From Application to Graduation: Uncovering Patterns of Venture Success

Abeba Turi, Dhruv Garg, Ethan Fang, Thamer Aldawood

Table of contents

Executive Summary

Creative Destruction Lab (CDL) is a non-profit startup accelerator that supports early stage science- and technology-based ventures through mentorship, resources, and investor networks to drive growth and commercialization. However, not all ventures that participate in the program graduate successfully. CDL aims to identify the factors contributing to a venture's success and failure in the program to improve its selection process and provide better support. Drawing on a rich dataset of over $5,000$ ventures across multiple cohorts, we analyzed various features from four different data sets, including application records, founder education data, admitted venture logs, and longitudinal program updates (e.g., revenue, funding, headcount) to develop predictive models of venture progression. Yet, substantial missing data posed significant challenges in modeling dropout from the program.

This work benchmarked multiple machine learning models, including logistic regression, random forest, and gradient boosting classifiers, alongside a conventional Cox Proportional Hazards (PH) model. Performance was evaluated using the F1 score for classifiers and the concordance index (C-index) for survival models. Our findings indicate that machine learning models underperformed in predictive power despite advanced tuning due to data quality limitations, including significant missing data and noisy categorical inputs. In the survival analysis, the Cox PH model achieved moderate predictive performance (C-index ≈ 0.61). In addition to modeling, we developed an interactive dashboard to surface descriptive insights: graduation trends by sector and session, cohort-level revenue trajectories, geographic venture distribution, a venture similarity tool for benchmarking new applicants, and feature attribution plots showing top survival-impacting variables through a tornado plot. While predictive power was limited due to data gaps and unmeasured qualitative factors, the dashboard provides an interpretable tool to support evidence-based decision-making for venture selection and monitoring.

Introduction

Creative Destruction Lab (CDL), an established mentorship-driven accelerator program, is a non-profit organization that runs a program focused on supporting early stage science- and technology-based ventures by connecting them with experienced mentors and potential investors, Creative Destruction Lab (2025). However, despite the program's support, a significant number of ventures do not make it to graduation, defined as completing all four mentorship sessions. Mentors currently determine whether ventures continue or are dropped from the program based on their progress across sessions. This raises an important question: what factors determine whether a venture will succeed in the CDL program?

Picking on this question, the project focuses on building a data-driven foundation for answering this critical question. To support our modeling and venture profiling efforts, we drew on four key datasets including applications data: self-reported characteristics of ventures at intake

(e.g., revenue, IP strategy, problem type), admitted data: internal program-level tags post-selection (e.g., stream, program year, geography), education data: detailing the academic background of founding teams, and Updates data: tracking revenue, funding, and team size across program sessions. In this work, a blend of machine learning and statistical models are applied (Hallen, Cohen, and Bingham (2020), Jongwoo Kim (2023), Lin and Wei (1989) and Mona Razaghzadeh Bidgoli (2024)). Given the modeling limitations resulting from data quality issues and unobserved factors like mentor impressions (Creative Destruction Lab (2025) and Nejad (2024)), pitch quality, or other qualitative signals, our dashboard heavily emphasizes on descriptive analytics and venture similarity comparisons within the dashboard.

The dashboard integrates all the analytical outputs in an interactive way that supports real-time venture profiling. This tool provides (1) granular venture profiling (through filters that let users drill down into individual ventures or specific cohorts, view key metrics like revenue over time, and analyze country-level distributions); (2) cohort comparisons (via dropdowns for session and stream, alongside visuals for graduation rate by Sector and global venture heatmap, which help contextualize venture performance across regions and industries); and (3) survival probability visualizations (which allows users to explore time-to-dropout predictions, highlighting venture-level risk profiles relative to historical patterns). This enables CDL stakeholders to move beyond binary predictions and instead leverage interpretable, data-informed guidance to support venture success across the mentorship lifecycle. CDL and stakeholders with access to the dashboard can use it to benchmark ventures, monitor cohort trends, and make informed decisions about early interventions and program design.

The remainder of this report is organized as follows: Section 2 outlines the data science methods highlighting the modeling techniques, feature engineering, evaluation metrics, and ethical considerations used to address the project's scientific objectives; Section 3 covers the data product and results presenting our interactive venture profiling tool, summarizes key findings, and reflects on the design choices, use cases, and areas for future improvement and; Section 4 covers the conclusion and recommendations evaluating how well our data product meets the partner's needs, and provides strategic guidance for potential enhancements and continued application.

Data Science Methods

Exploratory Data Analysis

Exploratory data analysis was conducted using the following methods:

- **Descriptive Statistics**: Mean, median, and standard deviation were calculated for numerical features to summarize their distributions.
- **Missing Data Analysis**: The proportion of missing values in each dataset was assessed to evaluate the extent of missingness and its potential impact on modeling.

- **Correlation Analysis**: Correlations between features and the target variable (graduation) were examined to identify potential predictors.
- **Cohort Analysis**: Data was analyzed across different cohorts to identify trends and patterns in venture performance over time.
- **Graduation Rate Analysis**: Graduation rates for different cohorts, sectors, and geographic locations were calculated to reveal trends and patterns in venture success.

Encoding features

The datasets comprised categorical, ordinal, and numerical features. The following preprocessing steps were applied to prepare the data for modeling:

1. **Categorical Feature Encoding**: One-hot encoding was employed to convert categorical features into numerical representations, enabling effective interpretation by the models. (Pedregosa et al. 2011)
2. **Ordinal Feature Encoding**: Label encoding was utilized for ordinal features to preserve the inherent order of categories during conversion to numerical values. (Pedregosa et al. 2011)
3. **Numerical Feature Handling**: Numerical features were retained in their original form, with appropriate scaling applied as required by the modeling process. (Pedregosa et al. 2011)
4. **Free Text Feature Processing**: Free text features in the updates dataset were processed using a combination of regular expressions and natural language processing techniques to extract relevant information such as revenue, funding, and headcount. This process involved:
    - The creation of a custom prompt to extract numerical values from the free text.
    - The use of the GPT-4.1 model via the OpenAI API to process the text and extract the required information. (Python Software Foundation 2023; OpenAI 2023)

Alternative encoding methods, such as target encoding or embeddings, could be explored to better capture information from high-cardinality categorical features. However, these approaches were not implemented due to their added complexity compared to these simpler methods which were fairly effective for our dataset.

Missing Data Imputation

Given the relatively small size of the data and the proportion of missingness, the removal of rows containing missing values was deemed unsuitable. Instead, two imputation techniques were explored:

1. **Simple Imputation**

- Missing values were filled using the mean, median, or mode of the respective feature, or with a constant value.
- While this approach is straightforward and computationally efficient, it may introduce bias if the missingness is not random (Little and Rubin 2002).

2. **Iterative Imputation**

   - Machine learning models were employed to estimate missing values based on the observed features in the dataset.
   - Although more complex and time-intensive, this method can yield more accurate imputations when the assumption of missing at random does not hold (Jerez et al. 2010; Buuren and Groothuis-Oudshoorn 2011).

Alternative advanced imputation methods, such as k-nearest neighbors (KNN) imputation and matrix factorization, could be considered, but were not implemented due to increased complexity and limited expected benefit given the dataset size.

Evaluation Metrics

To rigorously assess model performance in identifying the factors influencing venture success within the CDL program, evaluation metrics were selected based on their suitability towards our objectives and data characteristics. Given the binary classification task and the presence of class imbalance, the F1 score was adopted as the primary metric, as it effectively balances precision and recall, providing a more informative measure than accuracy alone (Lipton, Elkan, and Naryanaswamy 2014). For survival analysis, the concordance index (C-index) was utilized to quantify the model's ability to correctly rank event times and is well-suited for censored data (Uno et al. 2011). These metrics were chosen to help CDL minimize both false positives (ventures incorrectly expected to succeed) and false negatives (ventures overlooked despite high potential) in selection decisions, while also supporting session-level risk monitoring and targeted intervention through survival risk ranking.

Prediction Modeling

A series of machine learning models was implemented to predict venture graduation, with each model selected for its strengths in addressing the complexities of the dataset. Training was conducted exclusively on data collected prior to ventures being admitted into the CDL program. This approach was adopted to align with CDL's objective of identifying factors associated with venture success before program initiation.

The target variable was defined as the graduation status of a venture, where graduation was determined by the completion of all four mentorship sessions in the CDL program. The following models were implemented:

1. **Dummy Classifier**: Used as a baseline to establish the minimum performance level. It helps us determine if more complex models provide meaningful improvements over random or majority-class predictions. (Pedregosa et al. 2011)
2. **Logistic Regression**: Chosen for its simplicity, interpretability, and effectiveness in binary classification tasks. It provides a strong baseline and helps identify linear relationships between features and graduation outcomes. (Pedregosa et al. 2011)
3. **Hist Gradient Boosting Classifier**: Selected for its ability to handle missing values natively and capture complex, non-linear relationships in the data. Its ensemble approach often yields higher accuracy on structured datasets. (Pedregosa et al. 2011)
4. **XGBoost Classifier**: Included due to its reputation for high performance and efficiency in classification problems. It is robust to overfitting and can handle imbalanced datasets, making it suitable for our data characteristics. (Chen and Guestrin 2016)

Feature selection was performed using recursive feature elimination with cross-validation (RFECV) to identify the most informative predictors. This approach enabled the retention of features that contributed most to model performance, improving both interpretability and generalizability of the final model. (Pedregosa et al. 2011)

Advanced modeling techniques such as deep learning or stacking ensembles could potentially improve predictive performance, especially if more data were available. However, these approaches require larger datasets, increased computational resources, and careful tuning to avoid overfitting. Given the limited sample size and the need for interpretability, these improvements were not implemented in this analysis.

Similarity Modeling

Similarity modeling was implemented using the NearestNeighbors algorithm from scikit-learn, which identifies the most similar ventures by computing pairwise Euclidean distances on selected features. To ensure missing values did not distort similarity calculations, a simple imputer replaced them with a constant negative value that was chosen because it falls outside the range of valid feature values, clearly signaling missingness to the model. Numeric features were then standardized using a scaler to ensure fair distance comparisons. The fitted model was saved as a pickle file, enabling seamless integration and dynamic loading within the dashboard application. Instance-based learning algorithms, such as nearest neighbor approaches, have been widely used for comparative assessments in machine learning (Aha, Kibler, and Albert 1991).

Survival Analysis

Survival analysis was conducted using the Cox Proportional Hazards Model from the lifelines package to estimate the risk of venture dropout over time. After preprocessing and imputation, feature selection was performed by removing variables with low variance and high

multicollinearity, using variance thresholding and variance inflation factor (VIF) analysis. This approach ensured that only informative and independent features were retained for modeling. The model was then trained to predict time until dropout, enabling the estimation of survival probabilities for each venture throughout the program. This model was also saved as a pickle file for integration into the dashboard application. (Lin and Wei 1989; Davidson-Pilon 2019)

Dashboard

An interactive dashboard was implemented using the Dash framework to enable exploration of venture data and model outputs. Modular Python scripts were used to load processed data, apply models, and generate visualizations with Plotly. Callback functions allowed users to dynamically filter and view results by cohort or venture attributes. This approach ensured that analytical outputs could be explored interactively and consistently with the underlying data science workflow. (P. T. Inc. 2015; Parmer, Johnson, and Plotly team 2017)

Reproducibility

To facilitate reproducibility, a Docker container was created to encapsulate the entire data science workflow, including data preprocessing, feature engineering, model training, and evaluation. This container ensures that all dependencies and configurations are consistent across different environments, allowing for seamless execution of the analysis. This approach enables users to easily rerun the entire workflow and monitor changes in model performance or data characteristics over time. (D. Inc. 2013)

Results and Data Product

Exploratory Data Analysis

Analysis of 5,047 ventures admitted to CDL between 2013 and 2024 revealed that venture success lacks a single dominant predictive factor, as no feature has a correlation coefficient larger than 0.06 (Figure 1). It also revealed that missing data is significantly prevalent in many key features (Figure 2).

Predictive Modeling Performance

Table 1 summarizes the predictive modeling results. Logistic regression, HistGradientBoosting, and XGBoost were evaluated using both simple and iterative imputation strategies. HistGradientBoosting was also tested without imputation, leveraging its ability to handle missing data natively. The highest validation F1 score achieved was 0.32.

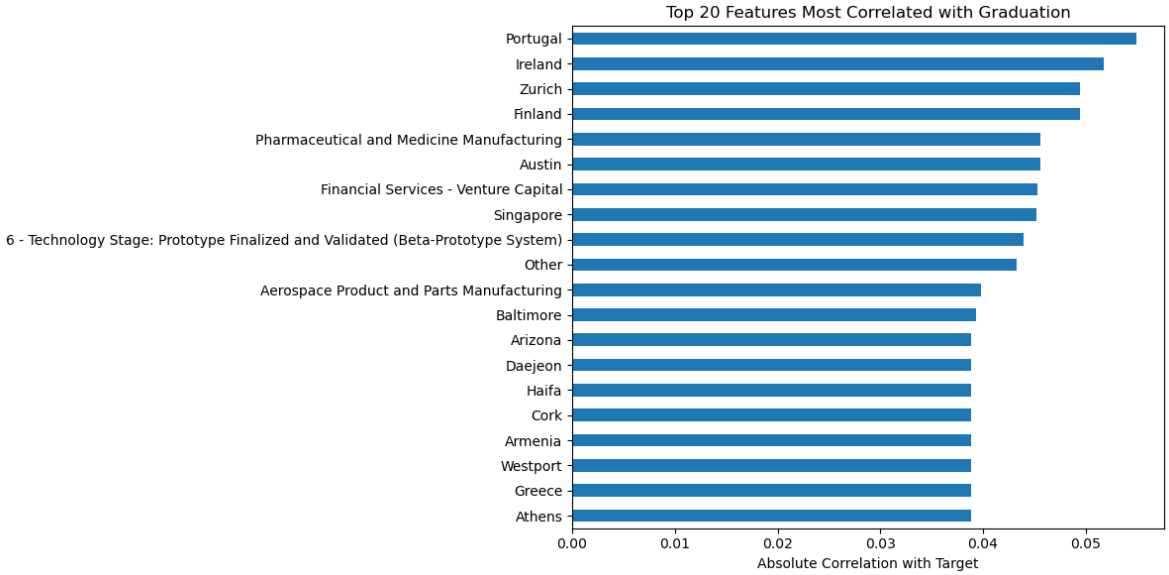

Figure 1: Top 20 features most correlated with graduation

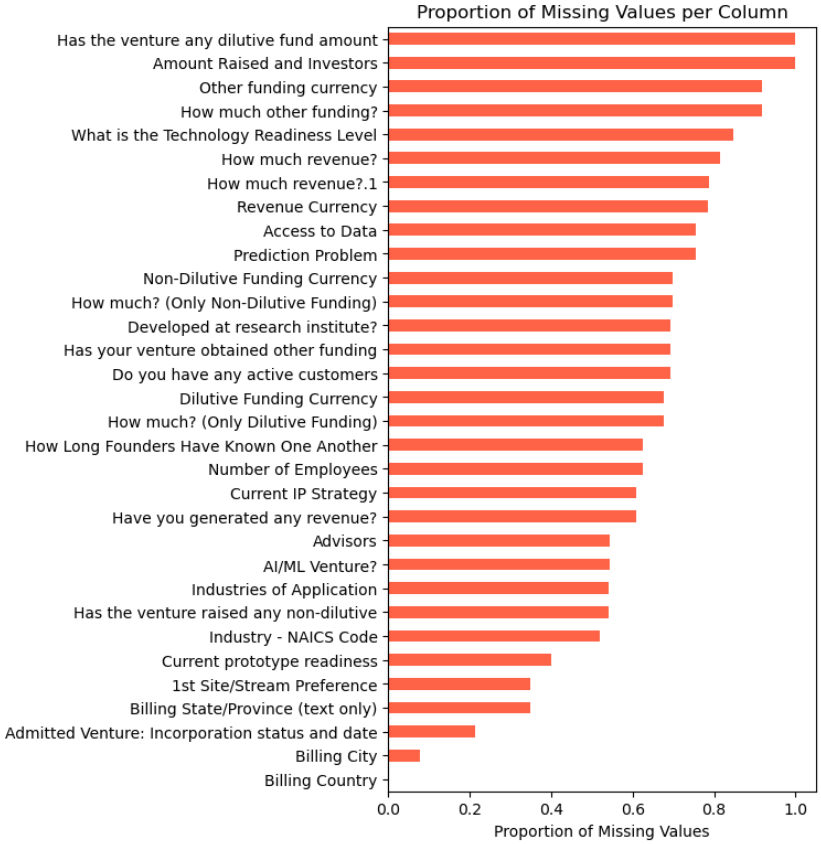

Figure 2: Proportion of missing values in the Applications dataset

Table 1: Model Performance Metrics

Model	Metric	Mean	Std
dummy_clf	train_accuracy	0.665	0
dummy_clf	validation_accuracy	0.665	0.001
dummy_clf	train_f1	0	0
dummy_clf	validation_f1	0	0
hgb_nonimpute	train_accuracy	0.849	0.005
hgb_nonimpute	validation_accuracy	0.645	0.026
hgb_nonimpute	train_f1	0.736	0.007
hgb_nonimpute	validation_f1	0.307	0.038
LogisticRegression_simple	train_accuracy	0.755	0.005
LogisticRegression_simple	validation_accuracy	0.635	0.012
LogisticRegression_simple	train_f1	0.516	0.014
LogisticRegression_simple	validation_f1	0.263	0.027
XGB_simple	train_accuracy	0.870	0.005
XGB_simple	validation_accuracy	0.631	0.015
XGB_simple	train_f1	0.776	0.010
XGB_simple	validation_f1	0.285	0.007
HGB_simple	train_accuracy	0.857	0.004
HGB_simple	validation_accuracy	0.645	0.013
HGB_simple	train_f1	0.748	0.009
HGB_simple	validation_f1	0.299	0.023
LogisticRegression_iter	train_accuracy	0.755	0.005
LogisticRegression_iter	validation_accuracy	0.634	0.014
LogisticRegression_iter	train_f1	0.517	0.012
LogisticRegression_iter	validation_f1	0.262	0.032
XGB_iter	train_accuracy	0.905	0.010
XGB_iter	validation_accuracy	0.622	0.018
XGB_iter	train_f1	0.843	0.019
XGB_iter	validation_f1	0.285	0.025
HGB_iter	train_accuracy	0.885	0.007
HGB_iter	validation_accuracy	0.632	0.007
HGB_iter	train_f1	0.806	0.013
HGB_iter	validation_f1	0.292	0.022
XGB_tuned	train_accuracy	0.966	0.003
XGB_tuned	validation_accuracy	0.634	0.013
XGB_tuned	train_f1	0.949	0.004
XGB_tuned	validation_f1	0.372	0.021

Venture Similarity Analysis

As seen in Figure 4, the similarity tool allows users to input seven key characteristics: industry, problem domain, IP strategy, technology readiness, team size, revenue, and funding status. Then, the tool returns the top 5 most similar ventures from the CDL portfolio and showcases the most similar features as well as the outcomes of the ventures.

Survival Analysis Results

Cox Proportional Hazards modeling identified Sessions 2-3 as the highest dropout risk period, with a concordance index of 0.61. Sector-specific survival patterns emerged, suggesting differential alignment between venture types and CDL's program structure.

Dashboard Architecture

Analytical outputs were integrated into an interactive web-based dashboard with three modules described below. The platform enables dynamic exploration supporting evidence-based decision-making while augmenting existing evaluation processes.

Portfolio Snapshot

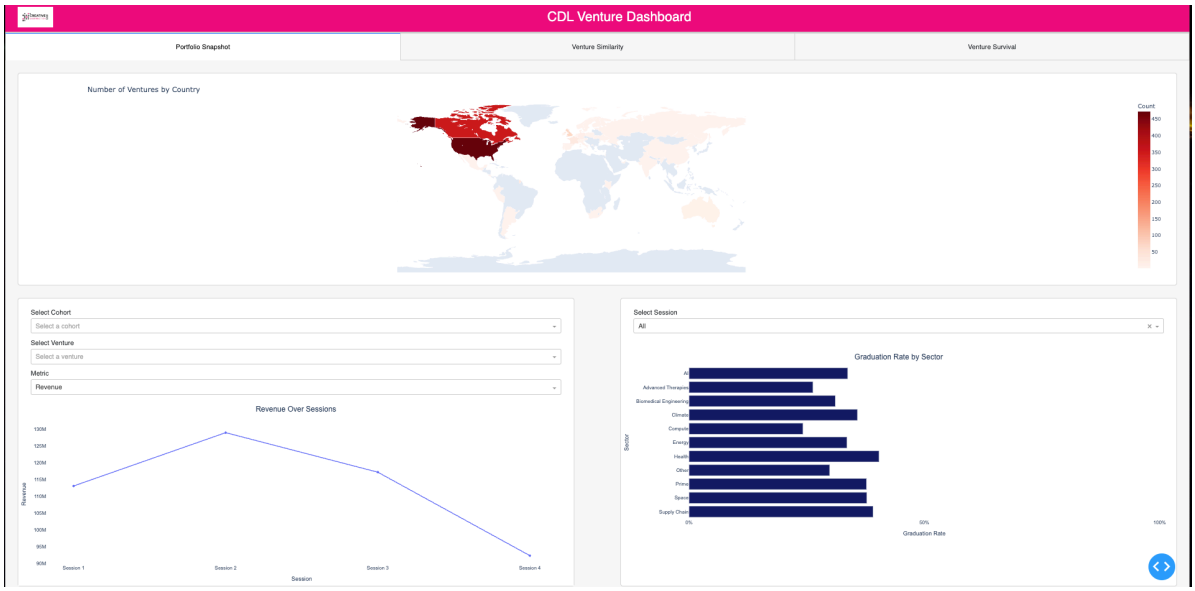

Figure 3: Portfolio Snapshot Dashboard

The Portfolio Snapshot module provides aggregate trend visualization across temporal, sectoral, and geographic dimensions with interactive filtering capabilities. The choropleth map showcases venture global distribution and enables filtering by country. The line chart showcases several key growth metrics (revenue, funding, employee count) over sessions. The bar chart visualizes graduation rates by sector.

Venture Similarity Tool

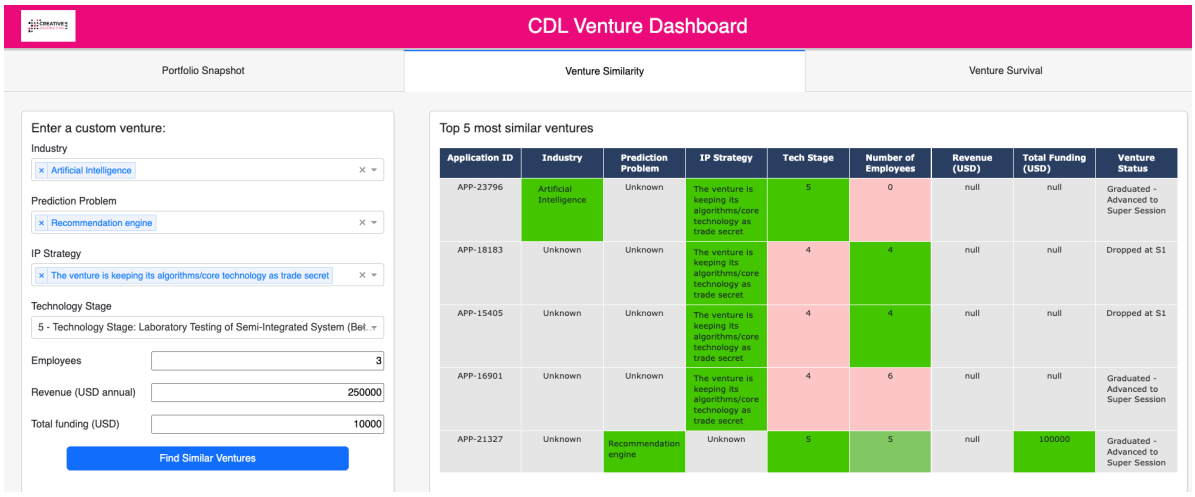

Figure 4: Venture Similarity Tool Dashboard

The Venture Similarity Tool enables real-time benchmarking by comparing new ventures against historical CDL portfolio data.

Survival Analysis Module

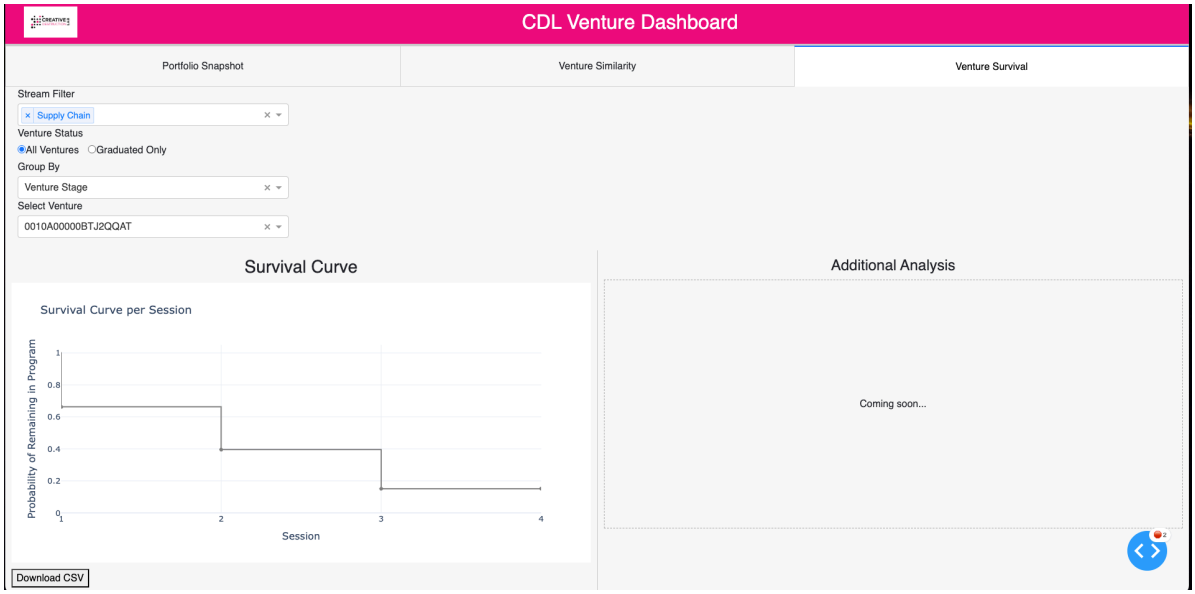

Figure 5: Survival Analysis Dashboard

The Survival Analysis module visualizes dropout risk profiles at both cohort and individual venture levels.

The data product was developed to integrate the most reliable analytical outputs specifically, similarity and survival analyses into a single tool that supports CDL's strategic goals of selection, monitoring, and risk management.

Discussion

The exploratory data analysis indicated that venture success within the CDL program is shaped by a complex combination of factors, with no single feature emerging as a dominant predictor of graduation. The diversity of the CDL portfolio, which includes ventures from fields such as artificial intelligence and quantum computing, further contributes to this complexity.

Data quality was identified as a significant limitation. Substantial missingness was observed in self-reported financial metrics, particularly for revenue, funding, and team size, as ventures were often reluctant to disclose sensitive information. In contrast, structured categorical variables—such as sector, geography, and prototype readiness—were found to be more reliable for modeling. Nevertheless, modeling results demonstrated that these structured features

alone were insufficient for accurately predicting graduation outcomes, highlighting the continued importance of qualitative assessments in CDL's decision-making process.

As summarized in Table 1, the best-performing prediction model achieved a validation F1 score of 0.372, indicating that accurate prediction of venture graduation remains challenging. The much higher F1 score observed on the training set (0.949) suggests overfitting rather than genuine predictive capability. Attempts to improve generalizability through model tuning and feature selection resulted in only marginal gains. Several factors likely contributed to these limitations. Venture graduation is influenced by numerous unobservable variables that are not captured in the available data. Graduation decisions are made by different mentors, each with their own biases and criteria, for whom no data are available. The dataset is relatively small, with 3,536 rows in the applications dataset, and a large proportion of columns contain missing data. Additionally, the available data are self-reported, raising concerns about accuracy.

Given these constraints, further model optimization was deemed unlikely to yield substantial improvements. As a result, similarity modeling was adopted as an alternative approach to support venture evaluation. This method may offer advantages because it does not rely on the same predictive assumptions as traditional models and can leverage CDL's qualitative knowledge in conjunction with structured data, enabling a more nuanced assessment of venture characteristics and outcomes.

In addition, survival analysis using the Cox Proportional Hazards Model was implemented to provide insights into venture retention and dropout risk over time. This approach allows for the analysis of time-to-event data (such as graduation) while accounting for covariates including industry, technology stage, and revenue, and can help identify factors influencing the likelihood of graduation at various stages of the program.

All models and analytical pipelines have been delivered in a reproducible format, with the expectation that their performance will improve as more high-quality data become available. At present, the dashboard offers CDL a comprehensive overview of venture characteristics, graduation rates, and dropout risks, supporting data-driven decision-making and strategic support for ventures throughout the program.

Conclusion

This capstone project analyzed the progression of ventures through the Creative Destruction Lab (CDL) program, from application to graduation. Data preprocessing, feature engineering, and machine learning models such as logistic regression and XGBoost were used to identify factors influencing venture success. However, predictive models showed limited generalization, with a maximum validation F1 score of 0.372, highlighting the complexity of predicting graduation outcomes.

To address these challenges, a similarity modeling framework was implemented, enabling comparative assessments between new and historical ventures. Survival analysis using the Cox

Proportional Hazards model further provided session-specific survival probabilities and identified key dropout risks.

All analytical results were consolidated into an interactive dashboard, offering users comprehensive venture analytics and comparative visualizations. This dashboard enhances CDL stakeholders' ability to monitor and support venture progress, while its effectiveness remains dependent on the quality and completeness of available data.

Recommendations

Based on the findings of this project, several recommendations are proposed to strengthen CDL's venture selection and monitoring processes. First, integrating the similarity modeling framework into mentor matching could enable more relevant pairings, increasing the likelihood of venture success through better aligned guidance. Survival analysis should also be incorporated into ongoing risk monitoring to help identify dropout risks early and support timely interventions. To address data quality challenges, CDL could consider redesigning application forms to standardize key fields particularly those related to financial metrics and make their completion mandatory. Building real-time data pipelines would ensure that dashboards reflect the most current venture metrics, improving the accuracy and timeliness of monitoring. Finally, collecting structured qualitative feedback, including mentor evaluations and founder pitch assessments, would provide additional context to enhance future predictive and comparative analyses.

References

Aha, David W., Dennis Kibler, and Marc K. Albert. 1991. "Instance-Based Learning Algorithms." *Machine Learning* 6 (1).

Buuren, S. van, and K. Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3).

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." ACM.

Creative Destruction Lab. 2025. "Programs." https://creativedestructionlab.com/program/.

Davidson-Pilon, Cameron. 2019. "Lifelines: Survival Analysis in Python." https://github.com/CamDavidsonPilon/lifelines.

Hallen, Benjamin L., Susan L. Cohen, and Christopher B. Bingham. 2020. "Do Accelerators Work? If so, How?" *Organization Science* 31 (2): 378–414. https://doi.org/10.1287/orsc.2019.1304.

Inc., Docker. 2013. "Docker: Enterprise Container Platform." https://www.docker.com.

Inc., Plotly Technologies. 2015. "Plotly: Python Graphing Library." https://plotly.com/python/.

Jerez, J. M., I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco. 2010. "Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem." *Artificial Intelligence in Medicine* 50 (2).

Jongwoo Kim, Youngjung Geum, Hongil Kim. 2023. "How to Succeed in the Market? Predicting Startup Success Using a Machine Learning Approach."

Lin, D. Y., and L. J. Wei. 1989. "The Robust Inference for the Cox Proportional Hazards Model." *Journal of the American Statistical Association* 84 (408): 1074–78. https://doi.org/10.1080/01621459.1989.10478874.

Lipton, Zachary C., Charles Elkan, and Balakrishnan Naryanaswamy. 2014. "Thresholding Classifiers to Maximize F1 Score." https://arxiv.org/abs/1402.1892.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. Wiley.

Mona Razaghzadeh Bidgoli, Mehdi Goodarzi, Iman Raeesi Vanani. 2024. "Predicting the Success of Startups Using a Machine Learning Approach."

Nejad, M. H. 2024. "A Structural Model of Mentorship in Startup Accelerators: Matching, Learning, and Value Creation."

OpenAI. 2023. "GPT-4 Technical Report." https://openai.com/research/gpt-4.

Parmer, Chris, Alex Johnson, and the Plotly team. 2017. "Dash: A Python Framework for Building Reactive Web Applications." https://dash.plotly.com/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12.

Python Software Foundation. 2023. "Python Language Reference, Version 3.11." https://www.python.org/.

Uno, Hajime, Lu Tian, Tianxi Cai, Isaac S. Kohane, and Lee-Jen Wei. 2011. "On the c-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored

Survival Data." *Statistics in Medicine* 30 (10).