Startup Success Prediction Report

Pau Chaves & Jan Aguiló

December 8, 2025

1 Problem statement

Venture capital firms continue to invest heavily in startups, yet choosing which ones to back remains a costly and uncertain process that often involves long evaluations and subjective decision-making. Despite technological progress and the availability of structured data, predicting which startups will succeed is still widely viewed as part science, part guesswork. As two students deeply interested in entrepreneurship and the startup world, we believe there may be recognizable patterns behind successful startups that data can help uncover. This project explores the possibility that historical startup data holds valuable signals that can inform or even anticipate success, laying the foundation for data-driven insights into early-stage company performance.

2 Dataset Overview

The dataset used in this project is the Startup Success Prediction Dataset from Kaggle (`https://www.kaggle.com/datasets/manishkc06/startup-success-prediction`). The dataset contains 923 startups founded between 1995 and 2013, with 49 initial variables describing various aspects of each startup.

A basic exploratory analysis reveals that the dataset includes information about startup location (latitude, longitude, state, city), founding and closure dates, funding history (total funding, number of rounds, funding dates), industry categories (software, web, mobile, enterprise, etc.), investment types (VC, angel, rounds A-D), relationships, milestones, and other characteristics. The target variable indicates whether each startup was acquired (success) or closed (failure), with approximately 65% of startups classified as successful and 35% as failures. The dataset exhibits typical startup ecosystem patterns, including right-skewed funding distributions and geographic concentration in major tech hubs such as California, New York, Massachusetts, and Texas.

3 Business questions and objectives

This project addresses three key business questions:

1. What factors historically influenced startup success or failure? The project aims to identify which startup characteristics from the 1995-2013 period were most predictive of success, enabling data-driven understanding of historical success patterns.

2. Do success patterns from the past still predict success today? While the primary model is trained on historical data, the project framework allows for validation on newer startups to assess whether historical patterns remain relevant.

3. Which features drive predictions and why? Through SHAP analysis, the project provides global and local interpretable explanations for each prediction, enabling stakeholders to understand the reasoning behind model outputs.

The **overall objective** of this project is threefold. First, we aim to **develop a reliable and interpretable machine learning model** trained on historical startup data, capable of predicting the probability of success for early-stage ventures. This model should identify the most influential features from a wide range of startup attributes and produce accurate, validated predictions that could support decision-making in the startup and investment landscape. Second, we intend to build a fully **interactive Streamlit web application** that allows users to explore the dataset, input startup characteristics, and receive real-time predictions about success likelihood. The app is designed to be intuitive and practical, providing a user-friendly interface for experimentation and analysis. Third, the project places **strong emphasis on explainability**: using SHAP, we will break down the contribution of each input variable to the model's predictions. This enables users to understand not just the output, but the reasoning behind it, providing both global insights into general success patterns and local explanations tailored to individual startups.

Together, these objectives aim to create a tool that is both analytically sound and accessible for stakeholders interested in the dynamics of startup success.

4 Methodology

4.1 Machine Learning Model

4.1.1 Data Preprocessing

The initial dataset contained 923 startups with 49 variables. Non-informative columns were removed, including identifiers, duplicate features, and the target leakage variable (*labels*). Date columns were converted to datetime format to enable temporal calculations. Missing values in milestone age variables (152 missing) were imputed with 0, indicating no milestone achieved. Invalid temporal data (48 rows with negative ages) was removed, resulting in a final dataset of 875 startups.

4.1.2 Feature Engineering

Six key engineered features were created: *company_age* (lifespan from founding to closure or last funding), *time_to_first_funding* (duration before securing first investment), *funding_duration* (active fundraising period length), and *has_milestones* (binary indicator for milestone achievement). *avg_funding_per_round* was initially created but later removed due to multicollinearity. Log-transformed funding variables were considered but removed to prevent data leakage.

4.1.3 Feature Selection and Multicollinearity Analysis

Correlation analysis revealed near-perfect correlation (0.99) between *avg_funding_per_round* and *funding_total_usd*. VIF analysis confirmed severe multicollinearity (VIF $\approx$ 56). *avg_funding_per_round* was removed, retaining *funding_total_usd* as the primary monetary feature. After removing leakage features and raw timestamps, the final feature set comprised 69 variables. *state_code* was one-hot encoded, while binary state indicators were retained.

4.1.4 Model Selection and Training

The dataset was split 80/20 (700 training, 175 test samples) using stratified sampling. Three models were evaluated using 5-fold stratified cross-validation: Logistic Regression (ROC-AUC: 0.93, Accuracy: 0.90), Random Forest (ROC-AUC: 0.88, Accuracy: 0.83), and LightGBM (ROC-AUC: 0.96, Accuracy: 0.92). LightGBM emerged as the superior model with higher predictive performance and lower variance across folds.

Hyperparameter tuning was performed using randomized search with 40 iterations and 5-fold cross-validation. Parameters optimized included n_estimators, learning_rate, num_leaves, max_depth, subsample, colsample_bytree, and min_child_samples. All models used balanced class weights to address class imbalance.

4.1.5 Model Evaluation

The tuned LightGBM model achieved test set performance of 94% accuracy and 0.983 ROC-AUC. Precision and recall for the success class were 91% and 100% respectively, while failure predictions achieved 100% precision and 82% recall.

Permutation importance analysis revealed *company_age* as the most influential feature ($\approx$30% importance), followed by *age_last_funding_year* ($\approx$16%), *relationships* ($\approx$3.5%), *funding_total_usd*, and *funding_duration*. Geographic and industry features contributed minimally, suggesting success is driven primarily by lifecycle and funding dynamics rather than location or sector.

4.2 Explainability

To address the project's third business question, *Which features drive predictions and why?*, SHAP was implemented to provide both global and local explanations. Global analysis was conducted through SHAP summary plots (bar and dot plots) to identify overall feature importance across all predictions. Dependence plots were generated to analyze how individual features interact with predictions and reveal non-linear relationships. Individual prediction explanations were implemented through waterfall plots

and feature contribution tables, which quantify how each feature contributes to specific predictions by showing how features push predictions away from the base value. This combination of global and local explanations enables stakeholders to understand both general success patterns and specific prediction reasoning.

4.3 Streamlit Application

The model was deployed in an interactive Streamlit web application with three main pages. The **Exploratory Analysis** page provides interactive visualizations including geographic distribution maps, success trends over time, feature distributions, correlation heatmaps, and success rates by category and state. Dynamic filtering capabilities allow users to filter by multiple criteria including success status, year range, state, industry category, funding range, company age, relationships, funding rounds, and milestones.

The **Predictions** page enables real-time success probability calculations based on user-input startup characteristics. Users can input comprehensive startup information and receive instant predictions with probability scores, probability distributions, and model confidence metrics. The page also displays model performance information (training samples, test accuracy, ROC-AUC score) for transparency.

The **Explainability** page integrates SHAP visualizations including global feature importance plots, feature dependence plots, and individual prediction explanations through waterfall plots and feature contribution tables. The page provides example explanations from the test dataset and summarizes key insights about factors driving predictions. The application uses caching for performance optimization and provides an intuitive interface for data exploration, predictions, and model interpretation.

5 Conclusions

This section highlights the summary of findings and key insights addressing the three business questions posed at the outset of the project.

What factors historically influenced startup success or failure? Permutation importance analysis and SHAP global feature importance reveal that *company_age* is the most influential factor (≈30% importance), with younger companies showing significantly higher success rates, consistent with venture capital patterns where successful exits typically occur early. *age_last_funding_year* ranks second (≈16% importance), indicating funding recency is a strong predictor. Together, these temporal factors suggest timing matters more than absolute funding amounts. *relationships* ranks third (≈3.5% importance), showing network effects matter. *funding_total_usd* and *funding_duration* contribute meaningfully, though with diminishing returns at higher funding levels. Geographic and industry features contributed minimally, suggesting success is driven primarily by lifecycle and funding dynamics rather than location or sector.

Do success patterns from the past still predict success today? The model was trained exclusively on historical data (1995-2013) and achieves excellent performance (94% accuracy, 0.983 ROC-AUC) on historical test data. However, its generalizability to modern startups (2020-2025) remains unvalidated. The dataset's time period may not capture recent trends such as remote work, new funding mechanisms, or evolving industry dynamics. Additionally, the binary success definition (acquisition vs. closure) excludes other outcomes like IPO or continued operation. While historical patterns are clearly identifiable, their relevance to today's startup ecosystem requires further validation with contemporary data.

Which features drive predictions and why? SHAP analysis provides interpretable explanations for both global patterns and individual predictions. Global analysis confirms temporal factors (company age, funding recency) have the strongest impact, followed by network effects (relationships) and funding amounts. Dependence plots reveal non-linear relationships: *funding_total_usd* shows diminishing returns, *relationships* exhibits positive correlation with success, and younger companies have higher success rates. Individual prediction explanations via waterfall plots and feature contribution tables show how each feature contributes to specific predictions, enabling stakeholders to understand why particular startups are predicted to succeed or fail.

Overall, the model and interactive Streamlit application provide actionable insights for investors, entrepreneurs, and researchers. SHAP explanations enable transparent, interpretable predictions, building trust and enabling data-driven decision-making. The framework successfully addresses the first and third business questions. The second question (temporal generalizability) requires future validation with modern startup data.