National College of Ireland

Exploring Startup Registration Trends in India: A Comprehensive Analysis and Regression Modeling with Scikit-Learn

MSc Research Project
Data Analytics

Pratik Shete
Student ID:x21229091

School of Computing
National College of Ireland

Supervisor: Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Pratik Shete
Student ID:	x21229091
Programme:	Data Analytics
Year:	2023-24
Module:	MSc Research Project
Supervisor:	Abdul Shahid
Submission Due Date:	25/04/2024
Project Title:	Startup Registration Trends in India: A Comprehensive Analysis and Regression Modeling with Scikit-Learn
Word Count:	8522
Page Count:	21

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:	Pratik Shete
Date:	27th May 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):	

Exploring Startup Registration Trends in India: A Comprehensive Analysis and Regression Modeling with Scikit-Learn

Pratik Shete
x21229091

Abstract

This in-depth analysis uses large registration data to explore the dynamic and constantly changing startup environment in India. By using rigorous data pretreatment, perceptive exploratory analysis, and sophisticated predictive modeling, we reveal a wealth of information about sector diversification, geographic dispersion, and the significant influence of foreign investment. We can observe that our models work exceptionally well: the Random Forest Regressor has an R-squared score of 0.96, the Gradient Boosting Regressor has an amazing score of 0.94, and the Bagging Regressor improves predictive accuracy considerably. Our research highlights the entrepreneurial energy fostering innovation and economic expansion in a variety of industries and geographical areas. Our report equips investors, entrepreneurs, and legislators with the necessary knowledge to see patterns, take advantage of chances, and confidently negotiate the intricacies of the startup environment.

1 Introduction

In India, a nation distinguished by its enormous diversity and organizational framework, it is imperative to recognize the existence of both states and union territories (UTs). Taken together, these organizations constitute the country's administrative structure, with each one making a distinct contribution to the socioeconomic environment. Because of this, it is essential to take into account both states and UTs while examining different facets of the Indian ecosystem, such as the startup scene.

To find trends and patterns that offer important insights into the Indian startup ecosystem, we thoroughly examine startup registration data in this research. Our goal is to provide light on many facets of this dynamic environment by employing an extensive dataset that includes data on business registrations. In India, startups play a crucial role in the country's economic development by generating jobs, advancing technology, and fostering general economic growth. By deciphering the subtleties of company registration data, we can take the pulse of the entrepreneurial environment and spot new trends and business prospects.

Government programs such as Made in India, Digital India, Startup India, and the Atal Creativity Mission help to facilitate the growth of India's startup scene. In keeping with these initiatives, this study examines the Indian startup scene to spot trends, comprehend regional distributions, appraise diversity, and gauge the influence of international investment. By offering insights, it hopes to help India's objectives of economic growth and

innovation through investment strategies, policy, and entrepreneurial endeavours. Their critical importance in social progress and economic revival is highlighted by the COVID-19 pandemic and the subsequent global upsurge in startup activity.(Singh et al.; 2021) The high failure rate requires efficient evaluation techniques, despite their potential. The goal of this article is to provide investors and stakeholders with strategic decision-making insights by utilizing machine learning models to forecast startup success with a particular focus on M and A outcomes.

1.1 Research Question:

1. Regional Influence on the Accuracy of Authorized Capital Predictions: What geographical differences exist in the machine learning models used to predict authorized capital's accuracy between different states? How do the observed variations in prediction performance arise from different factors?
2. Predictive Modeling for Foreign Subsidiary Authorized Capital Estimation:
When considering categorical and numerical parameters like state, category, class, company type, and activity description, can machine learning models estimate the authorized capital of overseas subsidiaries with any degree of accuracy?

1.2 objective

This study investigates startup registrations in different states and Union Territories (UTs) to give a thorough analysis of the Indian startup ecosystem. The study intends to map the geographic distribution of startups, identify important industries and emerging trends, and evaluate the diversity of startups about their concentration on different industries, legal structures, and business models. It also looks at the existence and functioning of foreign subsidiaries to assess how foreign investment affects Indian startups. In the end, the study aims to give interested parties data-driven insights and suggestions to direct investment plans, policy formulation, and company growth, ultimately promoting innovation, growth, and economic development in the nation.

1.3 Structure Of Document

The document is divided into various important sections: A contextual evaluation of previous research is given in Related Work (1). Methodology (2) describes data gathering and analysis procedures and their acknowledged limitations. Technical requirements, design criteria, and project goals are outlined in the Design Specification (3.9). The fourth implementation section explains the tools used, the problems faced, and the solutions put into practice. Conclusion and next Work (6) summarizes findings, addresses consequences, and suggests next study options. Evaluation (5) evaluates solution performance against preset metrics. This methodical approach guarantees a thorough review of the design, implementation, methodology, assessment, and future directions of the study.

2 Related Work

This review summarizes the body of literature on machine learning-based authorized capital prediction in finance. It explores a variety of techniques, such as ensemble methods,

neural networks, and regression. There is a discussion of evaluation criteria and data pretreatment approaches, with an emphasis on issues like model interpretability and data scarcity. The use of alternate data sources and the investigation of sophisticated machine learning algorithms are future directions. Financial prediction models for well-informed decision-making will progress if these issues are resolved.

2.1 Using Supervised Learning Regression to Forecast Indian Startup Growth

Investors and politicians need to know exact figures because of the explosion of Indian startups. Using supervised learning regression models, (Pandya et al.; 2023) predict the overall number of Indian startups by utilizing many data sources, such as startup databases and scholarly articles. Their study attempts to identify those critical elements—such as competition, consumer demand, regulatory environment, and capital availability—that impact the growth of startups. The report offers insightful information on the Indian market and startup growth potential by creating linear models based on historical data. This work adds to the body of research on regression models for predicting the expansion of new businesses and improves decision-making inside the Indian startup circuit.
(Dhochak et al.; 2024) This research uses data from 757 Indian startup deals from 2012 to 2019 to examine startup value through the prism of strategic management theories like RBV. The research predicts pre-money valuation by using Artificial Neural Networks (ANN) and comparing them with linear regression. Results point to the potential of artificial neural networks (ANNs) as an additional or substitute instrument for valuation, offering venture investors, entrepreneurs, and policymakers advantages in negotiations and ecosystem enrichment.
Because startups have a high failure rate, this study focuses on evaluating startup success through mergers or acquisitions (Pasayat et al.; 2022). Five models—Decision Trees, Random Forest, Gradient Boost, Logistic Regression, and MLP Neural Networks—are used in the study using historical startup data. Important characteristics such as investments, fundraising rounds, and valuations are used to train these models. Astoundingly, with a 92 percent accuracy rate overall, the study offers insightful information to stakeholders and possible investors, easing risk assessment and tactical decision-making.

2.2 Recognizing Startup Performance via Ensemble Machine Learning

This research addresses the issue of predicting successful enterprises, a crucial responsibility for venture investors. According to (Reddy et al.; 2023), it uses machine learning techniques—specifically, an ensemble approach—to identify possible unicorn pairs. In early-stage prediction, the study showcases the potential of machine learning algorithms, which highlights the benefits of the proposed ensemble method above traditional methods. Through improved performance metrics and a focus on innovation, this study offers valuable insights into identifying organizations that are likely to thrive in the global marketplace and support investment decision-making.

An important problem for venture capitalists that this study addresses is business performance predictions. In Mishra et al. (2023), Using machine learning techniques—especially an ensemble approach—it aims to identify future unicorns. Through the demonstration of

machine learning algorithms' early-stage prediction capabilities, the study demonstrates the benefits of the proposed ensemble method over traditional approaches. The insights this research offers into companies that are likely to compete globally and thrive with improved performance measurements help to foster innovation and investment decision-making.

2.3 Machine Learning for Startup Capital Selection

The increasing interest that angel investors and venture capital firms have in startups as possible investments is examined in this study. These investments have a dismal success rate, which emphasizes the need for extensive assessments right away, including behavioral analysis, founder's analysis, and product or prototype review (Singhal et al.; 2022). Anticipating the success of startups is crucial for governments and investors alike, especially in light of their possible exponential benefits. The success of an organization during its early stages is heavily influenced by the founders' history, whereas mid to late-stage enterprises require funds raised during the seed and series stages. Data-driven suggestions are produced with the help of machine learning algorithms applied to founder data. The project tackles important obstacles faced by investors in light of the thriving global startup scene and government programs encouraging entrepreneurship. Fundraising is essential to a company's survival until it turns a profit. This study examines Impact Tech Startups (ITS), a brand-new organizational category that is quickly taking shape at the nexus of technology and entrepreneurship. ITSs, which are frequently supported by private funding, use cutting-edge tactics to address social and environmental issues within a for-profit framework.ITSs have not been thoroughly studied, despite being a relatively recent category. In order to understand ITSs, the study provides a conceptual framework that blends elements of social businesses and startup companies.(Gidron et al.; 2021) It proposes an approach based on machine learning (ML) to identify ITSs in startup datasets by comparing them to the UN Sustainable Development Goals (SDGs). Impact category startups are chosen by parameters aligned with the 17 Sustainable Development Goals. An overview of the future directions for ML-based research into ITSs as a distinct organizational category is provided at the end of the study.

2.4 Prediction of Customer Churn for E-Commerce Businesses

Customer turnover is a significant risk and issue for e-commerce businesses, which drives attempts to reduce churn and retain customers. By focusing on the relationship between churn, client engagement, and complaint resolution, the study investigates a novel perspective on e-commerce customer attrition. the influence of (Batta and Kar; 1965) It introduces the "Customer Retention Through Support Attributes Model" and validates it using statistical testing and inferential modeling. The study uses inferential machine learning models and statistical tests for analysis, utilizing data from an Indian e-commerce firm from their online customer relationship management (CRM) platform. Findings show links between churn and characteristics like complaint handling and remuneration paid, giving managers practical insights to cut down on attrition. Moreover, in a multichannel setting, it reveals the consumers' dependence on particular support channels. The aforementioned results provide opportunities for ongoing investigation into other factors influencing employee turnover as well as approaches to mitigate this issue. Information technology (IT) is a key success factor in startup ecosystems, supporting the

full process from ideation to market leadership. IT infrastructure enables rapid prototyping and collaborative brainstorming, which in turn empowers startups to develop and ideate. Startups use IT solutions to expedite the release of novel products onto the market, guaranteeing their competitiveness and relevance through faster product development processes. IT-driven data analytics support customer acquisition and market research, allowing for more focused audience engagement and market penetration initiatives. IT also serves as a catalyst for securing crucial funding sources, providing venues for connecting with investors, and demonstrating the potential of the business.(Baragde; 2024) IT automation tools facilitate lean, flexible procedures that are conducive to growth and increase operational efficiency. Artificial intelligence (AI) and the Internet of Things (IoT) are examples of cutting-edge technology that give businesses a competitive edge. These technologies encourage innovation and allow firms to customize client experiences. Essentially, the mutually advantageous relationship that exists now between information technology and startups lays the foundation for long-term growth and success in the entrepreneurial environment.

2.5 An Effective Stacking Ensemble Method for Forecasting Indian Venture Success

A growing field of study within India's expanding startup ecosystem is the investigation of machine learning techniques to identify the factors that influence company success. Using web scraping techniques, a complete dataset about Indian startups—both unicorn and non-unicorn—has been compiled. This study uses a range of machine learning classifiers, including Naive Bayes, Decision Trees, Random Forests, and K-Nearest Neighbors, to try and forecast these companies' future success."(Abhinand and Poonam; 2022) In an intentional effort to enhance performance, a stacked ensemble model has been added to further increase forecast accuracy. The effectiveness of these models has been carefully evaluated using a thorough examination that includes measures like Accuracy, F1 scores, and AUC scores. With an amazing accuracy of 78.8 percent and an AUC score of 0.79, the Random Forest classifier using Python programming emerged as the winner. Further improvement was also obtained with the stacked ensemble model, which resulted in an accuracy of 80.1 percent. Interestingly, the prediction methodology was then used to project the success rates of firms that were part of the first season of Shark Tank India, providing useful information about how Indian entrepreneurship is changing.

3 Methodology

3.1 Collecting Information and Ensuring Integrity

The first step in the process is gathering data on Indian company registrations. Careful examination is then used to ensure data integrity and correct any missing values. This stage creates a solid dataset for analysis, which paves the way for further research. The approach guarantees the correctness and dependability of the data utilized in the study by meticulously gathering data and performing integrity checks.

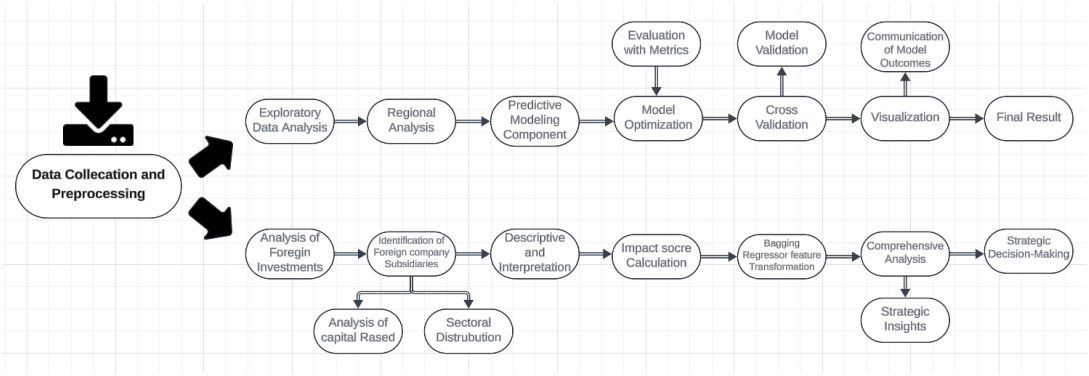

Figure 1: Design - Work Flow

3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the next stage that takes place to investigate the distribution of startups in more detail concerning different factors including states, categories, classes, and company kinds. To find regional differences in sectoral inclinations and registration patterns, this analysis focuses specifically on the top 5 states and Registrars of Companies (City). The dynamics of the startup ecosystem are better understood using EDA, which also throws light on the regional patterns and industry preferences that are common in the Indian startup scene.

3.3 A study of foreign investments within the Indian startup ecosystem

A thorough examination of foreign investments in the Indian startup ecosystem is part of the report. This entails locating overseas companies' subsidiaries and analyzing the money raised as well as the sectoral distribution. Through the application of sophisticated data visualization tools and descriptive statistics, the methodology allows for a comprehensive analysis of the effects of foreign investments on the Indian startup scene.

3.4 A Predictive Modeling Approach for Determining Allowed Capital for Enterprises

The Random Forest and Gradient Boosting regressors are integrated in the predictive modeling stage to estimate the allowable capital for firms. Before modeling, the data is prepared by separating the numerical and categorical columns and applying preprocessing techniques such one-hot encoding and normalization.

3.5 Hyperparameter Optimization and Model Evaluation

At its core, the technology optimizes the performance of both Random Forest and gradient-boosting regressors using hyperparameter adjustment via grid search with cross-validation. To ensure prediction accuracy and resilience, the top-performing models are

then trained using the optimized parameters and assessed on a test set using important metrics including mean squared error, mean absolute error, root mean square error, and R-squared.

3.6 Cross-Validation and Visualization of Model Performance

Through thorough performance validation, cross-validation strengthens the reliability of the model. In the meanwhile, scatter plots, which show actual vs expected values for specific states, provide a powerful visualization tool. This makes it easier to understand and convey the results of the model, which gives the study a fascinating new dimension.

3.7 Analyzing Foreign Company Subsidiaries and Computing Impact Scores

The methodology aims to improve analysis by identifying subsidiaries of foreign companies, retrieving pertinent data, and computing an effect score to measure the diversity of industry sectors among these entities. Furthermore, by utilizing one-hot encoding for feature transformation and training a Bagging Regressor model to estimate allowed capital for foreign subsidiaries, analytical depth is increased.

3.8 Feature transformation and Bagging Regressor Model Training

The approach comprises locating overseas corporate subsidiaries, gathering pertinent data, and calculating an effect score to measure industry sector variety across these entities in order to improve analysis. Furthermore, by using one-hot encoding for feature transformation and training a Bagging Regressor model to estimate allowed capital for foreign subsidiaries, analytical depth is increased.

3.9 Design Specification

The necessary methods and frameworks used in the implementation process are described in the design specification, along with any related requirements. It includes gathering data, prepossessing, analyzing exploratory data (EDA), analyzing foreign investments, conducting predictive modeling with regressors such as Random Forest and Gradient Boosting, optimizing and evaluating the model, visualizing the results, transforming features, and producing the final report. For example, data cleaning libraries are needed for prepossessing; sci-kit-learn is needed for modeling; and visualization tools are needed for interpreting results. Furthermore, each component calls for particular tools and procedures. In predictive modeling, parameter optimization, and model performance evaluation, the utilized algorithms—Random Forest and Gradient Boosting, among others—are essential. In general, the Design Specification offers an organized framework for carrying out the analytical process successfully and quickly. .

4 Implementation

The suggested method is finally put into practice by applying Gradient Boosting and Random Forest regressors for predictive modeling. Following the completion of foreign investment and exploratory data analysis, these algorithms are applied to the preprocessed dataset. To guarantee accurate predictions, the models are refined through the use of strategies like cross-validation and hyperparameter tuning. Performance measures, such as mean squared error and R-squared, are used to assess the models after training. Effective model outcome interpretation is achieved by using visualization approaches like scatter plots. The models are further enhanced in their predictive power by the use of the Bagging Regressor in feature transformation and ensemble learning. When the established system yields valuable information, thorough reports are prepared, and strategic decisions are made.

4.1 Modified Data:

To make sure the original dataset is appropriate for modeling, it goes through a pretreatment process. This includes activities like encoding categorical variables into a numerical format, feature engineering to generate new relevant features, and data cleaning to address missing values and discrepancies. To carry out these transformations effectively, Pandas, a potent Python data manipulation toolkit, is used.

4.2 Models Created

The main goal of this project is to create sophisticated machine learning models, namely the Gradient Boosting and Random Forest regressors. Because they provide precise forecasts and insights into the dataset being analyzed, these models are essential. Complex methods like hyperparameter tweaking and cross-validation are used to make sure these models operate as efficiently as possible. Cross-validation guarantees the models' robustness and generalizability, while hyperparameter tuning allows the models' parameters to be optimized for optimal performance. The utilization of Python libraries, specifically sci-kit-learn, enables the smooth execution and assessment of these models, consequently setting the stage for significant data-driven decision-making.

4.3 Metrics for Evaluation:

To evaluate the predicted accuracy and overall performance of the trained models, it is crucial to use rigorous evaluation measures. For this, metrics like mean absolute error, mean squared error, and R-squared are frequently used. These metrics provide important insights into the models' predictive accuracy and ability to identify underlying patterns in the data. The evaluation process can be streamlined and trustworthy assessments of model performance can be ensured by utilizing tools such as scikit-learn, which offers strong functions for computing these measures. Stakeholders can decide if the models are appropriate and useful for the purposes for which they are intended by carefully examining these evaluation measures.

4.4 Visualization Outputs

A variety of visualization approaches are used in the project's visualization phase to make it easier to interpret model results. To graphically depict the relationships in the data and reveal patterns and correlations, scatter plots are used. Our comprehension of the data dynamics is further improved by the application of impact score computation techniques, which quantify the impact of specific parameters on the model predictions. In order to make complex data understandable and to support the understanding of the model's predictions, interpretative visualization techniques are used. The use of these visualization techniques makes use of well-known Python libraries like Matplotlib and Seaborn, which provide flexible tools for producing informative visual representations of the data and model outputs.

5 Evaluation

5.1 Data Preprocessing and Quality Assurance

The first step of the workflow is to use the panda's library to import the dataset with data about Indian startups. Examining the dataset's first few rows and looking for any missing values constitutes an initial exploration of the dataset. Using a forward fill technique, missing data is handled by substituting the preceding non-null value in the same column for null values. Afterward, the interquartile range (IQR) approach is employed to identify and eliminate anomalies present in numerical columns. To preserve the integrity of the data, values that fall beyond the lower and higher ranges determined by utilizing the first and third quartiles are referred to as outliers and are filtered out. To comprehend the elements impacting startups' success in the Indian market, these stages set the stage for additional research and modeling.

5.2 Exploratory Data Analysis (EDA)

The project's analysis of startup registrations in India is the main goal of this section. It looks at the company's classification, industry sectors, regional distribution, and foreign investments. We seek to unearth insights into the startup ecosystem, trends, and patterns common to the nation's various states and industries through exploratory data analysis (EDA) tools and data visualization initiatives.

5.2.1 Analysis of Startup Distribution by State

The distribution of startups throughout India's states was one of the main topics we investigated for our study. This study can help stakeholders understand the geographic concentration of startups and provide insightful information on the regional landscape of entrepreneurial activity. We used a bar plot that displays the total number of startups in each state to visually represent this as shown in Figure 2. The number of startups registered in each of India's states is graphically represented in this graph. On the y-axis, each bar represents a state and shows the number of startups. The states are distinguished by color, making comparisons easier. Each bar has numerical values at the top that give exact counts. This map provides information about the geographic dispersion of startups, making it easier to analyze local entrepreneurship.

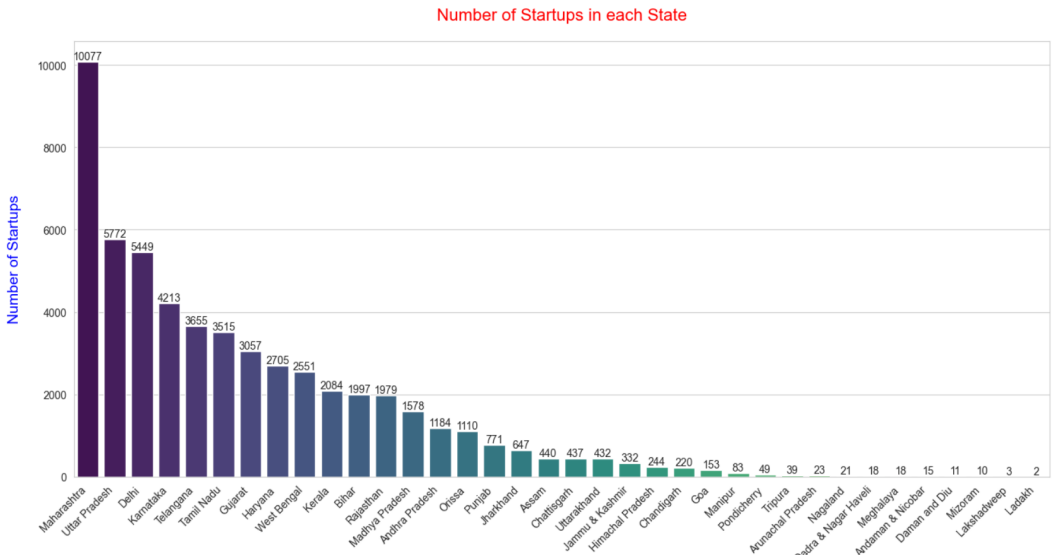

Figure 2: Distribution by State

5.2.2 Examination of Startup Categories by State

Analyzing the distribution of startup categories among India's various states was another important component of our analysis. Stakeholders can learn about the dominating industries propelling entrepreneurial activity by using the insights this report offers into the most common startup types in each state. We used a grouped bar plot that displays the startup category distribution inside each state to visualize this. An invaluable resource for understanding how startup categories are distributed around the states is this graphic. Examples of states where traditional company structures are still prevalent are Maharashtra, Uttar Pradesh, and Delhi, where the majority of startups are share-limited. Nonetheless, states such as Telangana and Karnataka also show a noteworthy fraction of businesses that are restricted by shares, indicating a varied entrepreneurial ecosystem that combines classic and novel company models as in Figure 3. The startup distribution among different states, classified as limited by shares or guarantees, is depicted in this graph. The heights of the bars, which stand in for the states, show how many startups there are. Distinct hues signify various categories for startups. The geographical distribution of these startup categories is briefly shown in this visualization.

5.2.3 Examination of State-specific Startup Classifications

Analyzing the categorization of startups across Indian states according to their legal structure (class) was a critical component of our approach. This analysis sheds light on how common it is for various startup entity types—like one-person businesses (OPCs), public limited corporations, and private limited companies—to be found in the entrepreneurial ecosystem of each state. The distribution of startup categories within each state was displayed in a grouped bar plot that we used to visualize in Figure 4. This graphic shows the breakdown of startups in each state by classifying them as either private, public, or private (OPC). The colors denote the different categories of startup classifications, and each bar represents a state and the number of startups within it. It provides an overview of the geographical distribution of different startup types.

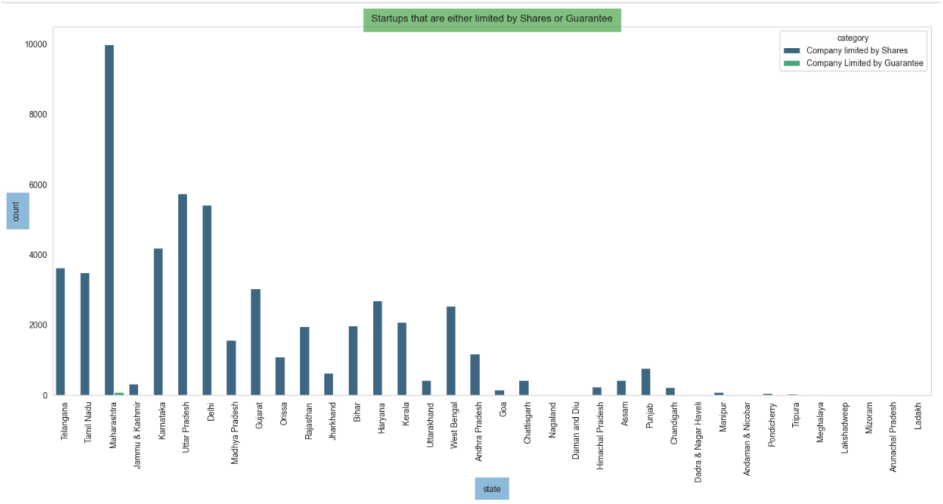

Figure 3: Startup Categories by State

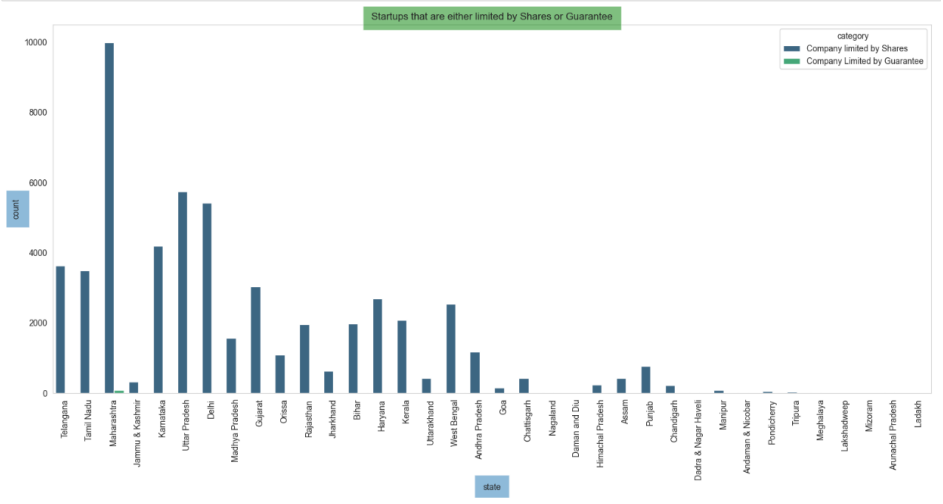

Figure 4: State-specific Startup Classifications

5.2.4 Examining the Different Types of Startup Companies

An important way to understand the variety of organizational structures found in the entrepreneurial environment is to analyze the different sorts of startup companies. The majority of startups are comprised of non-government firms, which are the most popular choice among entrepreneurs. This inclination is probably due to this organizational form's simplicity and flexibility. A growing trend of foreign investments and cooperation in the Indian startup ecosystem is indicated by the substantial role that foreign subsidiaries play. Furthermore, the existence of association and guarantee corporations is indicative of a dedication to social or community-focused projects. The fact that state and union government businesses are included further emphasizes how involved governmental organizations are in fostering innovation and entrepreneurship. Overall, this distribution of startup firm types shows how India's startup ecosystem is dynamic and multidimensional, impacted by both local and foreign factors, and serves a wide range of goals and ideals. Which is shown in Figure 5. This bar graph shows how various companies are distributed about the number of startups. The company kinds are represented by the x-axis, while the number of startups is shown by the y-axis. The colors distinguish each bar from the others, which each belongs to a particular sort of company. For every category, the precise number of startups is shown in the data labels at the top of each bar.

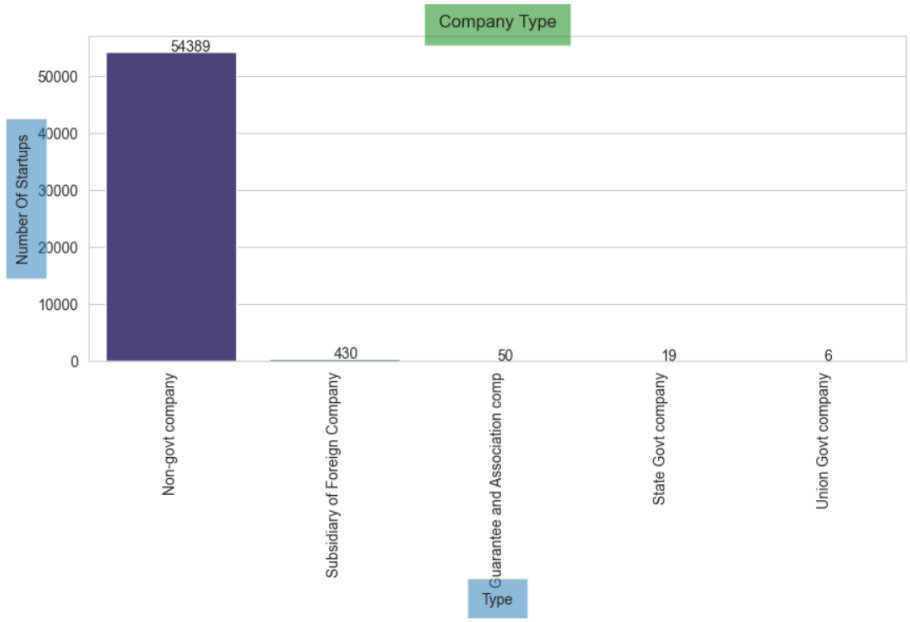

Figure 5: Different Types of Startup Companies

5.2.5 Analysis of Startup Business Descriptions

Innovation and new trends in the entrepreneurial ecosystem can be seen by examining the operations of startup companies. Startups in the technology sector are among the many business domains that may be found in the dataset. The presence of more conventional sectors, such as manufacturing, trade, and agriculture, nevertheless, also suggests a diverse economy. Business services and manufacturing are common among startups, as evidenced by the comparison with the top 5 Indian enterprises of 2021. Also, social media

and community projects mirror current entrepreneurship trends. This alignment shows how Indian startups are advancing innovation and expansion in the entrepreneurial scene by utilizing both contemporary and traditional business strategies.where it is in Figure 6. This bar graph shows how startups are distributed throughout various company categories. With the y-axis showing the number of startups and the x-axis indicating the type of business, each bar represents a certain business category. Colors designate distinct company categories, while data labels at the top of each bar give exact numbers of newcomers in each group.

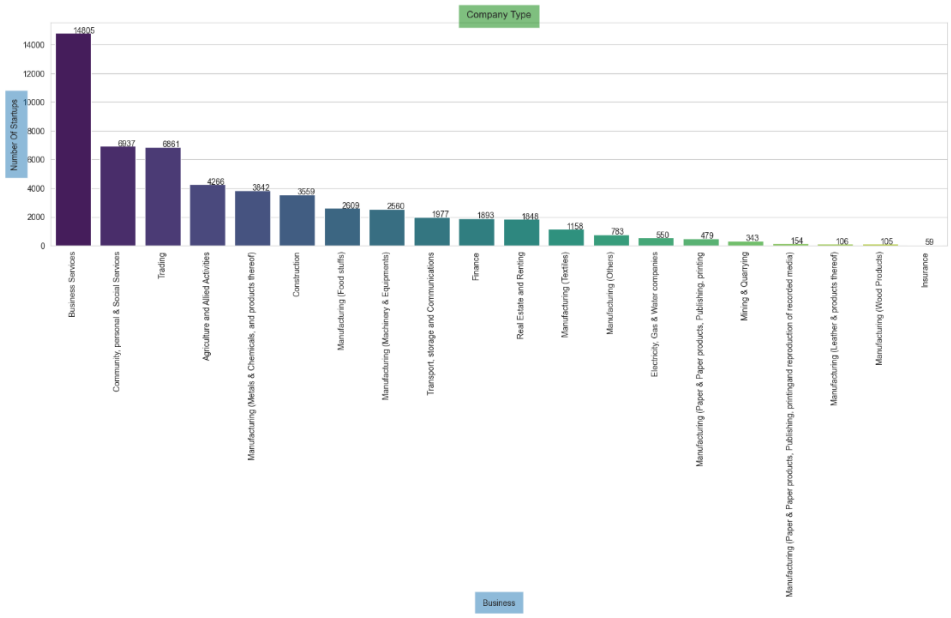

Figure 6: Business Descriptions

5.2.6 An Evaluation of Startup Trends Comparatively in the Top 5 States

The dynamics of local entrepreneurial ecosystems can be understood by contrasting national trends with startup trends in the top 5 states. The startup landscape's regional strengths, shortcomings, and new prospects are uncovered by this investigation.
The pie chart illustrates the leading startup industries in each of the top five states, emphasizing any divergences from the overall patterns. To comprehend regional preferences and opportunities, governments, investors, and entrepreneurs can benefit greatly from this knowledge.
A thorough analysis of the capital allocation among the top 5 states' industries is also provided by the point plot. By providing a more in-depth understanding of the dynamics of local startup ecosystems, it assists stakeholders in identifying industries with higher capital investment and prospective growth or innovation areas. In Figures 7 and 8 The distribution of registered companies in the top 5 states by sector is shown in the pie chart. A slice with percentage values shown inside represents each sector's proportionate representation. In the meantime, these states' capital allocation across industries is compared using a point plot. With the y-axis displaying the capital amount and the x-axis denoting sectors, each point indicates the capital allocation for a particular sector. Color is used to identify different types of capital, such as approved and paid-up capital.

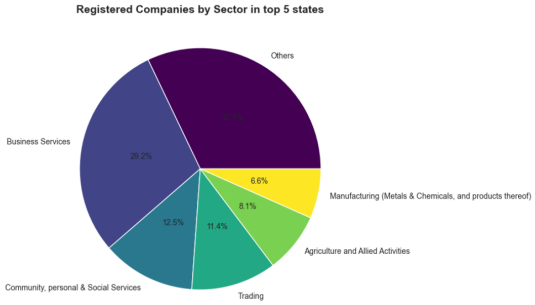

Figure 7: Registered Companies by section

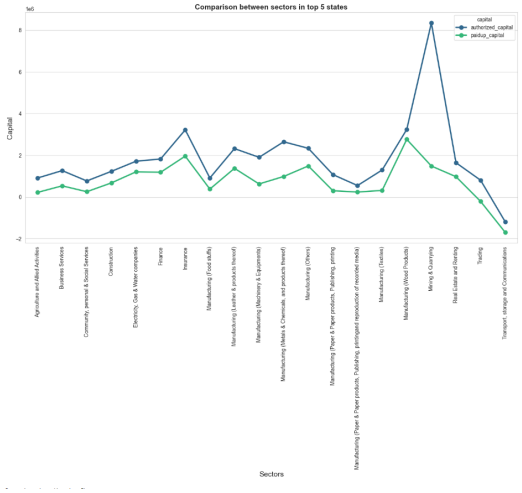

Figure 8: Comparison between Top 5 states

5.2.7 Comparing Firms Listed in Different Cities

Gaining knowledge about how registered companies are distributed throughout various cities can be quite beneficial in identifying areas of high entrepreneurial activity and regional economic activity. Stakeholders are assisted by this approach in identifying cities with high rates of entrepreneurship and prospective locations for capital investments and business expansion. In figure 9 The number of registered companies in each city is displayed visually in this bar plot. Every bar is a city, and the number of registered companies in each city is represented by the bar's height. The cities are shown on the x-axis, and the number of registered firms is shown on the y-axis.

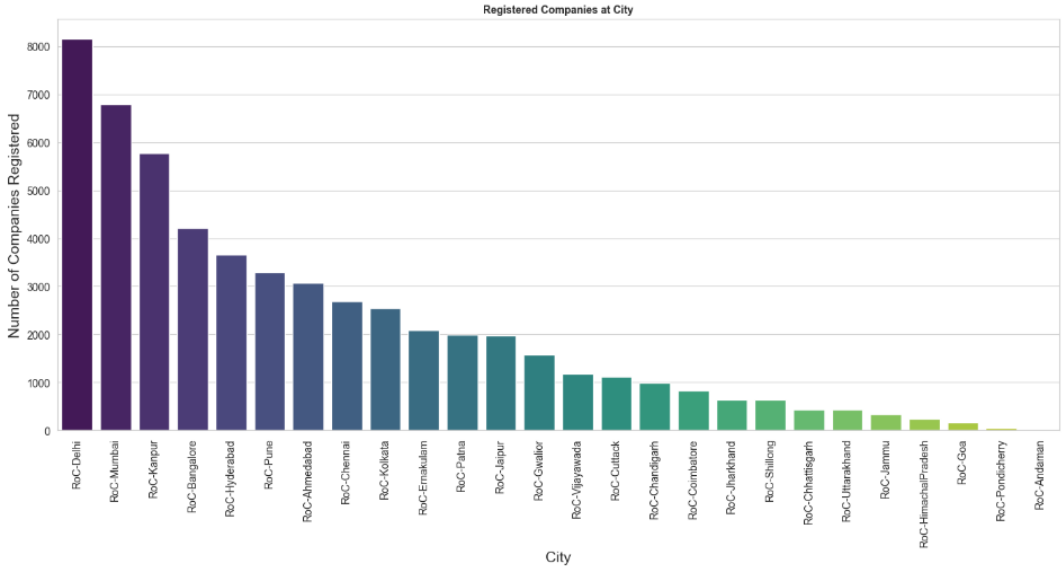

Figure 9: Comparing Firms Listed in Different City

5.2.8 Evaluation of International Subsidiaries

There are important ramifications for stakeholders in a variety of domains when examining the capital and activity of foreign subsidiaries in India. Policymakers can evaluate which sectors are drawing substantial foreign capital and evaluate the effects of foreign investment on the Indian economy by gaining valuable insights into the foreign investment landscape in India through an understanding of the distribution of foreign subsidiary companies across various sectors. Furthermore, by identifying industries that draw a lot of foreign investment, companies and investors can discover areas that show promise for growth and investment, coordinating their strategies to take advantage of these opportunities. In Figure 10 The distribution of enterprises across various sectors is depicted in this horizontal bar plot. The length of a bar denotes the number of companies registered in a particular sector, and each bar represents a sector. Axis: indicates the number of registered enterprises; y-axis: shows the industries. The quantity of businesses within each bar is indicated by the numbers on the right side.

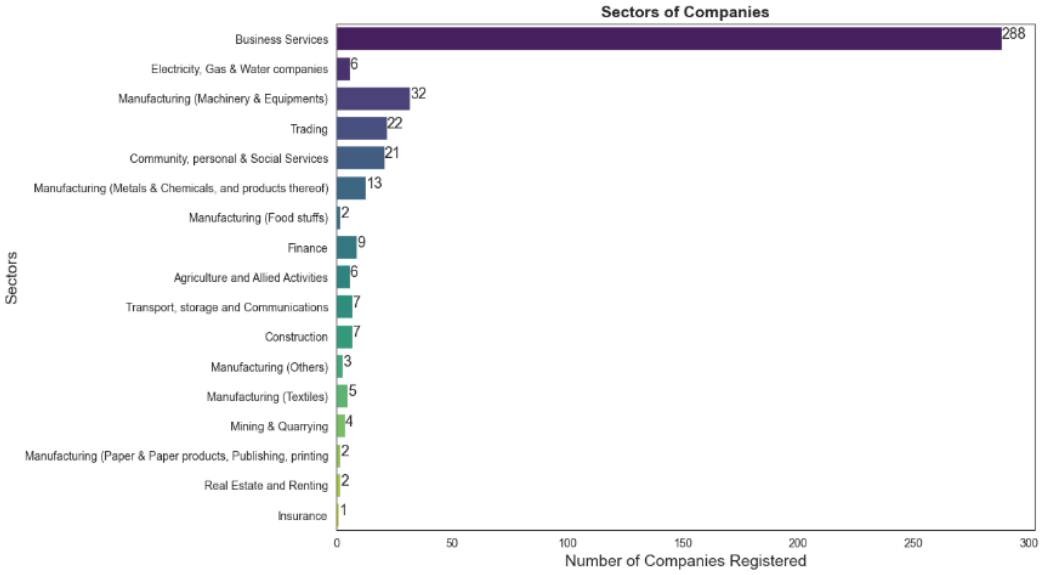

Figure 10: Sectors of company

Investment returns can be maximized and capital-intensive businesses can be better understood by investors thanks to the stripplot that shows the paid-up capital distribution across foreign subsidiary companies across sectors. Through the comparison of authorized and paid-up capital, stakeholders can expect future trends and adjust their strategies by gaining insights into investment patterns. To further promote economic growth and job creation, policymakers can use the insights gathered from this analysis to create policies that support foreign investment in strategic industries while resolving obstacles or issues that might prevent investment flows. In conclusion, examining the operations and financial commitments of foreign subsidiaries in India provides important information about the dynamics of foreign investment, helps to spot promising new ventures, and helps shape policy to support economic progress. in Figure 11The paid-up capital distribution across industries for different company types is shown in the first graph. With different colors denoting various firm types, each point reflects a company's capital in a certain

industry. in Figure 12 The capital distribution for international enterprises is compared across sectors in the second graph. Each sector's mean capital values are displayed using a point plot that is divided into various capital kinds.

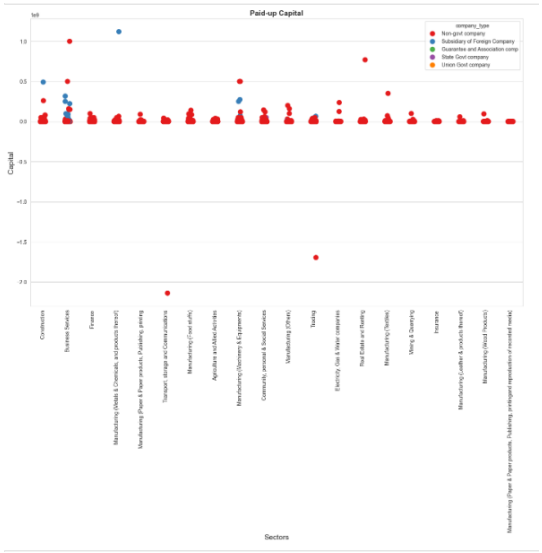

Figure 11: Paidup capital

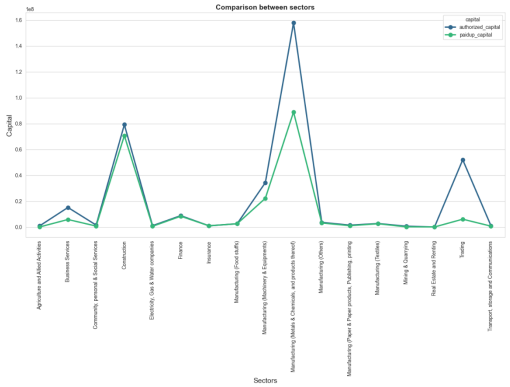

Figure 12: Comparing in sector

5.3 Top 10 state by Authorized capital

Key metrics including mean, standard deviation, minimum, maximum, and quartile values are highlighted in the allowed capital summary statistics, which span 54,894 entries and have an average authorized capital of about 1.23 million INR. States like Delhi and Maharashtra lead the way with approved capitals reaching 9 billion and 8 billion INR, respectively, when the data is aggregated by state. These make up the top 10 states in terms of total allowed capital, along with Tamil Nadu, Telangana, and Karnataka. A more focused study can be carried out by concentrating on these top states and taking into account characteristics like company UID, city, category, class, company type, and activity description. Understanding areas with large capital investments is aided by this analysis, which also provides information for strategic decision-making processes.

5.4 Comparing Gradient Boosting and Random Forest for Capital Prediction

To estimate permitted capital based on a given dataset, two regression models—the Random Forest Regressor and the Gradient Boosting Regressor—were evaluated. Grid search was used to adjust the models' hyperparameters, and a test set was used to evaluate the models' performance.

5.4.1 Random Forest regressor

x'max-depth': None, 'min-samples-leaf': 1, 'min-samples-split': 2, 'n-estimators': 100 were found to be the ideal values for the Random Forest Regressor. The model exhibited

an evaluation result of 1.47e+13 for the mean squared error (MSE), 55407.36 for the mean absolute error (MAE), 3829957.11 for the root mean squared error (RMSE), and 0.96 for the R-squared (R2) score. There was also a variety of MSE values obtained from cross-validation over 5 folds. in Figure 13 The comparison between the values that the Random Forest regression model produced as expected and actual is shown in this plot. With the x-coordinate signifying the actual value and the y-coordinate signifying the value the model predicted, each point represents a data instance. Perfect forecasts, when the expected and actual values fully align, are shown by the diagonal dashed line. Understanding the model's accuracy and variation from perfect prediction can be gained from looking at the distribution of scatter points around this line.

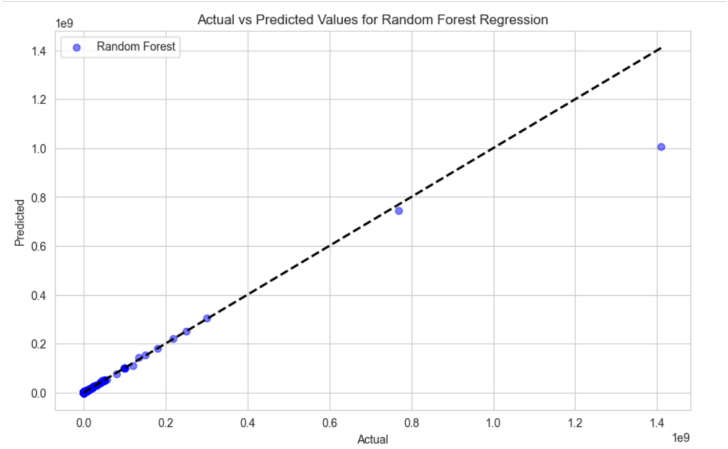

Figure 13: Random Forest regressor

5.4.2 Gradient Boosting Regressor

A set of optimum settings was applied to the Gradient Boosting Regressor: a maximum depth of 3, a learning rate of 0.2, a minimum of one sample per leaf, a minimum of five samples every split, and 300 estimators. This model yielded the following results when tested on the test set: mean squared error (MSE) = 2.32e+13, mean absolute error (MAE) = 63845.31, and root mean square error (RMSE) = 4814458.40. The dependent variable's predictable component based on the independent variables was indicated by the model's R-squared (R2) score, which came out at 0.94. With lower MSE and RMSE values suggesting better fit and higher R2 values indicating a tighter link between the anticipated and actual values, these metrics together show the predictive ability of the model.

Selecting the best model for permitted capital prediction is made easier by these assessments, which provide insightful information about the models' predictive capabilities. The Random Forest Regressor showed marginally better outcomes in terms of Mean Squared Error (MSE) and R-squared (R2) score, even though both models performed well. Here are some state-level results.

In Figure 14, using a RandomForestRegressor model, Because Random Forest can handle both numerical and categorical characteristics well and is stable when working with complicated datasets, it was selected. It can also capture non-linear correlations. Its ensemble approach reduces over-fitting and yields dependable forecasts, which makes it an excel-

lent choice for examining a variety of startup data from various states and sectors. this function plot-state-predictions creates scatter plots for each state, comparing the actual allowed capital levels to the expected values. The permitted capital is represented by the x-axis, while the expected values are represented by the y-axis. Perfect predictions are shown by a diagonal dashed line when actual and expected values fully align. For each state prediction, the distribution of points surrounding this line provides information about the accuracy and variance from a perfect prediction of the model.

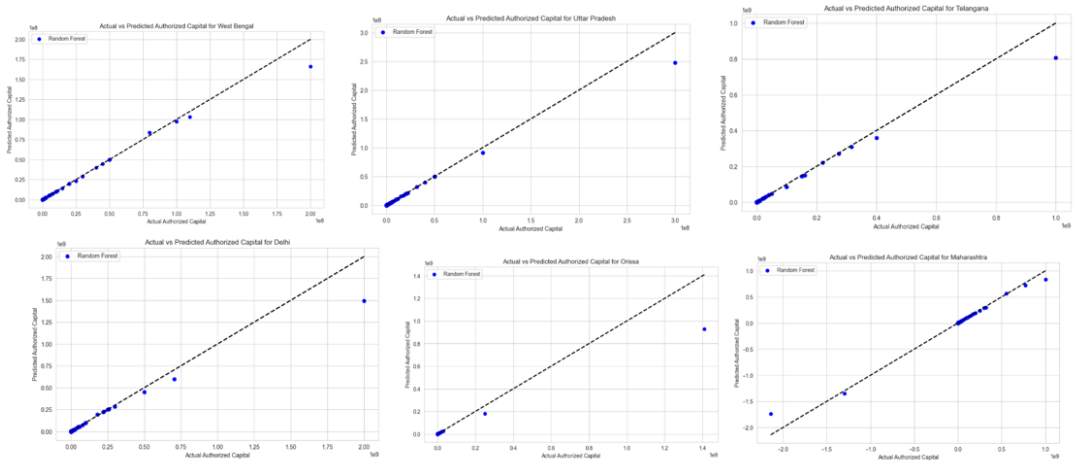

Figure 14: Comparing in sector

5.5 foreign companies

To assess how foreign subsidiaries are distributed throughout states, we limited the information to those that are owned by foreign corporations. The qualities that were chosen included important elements including the state, category, class, kind of firm, authorized capital, paid-up capital, and description of the activity. This in-depth investigation shed light on the environment around international company activities in India.
After distributing these subsidiaries throughout the states, we used a count plot to indicate how many international subsidiaries were in each state. The map, which was made with Seaborn, provides a clear picture of how subsidiaries of international corporations are distributed geographically throughout India's many regions. The concentration and dispersion of foreign business entities are better understood with the help of this graphic, which supports market analysis and strategic decision-making processes. In Figure 15 This graphic shows how foreign subsidiaries are distributed among the several states. The number of foreign subsidiaries in each bar indicates how many are present in that state. States are shown on the x-axis, while the number of overseas subsidiaries is shown on the y-axis. Quick evaluation of the states with a greater concentration of foreign subsidiaries is made possible by the distribution, which also offers information on the geographic distribution of foreign investment in India's startup ecosystem.

5.5.1 Bagging Regressor

To determine the industry diversification and overall impact of foreign businesses' subsidiaries, we conducted a thorough analysis as part of our evaluation. To measure the

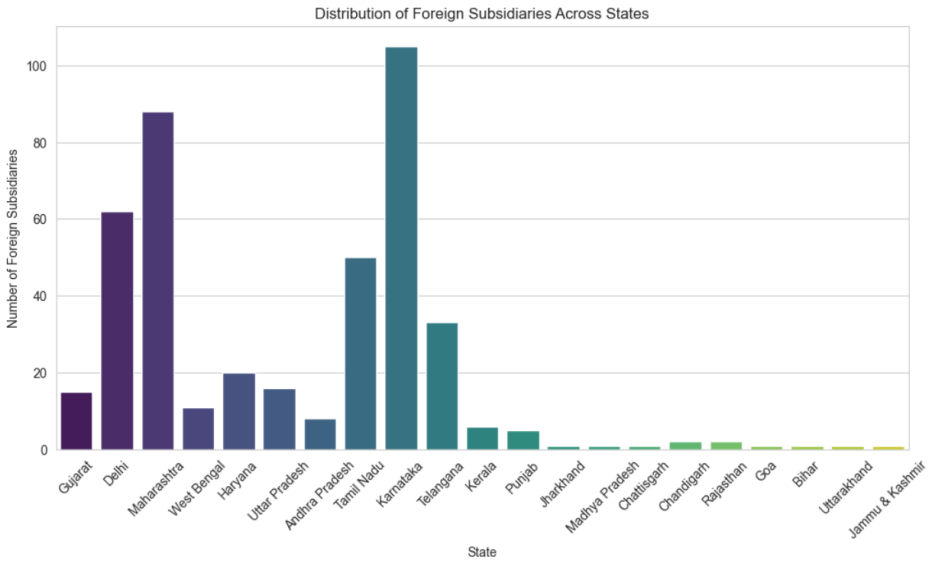

Figure 15: foreign companies distributed

scope of the operations carried out by these subsidiaries, we first identified the industry sectors that were represented in the dataset. From there, we computed a sector diversity score. Our impact score, which sheds light on the combined influence of these entities, was created by weighing and standardizing these ratings.(Bhattacharya; 2024)
We then switched to predictive modeling and used a Bagging Regressor to estimate allowed capital according to important attributes including state, category, class, kind of firm, and description of the activity. With an astounding R-squared score of 0.94 and a mean squared error of 2.36e+13, the model demonstrated strong performance. This highlights the model's capacity to accurately represent the variations in permitted capital among various subsidiaries.
We also plotted actual vs expected values to visually evaluate the model's prediction performance, and the results showed a close match between the observed and forecasted data. By showing the model's usefulness in predicting allowed capital for overseas subsidiaries, this visualization helps to bolster our faith in its dependability. All things considered, our assessment emphasizes the diverse methodology used to comprehend and forecast the behavior of these organizations in the corporate environment. In Figure 16 The target variable's actual values are contrasted with the values predicted by the Bagging Regressor model in this scatter plot. Every point on the plot corresponds to a test set data point. The dashed red line depicts the ideal situation in which the actual and anticipated values are precisely aligned, while the blue dots represent the predicted values. An evaluation of the model's performance is made possible by the plot, which offers a visual assessment of how well the model's predictions match the actual values.

6 Conclusion and Future Work

The report concludes by providing a thorough examination of the Indian startup ecosystem, illuminating its complex dynamics and the multitude of factors that shape it. A wealth of information about the challenges of promoting entrepreneurship in India has been gathered through studies of startup typologies, the spatial distribution of states,

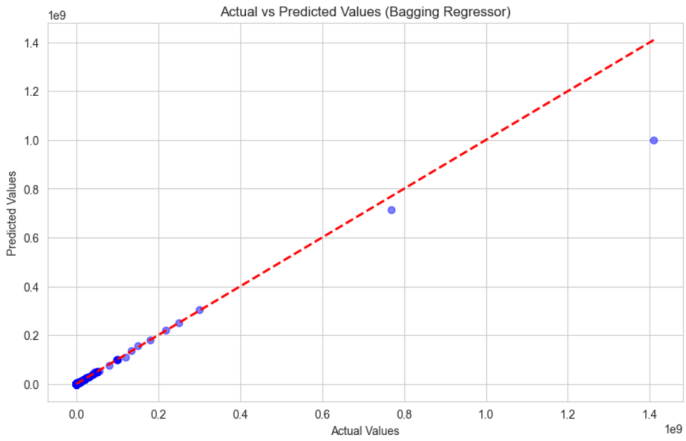

Figure 16: Bagging Regressor

and the effects of foreign investment. The research results emphasize the variety of startup activities across both traditional and innovative industries, and they also show how important foreign investment is in spurring innovation and growth. This research establishes a basis for well-informed decision-making by policymakers, investors, and entrepreneurs by clarifying these processes and advancing our understanding of the Indian startup scene.

According to our analysis, the startup scene in India is better understood, which also covers the impact of foreign investment, industry sectors, and regional distribution. Even though we didn't explicitly compare our results to those of Made in India or Startup India, our findings align with their goals. For example, we match Startup India's objectives by identifying areas of future support and highlighting thriving entrepreneurial hotspots. Similarly, Digital India's objective aligns with our emphasis on tech businesses. Our research provides data-driven recommendations to support policy decisions that support innovation and economic growth across the country.

Future studies could build on these findings in several intriguing directions. By monitoring the development of the startup ecosystem over time, longitudinal studies can identify new trends and obstacles. Additional research on the socio-economic effects of foreign investment and startups can yield important insights into how these factors contribute to social development, economic expansion, and job creation. Furthermore, investigating the impact of governmental policies and regulatory frameworks on the startup ecosystem can provide significant understanding of cultivating an environment that is favorable to entrepreneurship. Finally, the quality and dependability of the insights produced for stakeholders will be improved by continuous improvement and verification of predictive models for predicting trends and estimating capital allocation.

References

Abhinand, G. and Poonam, B. (2022). An efficient stacking ensemble technique for success prediction of indian ventures, *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, Vol. 1, IEEE, pp. 1–6.

Baragde, D. B. (2024). Role of information technology for startups in india, *Ecosystem Dynamics and Strategies for Startups Scalability*, IGI Global, pp. 114–132.

Batta, A. and Kar, A. K. (1965). Influence of complaint resolution mechanisms on customer engagement using machine learning based approach-insights from indian e-commerce startup, *Available at SSRN 4064508* .

Bhattacharya, D. (2024). Utilizing base machine learning models to determine key factors of success on an indian tech startup.

Dhochak, M., Pahal, S. and Doliya, P. (2024). Predicting the startup valuation: A deep learning approach, *Venture Capital* **26**(1): 75–99.

Gidron, B., Israel-Cohen, Y., Bar, K., Silberstein, D., Lustig, M. and Kandel, D. (2021). Impact tech startups: A conceptual framework, machine-learning-based methodology and future research directions, *Sustainability* **13**(18): 10048.

Mishra, A., Jat, D. S. and Mishra, D. K. (2023). An experimental study of machine learning algorithms for predicting start-up success, *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 1*, Springer, pp. 813–825.

Pandya, D. D., Patel, A. K., Purohit, J. M., Bhuptani, M. N., Degadwala, S. and Vyas, D. (2023). Forecasting number of indian startups using supervised learning regression models, *2023 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, pp. 948–952.

Pasayat, A. K., Mitra, A. and Bhowmick, B. (2022). Determination of essential features for predicting start-up success: an empirical approach using machine learning, *Technology Analysis & Strategic Management* pp. 1–19.

Reddy, S. H., Bathini, H., Ajmeera, V. N., Marella, R. S., Kumar, T. V. and Khari, M. (2023). Startup unicorn success prediction using ensemble machine learning algorithm, *International Conference on Intelligent Human Computer Interaction*, Springer, pp. 330–338.

Singh, K., Misra, M. and Yadav, J. (2021). Artificial intelligence and machine learning as a tool for combating covid-19: a case study on health-tech start-ups, *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, pp. 1–5.

Singhal, J., Rane, C., Wadalkar, Y., Joshi, M. and Deshpande, A. (2022). Data driven analysis for startup investments for venture capitalists, *2022 International Conference for Advancement in Technology (ICONAT)*, IEEE, pp. 1–6.