AHMED WALEAD MOSAAD

Multiple Objective Optimization With Prediction Of Startup Success and Company Age

Drawing inspiration from a comprehensive analysis of startup ventures through advanced machine learning techniques, this research enhances the predictive framework for startup success by incorporating multi-objective optimization. This study builds on foundational research that analyzed a diverse dataset of 3,160 companies, recognized for their innovation and technological potential. By extending the predictive model, we now simultaneously consider crucial dimensions of overall company success and organizational longevity. Utilizing the enriched dataset, which includes additional details on company attributes, team dynamics, and financial awards, we apply sophisticated multi-objective optimization techniques to improve the predictive accuracy regarding a startup's potential for success and its projected lifespan. Our methodological advancements aim to not only predict these critical aspects but also to balance often-competing objectives, enhancing strategic decision-making within the dynamic startup ecosystem. The expected outcomes of this research include a deeper understanding of how various startup characteristics interact and their collective impact on achieving sustained success and longevity. This study offers valuable insights for investors, policymakers, and entrepreneurs, optimizing resource allocation and strategic planning in the vibrant ecosystem of startup ventures.

ACM Reference Format:
Ahmed Walead Mosaad. 2024. Multiple Objective Optimization With Prediction Of Startup Success and Company Age. 1, 1 (April 2024), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the evolving landscape of startup ventures, the path to success is multifaceted, encompassing a blend of innovation, strategic foresight, navigating market complexities, and a stroke of good fortune. Central to this journey is the identification and optimization of key factors that forecast the overall success and longevity of startups. Building upon the foundational analysis presented in the seminal study by Thirupathi, Alhanai, and Ghassemi (2021), this research extends the predictive modeling of startup success to include not only the likelihood of sustained market presence but also the longevity of the organization. Utilizing a rich dataset derived from a heterogeneous set of 3,160 startup ventures, all beneficiaries of Small Business Innovation Research (SBIR) or Small Business Technology Transfer (STTR) awards, this study employs advanced machine learning and statistical techniques to unravel the intricate dynamics between startup attributes and their objectives of long-term success and enduring operation.

The premise of this research is rooted in the notion that the entrepreneurial journey is not merely about launching a business but doing so in a manner that ensures sustained growth and viability. Thus, we pivot the predictive focus towards a multi-objective framework that seeks to balance the aspirational goal of market success with the pragmatic objective of longevity. Utilizing Python for data pre-processing and feature engineering, we leverage statistical methods to model the complex interplay of factors that contribute to these outcomes. This approach not only underscores the

Author's address: Ahmed Walead Mosaad, awm5999@psu.edu, The Pennsylvania State University.

Manuscript submitted to ACM

multifaceted nature of startup success but also provides a nuanced understanding of how startups can strategically position themselves for optimal outcomes in the competitive marketplace.

This paper is structured to first revisit the methodology of the original study, highlighting the data preparation pipeline, feature extraction, and the statistical modeling techniques employed. We then detail the adaptation of this methodology to incorporate the new targets, elucidating the modifications to the outcome specification and model training to accommodate multi-objective optimization. Through this research, we aim to offer valuable insights into the predictors of startup success and longevity, offering a roadmap for entrepreneurs, investors, and policymakers to foster the growth of robust and enduring ventures in the dynamic startup ecosystem.

2 METHODOLOGY OF PARENT PAPER

This section outlines the methodology employed in our study, including the data preparation, model specification, and evaluation metrics.

2.1 Data Preparation

Our data preparation process involved multiple steps to extract and pre-process information from various sources such as U.S. federal small business RD seed-fund databases and company profiles from Crunchbase. Key steps included the collection of award data, retrieval and cleaning of company profile information, and web scraping for additional data. After pre-processing, the dataset was comprised of continuous, nominal, and text features which underwent further processing, including z-score normalization grouped by the company's founding year.

2.2 Normalization and Feature Selection

To mitigate any potential bias from the age of companies, we performed normalization procedures to ensure feature values were comparable across different company ages. This process was vital for developing a model that captures attributes of successful companies independent of their age.

2.3 Model Specification

To address missing values in the processed dataset, we utilized XGBoost, known for its effectiveness with sparse data. The model was specified with hyper-parameters selected to optimize performance on the diverse data features.

2.4 Performance Metrics and Validation

We measured the model's performance through leave-one-out cross-validation and a variety of metrics such as AUC, accuracy, precision, recall, F-score, and Brier score. These provided a comprehensive view of the model's classification capabilities and its statistical calibration.

2.5 Data Sharing

In the spirit of reproducibility and further research, a de-identified version of the data and the codebase have been made available in a public repository.

3 CONTRIBUTIONS OF THIS RESEARCH

3.1 Methodological Shift to Multi-Objective Optimization

This research introduces a paradigm shift by employing multi-objective optimization to simultaneously predict the age of startups and their likelihood of achieving an Initial Public Offering (IPO) or an Merge and Acquisition, which is the dentition of 'success' in this context. This approach departs from traditional methods by integrating multiple targets into a single predictive framework, enhancing the model's ability to capture inter-dependencies between different startup success metrics.

3.2 Model Design Incorporating Dual Objectives

Our model utilizes a custom neural network architecture implemented with the PyTorch framework, designed to accommodate dual objectives: classification and regression.

- **Feature Engineering:** We pre-process our features through Z-score normalization and logarithmic transformations to stabilize variance, particularly grouping by the 'Years Since Founded' feature for normalization.
- **Neural Network Architecture:** The neural network, named SimpleNN, is structured with fully connected layers. The architecture consists of an input layer that connects to a sequence of three hidden layers with 128, 64, and 32 neurons respectively. Each hidden layer employs a ReLU activation function to introduce non-linearity. The network splits into two output layers: a classifier and a regressor, each consisting of a single neuron. The classifier uses a sigmoid activation function to predict IPO success, while the regressor is designed to predict the age of the company.
- **Training Process:** During training, we employ a hybrid loss function that combines categorical cross-entropy for the classification output and mean squared error for the regression output. This combination allows simultaneous optimization for both IPO prediction and age estimation. The training proceeds for a predefined number of epochs, with loss computed separately for each task before being summed for back-propagation. We utilize the Adam optimizer for its efficiency in handling sparse gradients and adaptively adjusting learning rates.

3.3 Advanced Multi-Objective Optimization Techniques

- **Optimization Strategy:** The Adam optimizer is utilized to update network weights iteratively. This optimizer is well-regarded for its handling of sparse gradients and adaptive learning rate management, leading to more efficient convergence during training.
- **Loss Functions:** A dual loss function is employed to optimize both objectives simultaneously. For IPO success prediction, we use categorical cross-entropy, which is suitable for binary classification problems. For predicting the age of the company, we use mean squared error, which is standard for regression tasks.

3.4 Loss Objective and Strategic Implications

- **Model Evaluation:** We assess model performance using accuracy for the classification task and mean squared error for the regression task. These metrics provide a comprehensive evaluation of the model's effectiveness across both objectives.

- **Visualization and Analysis:** The training and validation losses are plotted over each epoch to visualize the learning progression. Additionally, scatter plots comparing actual versus predicted company age offer insights into the model's regression accuracy.

3.5 Conclusion

This methodological innovation not only increases the analytical depth of startup success predictors but also provides a more comprehensive framework for strategic decision-making within the startup ecosystem. By illustrating the dynamic relationship between a startup's developmental timeline and its achievement of significant financial milestones, this research contributes significantly to the fields of entrepreneurial success prediction and strategic business analysis.

4 RESULTS

4.1 Model Performance

The performance of the model on the test dataset is summarized by the precision, recall, F1 score, and accuracy. The precision of the model, which indicates the proportion of true positive predictions out of all positive predictions, was measured at 63.38%. The recall, reflecting the proportion of actual positives that were correctly identified, was 74.39%. The F1 score, which is the harmonic mean of precision and recall, was calculated to be 68.44%. This indicates a reasonably balanced model in terms of precision and recall. The overall accuracy of the model was 64.40%, suggesting that the model correctly predicted the outcome for approximately two-thirds of the cases in the test set.

4.2 Regression Performance

The regression aspect of the model was evaluated using the Mean Squared Error (MSE), which measures the average squared difference between the estimated values and the actual value. An MSE of 195.7349 was obtained, which suggests that there is room for improvement in the model's predictive accuracy for the continuous target variable representing the age of the company.

4.3 Training and Validation Losses

During the training process, both the loss on the training data and the validation data were recorded for each epoch to monitor the learning progress and to detect overfitting. The initial training loss was 1.6779, which decreased consistently to 1.1787 by the tenth epoch. Similarly, the validation loss started at 1.4438 and showed a fluctuation with the lowest being 1.3755 at epoch five and increasing thereafter, ending with a loss of 1.4549 in the last epoch. The validation loss pattern suggests the model could be beginning to overfit as the training progresses, indicating a need for potential adjustments in the training regimen or model architecture.

These results provide insights into the current capabilities of the model and highlight areas for future improvement, especially in refining the model to address the variance in validation loss and to reduce the Mean Squared Error for the regression predictions.

5 FUTURE WORK

While the current implementation of the multi-objective optimization (MOO) model demonstrates a promising direction for startup success prediction, the results indicate several areas for potential enhancement. Future iterations of this

research could benefit from a comparative analysis between single-objective models and the MOO model to better understand the advantages or limitations introduced by the MOO approach.

5.1 Comparative Analysis with Single-Objective Models

A comprehensive study involving single-objective models dedicated to either classification or regression tasks could provide a benchmark for evaluating the performance gains of the MOO model. Such an analysis would clarify the trade-offs between specialized single-task models and our multi-faceted MOO approach, particularly in terms of precision, recall, and mean squared error.

5.2 Hyperparameter Optimization

Further exploration into hyperparameter tuning is warranted. The current model could be improved by employing techniques such as grid search, random search, or Bayesian optimization to find an optimal set of hyperparameters. This process could lead to better model performance and generalization by fine-tuning aspects such as the number of neurons in each layer, learning rates, and regularization parameters.

5.3 Advanced Regularization Techniques

To address the potential overfitting observed as training progresses, advanced regularization techniques beyond dropout could be investigated. Methods such as L1 and L2 regularization, early stopping, or the use of more sophisticated dropout variants might yield a more robust model capable of maintaining performance consistency across training and validation phases.

5.4 Ensemble Methods and Model Stacking

Employing ensemble methods or model stacking could also be a promising area of development. Combining the predictions from multiple models, potentially trained on different aspects of the data, could enhance the predictive power and reliability of the overall system.

5.5 Extended Feature Engineering

There is also scope for expanding the feature set used for training the model. Future work could involve more complex feature engineering strategies or the incorporation of additional data sources that could provide new insights into the factors affecting startup success and longevity.

5.6 Cross-Domain Applicability

Lastly, exploring the cross-domain applicability of the MOO model could provide additional validation of its utility. Testing the model on datasets from various sectors and startup stages would not only confirm its robustness but also its adaptability to different market dynamics.

In conclusion, these prospective research directions aim to not only refine the performance of the existing model but also to broaden the understanding and applicability of machine learning techniques in the context of startup success prediction.

6 RELATED WORKS

The evolution of predictive modeling for startup success has been significantly influenced by advancements in data analytics and machine learning techniques. This section explores notable contributions to the field, particularly focusing on the use of multi-objective optimization and data mining to analyze startup performance metrics. This examination provides a foundation for our study's approach, emphasizing the integration of these methodologies for predicting both IPO success and the age of a company.

6.1 Predictive Modeling for Startup Success

Krishna, Agrawal, and Choudhary (2016) conducted a study to predict startup outcomes using various data mining classification techniques, such as Random Forest and Bayesian Networks. Their dataset included both operational and closed or acquired companies, highlighting the potential of analytical models to identify factors contributing to startup success or failure. Their research stressed the importance of seed funding and strategic financial management in guiding startups toward achieving their goals.

6.2 Multi-objective Optimization in Evaluating Startup Success

Mostaghim, Presse, and Terzidis (2013) developed a framework using multi-objective optimization to assess startups based on employee growth, sales, and earnings before interest and taxes (EBIT). Their methodology focused on the impact of research and development expenditures and the proportion of full-time employees in RD, suggesting that while market growth is necessary, it is not solely sufficient for startup success.

6.3 The Startup Compass as a Benchmarking Tool

The Startup Compass serves as a significant tool for startups, particularly in high-tech sectors, to benchmark their performance against a broad set of industry data. This initiative demonstrates the need for reliable, data-based frameworks to assist startups in navigating the complexities of early-stage growth and development.

6.4 Synthesis and Research Gap

The review of related works highlights a gap in the literature: the need for an integrated approach that combines predictive modeling and multi-objective optimization for a detailed analysis of startup performance. Our study addresses this gap by proposing a framework that predicts the likelihood of IPO success and assesses company age, allowing startups to align their objectives with industry benchmarks and success metrics. This contribution adds to the ongoing discussion on startup evaluation methodologies, providing insights for entrepreneurs, investors, and policymakers.

6.5 Multiple Objective Optimization

In the study "Multi-Task Learning as Multi-Objective Optimization" by Ozan Sener and Vladlen Koltun from Intel Labs, multi-task learning is reinterpreted as a multi-objective optimization problem aimed at finding Pareto optimal solutions. This approach, which does not assume non-competing tasks, addresses scalability issues related to high-dimensional gradients and multiple tasks in deep learning contexts. Their methodology, validated across various tasks, including digit classification and scene understanding, demonstrates improved performance over existing multi-task learning frameworks, offering a nuanced approach to managing conflicting learning objectives.

CITATION AND BIBLIOGRAPHY

(1) Thirupathi, A. N., Alhanai, T., & Ghassemi, M. M. (2017). A Machine Learning Approach to Detect Early Signs of Startup Success. Proceedings of the ACM Conference, East Lansing, Michigan, USA; Abu Dhabi, UAE.
(2) Sener, O., & Koltun, V. (2017). Multi-Task Learning as Multi-Objective Optimization. Proceedings of the ACM Conference, Intel Labs.
(3) Bar-Haim, R., Kantor, Y., Lahav, D., Eden, L., Friedman, R., & Slonim, N. (2017). Quantitative Argument Summarization and Beyond: Cross-Domain Key Point Analysis. Proceedings of the ACM Conference, IBM Research. Email: roybar, yoavka, lilache, roni.friedman-melamed.
(4) Krishna, A., Agrawal, R., & Choudhary, A. (2016). Predictive modeling for startups: Analyzing data mining techniques for startup success predictions.
(5) Mostaghim, S., Presse, A., & Terzidis, O. (2013). Multi-objective optimization for startup evaluation based on growth metrics.
(6) Sener, O., & Koltun, V. (2017). Multi-task learning as multi-objective optimization. Intel Labs.
(7)

Received 8 April 2024; revised 16 April 2024; accepted 16 April 2024