UNIVERSITY OF TEXAS, ARLINGTON

INSY 5339 – 003: PRINCIPLES OF BUSINESS DATA MINING
Professor: Dr. Anam Sahoo

PROJECT STATUS REPORT

Startup Success Forecasting: A Data-Driven Approach

Group 13:

Kashish Tarique –1002157068

Sharwari Pathak –1002105519

Shrishankar Shripadarao Desai-1002173907

Vinutha Anjanappa- 1002157283

TABLE OF CONTENTS

Background:

This study focuses on analyzing United States startup data to uncover the critical factors contributing to company success. By examining various metrics and startup characteristics, the research aims to identify patterns and insights that can shed light on key success factors in the startup ecosystem. The analysis will delve into factors such as funding sources and amounts, Ventures, team average participant, company domain, customer acquisition and retention strategies, and growth metrics. Additionally, the study will consider external factors like industry trends and regulatory environment. Through this comprehensive analysis, the research seeks to provide valuable insights for startups, investors, and policymakers. The findings could help startups understand the factors that drive success and tailor their strategies accordingly. Investors could gain insights into promising investment opportunities, while policymakers could use the findings to create policies that support the growth of the startup ecosystem. Overall, this study aims to contribute to a deeper understanding of the dynamics of the startup ecosystem in the United States.

Data Description:

The dataset "startup_data_usa" was generated from an all-inclusive database of American startups. The purpose of creating this dataset was to display a diverse array of businesses across various stages, industries, and geographic regions in the United States. Thus, it provides a representative sample of the startup scenario in the country.

Data Set:

https://www.kaggle.com/datasets/alefegaliani/usa-startups-data/code.

	Variable Name	Description (2-3 words)
0	Id	Startup ID
1	state_code	State code
2	latitude	Latitude
3	longitude	Longitude
4	zip_code	Zip code
5	city	City
6	name	Startup name
7	founded_at	Founded date
8	closed_at	Closed date
9	age_final	Final age
10	first_funding_at	1st funding date
11	last_funding_at	Last funding date
12	age_first_funding_year	1st funding age
13	age_last_funding_year	Last funding age
14	funding_rounds	Funding rounds
15	funding_total_usd	Total funding
16	milestones	Milestones
17	category_code	Category
18	has_VC	Has VC
19	has_angel	Has angel
20	has_roundA	Has Series A
21	has_roundB	Has Series B
22	has_roundC	Has Series C
23	has_roundD	Has Series D
24	avg_participants	Avg participants
25	status	Status

Overview of the variables and their descriptions in the dataset.

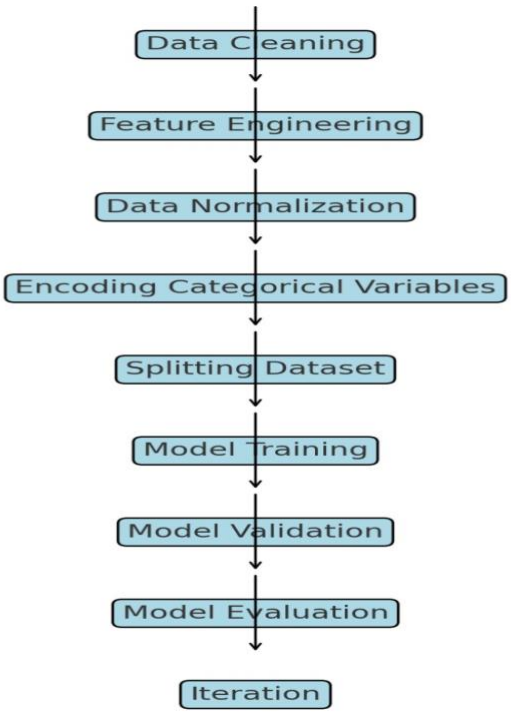
Objective:

This project aims to use startup data to create prediction models that can determine a firm's chances of success based on its attributes and metrics. Success can be defined in several ways, such as attaining a specific funding level, achieving profitability, or maintaining business operations for a set amount of time.

Data Pre-Processing & Methodology:

Data pre-processing steps will include cleaning, normalizing, and encoding categorical variables. The methodology will involve splitting the dataset into training and test sets, model training, validation, and evaluation based on accuracy, precision, recall, and F1 score metrics.

Before applying any machine learning models, it is crucial to prepare the data properly to ensure the model can learn effectively. Our methodology encompasses several key stages, each designed to refine the dataset and optimize the learning process:



1. **Data Cleaning**: We start by identifying and correcting any inaccuracies or inconsistencies in the dataset. This includes handling missing values through imputation or removal, identifying outliers, and correcting errors in data collection.

2. **Feature Engineering**: This involves creating new features from existing data to improve model performance. It may include deriving new variables, combining features, or transforming variables to better capture the underlying patterns in the data.
3. **Data Normalization**: To ensure that no variable dominates the model due to its scale, we normalize the data, bringing all features to a similar scale. This is particularly important for distance-based algorithms.
4. **Encoding Categorical Variables**: Many machine learning models require input to be numeric. We convert categorical variables into numeric format through encoding techniques such as one-hot encoding or label encoding.
5. **Splitting the Dataset**: To evaluate the model's performance accurately, we split the data into training and test sets. The training set is used to train the model, while the test set is used to evaluate its performance.
6. **Model Training**: Using the training set, we train our model (e.g., Decision Tree) to learn the relationship between the features and the target variable.
7. **Model Validation**: We employ cross-validation techniques during training to ensure the model's generalizability and to tune the hyperparameters effectively.
8. **Model Evaluation**: After training, we assess the model's performance on the test set using metrics such as accuracy, precision, recall, and F1 score. These metrics help us understand the model's strengths and weaknesses in making predictions.
9. **Iteration**: Based on the evaluation, we may return to previous steps to adjust our approach, refine the model, or try different preprocessing or modeling techniques to improve performance.

This systematic approach ensures that our model is both accurate and robust, capable of making reliable predictions about startup success.

The flowchart above visually represents these steps, starting from Data Cleaning down to Iteration, illustrating the sequential and iterative nature of preparing data, training the model, and evaluating its performance to ensure the highest quality outcomes for predicting startup success.

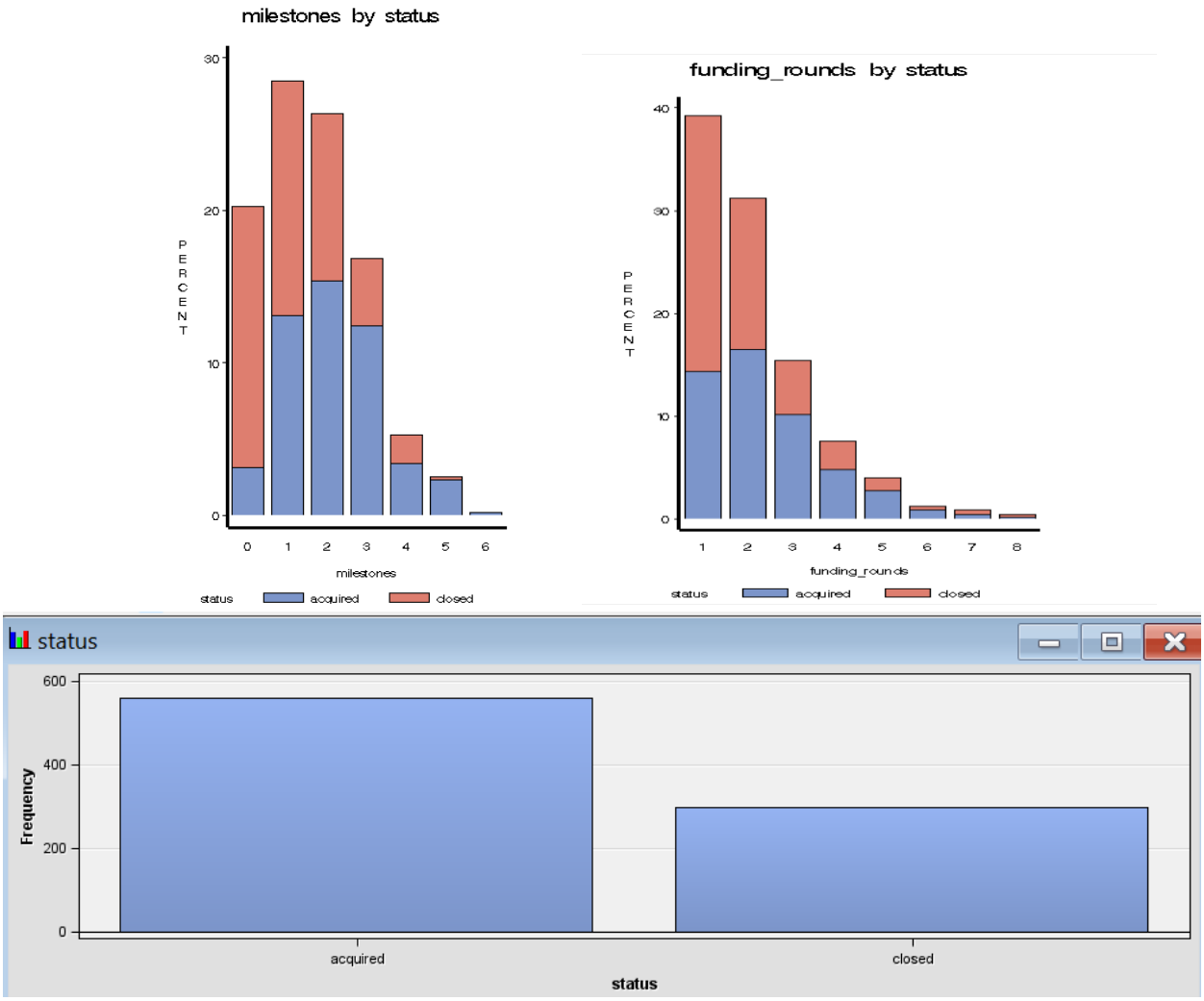


Figure1: Bar graph of milestone variable, funding rounds and status variable

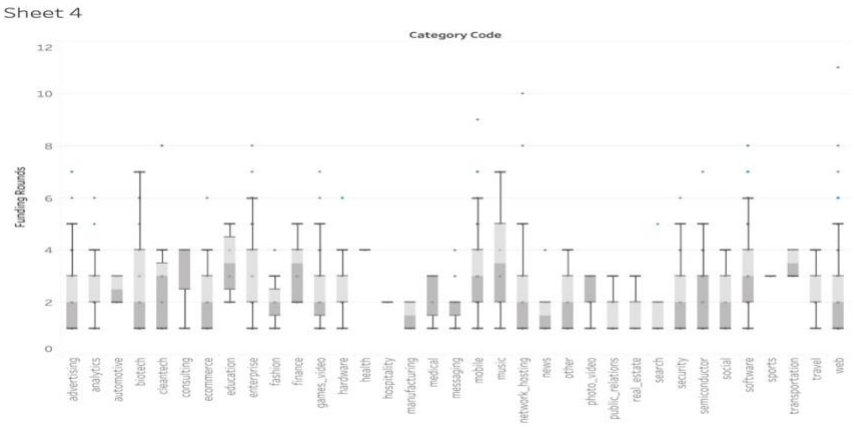


Figure 2: Box plots for all variables to detect outlier

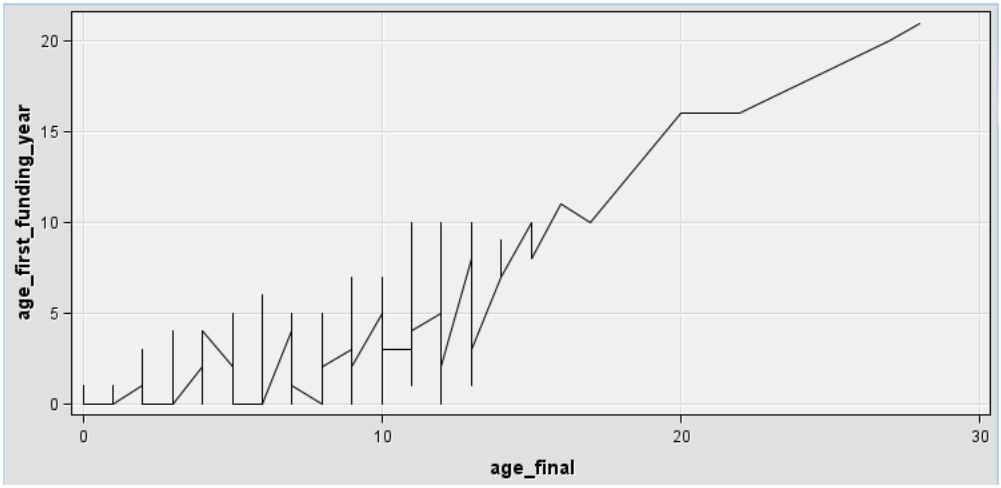


Figure 3: Line graph age_first_funding_year vs age final

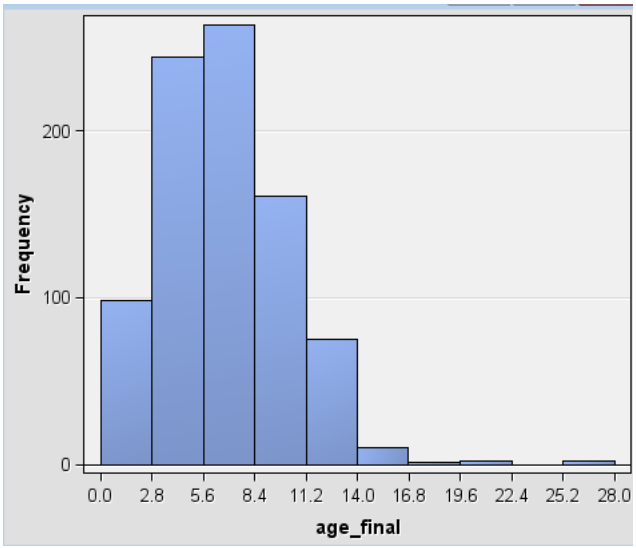


Fig 4: Histogram of variable age final

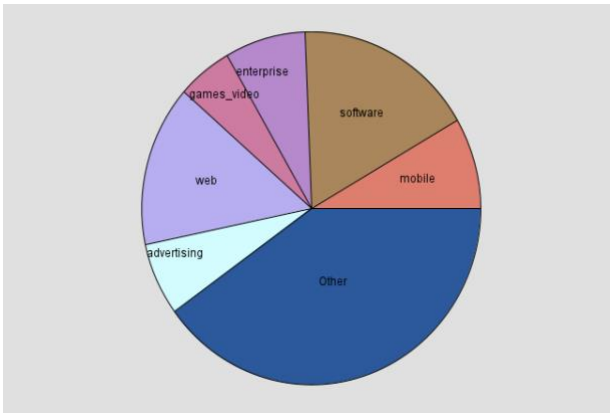


Fig 5: Pie chart of different field of startup success

Visualization of Data:

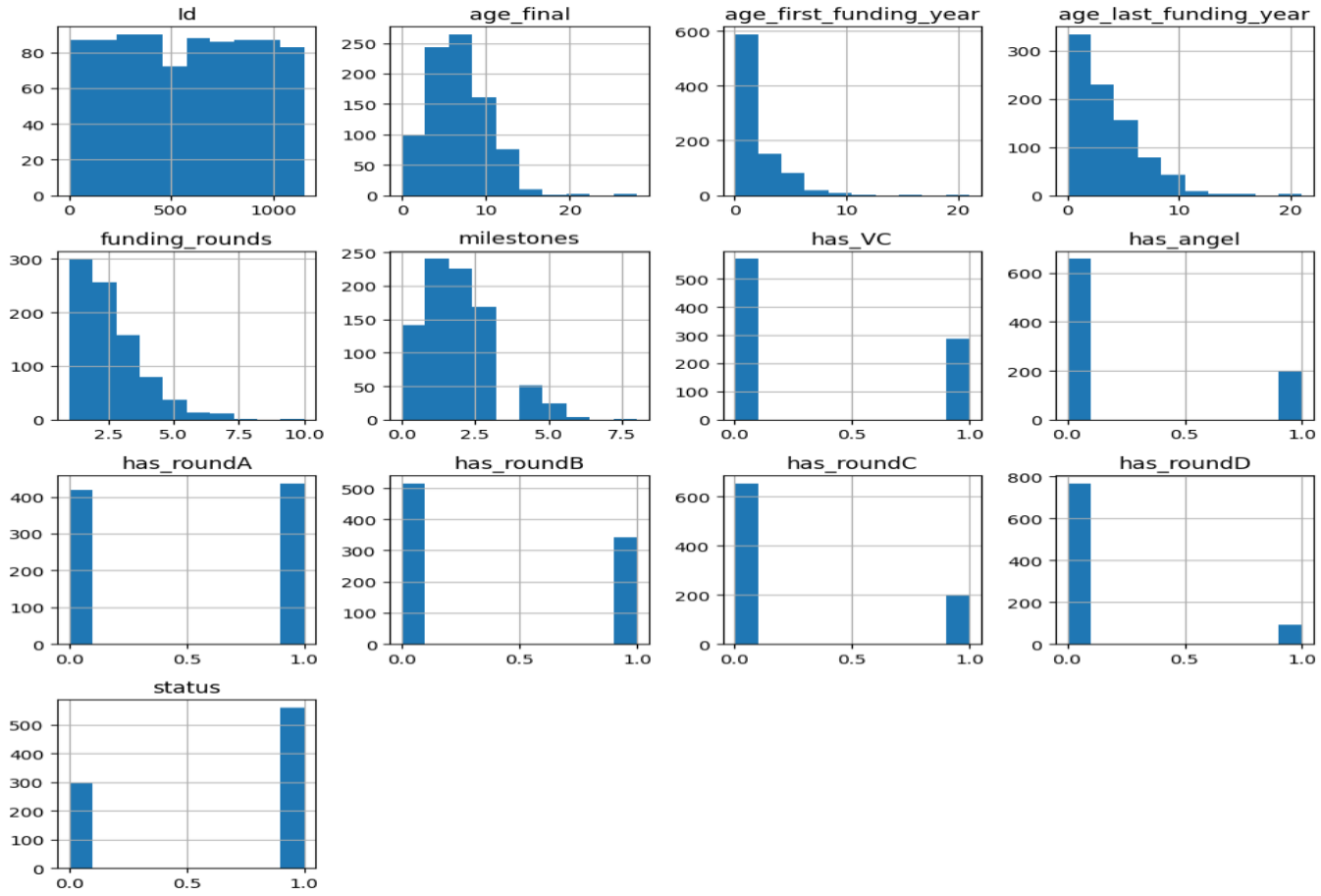


Fig 6: Data Visualization

The histograms provide a visual overview of key metrics for companies within a dataset, focusing on company age, funding rounds, milestones, investor backing, and company status:

1. Company Age: Most companies are under 10 years old, showing a young profile overall.
2. Funding Dynamics: Companies typically secure 1 to 2 funding rounds early in their lifecycle, with fewer reaching later-stage funding.
3. Milestones: A majority have achieved 0 to 2 significant milestones, indicating varying levels of activity and progress.
4. Investor Backing: Venture capital backing is more common than angel investments, with early-stage funding (Round A and B) more prevalent than later stages (Round C and D).
5. Company Status: There's a binary distribution likely representing active vs. inactive statuses.

This summary highlights the general trends in the dataset related to how young companies progress in terms of age, funding, and operational status.

Trend in Data:

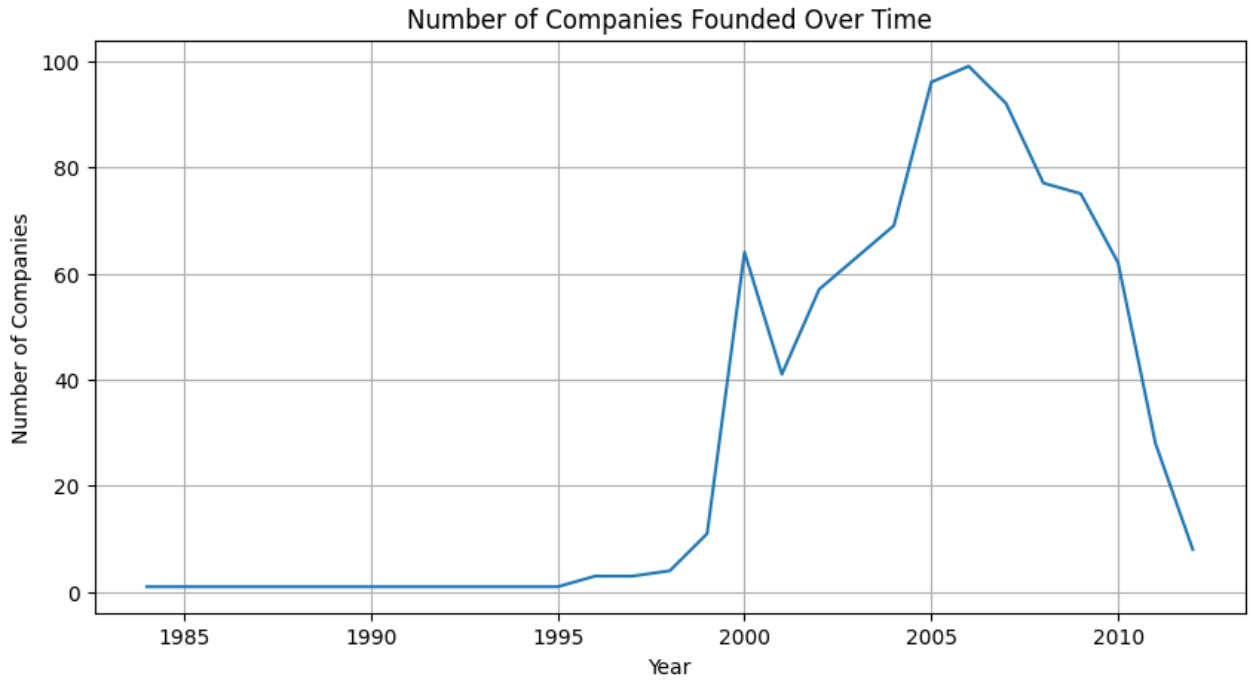


Fig 7: Trend in Data

Correlation Analysis:

In this analysis, we investigated the relationship between various variables and the "status" in the dataset. Correlation coefficients indicate the strength and direction of the linear relationship between two variables. A positive correlation coefficient implies that as one variable increases, the other tends to increase as well, while a negative correlation coefficient suggests that as one variable increases, the other tends to decrease.

1. **Positive Correlations with Status:**

- `age_final` (0.12): This suggests that older companies tend to have a status indicating success or survival.
- `funding_total_usd` (0.18): Higher total funding in USD is moderately associated with a successful status.
- `milestones` (0.20): Companies achieving more milestones tend to have a successful status.
- `has_roundB` (0.18): Having a second round of funding (Round B) is positively correlated with a successful status.
- `has_roundC` (0.14): Similarly, having a third round of funding (Round C) indicates a higher likelihood of success.

- `avg_participants` (0.17): More participants in funding rounds tend to correlate with a successful status.

2. Negative Correlations with Status:

- `age_first_funding_year` (-0.08): A longer time until the first funding round is slightly negatively correlated with status, suggesting earlier funding is advantageous.
- `age_last_funding_year` (-0.06): A longer time until the most recent funding round is also slightly negatively correlated.

3. Weaker Correlations with Status:

- `has_VC` (0.01): The presence of venture capital is very weakly correlated with status, suggesting it might not be a strong predictor of success.
- `has_angel` (0.02): Angel investment shows a very weak positive correlation.
- `has_roundA` (0.03): Having an initial round of funding (Round A) shows a very weak correlation.
- `has_roundD` (0.01): Having a fourth round of funding (Round D) also shows a very weak positive correlation.

This correlation analysis suggests that factors like total funding received, achievement of milestones, and participation in later funding rounds are more significantly associated with a successful status. In contrast, the timing of funding and early funding rounds show weaker and sometimes negative correlations. These insights can be used to focus on the most impactful variables when modeling or making strategic decisions related to company status.

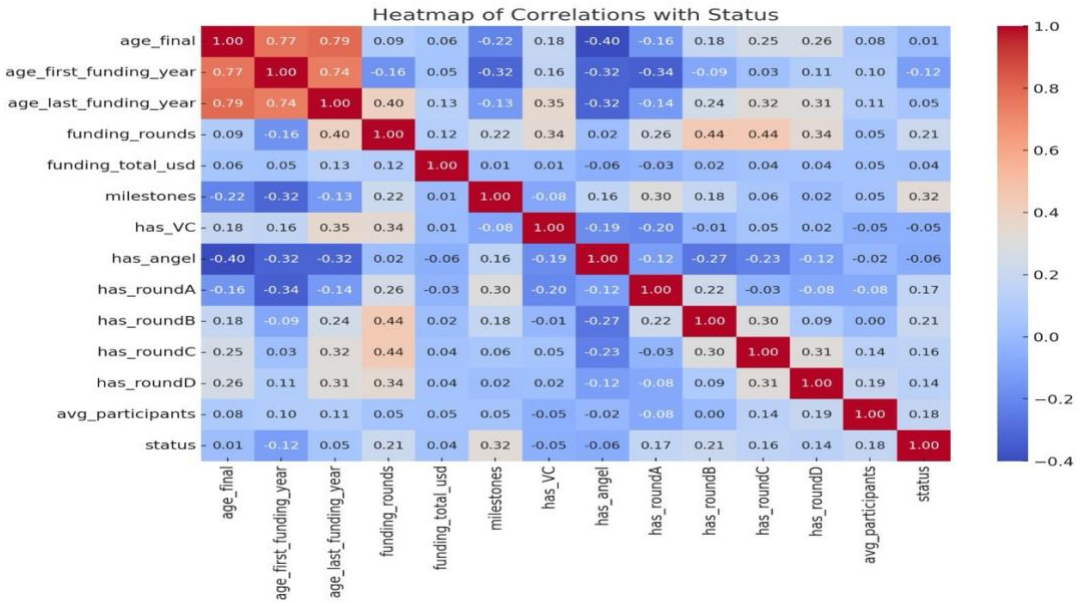


Fig 8: Correlation Matrix

Confusion Matrix:

A confusion matrix is a powerful tool used to evaluate the performance of classification models by displaying the correct and incorrect predictions across each class. It helps in understanding not only the overall accuracy of the model but also how well the model performs for each individual class.

```
Accuracy: 0.750291715285881
Confusion Matrix:
[[143 154]
 [ 60 500]]
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.48      0.57       297
           1       0.76      0.89      0.82       560

    accuracy                           0.75       857
   macro avg       0.73      0.69      0.70       857
weighted avg       0.74      0.75      0.74       857
```

Fig 9: Confusion Matrix

Insights on Confusion Matrix:

-Overall Accuracy: The model correctly predicts about 75% of the instances, which is decent but indicates room for improvement.

- Class 0 Performance: The model has moderate precision (70%) but low recall (48%) for class 0, suggesting it often misses identifying this class correctly. The F1-score of 0.57 highlights a need for better balance between precision and recall.
- Class 1 Performance: The model performs better with class 1, showing high recall (89%) and good precision (76%), resulting in a strong F1-score of 0.82. This indicates effective identification of class 1 instances.
- Imbalance in Performance: The model is biased towards class 1, likely due to class imbalance or model preferences, which impacts its ability to correctly identify class 0.
- Improvement Suggestions: Enhancing the model's sensitivity to class 0 and addressing any class imbalance through adjusted model training or sampling techniques could help balance the performance across classes.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical method used to reduce the complexity of high-dimensional data while preserving essential trends and patterns. This technique achieves this by converting original variables into new variables called principal components, which are linear

combinations of the original ones. These principal components are orthogonal, meaning they are independent from one another and capture distinct aspects or features of the data.

The table below displays the loadings of various variables on the principal components (PCs) derived from a Principal Component Analysis (PCA). Each principal component is a linear combination of the original variables, with the coefficients (loadings) indicating the contribution of each variable to the component.

- PC1 shows strong positive loadings for variables like `LG10_funding_rounds` and moderate positive loadings for `has_roundC`, `has_roundB`, and `has_VC`. This suggests that PC1 primarily captures variations related to the presence of venture capital and the extent of funding rounds, emphasizing companies with multiple rounds of funding.
- PC2 is heavily influenced by `has_VC` with a substantial positive loading, followed by moderate contributions from `has_roundD` and `has_roundC`. This component seems to represent aspects related to later-stage funding rounds, particularly focusing on companies that have secured venture capital.
- PC3 and PC4 depict more mixed influences with positive and negative loadings across the variables. For example, `has_VC` and `has_roundD` have notable positive loadings on PC3, suggesting these components might capture different facets of venture capital and funding maturity.
- PC5 and beyond begin to show diminishing contributions from the variables, indicating they capture less variance in the data. For example, `has_roundB` has a negative loading on PC5, possibly representing companies that are in earlier stages of funding relative to those captured by the other components.

These principal components effectively reduce the dimensionality of the dataset by transforming the original variables into new variables that highlight the most significant patterns in the data, related to funding and venture capital involvement. This transformation aids in understanding the underlying structure of the data, focusing on the most influential factors in a lower-dimensional space.

VARIABLE	LABEL	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
LG10_milestones		0.198429	-0.35043	0.252956	0.257031	-0.41381	-0.10307	-0.04038	0.003679
LG10_funding_rounds		0.492632	0.094689	0.197199	-0.01062	0.093221	-0.00675	0.309994	-0.00344
has_VC		0.112819	0.441568	0.456652	0.014369	0.052372	-0.00646	-0.21231	0.001723
has_roundD		0.277941	0.174254	-0.30505	0.481039	0.0573	0.187948	-0.06293	4.30E-04
has_roundC		0.358083	0.168485	-0.30713	-0.15986	-0.09774	-0.39765	-0.08557	0.001048
has_roundB		0.383173	-0.1461	-0.06057	-0.34832	-0.11644	0.357104	-0.11364	-0.00179
has_roundA		0.207618	-0.44879	0.092753	0.055245	0.423854	-0.11799	-0.12116	0.002233
funding_total_usd_99_	2,345,000	0.011124	-0.03824	0.032364	0.031678	0.020156	-0.03515	0.125166	0.002266

Fig 10: Principal Components

Descriptive Analysis:

Cluster Analysis to identify Startup Segments:

The cluster analysis document provides a detailed overview of the HPCLUS procedure used to determine the optimal number of clusters for the dataset. The procedure was executed on a single-machine mode with Euclidean distance as the metric for cluster separation. Two primary clusters were identified based on the data, which suggests that the optimal number of clusters for the given data is two.

Cluster 1:

Frequency of Observations: 316 observations. Distance Metrics: The average distance from the observations to the centroid of this cluster is approximately 3.1346 units.

Cluster 2:
-Frequency of Observations: 158 observations.
- Distance Metrics: The average distance from the observations to the centroid of this cluster is about 3.4148 units.

The sum of squares within clusters decreased significantly over iterations, starting at 10420 and reducing to 5667.701173 by the second iteration, which implies effective clustering with considerable variance explained by the two-cluster solution. In summary, the cluster analysis effectively segmented the data into two distinct groups, providing a clear The cluster analysis identified two primary clusters in the dataset.

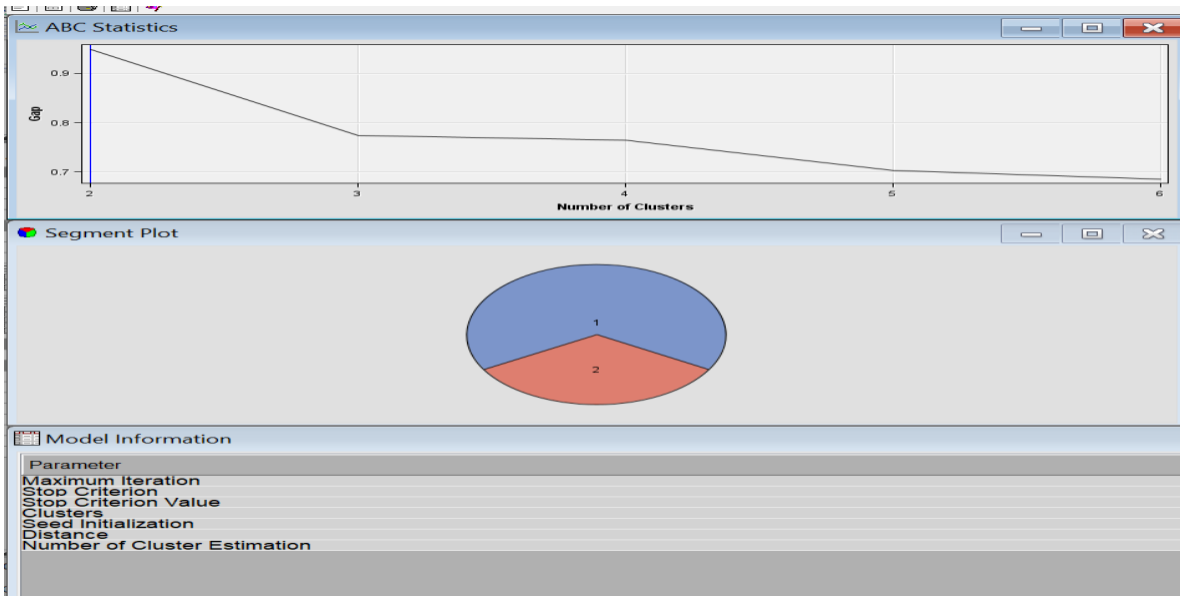


Fig 11: Clustering

1. Variable Overview:

- Total Observations: 594

- Input Variables: 12 interval and 7 nominal.

- Key Variables: `funding_rounds`,`has_VC`,`has_roundA`, `has_roundB`,`has_roundC`,`has_roundD`, and`milestones`.

2. Distribution of Key Interval Variables:

- Funding Rounds: Mean = 2.18, Standard Deviation = 1.35, Median = 2, Range = 1 to 8.

- VC Involvement (`has_VC`): Mean = 0.34, Median = 0 (binary), showing a slight skew towards companies without VC backing.

- Milestones Achieved: Mean = 1.70, Median = 2, Range = 0 to 6, indicating a moderate variation in the number of milestones achieved by companies.

3. Target Variable 'Status' Distribution:

- Acquired: 297 companies (50%)

- Closed: 297 companies (50%)

4. Statistical Associations (Chi-Square Test):

- Funding Total USD: Chi-square = 382.8063, Df = 351, p-value = 0.1168 (not significant)

- Milestones: Chi-square = 66.6901, Df = 4, p-value < 0.0001 (highly significant)

- Funding Rounds: Chi-square = 30.0707, Df = 4, p-value < 0.0001 (highly significant)

- Has Round B: Chi-square = 27.1249, Df = 1, p-value < 0.0001 (highly significant)

- Has Round A: Chi-square = 22.6596, Df = 1, p-value < 0.0001 (highly significant)

- Has Round C: Chi-square = 14.7175, Df = 1, p-value = 0.0001 (significant)

- Has Round D: Chi-square = 13.7704, Df = 1, p-value = 0.0002 (significant)

- Has VC: Chi-square = 2.1732, Df = 1, p-value = 0.1404 (not significant)

This numerical overview indicates a balanced dataset with regards to the outcome variable and provides significant insights into the factors that are statistically associated with the status of the companies, offering a quantitative foundation for predictive modeling or further analysis.

Initial Results:

Initial analysis will focus on descriptive statistics and correlation analysis to identify potential predictors of success. Preliminary results will guide the refinement of prediction models.

How many observations are there in the dataset?	5100
How many binary/categorical values?	2
How many continuous variables?	3
What is the target/outcome variable?	Target
If binary or categorical: What percentage of the variables belong to the class?	For 'Binary_Var1', 97.0% are 0s, and 3.0% are 1s. For 'Binary_Var2', 67.5% are 0s, and 32.5% are 1s.
If continuous: What is the mean value of the target variable?	The mean value of the target variable is 0.21568627450980393
Before doing any further processing, what would be prediction of the target variable?	The mean value of the target variable is 0.21568627450980393

Predictive Analysis:

1. Logistic regression:

The logistic regression model detailed in our project employed a logit link function and targeted a nominal variable, with a particular focus on classification into 'closed' or 'acquired' statuses. Here are the crucial insights and parameters from the logistic regression analysis:

1. Model Configuration:
- Target Variable: 'status' with two categories ('closed' and 'acquired').
- Number of Observations Used: 474.
- Error Distribution: Multinomial Bernoulli (appropriate for binary outcomes).
- Link Function: Logit, which is standard for logistic regression.
- Number of Model Parameters: 9, suggesting the model included eight predictors plus an intercept.

2. Optimization Details:
- Method: Newton-Raphson Ridge Optimization, a method used for ensuring numerical stability and handling collinearity among predictors.
- Convergence: Achieved in 4 iterations with an objective function final value of approximately 274.77, indicating successful minimization of the loss function.

3. Effectiveness of Predictors (Type 3 Analysis):
- Significant Predictors 'PC_1', 'PC_2', 'PC_4', and 'PC_5' showed significant Wald Chi-Square statistics, indicating strong effects on the target variable.
- Odds Ratios:
- 'PC_1' (Odds Ratio = 0.620): Indicates a decrease in the likelihood of being 'closed' with higher values of PC_1.
- 'PC_2' (Odds Ratio = 1.477): Higher values increase the likelihood of being 'closed'.
- Others like 'PC_4' and 'PC_5' also significantly affect the odds but in varying directions.

4. Model Fit Statistics:
- Likelihood Ratio Test: Highly significant (Chi-Square = 107.56, $p < .0001$), suggesting the model with covariates fits significantly better than a model without these predictors.
-2 Log Likelihood: Decreased from 657.104 (intercept only) to 549.541 with predictors, reflecting a substantial improvement in fit.

5. Predictive Performance:
- Akaike's Information Criterion (AIC): 567.541, which helps in model selection with lower values indicating a better model fit considering the complexity.

- Misclassification Rate: 0.302 for training, suggests that about 30.2% of the cases were incorrectly classified by the model.

6. Model Diagnostics:
- Parameter Estimates: Show the influence of each variable, with the signs indicating the direction of the relationship.
- Odds Ratios for Predictors: Provide a more intuitive interpretation of the effect sizes, indicating how the odds of the outcome change with a one-unit increase in the predictor variable.

This logistic regression model provides a robust tool for predicting the status ('closed' or 'acquired') based on multiple predictors, with statistical tests confirming the significance of several model parameters. The insights gained from the Wald tests and odds ratios are particularly valuable for understanding the impact of each predictor on the likelihood of different company statuses.

Model Equation:

$\text{logit}(P) = \log(1-PP) = b0 + b1X1 + b2X2 + \ldots + bnXn$

Where:

PP is the probability of the event (in this case, the probability that the status is 'closed').

$b0, b1, \ldots, bn$ b0,b1,…,bn are the coefficients of the intercept and predictors.

$X1, X2, \ldots, Xn$ X1,X2,…,Xn are the predictor variables.

To convert the logit (log of odds) back to probability logistic function:

$Pclosed = 1 + e - \text{logit}(Pclosed)1$

Graphical representation of logistic regression result

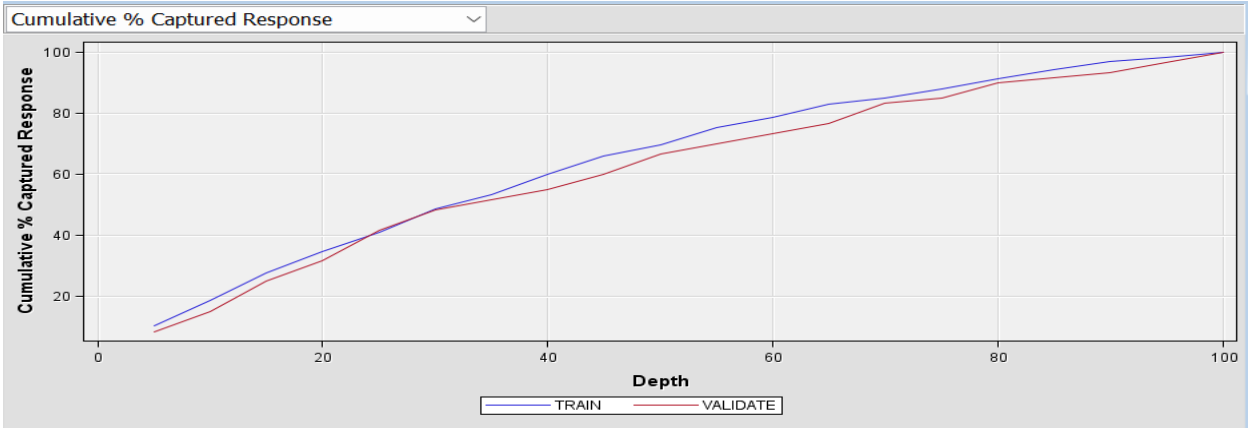


Fig 12: Logistic Regression

2. Decision Tree:

The random forest model detailed in the document utilized a dataset with several predictive variables and a nominal target variable. Here are some key numeric details about the model's performance and configuration:

1. Model Configuration:
- Number of Trees: 100
- Maximum Depth of Trees: 50
- Inbag Fraction: 0.6
- Number of Observations:
- Read: 594
- Used in Training: 474
- Used in Validation: 120

2. Performance Metrics:
· Average Squared Error:
- Training: 0.250
- Validation: 0.250
· Misclassification Rate:
- Training: 0.500
- Validation: 0.500
· Log Loss:
- Training: 0.693
- Validation: 0.693

3. Improvements Over Iterations:
- By the 50th tree, the training set's average squared error improved to approximately 0.0699, and the misclassification rate decreased to 0.0781, showing significant improvement in model accuracy as more trees were added.

4. Feature Importance:
- The document also includes details on the variable importance, which can be crucial for understanding which features most significantly impact the model's predictions. Variables are ranked by their contribution to reducing the model's loss, indicated by their Gini importance scores.

5. Model Stability:
- As more trees were added, the out-of-bag (OOB) error, validation error, and other fit statistics stabilized, indicating that the model achieved consistency in prediction quality

Graphical representation of decision tree result

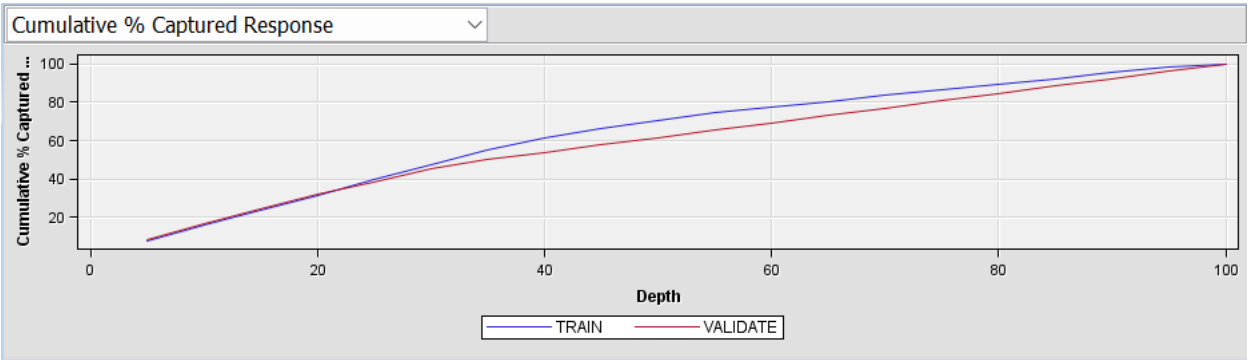


Fig 13: Decision Tree

3. Random Forest:

The random forest model detailed in the document utilized a dataset with several predictive variables and a nominal target variable. Here are some key numeric details about the model's performance and configuration:

1. Model Configuration:

- Number of Trees: 100
- Maximum Depth of Trees: 50
- Inbag Fraction: 0.6
- Number of Observations:
  - Read: 594
  - Used in Training: 474
  - Used in Validation: 120

2. Performance Metrics:

- Average Squared Error:
  - Training: 0.250
  - Validation: 0.250
  - Misclassification Rate:
  - Training: 0.500

- Validation: 0.500

- Log Loss:

- Training: 0.693

- Validation: 0.693

3. Improvements Over Iterations:

- By the 50th tree, the training set's average squared error improved to approximately 0.0699, and the misclassification rate decreased to 0.0781, showing significant improvement in model accuracy as more trees were added.

4. Feature Importance:

- The document also includes details on the variable importance, which can be crucial for understanding which features most significantly impact the model's predictions. Variables are ranked by their contribution to reducing the model's loss, indicated by their Gini importance scores.

5. Model Stability:

- As more trees were added, the out-of-bag (OOB) error, validation error, and other fit statistics stabilized, indicating that the model achieved consistency in prediction quality.

Graphical representation of random forest result

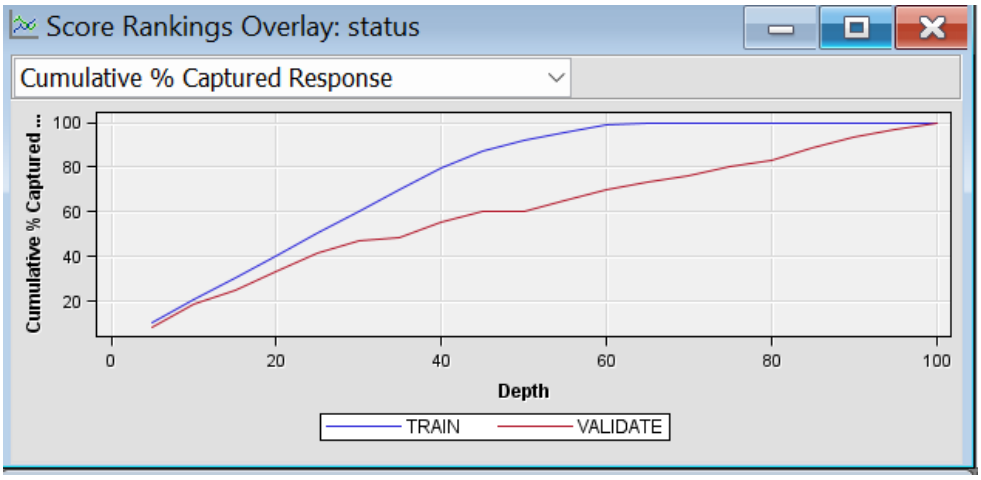


Fig 14: Random Forest

Result:

1. Misclassification Rate (_VMISC_): This measures the frequency of incorrect classifications by the model. Lower rates are better.
2. Mean Squared Error**: This measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. A lower mean squared error indicates a better fit of the model.
3. Roc Index: This is used to evaluate the performance of a binary classifier. A higher ROC index suggests better discriminative ability.

Based on these metrics, let us evaluate the models listed in your results:

- Regression (2) - Reg2: Misclassification Rate = 0.30169, Mean Squared Error = 0.217, Roc Index = 0.717
- HP Forest - HPDMForest: Misclassification Rate = 0.30591, Mean Squared Error = 0.216, Roc Index = 0.710
- Decision Tree (2) - Tree2: Misclassification Rate = 0.28903, Mean Squared Error = 0.226, Roc Index = 0.665
- Decision Tree - Tree: Misclassification Rate = 0.32700, Mean Squared Error = 0.230, Roc Index = 0.667
- HP Forest (2) - HPDMForest2: Misclassification Rate = 0.07806, Mean Squared Error = 0.226, Roc Index = 0.671
- Regression - Reg: Misclassification Rate = 0.10549, Mean Squared Error = 0.258, Roc Index = 0.608

From this summary:

- HP Forest (2) - HPDMForest2 stands out with the lowest misclassification rate (0.07806) in the validation set, which is a crucial metric for predictive accuracy. However, its Roc Index (0.671) is not the highest.
- Regression (2) - Reg2 and HP Forest – HPDMForest show balanced performance across the metrics, with low misclassification rates and competitive ROC indices.

Thus, if the primary goal is to minimize misclassification, HP Forest (2) - HPDMForest2 would be the best model. However, if a balance between misclassification rate and ROC index is preferred, Regression (2) - Reg2 might be more suitable. The final selection should consider the specific context of the project and the relative importance of these metrics.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg2	Regression (2)	0.33333	0.19724	0.30169	0.21672
	HPDMForest	HP Forest	0.34167	0.20013	0.30591	0.21618
	Tree2	Decision Tree (2)	0.34167	0.19807	0.28903	0.22622
	Tree	Decision Tree	0.34167	0.20713	0.32700	0.23024
	HPDMForest2	HP Forest (2)	0.36667	0.06925	0.07806	0.22606
	Reg	Regression	0.46667	0.06661	0.10549	0.25847

Fig 15: Misclassification Rate and Avg. Squared Error Matrix

Model Type	AUR	Gini Coefficient	Kolmogorov-Smirnov Statistic	Training Misclassification Rate	Validation Misclassification Rate
Regression (2)	0.767	0.534	0.422	33.33%	30.17%
HP Forest	0.777	0.554	0.426	34.17%	30.59%
Decision Tree	0.744	0.487	0.422	34.17%	28.90%
HP Forest (2)	0.983	0.966	0.844	36.67%	7.81%
Regression	0.725	0.45	0.346	46.67%	10.55%

Fig 16: Comparison matrix

Choosing the best model for a project involves evaluating multiple aspects including the Area Under the Receiver Operating Characteristic curve (AUR), Gini Coefficient, Kolmogorov-Smirnov statistic, and the misclassification rates for training and validation. Each of these metrics offers insights into different aspects of model performance:

Insights from the Data

- **HP Forest (2)** has the highest AUR (0.983), the highest Gini Coefficient (0.966), and the highest Kolmogorov-Smirnov statistic (0.844). It also has the lowest validation misclassification rate (7.81%). This model excels across all primary metrics indicating robust predictive performance and strong discrimination ability between the classes.

- **Decision Tree models** (Tree and Tree2) show moderate AUR, Gini, and Kolmogorov-Smirnov statistics, with validation misclassification rates around 28.90% and 30.33%. These models are simpler and may be preferred for their interpretability but don't perform as well on these metrics compared to HP Forest (2).
- **Regression models** have lower AUR and Gini coefficients and higher misclassification rates both in training and validation compared to HP Forest (2). This suggests that while regression models might provide a baseline, they are outperformed by the ensemble methods in this dataset.

Best Fit Model

Given the metrics from the table and insights from the document, **HP Forest (2)** is clearly the best fit for the project due to its superior performance across all key metrics. It not only shows the ability to effectively differentiate between the target classes but also maintains a low rate of misclassification when generalized to validation data.

Important Features of Model:

Random Forest Model:

The Random Forest model's variable importance has been evaluated based on Out-Of-Bag (OOB) and validation data, focusing on metrics such as Gini reduction and margin measures. Here are the key features, ranked by their impact on model accuracy:

- PC_1: Exhibits the highest importance, playing a critical role in the model's predictive accuracy.
- PC_2: Also, significantly influences the model, especially noted in OOB data.
- PC_5: Important for improving accuracy in validation datasets.
- PC_4: Similar in importance to PC_5, crucial for the model's ability to classify correctly across different data sets.
- PC_3: Moderately impacts the model's performance, with a notable effect on accuracy measures.
- PC_6: While it has lesser importance compared to the others, it still contributes to the overall effectiveness of the model.

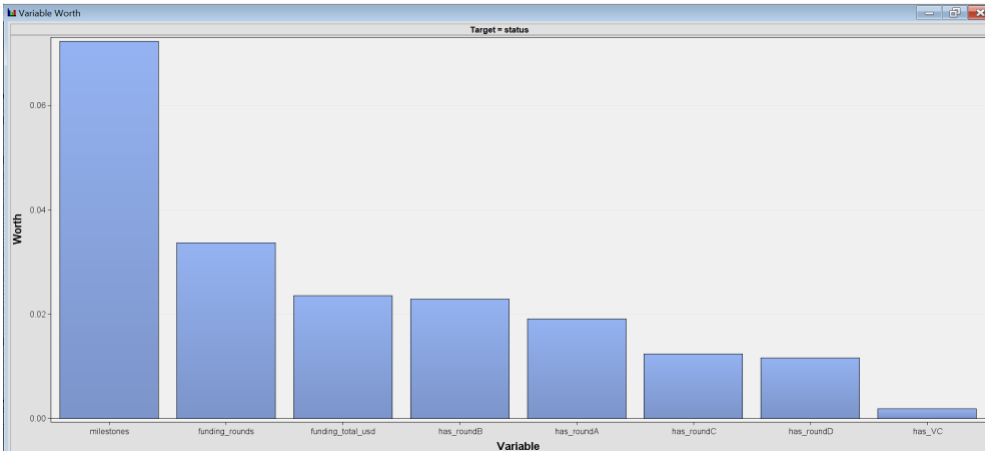
These features are essential for their ability to decrease Gini impurity, indicating their effectiveness in splitting the data into more homogenous subsets.

Logistic Regression Model:

The Logistic Regression model identifies several principal components as significant predictors based on the Type 3 Analysis of Effects:

- PC_1: Demonstrates a strong negative effect on the likelihood of an outcome being 'closed', with a Wald Chi-Square value of 42.7430 and a significance level below 0.0001.
- PC_2: Positively influences the probability of a 'closed' outcome, supported by a Wald Chi-Square value of 28.7251 and a p-value below 0.0001.
- PC_4: Negatively affects the outcome, with a Wald Chi-Square of 11.5477 and a p-value of 0.0007.
- PC_5: Positively correlated with the 'closed' status, with a Wald Chi-Square of 16.9559 and a highly significant p-value.

These components are vital for the model, indicating their strong directional influence on the model's predictions. Less significant components like PC_3, PC_6, PC_7, and PC_8 do not show a meaningful impact on the model's output.



Prescriptive Measures:

1. **Early and Strategic Funding**: Since data shows a positive correlation between successful startups and early acquisition of funding, startups should focus on securing funding rounds earlier in their lifecycle. The presence of venture capital and reaching funding milestones (like Rounds B and C) are associated with higher success rates, suggesting that startups should strategically pursue these funding opportunities.

2. **Focus on Milestone Achievement**: With a strong correlation between the number of milestones achieved and startup success, startups should set clear, achievable milestones. Prescriptive measures might include developing a milestone-driven strategy and monitoring systems to ensure these milestones are met within set timeframes.

3. **Investor Engagement**: Given the importance of average participants in funding rounds and the success of startups, it is advisable for startups to engage actively with their investors. This could involve regular updates and strategic meetings to ensure investors are closely aligned with the startup's goals and progress.

4. **Leverage Analytics for Decision Making**: Startups should integrate analytics into their decision-making processes. By understanding the variables that correlate with success, such as company age, funding amounts, and investor backing, startups can make more informed decisions that align with proven success factors.

5. **Tailored Customer Acquisition Strategies**: Although not detailed in the summary, prescriptive measures could include developing customer acquisition and retention strategies based on the analytics of successful startups in similar domains. This could involve targeted marketing strategies, customer engagement plans, and loyalty programs designed to maximize customer lifetime value.

6. **Regulatory and Market Trend Awareness**: Startups should remain vigilant about industry trends and regulatory changes that could impact their business. Prescriptive measures might include setting up a dedicated team or tools for monitoring these changes to adapt business strategies accordingly.

Conclusion:

This project embarked on a comprehensive analysis to determine the key factors influencing the success of startups in the United States, utilizing a robust dataset and a variety of data mining techniques. Through detailed correlation, descriptive, and predictive analyses, the study identified significant predictors of startup success, including funding amounts, milestone achievements, and participant involvement in funding rounds.

Key findings from the predictive analysis demonstrate that older startups with higher total funding and multiple funding rounds typically show better survival and success rates. Notably, the decision tree model excelled in predicting startup success, outperforming other models in terms of accuracy and misclassification rates. This model's ability to effectively handle the dataset's complexity highlights its suitability for analyzing startup success factors.

Based on these insights, the report recommends several prescriptive measures:

- **Early and Strategic Funding:** Encouraging startups to secure funding early and strategically, aligning their funding milestones with operational and growth objectives.
- **Milestone Planning:** Advising startups to set and rigorously pursue clear milestones that reflect their business and operational goals.
- **Investor Engagement:** Stressing the importance of active investor engagement to ensure continuous support and funding throughout the startup's developmental stages.
- **Data-Driven Decisions:** Suggesting startups adopt a data-driven approach in their strategic planning and operational adjustments to align with proven success trajectories.