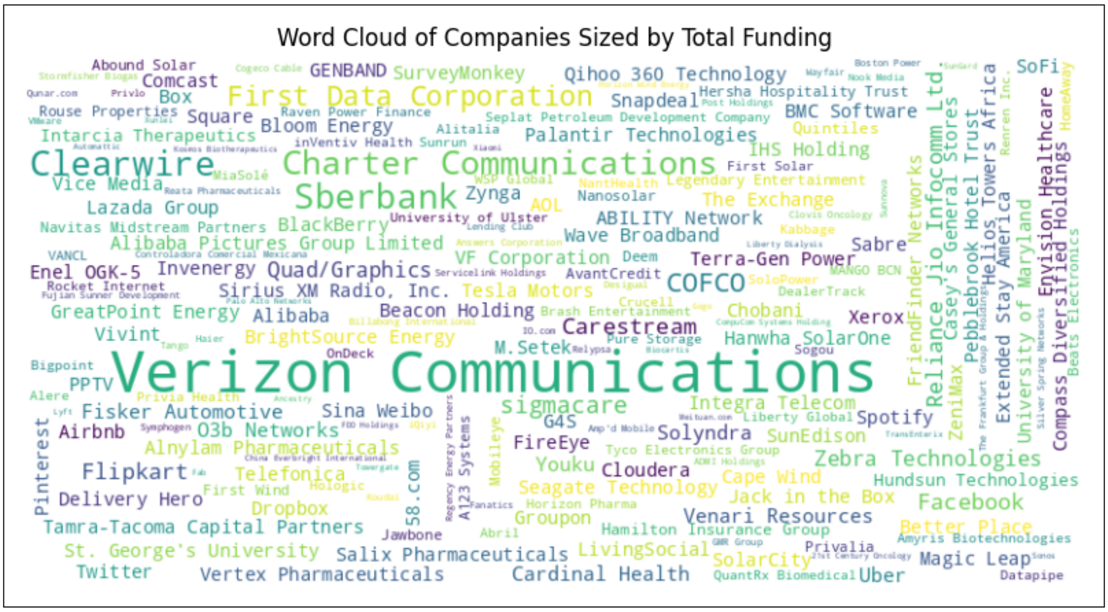

Figure 2: Word cloud of companies

Figure 3 shows the total investment over time for different market sectors. The graph displays multiple lines, each representing a different market sector, such as Commerce and Retail, Education and Training and Energy and Environment, etc. plotted over a timeline from 1900 to a point beyond 2015. The Y-axis represents the "Total Investment" on a logarithmic scale (as indicated by the "1e10" notation), which allows for a wide range of values to be displayed clearly. This is useful for visualizing data that has large variations in magnitude. There is a significant increase in funding for the Technology and Software market sectors as the first personal computer (PC) was introduced in the 1980s (Haddon; 1988).

Figure 4 shows the spread of funding total bins in USD, which suggests that investments of all sizes exponentially increased from the 1980s. This might be due to the data in Crunchbase where many companies' information might be available before the 1980s.

3.3 Data Preprocessing

Ensuring the datasets' quality and significance in data selection is one of the big challenges. The data is thoroughly preprocessed including normalization, cleaning and transformation to make it suitable for the machine learning models. Special attention was

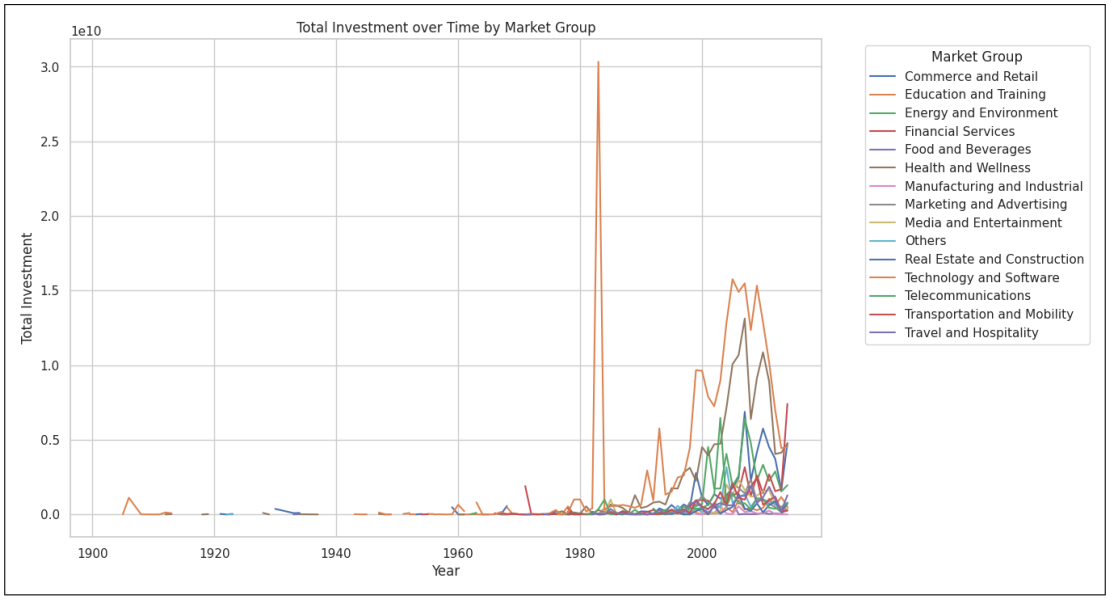

Figure 3: Total Investment over Time by Market Sector

given to feature selection and engineering to make sure the models have access to the most predictive and related information due to given diversity and complexity of the data.

A series of preprocessing steps were taken to ensure the data quality and relevance for analysis due to the complexity and volume of data. After the dataset was loaded, a series of detailed cleaning and transformation processes were applied to ensure suitability for advanced machine learning analyses in the data preprocessing phase of the research. Firstly, the dataset was loaded with Latin encoding to adapt to special characters which were followed by an important step of cleaning and transforming the total funding total in the USD column which would be the target variable and this column is in non-numeric type so it was converted to numeric type by removing all non-numeric characters along with that missing values was also addressed appropriately. The dataset was refined further by removing irrelevant columns like 'permalink' and 'name' which are not important for the prediction of total funding total in USD and also to focus more on impactful features. Features with temporal aspects like the first funding date and last funding date were converted to date time objects which will be helpful for temporal analysis.

Missing values in important columns like state code and city were intelligently imputed using mappings from related columns and external mapping file [3] while ensuring data integrity and consistency. Also, an important step in preprocessing was categorizing the 'market' column into logical and broader groups which in turn helps in analyzing the trends in different market sectors. The final step in data preprocessing included removing rows with missing values in key columns and removing duplicate and null value rows. This phase helps in improving and maintaining the data quality which will be subsequently used in machine learning models to predict startup total funding accurately.

[3] `https://github.com/dr5hn/countries-states-cities-database`

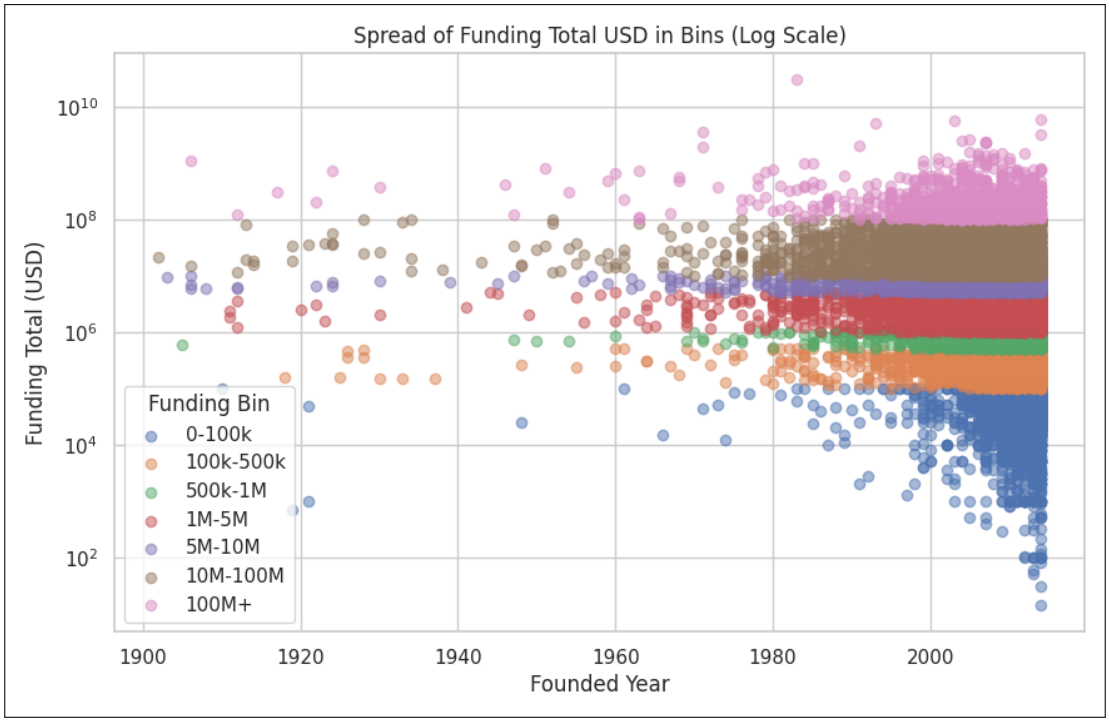

Figure 4: Funding distribution

3.4 Modelling

In this research, the focus is on evaluating prominent ML algorithms using a GP approach to select the features important for prediction and to optimize these models for enhanced performance. Random Forest and KNN are established algorithms used in many previous works on Crunchbase data Pan et al. (2018).

3.4.1 Symbolic Regression

Symbolic regression Mousavi Astarabadi and Ebadzadeh (2019) using GP is a useful instrument in machine learning for addressing complex regression tasks especially effective in modeling nonlinear and complex relationships in data. GP is an evolutionary algorithm in which models are represented as a tree-like mathematical expression that allows them to adapt and discover underlying data patterns without predefined model constraints. Also, its ability to produce interpretable models makes it an important asset in fields requiring in-depth data analysis like financial forecasting. Figure 5 shows how GP constructs the structure [4].

3.4.2 Random Forest

Random Forest (Li; 2021) is an ensemble method consisting of numerous individual decision trees. Each tree in the forest outputs a class prediction, and the final output of the Random Forest model is the class that receives the majority of votes from individual trees. This method typically yields high accuracy and is robust against overfitting.

[4] `https://astroautomata.com/paper/symbolic-neural-nets/`

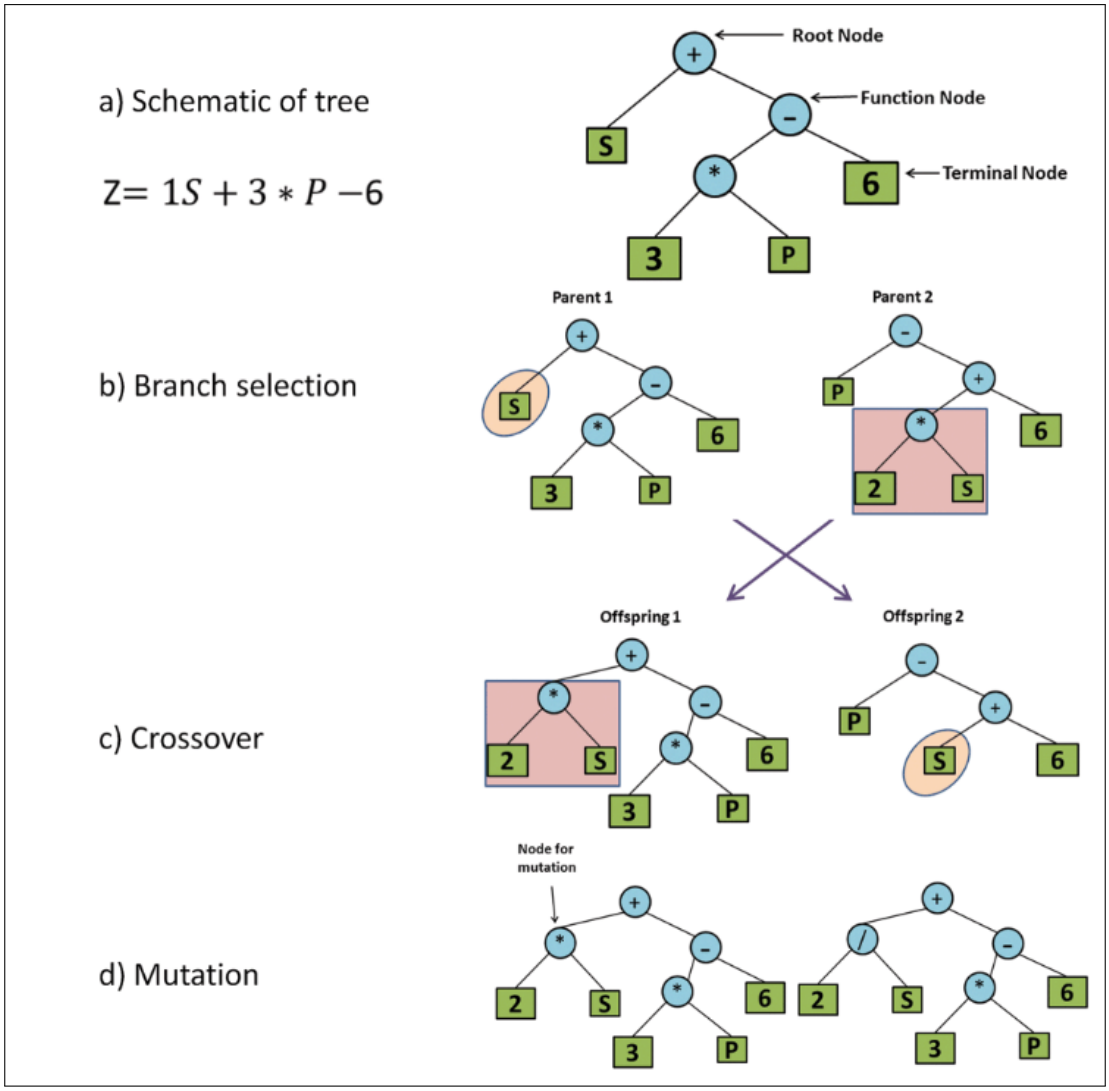

Figure 5: Genetic Programming Vyas et al. (2018)

3.4.3 Linear Regression

Linear Regression (Chiou et al.; 2016) is an essential statistical technique in predictive modeling and machine learning that is used for understanding and predicting relationships between variables. It is important for scenarios where the relationship between the independent variables (predictors) and the dependent variable (outcome) is linear. Relationship in Linear Regression is exhibited through a linear equation where each predictor has a related coefficient that represents its relationship with the outcome variable.

3.4.4 XGBoost

XGBoost regression (Wang et al.; 2022) is an application of the XGBoost algorithm which is a powerful tool for startups due to its efficient nature in handling complex and large datasets. Due to this, it is ideal for regression tasks with complex variable relationships. XGBoost's efficiency and scalability align well with the frequently resource-constrained nature of startups. Also, its robustness against overfitting and feature importance analysis are vital for startups enabling them to make data-driven decisions with greater confidence

and precision and to concentrate their efforts on the most impactful aspects of their business.

3.4.5 KNN

The KNeighbors regression Yu et al. (2008) leverages the KNN approach and it is mainly valuable for startups due to its flexibility, effectiveness and simplicity in handling non-linear data. The values of 'k' closest neighbors are used to make predictions making it easy to implement. Its key strength lies in its capability to work without making assumptions about data distribution which is crucial for startups dealing with diverse and complex datasets. Also, it can be fine-tuned by adjusting the number of neighbors with computational intensity with large datasets.

3.4.6 Genetic Algorithm for Model Optimization

A Genetic algorithm is used to enhance the performance further for these models. The evolutionary algorithm will optimize the hyperparameters for each model by simulating the process of natural selection. By To further enhance the performance of these models, a genetic algorithm is employed. This evolutionary algorithm optimizes the hyperparameters of each model by simulating the process of natural selection. It selects the best-performing models iteratively, combines their features and alters them to search the hyperparameter space efficiently. This approach aims to find the ideal set of hyperparameters that produce the best predictive performance.

3.5 Evaluation Metrics

The evaluation of machine learning models in this research is tackled in widely established metrics within the domain Majumder et al. (2022), focusing on Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R^2). These metrics provide a complete understanding of the models' performance.

3.5.1 MSE

MSE (Varner et al.; 2022) is a common metric used to measure the average of the squares of the errors that is the difference between the estimator and what is estimated. It quantifies the average squared difference between the estimated values and the actual value. MSE is used in regression analysis to verify the efficiency of the estimator. It's given by the formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \tag{1}$$

where
Y_i = actual value of the i-th observation,
$\hat{Y_i}$ = predicted value of the i-th observation,
n = number of observations.

3.5.2 MAE

MAE (Hao and Li; 2020) is another regression metric that measures the average magnitude of errors in a set of predictions without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. It's expressed as below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{2}$$

MAE provides a simple measure of prediction accuracy with a lower MAE indicating better model performance.

3.5.3 R^2

R^2 (Colin Cameron and Windmeijer; 1997) which is also known as the coefficient of determination is a statistical measure in a regression model that decides the proportion of variance in the dependent variable that can be explained by the independent variables. It indicates the goodness of fit of a set of predictions to the actual values. In easy terms, it tells how close the data points are to the fitted regression line. The formula for R-squared is:

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}} \tag{3}$$

R^2 values range from 0 to 1 with higher values indicating a better fit between the model and the data.

4 Design Specification

4.1 Environment Setup

Python Packages: Renowned for its versatility and wide-ranging applicability in data science was the primary programming language used. The Python ecosystem is rich with libraries that cater to various aspects of data science, such as data manipulation, modeling, and machine learning. Key libraries employed in this project included:

4.1.1 Packages and Libraries

Library	**Purpose**
Pandas	Efficient data manipulation and analysis.
NumPy	Numerical computations and array-based data handling.
Matplotlib and Seaborn	Data visualization.
Scikit-learn	Implementing various machine learning algorithms.
Plotly	Interactive, publication-quality graphs.
DEAP	Evolutionary algorithms and genetic programming.

Table 1: Packages and Libraries

Additional libraries like SciPy for scientific computing and scikit-learn extensions for advanced machine learning functionalities.

4.1.2 Collaborative and Cloud Tools

- Google Colab: Leveraging cloud computing, Google Colab was used for executing more resource-intensive tasks. It provided an easy-to-use interface and access to powerful computational resources, including GPU acceleration, which is crucial for complex data processing and machine learning tasks.
- Google Drive: For data storage and sharing, Google Drive was utilized. It ensured secure and convenient access to datasets and project files, enabling seamless collaboration and data management.

5 Implementation

5.1 Symbolic Regression Model

The Symbolic regression configuration is based on ideas that aim to strike a compromise between overfitting prevention, computational efficiency, and solution space exploration [5]. To enable a quick but thorough investigation of the solution space, a relatively small population size of 30 is initially selected. Then it is increased to 100 in future runs to amplify the search and possibly find more specialized solutions. Although there is a chance of overfitting as generations go by, the number of generations is initially set at 20 and then increased to 50, providing the evolving solutions more chances to get better over time. The stopping criterion is aggressively placed at a low threshold of 0.01 to end the evolution early if a superior solution is found thereby saving computational resources.

Symbolic regression has a set of parameters like below to configure [6].

- Population Size: Determines how many individual candidate solutions can exist in each generation.
- Number of Generations: Number of iterations over which the population can evolve. More generations give the population more prospects to evolve towards better solutions. However, too many generations can lead to overfitting if the model becomes too complex.
- Stopping Criteria: Determines the threshold for when the algorithm should terminate early if a certain level of fitness is achieved.
- Crossover Probability: The probability of two programs "mating" to produce offspring, which is a primary mechanism of genetic algorithms to combine and propagate successful traits.
- Subtree Mutation Probability: Chance that a randomly selected part of a program is replaced with a new randomly generated subtree.
- Hoist Mutation Probability: Mutation where a randomly chosen subtree is "hoisted" to replace its parent tree, effectively simplifying the program.
- Point Mutation Probability: Probability of randomly altering parts of a program.

[5] `https://deap.readthedocs.io/en/master/examples/gp_symbreg.html`
[6] `https://gplearn.readthedocs.io/en/stable/reference.html`

- Maximum Samples: Fraction of samples used to evaluate each candidate.
- Parsimony Coefficient (Burlacu et al.; 2019): Penalty factor that discourages overly complex models to help prevent overfitting.

Promoting the emergence of robust offspring when a high crossover probability (0.7) guarantees a rich trait exchange between solutions. The subtree, hoist, and point mutations have mutation probabilities of 0.1, 0.05, and 0.1, respectively, which are chosen to promote variation and avoid an early convergence to suboptimal solutions. The algorithm can explore new regions of the solution landscape and break out of local optima by permitting mutations. While allowing complexity when it greatly improves model performance, the parsimony coefficient is kept low (0.01) to gently penalize complexity and therefore lean towards simpler models that are less likely to overfit. When the maximum samples parameter is set to 0.9, it indicates that 90% of the data were used to train the model. This robust sample size aids in generalization and permits implicit validation using the remaining data. Using a fixed random state guarantees repeatability and consistency in the outcomes across runs. By carefully weighing the trade-offs included in genetic programming, these parameter selections help to build a model that is both precise and broadly applicable.

After trying with different population and generation sizes as mentioned above in this research a population size of 30 and generation of 20 gave a better fitness function score. Due to this, all other models are configured with these values.

5.2 Linear Regression Model using GP for Feature selection

For the Linear regression model, minimal hyperparameter tuning was required, as the primary focus was on feature selection. The model inherently assumes a linear relationship between input variables and the target. In our implementation, after the genetic programming-based feature selection, Linear regression was used without modifying the default parameters. The default configuration includes ordinary least squares regression, which minimizes the sum of squared differences between observed and predicted values. This choice was made because of the model's nature as it is not prone to overfitting with a large number of features, and its simplicity provides a transparent baseline for comparing with more complex models.

5.2.1 Genetic Programming Configuration

- Feature Representation: Each individual in the genetic algorithm population represented a different combination of features, encoded as a binary string. Each bit in the string corresponds to the presence (1) or absence (0) of a feature.
- Fitness Function: The fitness of each individual was evaluated based on the performance of the Linear regression model, using the selected features. The fitness metric was the mean squared error (MSE) obtained from cross-validation of the training data. Lower MSE values indicated better fitness.
- Population Size: Set at 30, to maintain diversity in feature combinations.

- Crossover and Mutation: Standard two-point crossover and flip-bit mutation [7] were used, with probabilities of 0.7 and 0.2, respectively. These operators facilitated the exploration and exploitation of the feature space.
- Selection: Tournament selection was employed to choose the best-performing individuals for the next generation.
- Outcome: The best features that significantly contributed to the predictive power of the model while excluding redundant or irrelevant features.

5.3 K-Nearest Neighbors (KNN) Regression Model

For the KNN regression, the key hyperparameter was the number of neighbors. Typically, values ranging from 3 to 10 are a good starting point, as smaller values can lead to high variance, and larger values might smooth out the predictions too much. However, through GP, it was identified that a value of 5 clusters provided the best balance between bias and variance for our specific dataset. The weights parameter was set to distance which assigns greater weight to the nearest neighbors for more localized predictions. The distance metric used was the default Minkowski (Maruf and Laksito; 2020) distance. This setup was chosen because it tends to perform well in scenarios where the relationship between variables is complex and not necessarily linear.

5.3.1 Genetic Programming Configuration

- Parameter Encoding: Individuals in the population represented different values of clusters, encoded as integers.
- Fitness Function, Selection, Population Size and Genetic Operators: Similar to those used for Linear regression.
- Outcome: Through this process, the genetic algorithm identified an optimal cluster value that minimized the MSE, enhancing the KNN model's accuracy. This optimized cluster value reflected the best trade-off between underfitting and overfitting for the given dataset.

5.4 Random Forest Regression Model

For the Random forest regression model hyperparameters like the number of trees in the forest and the number of splits that each decision tree were searched with different combinations through multiple generations. Through GP multiple generations six trees were found to be a suitable parameter between computational efficiency and model performance. Typically, more trees in the forest lead to better performance but at the cost of increased computation. Deeper trees can model more complex relationships but also increase the risk of fitting to noise in the training data. Additionally, parameters like maximum number of features were set to auto, allowing each tree to consider a subset of features at each split, thereby increasing diversity among the trees and contributing to the robustness of the model.

[7] `https://deap.readthedocs.io/en/master/api/tools.html`

5.4.1 Genetic Programming Configuration

- Parameter Encoding: Each individual represented a set of hyperparameters for the Random forest model. The parameters were encoded as integers and floats, corresponding to the number of trees and the maximum depth.
- Fitness Function, Selection, Population Size and Genetic Operators: Similar to those used in previous models.
- Crossover and Mutation: A blend of crossover and mutation specific to integer and float representations was used. These operators were tailored to explore the hyperparameter space effectively.

5.5 XGBoost Model

XGBoost model (Sagi and Rokach; 2021) which is a powerful and widely used ML algorithm was used as the next model. The learning rate parameters in the range 0.01, 0.05, 0.1, 0.2 and 0.3 were chosen for controlling the speed of the model's learning. These values provide a good spread from slower, more precise updates to faster, more aggressive ones. Several estimators were chosen from 50 to 300 which dictates the number of boosting stages the model will undergo. A higher number of estimators will lead to better performance also it increases the possibility of overfitting (Pan et al.; 2022). And maximum depth that controls the depth of each tree was set to vary between 3 to 10.

5.5.1 Genetic Algorithm Configuration

- Subsample: This parameter, with values 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0, specifies the fraction of samples to be used for fitting individual base learners. It helps in reducing overfitting.
- Crossover Probability: A probability of 0.7 for crossover was chosen, encouraging a healthy mix of genetic material from different individuals.
- Parallel Processing: To enhance computational efficiency, a multiprocessing pool was initialized, allowing parallel evaluation of individuals. This approach significantly reduced the overall computation time, crucial for handling the computationally expensive task of training and evaluating XGBoost models.
- Outcome: Upon completion of the genetic algorithm, the best individual, representing the optimal combination of hyperparameters for the XGBoost model was identified. The fitness of this individual was indicated by the lowest negative MSE which reflected its superiority in terms of prediction accuracy on the training data. This set of hyperparameters was then considered the optimal configuration for the XGBoost model in the context of this study.

6 Evaluation

All algorithms were run through a set number of generations, continually improving the fitness of the population with parameters specified in section 5 for each model. For Linear regression and KNN the best individual in each generation represented the selected

feature subset. The feature selection process successfully identified a subset of features that maximized the predictive performance. The selected features were used to train and evaluate the model. After the optimal feature subset was decided, a linear regression model was trained using this reduced set of features.

Table 2: Regression Model Performance Metrics

Model	**MSE**	**MAE**	**R^2**	**Median AE**
Symbolic Regression	9760.18	76.36	-0.0202	58.22
Linear Regression	9278.68	74.96	0.0302	59.61
Random Forest	5347.05	56.93	0.4411	45.83
XGBoost	9568.82	77.67	-0.0002	62.96
KNN	5615.69	53.84	0.4130	36.30

The model was evaluated using various regression metrics to assess its predictive accuracy and generalization ability.

For Symbolic regression, Random Forest and XGBoost various hyperparameters were found similarly by running the same set of generation and population as used for Linear regression and KNN. Using these hyperparameters all three models were trained on the training dataset and the prediction was performed on the test dataset split. The performance of these models is represented using Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2) and Median Absolute Error.

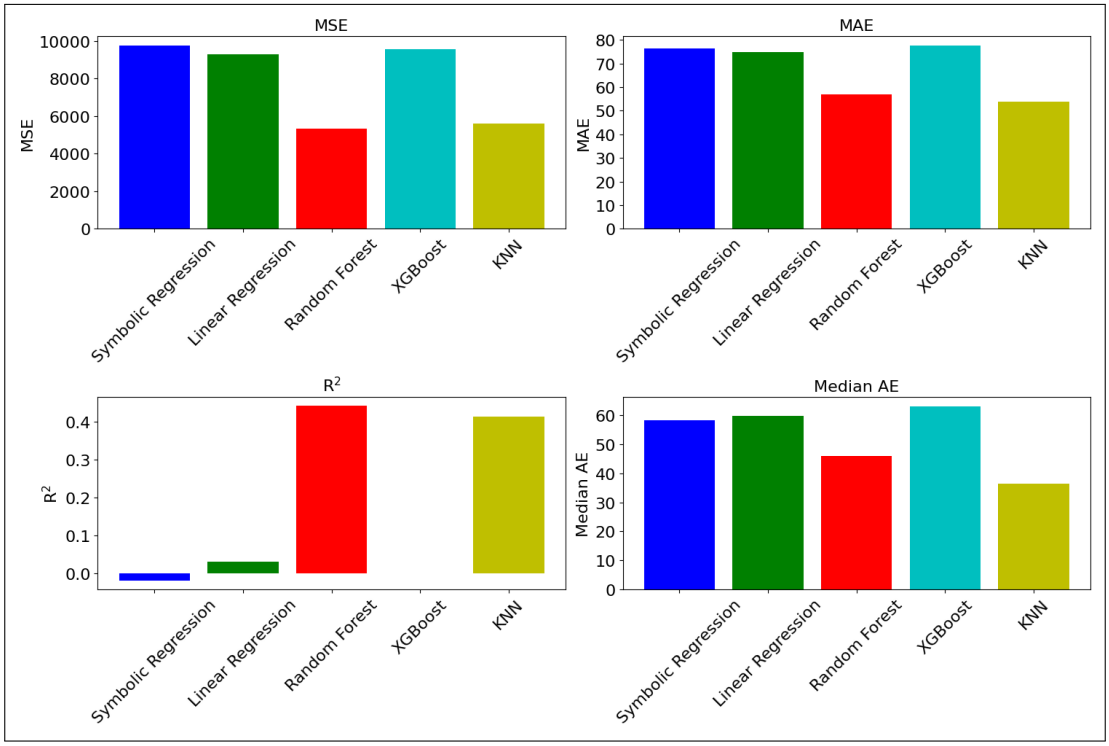

Figure 6: Performance Evaluation of Machine Learning Models Across Key Metrics

These metrics offer a thorough assessment of the regression models' effectiveness with the test set of data and results as displayed in table 2. From figure 6 promising findings are displayed by the Random Forest and KNN models, which demonstrate superior prediction

accuracy and a stronger fit to the data with lower MSE and MAE values and relatively higher R-squared values. The Symbolic regression and XGBoost models, on the other hand, show lower R-squared values and higher MAE and MSE values, indicating that they might not perform as well in this particular situation.

The integration of GP in feature selection for startup funding prediction offers a robust method for identifying key predictors. However, the true potential for improvement is present in incorporating finance domain factors like macroeconomic indicators and market sentiment. For business applications, the balance between interpretability and model complexity is crucial but simple models could improve transparency without compromising significantly on predictive strength. The research when compared to existing literature emphasizes that the field is still evolving by techniques like GP, Random forest and KNN. These improvements suggest a promising trend for future research, a balanced approach to model complexity and the importance of domain-specific tailoring.

7 Conclusion and Future Work

The goal of this study was to tackle the difficult task of accurately estimating the overall amount of investment financing. With the ultimate goal of improving funding prediction accuracy, model resilience, and model robustness, the challenges and constraints related to this study were examined through the integration of genetic programming for feature and hyperparameter selection with different machine learning models. Experiments with several regression models, including Symbolic regression, Linear regression, Random forest, XGBoost and KNN research have effectively uncovered a variety of prediction skills. Random forest and KNN performed well, with lower MSE and MAE and higher R^2 values indicating better predicted accuracy and a better fit to the data. These models show the potential for accurate funding estimates in the context of startup finance. Applying symbolic and XGBoost regression to this specific prediction challenge yielded less desirable results. The research highlights the significance of model selection and hyperparameter tweaking in achieving optimal prediction performance.

Future research in this field should concentrate more on the integration of advanced machine-learning techniques with domain-specific knowledge to enhance startup finance estimates. A greater emphasis on feature engineering and the integration of a larger range of factors, including market sentiment research, industry-specific information, and macroeconomic data, can improve prediction accuracy. Further research into ensemble learning strategies that leverage the complementary strengths of many models could further enhance robustness and reliability. Furthermore, because startup finance dynamics are inherently temporal, time series analysis and real-time data integration approaches will facilitate adaptive forecasts and improve the model's responsiveness to shifting market conditions. Fairness evaluations in prediction algorithms and other ethical considerations should be given high importance when making decisions in the complex world of startup finance.

References

Arroyo, J., Corea, F., Jimenez-Diaz, G. and Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments, *IEEE Access* **7**: 124233–124243.

Burlacu, B., Kronberger, G., Kommenda, M. and Affenzeller, M. (2019). Parsimony measures in multi-objective genetic programming for symbolic regression, *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '19, Association for Computing Machinery, New York, NY, USA, p. 338–339.

Chen, Q., Zhang, M. and Xue, B. (2017). Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression, *IEEE Transactions on Evolutionary Computation* **21**(5): 792–806.

Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. (2016). Multivariate functional linear regression and prediction, *Journal of Multivariate Analysis* **146**: 301–312. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Colin Cameron, A. and Windmeijer, F. A. (1997). An r-squared measure of goodness of fit for some common nonlinear regression models, *Journal of Econometrics* **77**(2): 329–342.

Etemadi, H., Anvary Rostamy, A. A. and Dehkordi, H. F. (2009). A genetic programming model for bankruptcy prediction: Empirical evidence from iran, *Expert Systems with Applications* **36**(2, Part 2): 3199–3207.

Gandomi, A. H. and Roke, D. A. (2015). Assessment of artificial neural network and genetic programming as predictive tools, *Advances in Engineering Software* **88**: 63–72.

Haddon, L. (1988). The home computer: The making of a consumer electronic, *Science as Culture* **1**(2): 7–51.

Hao, S. and Li, S. (2020). A weighted mean absolute error metric for image quality assessment, *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 330–333.

Hautsch, N. and Hess, D. (2007). Bayesian learning in financial markets: Testing for the relevance of information precision in price discovery, *Journal of Financial and Quantitative Analysis* **42**(1): 189–208.

Hayat Suhendar, M. T. and Widyani, Y. (2023). Machine learning application development guidelines using crisp-dm and scrum concept, *2023 IEEE International Conference on Data and Software Engineering (ICoDSE)*, pp. 168–173.

Hotz, N. (2023). What is CRISP DM? - Data Science Process Alliance — datascience-pm.com, `https://www.datascience-pm.com/crisp-dm-2/`. [Accessed 26-01-2024].

Kumari, A. and Thakar, U. (2017). Hellinger distance based oversampling method to solve multi-class imbalance problem, *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 137–141.

Lambora, A., Gupta, K. and Chopra, K. (2019). Genetic algorithm-a literature review, *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, IEEE, pp. 380–384.

Li, J. (2021). Prediction of the success of startup companies based on support vector machine and random forest, *2020 2nd International Workshop on Artificial Intelligence and Education*, WAIE 2020, Association for Computing Machinery, New York, NY, USA, p. 5–11.

Lobo, F. and Goldberg, D. (1997). Decision making in a hybrid genetic algorithm, *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC '97)*, pp. 121–125.

Majumder, A., Rahman, M. M., Biswas, A. A., Zulfiker, M. S. and Basak, S. (2022). Stock market prediction: A time series analysis , *in* A. K. Somani, A. Mundra, R. Doss and S. Bhattacharya (eds), *Smart Systems: Innovations in Computing*, Springer Singapore, Singapore, pp. 389–401.

Maldonado, S., Weber, R. and Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines, *Information Sciences* **286**: 228–246.

Mao, K. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **34**(1): 629–634.

Marcu, D. C. and Grava, C. (2021). The impact of activation functions on training and performance of a deep neural network, *2021 16th International Conference on Engineering of Modern Electric Systems (EMES)*, pp. 1–4.

Maruf, Z. R. and Laksito, A. D. (2020). The comparison of distance measurement for optimizing knn collaborative filtering recommender system, *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 89–93.

Misra, A. K., Jat, D. S. and Mishra, D. K. (2023). Startup success and failure prediction algorithm using k-means clustering and artificial neural network, *2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, pp. 190–195.

Mousavi Astarabadi, S. S. and Ebadzadeh, M. M. (2019). Genetic programming performance prediction and its application for symbolic regression problems, *Information Sciences* **502**: 418–433.

Muni, D., Pal, N. and Das, J. (2006). Genetic programming for simultaneous feature selection and classifier design, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **36**(1): 106–117.

Pan, C., Gao, Y. and Luo, Y. (2018). Machine learning prediction of companies' business success, *CS229: Machine Learning, Fall 2018, Stanford University, CA* .

Pan, S., Zheng, Z., Guo, Z. and Luo, H. (2022). An optimized xgboost method for predicting reservoir porosity using petrophysical logs, *Journal of Petroleum Science and Engineering* **208**: 109520.

Sagi, O. and Rokach, L. (2021). Approximating xgboost with an interpretable decision tree, *Information Sciences* **572**: 522–542.

Sandin, I., Andrade, G., Viegas, F., Madeira, D., Rocha, L., Salles, T. and Gonçalves, M. (2012). Aggressive and effective feature selection using genetic programming, *2012 IEEE Congress on Evolutionary Computation*, pp. 1–8.

Schröer, C., Kruse, F. and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model, *Procedia Computer Science* **181**: 526–534.

Sette, S. and Boullart, L. (2001). Genetic programming: principles and applications, *Engineering Applications of Artificial Intelligence* **14**(6): 727–736.

Singh, A. K. and Ashraf, S. N. (2020). Association of entrepreneurship ecosystem with economic growth in selected countries: An empirical exploration, *Journal of Entrepreneurship, Business and Economics* **8**(2): 36–92.

Varner, M. A., Mitchell, F., Wang, J., Webb, K. and Durgin, G. D. (2022). Enhanced rf modeling accuracy using simple minimum mean-squared error correction factors, *2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pp. 1–5.

Viegas, F., Rocha, L., Gonçalves, M., Mourão, F., Sá, G., Salles, T., Andrade, G. and Sandin, I. (2018). A genetic programming approach for feature selection in highly dimensional skewed data, *Neurocomputing* **273**: 554–569.

Vyas, R., Bapat, S., Goel, P., Karthikeyan, M., Tambe, S. S. and Kulkarni, B. D. (2018). Application of genetic programming (gp) formalism for building disease predictive models from protein-protein interactions (ppi) data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15**: 27–37.

Wang, L., Zhang, T.-Z., Chen, Y., Huang, Y., Yin, X., Liu, X. F. and Hu, D. (2022). Machine learning-based start-up company lifespan prediction: the chinese market as an example, *2022 6th International Conference on Universal Village (UV)*, pp. 1–7.

Wu, S., Yang, J., Wang, X. and Li, X. (2022). Iou-balanced loss functions for single-stage object detection, *Pattern Recognition Letters* **156**: 96–103.

Yap, B. C.-F., Yong, D. G.-F. and Poon, W.-C. (2010). How well do financial ratios and multiple discriminant analysis predict company failures in malaysia, *International Research Journal of Finance and Economics* **54**(13): 166–175.

Yin, L., Ge, Y., Xiao, K., Wang, X. and Quan, X. (2013). Feature selection for high-dimensional imbalanced data, *Neurocomputing* **105**: 3–11. Learning for Scalable Multimedia Representation.

Yu, Q., Sorjamaa, A., Miche, Y., Lendasse, A., Séverin, E., Guillen, A. and Mateo, F. (2008). Optimal pruned k-nearest neighbors: Op-knn application to financial modeling, *2008 Eighth International Conference on Hybrid Intelligent Systems*, pp. 764–769.

Żbikowski, K. and Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using crunchbase data, *Information Processing Management* **58**(4): 102555.