Predictive Modeling of Company Success

Shiva Kumar Pendem

Introduction

Predictive modeling holds an intriguing and significant place in venture capital (VC) decision-making, particularly due to its prowess in sifting through extensive datasets to identify trends and patterns crucial for forecasting the success of companies. Our choice to focus on this topic stems from a deep interest in startups and a keenness to understand how data-driven approaches can revolutionize their evaluation and funding. Our project aims to explore predictive modeling not just as a theoretical concept, but as a practical tool that adds a layer of data-backed insight to the traditional, intuition-led processes in VC.

However, the effectiveness of predictive modeling is not without its caveats. The reliability of these models is heavily reliant on the quality and pertinence of the data they process. In the unpredictable and ever-evolving world of startups, certain aspects of success remain elusive to even the most sophisticated algorithms. Furthermore, the rapid shifts in market dynamics and technological advances can sometimes outpace the adaptability of these models.

Data

Our dataset was obtained from Kaggle. The specific dataset used is 'Startup Success Prediction,' which can be accessed at Kaggle Dataset. The data has 923 rows and 49 columns.

Below are the names of the features present in the dataset

state_code	**latitude**	**longitude**	**zip_code**	**id**
city	**Unnamed: 6**	**name**	**labels**	**founded_at**
closed_at	**first_funding_at**	**last_funding_at**	**age_first_funding_year**	**age_last_funding_year**
age_first_milestone_year	**age_last_milestone_year**	**relationships**	**funding_rounds**	**funding_total_usd**
milestones	**state_code.1**	**is_CA**	**is_NY**	**is_MA**
is_TX	**is_otherstate**	**category_code**	**is_software**	**is_web**
is_mobile	**is_enterprise**	**is_advertising**	**is_gamesvideo**	**is_ecommerce**
is_biotech	**is_consulting**	**is_othercategory**	**object_id**	**has_VC**
has_angel	**has_roundA**	**has_roundB**	**has_roundC**	**has_roundD**
avg_participants	**is_top500**	**status**		

Data Preprocessing Summary

Challenges Encountered: Our initial dataset presented multiple challenges characteristic of raw data. We encountered columns without names or with redundant information, inconsistent formats in date entries, and a mix of data types, which included some columns with missing values.

Cleaning Actions Undertaken: To address these issues, we undertook a series of cleaning actions:

- We removed unnecessary columns, specifically those unnamed or duplicative, such as 'Unnamed: 0', 'Unnamed: 6', and 'state_code.1'.
- We standardized date columns to a consistent 'YYYY-MM-DD' format to facilitate chronological analysis.
- We corrected data types for fields containing numerical data to ensure computational accuracy.
- We carefully addressed missing data, considering the significance of each field to determine the most appropriate handling method.

Composition of the Refined Dataset: Post-cleaning, our dataset was distilled down to 37 well-defined columns. These included critical variables such as geographic location, funding details, operational timelines, company status, and industry categories. We also ensured that dates were consistently formatted, categorical variables were suitably encoded, and numerical data was appropriately scaled for analysis.

Exploratory Data Analysis

```{r}
data
```

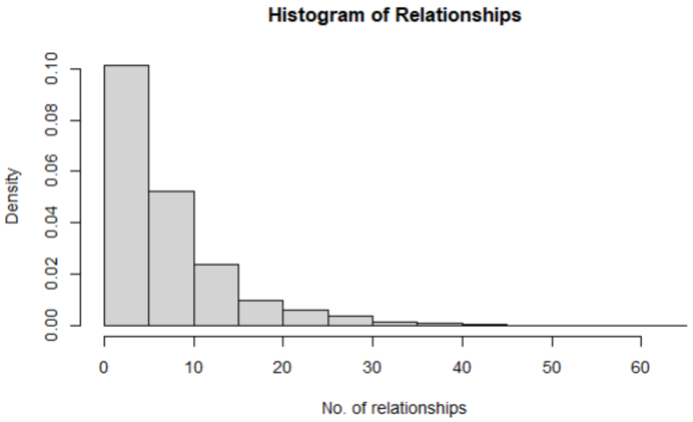
Description: df [923 × 45]

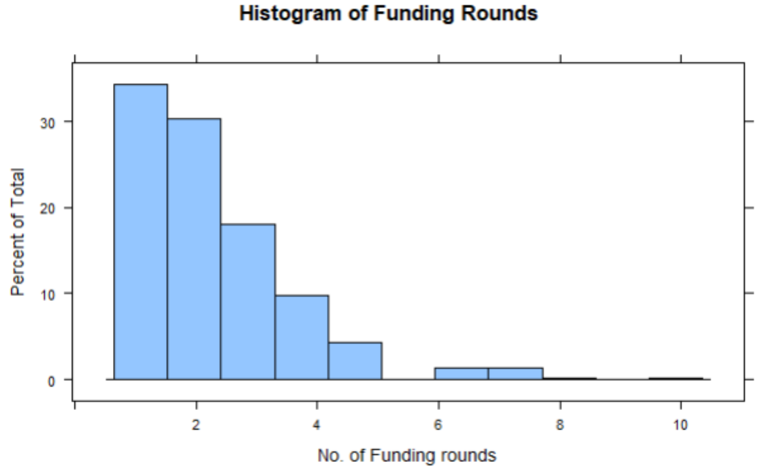
state_code <chr>	latitude <dbl>	longitude <dbl>	zip_code <chr>	city <chr>	name <chr>	labels <int>	founded_at <chr>	closed_at <chr>	first_funding_at <chr>
CA	42.35888	-71.056820	92101	San Diego	Bandsintown	1	2007-01-01		2009-04-01
CA	37.23892	-121.973718	95032	Los Gatos	TriCipher	1	2000-01-01		2005-02-14
CA	32.90105	-117.192656	92121	San Diego	Plixi	1	2009-03-18		2010-03-30
CA	37.32031	-122.050040	95014	Cupertino	Solidcore Systems	1	2002-01-01		2005-02-17
CA	37.77928	-122.419236	94105	San Francisco	Inhale Digital	0	2010-08-01	2012-10-01	2010-08-01
CA	37.40691	-122.090370	94043	Mountain View	Matisse Networks	0	2002-01-01	2009-02-15	2006-07-18
CA	37.39156	-122.070264	94041	Mountain View	RingCube Technologies	1	2005-01-01		2006-09-21
CA	38.05711	-122.513742	94901	San Rafael	ClairMail	1	2004-01-01		2005-08-24
MA	42.71221	-73.203599	1267	Williamstown	VoodooVox	1	2002-01-01		2005-08-02
CA	37.42724	-122.145783	94306	Palo Alto	Doostang	1	2005-06-01		2007-02-01

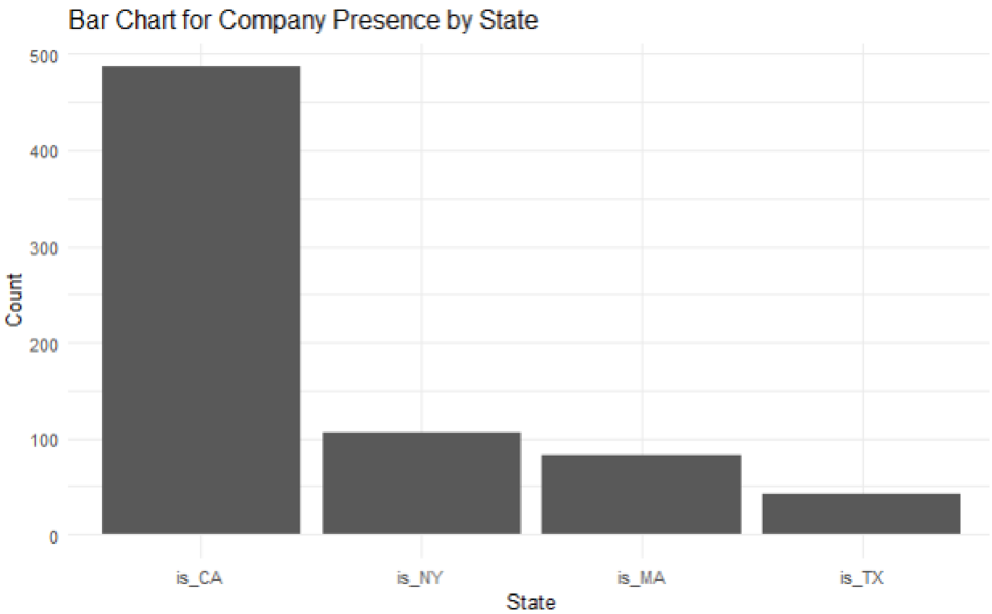
```{r}
summary(data)
```

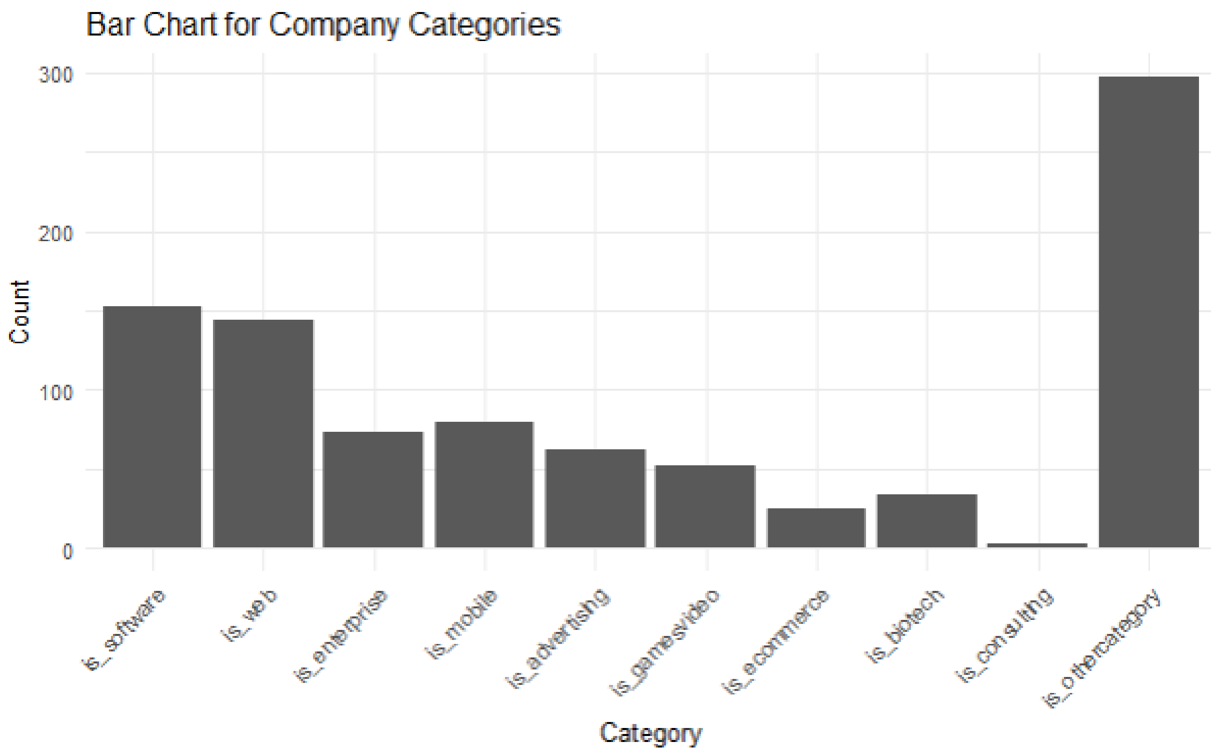
```
  state_code           latitude        longitude          zip_code             city               name               labels          founded_at          closed_at         first_funding_at   last_funding_at    age_first_funding_year
 Length:923         Min.   :25.75   Min.   :-122.76   Length:923         Length:923         Length:923         Min.   :0.0000   Length:923         Length:923         Length:923         Length:923         Min.   :-9.0466
 Class :character   1st Qu.:37.39   1st Qu.:-122.20   Class :character   Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character   Class :character   Class :character   1st Qu.: 0.5767
 Mode  :character   Median :37.78   Median :-118.37   Mode  :character   Mode  :character   Mode  :character   Median :1.0000   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 1.4466
                    Mean   :38.52   Mean   :-103.54                                                            Mean   :0.6468                                                                                Mean   : 2.2356
                    3rd Qu.:40.73   3rd Qu.: -77.21                                                            3rd Qu.:1.0000                                                                                3rd Qu.: 3.5753
                    Max.   :59.34   Max.   :  18.06                                                            Max.   :1.0000                                                                                Max.   :21.8959
 age_last_funding_year age_first_milestone_year age_last_milestone_year relationships    funding_rounds   funding_total_usd     milestones     state_code.1           is_CA             is_NY             is_MA
 Min.   :-9.047        Min.   :-14.170          Min.   :-7.005          Min.   : 0.000   Min.   : 1.000   Min.   :1.100e+04   Min.   :0.000   Length:923         Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.: 1.670        1st Qu.:  1.252          1st Qu.: 2.930          1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.:2.725e+06   1st Qu.:1.000   Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Median : 3.529        Median :  2.521          Median : 4.477          Median : 5.000   Median : 2.000   Median :1.000e+07   Median :2.000   Mode  :character   Median :1.0000   Median :0.0000   Median :0.00000
 Mean   : 3.931        Mean   :  2.967          Mean   : 4.709          Mean   : 7.711   Mean   : 2.311   Mean   :2.542e+07   Mean   :1.842                      Mean   :0.5276   Mean   :0.1148   Mean   :0.08992
 3rd Qu.: 5.560        3rd Qu.:  4.003          3rd Qu.: 6.040          3rd Qu.:10.000   3rd Qu.: 3.000   3rd Qu.:2.472e+07   3rd Qu.:3.000                      3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
 Max.   :21.896        Max.   : 24.685          Max.   :24.685          Max.   :63.000   Max.   :10.000   Max.   :5.700e+09   Max.   :8.000                      Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
     is_TX          is_otherstate   category_code        is_software         is_web          is_mobile        is_enterprise     is_advertising     is_gamesvideo      is_ecommerce        is_biotech       is_consulting
 Min.   :0.0000   Min.   :0.000   Length:923         Min.   :0.0000   Min.   :0.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.000   Class :character   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :0.0000   Median :0.000   Mode  :character   Median :0.0000   Median :0.000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
 Mean   :0.0455   Mean   :0.221                      Mean   :0.1658   Mean   :0.156   Mean   :0.08559   Mean   :0.07909   Mean   :0.06717   Mean   :0.05634   Mean   :0.02709   Mean   :0.03684   Mean   :0.00325
 3rd Qu.:0.0000   3rd Qu.:0.000                      3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.000                      Max.   :1.0000   Max.   :1.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
 is_othercategory     has_VC          has_angel         has_roundA        has_roundB        has_roundC        has_roundD      avg_participants    is_top500          status
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   : 1.000   Min.   :0.0000   Length:923
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.: 1.500   1st Qu.:1.0000   Class :character
 Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000   Median :0.00000   Median : 2.500   Median :1.0000   Mode  :character
 Mean   :0.3229   Mean   :0.3261   Mean   :0.2546   Mean   :0.5081   Mean   :0.3922   Mean   :0.2329   Mean   :0.09967   Mean   : 2.839   Mean   :0.8093
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.: 3.800   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :16.000   Max.   :1.0000
```

Histograms & Bar charts

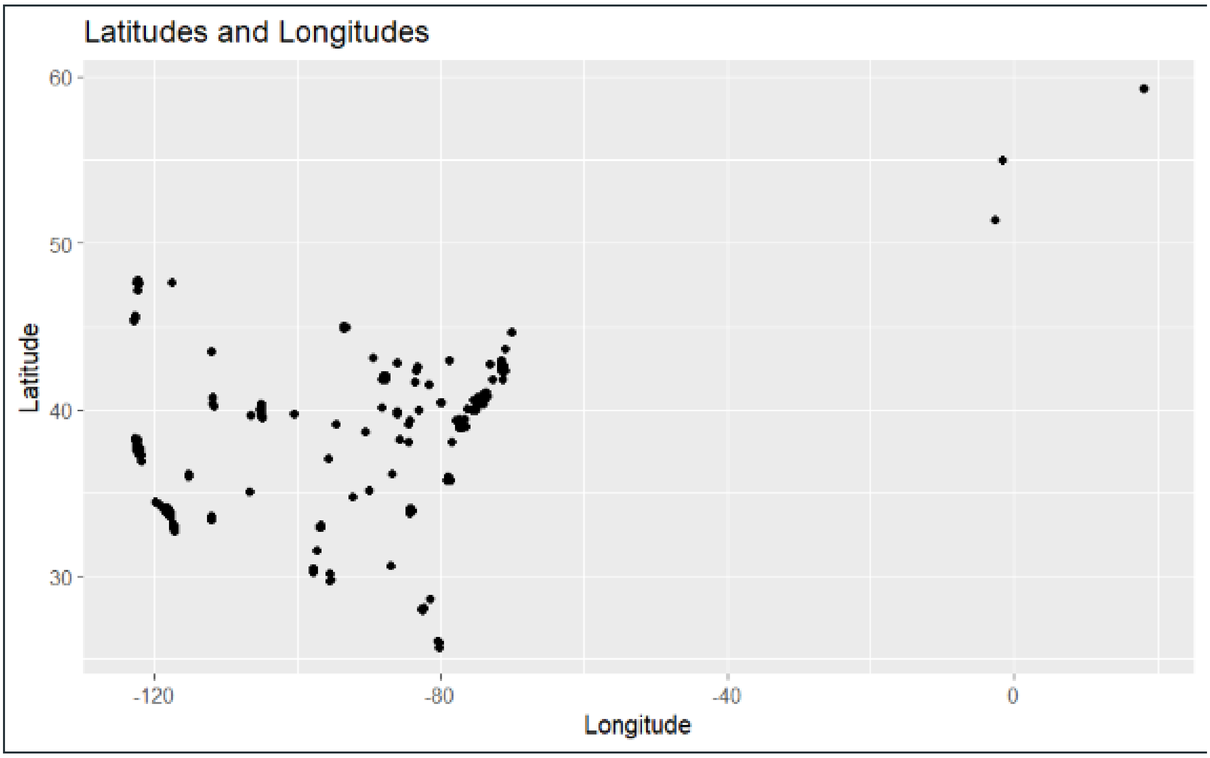




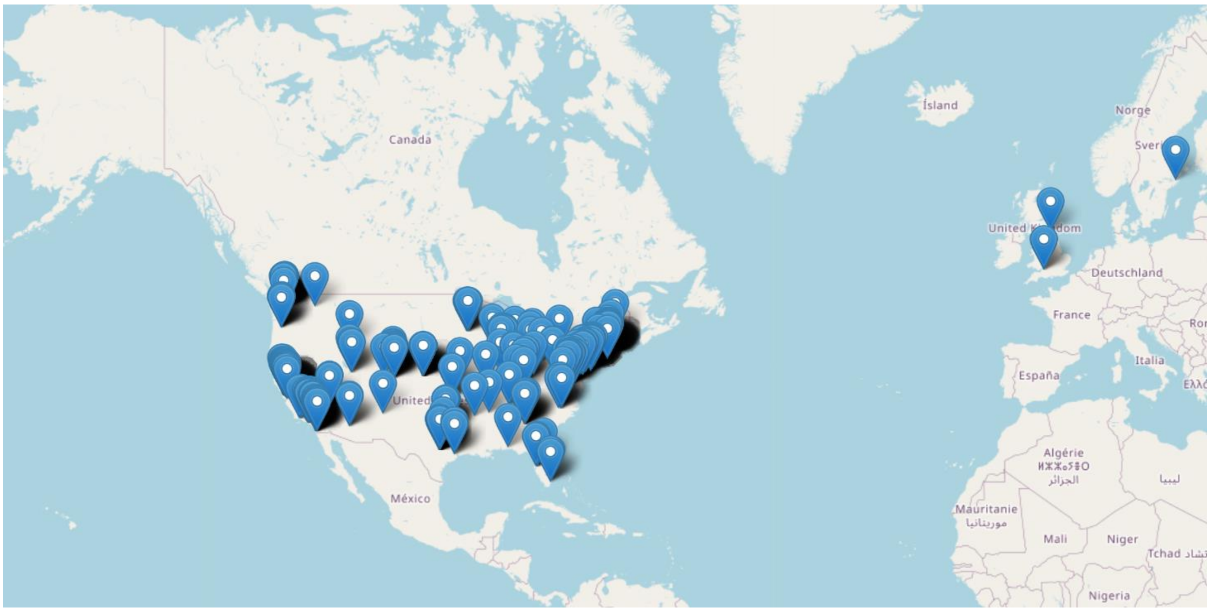




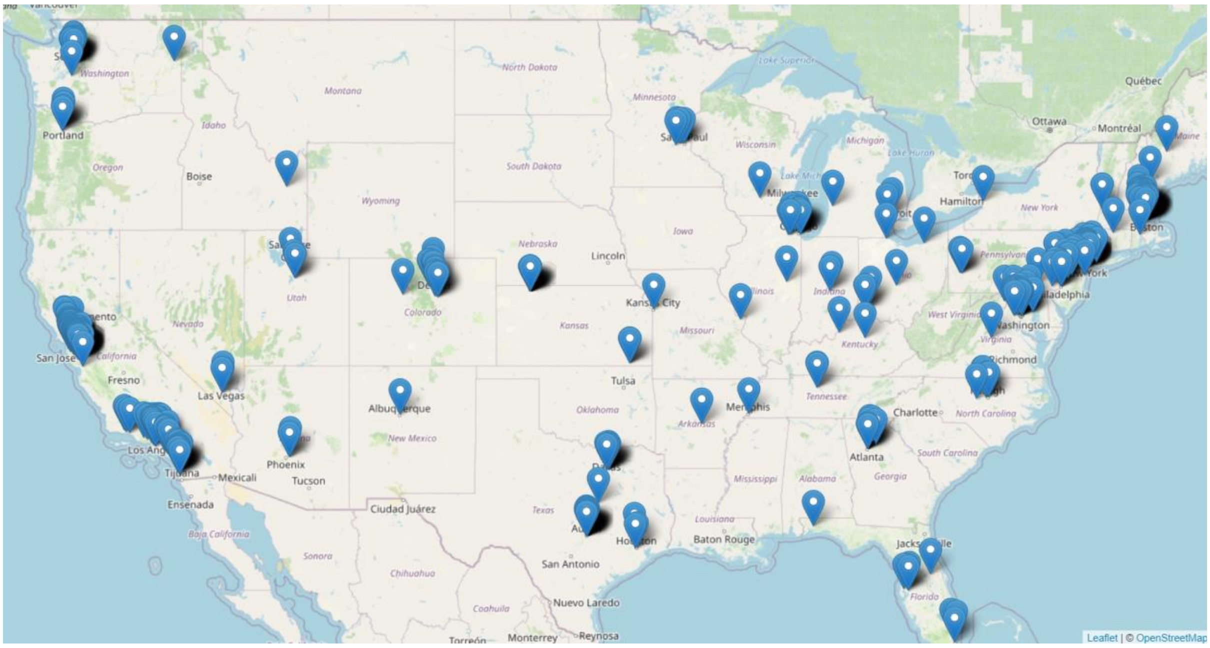
Geographical visualization



Using Leaflet Library



The predominant share of our dataset originates from companies based in the United States, with only minimal representation of approximately three companies operating outside of the U.S.

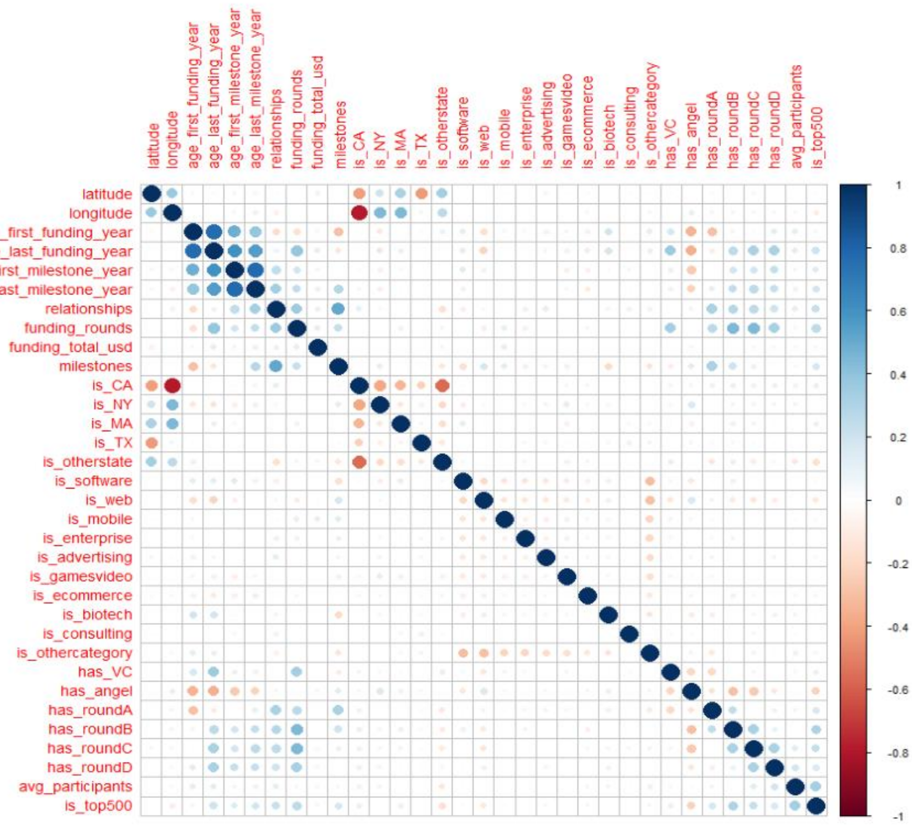


How the companies are present in the US based are presented above where each marker indicates the location of the company.

Libraries Used

Package	Description
ggplot2	A versatile data visualization package for creating complex, multi-layered graphics.
corrplot	Specializes in visualizing correlation matrices, making it easier to identify relationships between variables.
pastecs	Offers tools for statistical analysis, particularly in time series data, enhancing trend analysis and data decomposition.
coefplot	Simplifies the presentation of model coefficients, aiding in the interpretation and comparison of model estimates.
FSelector	Provides functions for feature selection, helping to identify the most informative variables for predictive models.
caret	An all-in-one package for training and evaluating machine learning models, supporting various methods and processes.
dplyr	A toolkit for data manipulation, making data cleaning and transformation more straightforward and readable.
pROC	An essential tool for analyzing and visualizing receiver operating characteristic (ROC) curves, crucial in evaluating binary classifiers.
glmnet	Implements regularized regression models like lasso and elastic-net, useful for handling collinear data and variable selection.
rpart	A package for constructing decision trees, an intuitive and interpretable modeling approach.
randomForest	Implements the Random Forest algorithm, great for both classification and regression with high accuracy.

Correlation Analysis



Correlation Analysis Overview

The correlation heatmap highlights crucial relationships within our dataset:

- **Funding Over Time**: There's a robust link between the timing of initial and subsequent funding rounds, indicating a trend where earlier funding leads to quicker follow-up investments.
- **Milestone Achievement**: Milestone timings are closely aligned, with initial and subsequent milestones often occurring in rapid sequence.
- **Networking and Success**: A positive correlation between the number of relationships and milestones suggests that a wider network is beneficial in reaching key company milestones.
- **Diverse Funding Sources**: The varying relationships between company funding stages and investor types ('VC', 'angel', etc.) illustrate the complex funding landscape startups navigate.
- **State-Specific Trends**: Geographic correlations reflect the distinct startup ecosystems in California and New York, with longitude serving as a clear differentiator.

Linear Model on All features as a base model

Key Findings:

- **State Variables**: The dummy variables for states (*is_CA, is_NY, is_MA, is_TX, is_otherstate*) were significant predictors, with positive coefficients indicating a higher likelihood of the target variable increasing in these states.

- **Relationships and Milestones**: Both 'relationships' and 'milestones' showed significant positive coefficients, suggesting that these features are influential predictors. The number of relationships a company has is notably associated with the target, emphasizing the importance of networks.
- **Top 500 Rank**: The *is_top500* variable had a substantial positive effect, indicating that companies within the top 500 ranking are more likely to have an increased target variable.

Model Metrics:

- The residual standard error of 0.4193 on 890 degrees of freedom indicates the average distance of the data points from the fitted line.
- The **Multiple R-squared** value of 0.258 suggests that approximately 25.8% of the variability in the target variable can be explained by the model. Although this indicates some level of fit, there is still a substantial amount of variability that the model does not account for.
- The **Adjusted R-squared** of 0.2314 adjusts this figure for the number of predictors in the model, providing a more conservative estimate of the model's explanatory power.
- The overall **F-statistic** is significant ($p < 2.2e-16$), meaning that the model is a better fit than an intercept-only model.

Logistic Model

Building upon our initial linear regression analysis, we acknowledged the binary nature of our target variable—representing whether a company is successful or not—and thus opted for logistic regression to better cater to our modeling needs. Logistic regression is more apt for binary outcomes and enables the estimation of probabilities, which is crucial for classification tasks such as ours.

To refine our predictive model, we selected key features that showed significant relationships with the target variable. These included *relationships*, which may reflect a company's network; *milestones*, indicative of tangible achievements; and *is_top500*, suggesting a correlation between market recognition and company success. Additionally, we incorporated various state indicators and *avg_participants* to account for regional influences and the average number of participants in funding rounds, respectively.

The logistic regression was then executed on these important features. Preliminary checks for multicollinearity, using VIF scores, signalled high values for state variables. However, given their conceptual relevance, we retained them in the model. Notably, while the state indicators showed high multicollinearity, they did not significantly affect the target in the logistic model, possibly due to the dominance of other variables with stronger predictive power.

The logistic model's performance metrics were compelling: an accuracy rate of 75%, precision at 78%, and an exceptional recall of 87%, which, combined with an F1 score of 82%, confirmed the model's robustness in classification. The AUC-ROC value of 0.81 further attested to the model's strong discriminative capacity.

In conclusion, the logistic regression model *solidified the importance of networking, achievements, and recognition in influencing a company's success*. These insights, derived from a statistically rigorous approach, provide a quantitative backbone for strategic business decisions and future analyses focused on enhancing company performance.

Lasso Selected Features

Continuing our model refinement, we employed Lasso regression to further streamline our feature set. This approach led to the selection of the most impactful predictors, *including age_last_milestone_year, relationships, milestones, is_MA, is_otherstate, is_enterprise, has_VC, has_roundB, has_roundD, avg_participants*, and *is_top500*. Notably*, is_top500* emerged as a significant positive predictor, while *is_otherstate* was negatively associated with the target variable. Variables such as *age_first_funding_year* and state indicators like *is_CA* were excluded by Lasso, suggesting their lesser relevance in the presence of other variables. The refined model offers a concise set of features, which will be tested in subsequent logistic regression analyses to ensure robustness and accuracy.

Logistic Model on Lasso Selected Features

After refining our approach with a Lasso regression that pinpointed the most significant predictors, we implemented a logistic regression model to align with the binary nature of our target variable. The logistic model was applied using selected features including *age_last_milestone_year, relationships, milestones, is_MA, is_otherstate, is_enterprise, has_VC, has_roundB, has_roundD, avg_participants, and is_top500.*

This targeted model yielded an accuracy of 75.73%, with a precision of 78.65% and a recall of 85.76%, leading to an F1 score of 82.05%. The model's AUC-ROC score was 0.8129, indicating a strong ability to differentiate between the successful and unsuccessful companies.

These results build on our earlier linear model, which identified state variables, relationships, milestones, and top 500 ranking as significant predictors. The logistic regression confirms the importance of networks *(relationships and milestones*) and market recognition (*is_top500*). The negative coefficient for *is_otherstate* implies that companies outside the major hubs may face a disadvantage, while *is_MA* shows a positive influence.

The performance metrics of the logistic model demonstrate a marked improvement from the linear regression, particularly in classification accuracy. This suggests that the predictors retained through Lasso selection carry substantial weight in defining company success, reinforcing the value of strategic networking, milestone achievement, and the benefits of being among the top-ranked companies.

Lasso and Ridge Regression Insights

Continuing the narrative from the Lasso regression analysis, we proceeded to scale our data and apply both Lasso and Ridge regression models for logistic regression, aiming to enhance our predictive accuracy and ensure robustness in our findings.

Scaled Logistic Regression Analysis

Upon scaling our features to account for any disparities in measurement scales, we conducted logistic regression with both Lasso and Ridge regularization. The purpose of these techniques is to improve the model's generalizability and prevent overfitting, with Lasso having the added benefit of feature selection by shrinking some coefficients to zero.

Model Performance Metrics

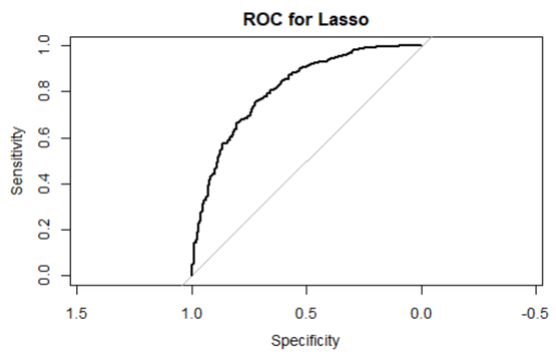
The Lasso and Ridge models demonstrated similar classification effectiveness:

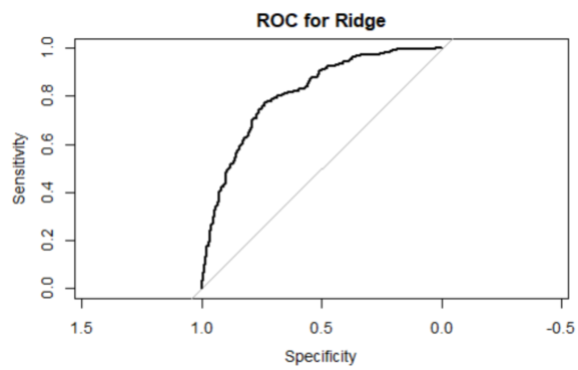
- **Accuracy**: Both models achieved over 76% accuracy, indicating a high level of correct predictions over the total number of cases.
- **Precision**: The Lasso model showed slightly higher precision than Ridge (78.65% vs. 78.22%), suggesting it was marginally better at predicting true positives.

- **Recall**: The recall for both models was high, over 85%, indicating a strong ability to identify all relevant cases.
- **F1 Score**: At just over 82%, the F1 scores for both models were robust, reflecting a balanced measure of precision and recall.
- **AUC-ROC**: The area under the ROC curve was approximately 0.81 for both models, signifying a strong discriminative ability.

ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves for both Lasso and Ridge (as shown in the accompanying figures) illustrate the models' true positive rate (sensitivity) against the false positive rate (1-specificity). Both curves ascend rapidly towards the top-left corner, indicating excellent model performance.





Feature Optimization with Recursive Feature Elimination (RFE)

In our ongoing efforts to refine our predictive model, we implemented the RFE methodology, a feature selection technique designed to identify the most significant predictors for our model. RFE systematically considers smaller and smaller sets of features, recursively removing the least significant features to optimize model performance.

RFE Outcomes:

Note: Due to challenges encountered with R, we employed Python for RFE

The execution of RFE on our dataset resulted in the selection of a robust subset of features deemed most informative for our predictive model. These features, listed below, represent the variables that RFE identified as having the highest predictive power for our target variable are latitude,longitude, age_last_funding_year,age_first_milestone_year,age_last_milestone_year,relationships, funding_total_usd, milestones, avg_participants

Refined Predictive Analysis with Lasso and Ridge Regression

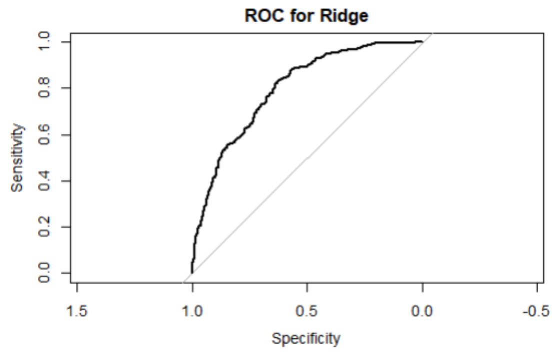
Upon identifying crucial features through Recursive Feature Elimination (RFE), we proceeded with Lasso and Ridge regression models to predict our binary target variable. The cross-validation process optimized the regularization strength via the lambda parameter, ensuring the models were neither overfitting nor underfitting.

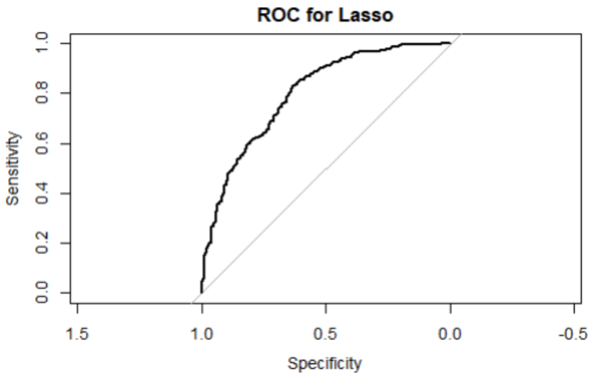
Model Performance Highlights:

- **Lasso Regression**: Achieved an Area Under the Curve (AUC) of 0.7992, reflecting a strong ability to distinguish between the classes.

- **Ridge Regression**: Slightly lower AUC of 0.7966, yet close to Lasso's performance, indicating robustness across regularization techniques.
- **Classification Metrics**: The Lasso model demonstrated a balanced accuracy of 72.26%, with a sensitivity of 56.75% and specificity of 87.77%.

The Receiver Operating Characteristic (ROC) curves for both models, illustrating their sensitivity and specificity, validated the discriminative power of the models, as seen in the attached figures.





Following our feature selection through RFE, Lasso, and Ridge regression, we conducted a Random Forest classification to evaluate performance across the entire feature set.

Random Forest Model Execution

Using the **randomForest** and **caret** libraries, we partitioned our scaled dataset into training and test sets with an 80/20 split. A Random Forest model was then trained on the training set.

Model Performance Metrics:

- **Accuracy**: The model achieved an accuracy of 81.52%, demonstrating a high level of correct predictions.
- **Precision**: With a precision of 81.02%, the model reliably predicted the positive class.
- **Recall**: The recall rate was 93.28%, indicating the model's strength in identifying true positives.
- **F1 Score**: The F1 score, stood at 86.72%, suggesting a balanced model.

The confusion matrix generated from the model's predictions revealed a strong predictive capability, particularly in identifying true positives (111 out of 119 actual positives).

Focused Random Forest Classification on Selected Features

Transitioning from a comprehensive Random Forest classifier, our methodology evolved to concentrate on pivotal features ascertained through Recursive Feature Elimination (RFE) and refined by Lasso and Ridge regression techniques.

Model Metrics with Selected Features:

- **Accuracy**: Achieved an impressive 79.89%, signifying reliable prediction capabilities.
- **Precision**: At 81.06%, indicating the model's precision in predicting positive outcomes remains high.

- **Recall**: 89.92%, showcasing the model's strength in identifying a large proportion of actual positive cases.
- **F1 Score**: 85.26%, reflecting a strong predictive balance between precision and recall.

Comparative Assessment:

Upon comparing this focused model to the base Random Forest model—which utilized the entire suite of features—we observed the following:

- The base model demonstrated marginally superior accuracy and recall, potentially benefiting from a wider breadth of data, albeit at the risk of increased complexity and overfitting.
- Precision was consistent between the two models, underscoring that the essential predictive quality is preserved even after feature reduction.
- The slight dip in the F1 score for the refined model indicates a small trade-off in achieving a perfect balance between precision and recall, which is often the case when the model simplifies to focus on key features.

Decision Tree Performance on Complete Feature Set

In our exploratory analysis, a Decision Tree model was applied to the entire set of features to benchmark its performance against more complex models like Random Forest.

Decision Tree Model Evaluation:

Accuracy: The model achieved a solid 78.80% accuracy, indicating its effectiveness in classifying the data.

Precision: It demonstrated precision at 78.17%, reflecting its capability to correctly identify positive outcomes.

Recall: The recall was high at 93.28%, illustrating the model's strength in capturing most actual positive instances.

F1 Score: With an F1 score of 85.06%, the model showed a balanced measure of precision and recall.

Model Metrics Comparison:

The Decision Tree model's metrics were commendably close to those of the Random Forest classifier, despite the latter's inherent complexity and ensemble approach.

Summary

In this project, we investigated the predictive modeling of company success in venture capital, using advanced statistical techniques like linear and logistic regression, Lasso and Ridge regression, and Random Forest classification. Our findings highlight the significant impact of early funding, milestone achievement, robust networking, and diverse funding sources on startup success. Particularly, the importance of state-specific trends, such as those in California and New York, emerged as crucial factors. Our models, especially the Random Forest, demonstrated high accuracy and recall, emphasizing the predictive power of relationships, milestones, and top 500 rankings. This project not only sheds light on the complex dynamics of venture capital decision-making but also provides actionable insights for investors and startups to strategize their growth and investment approaches effectively.

Future Scope

Future predictive models of company success will likely integrate real-time data, leverage advanced NLP to analyze diverse textual sources, and employ sophisticated AI algorithms for uncovering complex patterns, ultimately leading to more accurate and dynamic predictions that adapt to the ever-evolving startup landscape.