A DEA and Machine Learning based platform for investigating success of Initial Coin Offerings (ICOs)

Ramona Jaberian[a], Hamidreza Ahady Dolatsara[b,*]

a: rjaberian@clarku.edu

b,*: hamid@clarku.edu

a, b School of Management, Clark University, Worcester, MA, USA, 01610

Abstract

Initial coin offerings (ICOs) have become a popular way for startups to raise capital. However, the ICO market is also highly volatile and risky, with many ICOs failing. In this paper, we use a combination of data envelopment analysis (DEA) and machine learning to identify successful ICOs and the factors that are associated with their success.

First, we use DEA to identify a set of efficient ICOs. Once we have identified a set of efficient ICOs, we use machine learning to identify the factors that are most associated with their success. Our results show that DEA and machine learning can be used to identify successful ICOs and the factors that are associated with their success. We find that the most important factors associated with ICO success are existence of a legal entity backing up the coin, social media activities on reddit, availability of roadmap, being in finance industry, and a triable product.

Key Words: DEA, Machine Learning, ICO

1- Introduction

Initial coin offerings (ICOs) have become a popular way for startups to raise capital. In an ICO, a startup creates a new cryptocurrency and sells it to investors in exchange for other cryptocurrencies, such as Bitcoin and Ethereum. ICOs have raised billions of dollars in recent years, but they have also been controversial. Some ICOs have been scams, and others have failed to deliver on their promises [1]. Despite the risks, ICOs offer a number of advantages over traditional fundraising methods. ICOs are faster and cheaper than traditional fundraising methods, and they allow startups to raise capital from a global audience. ICOs also give investors the opportunity to invest in early-stage startups with high growth potential [2].

However, the ICO market is also highly volatile and risky. Many ICOs fail, and investors can lose their entire investment. In order to reduce the risk of investing in ICOs, investors need to be able to identify successful ICOs [3,4,5].

In this paper, we use a combination of data envelopment analysis (DEA) and machine learning to identify successful ICOs and the factors that are associated with their success.

DEA is a non-parametric technique that can be used to measure the relative efficiency of a set of decision-making units (DMUs) that consume multiple inputs to produce multiple outputs. In our case, the DMUs are the ICOs and the inputs and outputs are various factors that are associated with ICO success, such as the amount of capital raised, the team behind the ICO, and the project's whitepaper.

Data envelopment analysis (DEA) is a non-parametric technique that can be used to measure the relative efficiency of a set of decision-making units (DMUs) that consume multiple inputs to produce multiple outputs. DEA has been widely used in financial analysis, but there is a limited number of studies that have used DEA to analyze ICOs. For example, some researchers have used DEA to identify the most efficient ICOs in terms of their fundraising ability and project quality. Others have used DEA to assess the impact of various factors, such as the team's experience and the project's whitepaper quality, on ICO success.

Pierluigi and Ripamonti [6] utilized a two-stage DEA model to investigate efficiency of ICOs. They found there is a positive correlation between early bird bonus schemes and efficiency; ICOs with longer durations are more likely to offer early bird bonus schemes; and a higher percentage of tokens made available to the crowdsale is associated with a higher efficiency score. Findings of a study by Abraham [7] using DEA and propensity score matching (PSM) suggest that DEA-efficient crypto coins and crypto tokens with a whitepaper and expert credit ratings are likely to be good investments. Additionally, DEA can be used to benchmark crypto coins, and ICOs can be seen as an equitable and accessible way for startups to raise capital.

Overall, DEA is a powerful tool that can be used to investigate ICOs in a number of ways. By using DEA, researchers can identify efficient ICOs, assess the impact of various factors on ICO success, and analyze the relationship between ICO success and cryptocurrency market conditions.

Machine learning (ML) is a type of artificial intelligence (AI) that allows computers to learn without being explicitly programmed. Machine learning algorithms can be used to predict outcomes based on historical data. In our case, we train machine learning algorithms to predict whether an ICO will be successful or not.

ML has the potential to be a powerful tool for investigating ICOs in a number of ways. ML algorithms can be trained to identify ICOs that are likely to be successful. Chursook et. al. [8] investigated the use of market sentiment and expert ratings to predict the success of initial coin offerings (ICOs) in the Australian and Singapore markets. Their results showed that sentiment analysis of tweets and expert ratings can be used to predict the success of an ICO. Wang et. al. [9] developed a deep learning model to investigate white papers of crypto coins and predict their success in ICOs. Gihan et. al. [10] proposed a new model for predicting the success of initial coin offerings (ICOs). The model combines three different techniques: Information Gain Directed Feature Selection (IGDFS), Genetic Algorithm (GA), and Fuzzy Support Vector Machine for Class

Imbalance Learning (FSVM-CIL). Karimov and Wójcik [11] utilized various machine learning algorithms together with new tools that explain artificial intelligence predictions (XAI tools) to find the most important factors that predict whether an ICO will fail and to understand how these factors relate to each other. The results show that it is possible to predict with 65-70% accuracy whether an ICO is a scam based on information that is known before the ICO starts. Additionally, nonlinear machine learning models perform better at this task than traditional logistic regression models and its regularized extensions.

Our results show that DEA and machine learning can be used to identify successful ICOs and the factors that are associated with their success. We find that the most important factors associated with ICO success are elements that developers could invest before an ICO to expect a higher chance of success.

2- Methodology

In this paper, we propose a data-driven platform for identifying efficient initial coin offerings (ICOs) and predicting their future performance. The platform consists of two main components: a data envelopment analysis (DEA) module and a machine learning module.

The DEA module is used to identify efficient ICOs, which are those that are able to raise the most capital with the least amount of effort and have the highest quality projects. The machine learning module is used to predict the future performance of efficient ICOs. In addition, by developing machine learning algorithms, we reported the importance of features in prediction of the outcome (i.e. efficiency of ICOs). It gives us an insight how prepare steps before ICO window of time to ensure their success. The machine learning module has three sub-sections: data gathering, feature selection, and prediction.

The feature selection sub-section is particularly comprehensive, consisting of two levels. This helps to reduce the risk of overfitting, which is when the model learns the features of the training data too well and is unable to generalize to new data.

In the first level, we use three different feature selection methods: Fast Correlation Based Feature Selection (FCBF), LASSO, and Random Forest. Each of these methods identifies a different subset of features that are important for predicting ICO success. We then take the union of all three subsets to create a comprehensive set of features. In the second level of feature selection, we use a simulated annealing algorithm to further refine the set of features and identify the most important ones. Our two-level feature selection approach addresses this limitation by using three different methods in the first level and a simulated annealing algorithm in the second level. This helps to ensure that the most important features are identified for predicting ICO success.

2-1 Data Gathering

We collected data on token sales from primary sources for two reasons: concerns about data quality and the amount of data items available from secondary sources.

To construct our sample, we first created a list of completed ICOs using publicly available data sources. We retained only records for which the total ICO funding exceeded $1 million. The reason for truncating the sample in this manner is that primary source data on the smaller ICOs are frequently scarce or unavailable.

To collect data on the characteristics of those ICOs, we relied exclusively on primary sources such as whitepapers, other documents published by issuers, archived issuer websites kept by the Internet Archive (web.archive.org), company announcements on social media (primarily on Reddit and Twitter), and source code on GitHub. Various reference websites (e.g., www.coinmarketcap.com and www.coingecko.com) were also used to supplement the data.

To ensure that we always used the original version of whitepapers and other documentation available during the fundraising, we used the Internet Archive (web.archive.org) to retrieve the last version of the whitepaper published before the ICO. Our final sample consisted of 129 ICOs between January 2017 and March 2018.

2-2 Data Envelopment Analysis (DEA)

Input-oriented DEA is a non-parametric technique that can be used to measure the relative efficiency of a set of decision-making units (DMUs) that consume multiple inputs to produce multiple outputs. DEA is a deterministic approach, which means that it does not make any assumptions about the distribution of the data.

To measure the efficiency of a DMU using input-oriented DEA, we first need to identify the DMU's inputs and outputs. The inputs are the resources that the DMU consumes. The outputs are the products or services that the DMU produces (here ICO capital). Once we have identified the DMU's inputs and outputs, we can calculate the DMU's efficiency score using the following formula:

Efficiency score = 1 - (Weighted sum of inputs) / (Weighted sum of outputs)

The weighted sums of inputs and outputs are calculated using the following formulas:

Weighted sum of inputs = Σ_i w_i * x_i

Weighted sum of outputs = Σ_j v_j * y_j

where:

w_i is the weight of the i-th input

v_j is the weight of the j-th output

x_i is the amount of the i-th input consumed by the DMU y_j is the amount of the j-th output produced by the DMU

The weights w_i and v_j are determined using the following optimization problem:

Maximize: 1 - Σ_i w_i * x_i_0

Subject to:

Σ_i w_i * x_i <= 1 for all DMUs

Σ_j v_j * y_j >= 1 for all DMUs

w_i >= 0 for all i

v_j >= 0 for all j

The subscripted 0 indicates that the weights are being calculated for the DMU whose efficiency is being measured. In this study we considered the same weight for all inputs and the output.

2-3 Feature Selection

In this paper, for the first level of feature selection, we used three different feature selection methods to select the most important features for predicting the success of ICOs: Fast Correlation Based Feature Selection (FCBF), LASSO regression, and Random Forest.

2-3-1 Fast Correlation Based Feature Selection (FCBF)

FCBF is a filter-based feature selection method that selects features based on their correlation with the target variable and their mutual information with each other. Correlation is a measure of how strongly two variables are related to each other. Mutual information is a measure of how much information one variable contains about another variable.

FCBF works by first calculating the correlation between each feature and the target variable. It then calculates the mutual information between each pair of features. Finally, it iteratively removes features until the remaining features are maximally correlated with the target variable and minimally correlated with each other.

FCBF is a relatively simple and efficient feature selection method. However, it is important to note that FCBF is a linear feature selection method. This means that it can only identify linear relationships between the features and the target variable.

2-3-2 Lasso Regression

Lasso regression is a regularized regression method that selects features by shrinking the coefficients of less important features to zero. Regularization is a technique that is used to prevent overfitting in machine learning models. Overfitting occurs when a model learns the training data too well and is unable to generalize to new data.

Lasso regression uses an L1 penalty term to shrink the coefficients. The L1 penalty term is a measure of the absolute value of the coefficients. Lasso regression shrinks the coefficients of less important features to zero by increasing the penalty term for those features.

Lasso regression is a powerful feature selection method that can be used to identify both linear and non-linear relationships between the features and the target variable. However, Lasso regression can be computationally expensive to train, especially for large datasets.

2-3-3 Random Forest

Random forest is an ensemble learning method that uses multiple decision trees to make predictions. Ensemble learning methods combine multiple models to improve the overall performance of the model. Random forest is a particularly powerful ensemble learning method because it is able to reduce overfitting.

Random forest also has a built-in feature selection mechanism that ranks features based on their importance to the model. Random forest calculates the Gini importance of each feature. The Gini importance of a feature is a measure of how much the feature contributes to the purity of the decision tree nodes. Random forest then selects the top K features, where K is a hyperparameter that can be tuned to optimize the performance of the model.

Random forest is a versatile feature selection method that can be used to identify both linear and non-linear relationships between the features and the target variable. Random forest is also relatively efficient to train, even for large datasets.

2-3-3 Simulated annealing (SA)

In the second level of feature selection we utilized Simulated annealing (SA) to refine the previously selected features from the first level of feature selection and identify the most important features. is a stochastic optimization algorithm that can be used for feature selection. SA is inspired by the annealing process in metallurgy, where a metal is heated and then slowly cooled to produce a desired structure. In SA, we start with a random subset of features and then iteratively add and remove features to improve the performance of the model.

SA is a powerful feature selection method that can be used to identify both linear and non-linear relationships between the features and the target variable. SA is also relatively efficient to train, even for large datasets.

2-3-4 Prediction

To predict the efficient ICOs, we used a variety of well-known whitebox and blackbox machine learning algorithms, including decision trees, random forest, support vector machines (SVMs), Naive Bayes, and XGBoost. We evaluated the performance of each algorithm using the area under the curve (AUC) metric. The algorithm with the highest AUC was SVM, so we selected SVM as our final model.

Once we had selected SVM as our final model, we investigated the importance of the features in the model. The importance of a feature is a measure of how much the feature contributes to the prediction of the model. We used the following formula to calculate the importance of a feature:

Importance = |(Mean prediction with feature - Mean prediction without feature) / Standard deviation of prediction with feature|

The higher the importance of a feature, the more important the feature is to the prediction of the model.

3- Results

In this section of the study, the results of DEA, feature selection, and prediction are presented. Finally, the importance of feature in the most accurate predictive models is discussed.

We used data envelopment analysis (DEA) to identify the efficient initial coin offerings (ICOs). DEA is a non-parametric technique that can be used to measure the relative efficiency of a set of decision-making units (DMUs) that consume multiple inputs to produce multiple outputs.

In the context of ICOs, the DMUs are the ICO projects and the inputs and outputs are as follows:

Inputs: Existence of a roadmap, availability of a product prototype, availability of a triable product, availability of a business model, availability of project codes, team size, days on Reddit, and days on Twitter (day further than first 30 days).

Output: Crowd sale

We used DEA to calculate the efficiency score for each ICO project. The efficiency score is a measure of how well an ICO project is able to convert its inputs into outputs. A higher efficiency score indicates that an ICO project is more efficient.

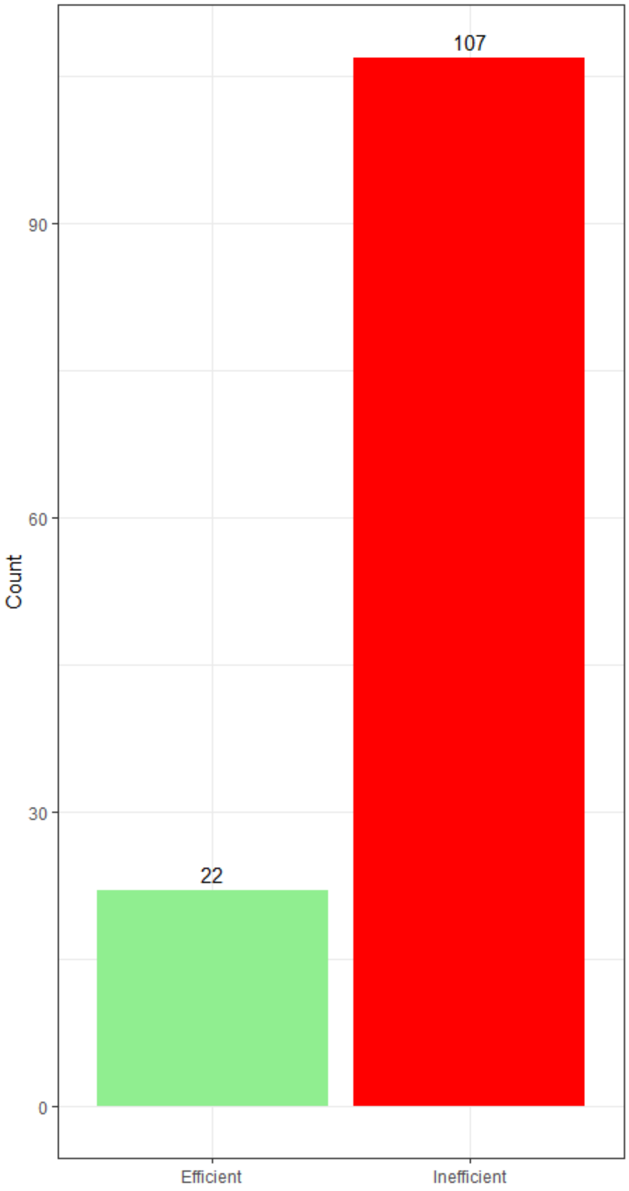

Figure 1: Distribution of efficiency for ICOs

We found that 22 out of 129 ICOs were efficient, which means that they were able to raise the most crowd sale funds with the least amount of inputs. The remaining 107 ICOs were not efficient, which means that they could have improved their efficiency by using their inputs more effectively.

The ICOs in two groups of efficient and non-efficient, this added information could be used as a target for machine learning (ML) algorithms to get trained on their data and predict success of future ICOs. To optimally train the ML algorithms, we picked the most important features. In the first level of feature selection, FCBF, LASSO, and Random Forest identified 2, 18, and 3 features, respectively. The union of these features resulted in 18 distinct features. Using simulated annealing, 13 of these features were selected for the final set of features.

Table 1: Prediction Performance of ML Algorithms

		Decision Trees	Random Forest	SVM	Naïve Bayes	XGBoost
Level 1 Feature Selection	AUC	0.761	0.741	0.800	0.768	0.786
	Sensitivity	0.800	0.886	0.657	0.171	0.857
	Specificity	0.500	0.375	0.625	1.000	0.750
	Accuracy	0.744	0.791	0.651	0.326	0.837
Level 2 Feature Selection	AUC	0.716	0.759	**0.843**	0.775	0.754
	Sensitivity	0.771	0.886	0.743	0.600	0.829
	Specificity	0.500	0.250	0.750	0.625	0.625
	Accuracy	0.721	0.767	0.744	0.605	0.791

The results of predicting the efficiency category of ICOs show that the SVM model trained on the second level of features has the highest AUC, as well as a balanced sensitivity and specificity, indicating that this model can predict both efficient and non-efficient ICOs with relatively high accuracy.

Table 2: Feature Importance of Variables

variables	Importance
legal_entity	100
reddit_days	95.69536
road_map	90.06623
industry_finance	87.41722
product	65.56291
business_model	49.66887
backed_by_VC	39.7351
country_asia	39.07285
Team_member_with_business_background	37.08609
decentralised	34.43709
cryptocurrency_token	13.90728
new_blockchain	13.90728
twitter_days	0

To elucidate the importance of features in predicting the success of ICOs, the following table presents the importance of features in the best predictive algorithm, which is SVM. The most important features are the existence of a legal entity backing the coin, social media activity on Reddit, and the availability of a roadmap, being in finance industry, and a triable product.

Having a legal entity backing the ICO token can help to build trust with potential investors and increase the chances of success. This is because a legal entity is subject to certain regulations and requirements, which can provide investors with some assurance that the ICO is being conducted in a fair and transparent manner. Social media activity on Reddit can be a good indicator of the level of interest and support for an ICO. A high level of social media activity suggests that there is a strong community behind the ICO, which can help to attract investors and generate hype. Having a clear and well-defined roadmap is important for any ICO, as it shows investors that the team has a plan for how they will use the funds raised. Additionally, having a triable product can help to demonstrate the value of the ICO and make it more appealing to investors.

4- Conclusion

In this paper, we have presented a methodology for predicting the efficiency of ICOs. Our methodology is based on data envelopment analysis (DEA), feature selection, and machine learning. Our results suggest that there is a significant opportunity for ICO projects to improve their efficiency and raise more funds. By focusing on the inputs that are most important for success, such as having a strong roadmap and a triable product, ICO projects can increase their chances of success.

There are several directions for future work. First, we would like to collect data on a larger number of ICOs to improve the accuracy of our model. Second, we would like to explore other feature selection methods to see if we can identify a more effective subset of features for predicting the efficiency of ICOs. Finally, we would like to develop a model that can predict the success of ICOs in terms of other metrics.

We believe that our work has the potential to make a significant contribution to the field of ICO prediction. By providing a methodology for predicting the efficiency of ICOs, we can help investors to make more informed investment decisions.

References

[1] Zetzsche, Dirk A., et al. "The ICO Gold Rush: It's a scam, it's a bubble, it's a super challenge for regulators." University of Luxembourg Law Working Paper 11 (2017): 17-83.

[2] Schückes, Magnus, and Tobias Gutmann. "Why do startups pursue initial coin offerings (ICOs)? The role of economic drivers and social identity on funding choice." Small Business Economics 57.2 (2021): 1027-1052.

[3] Šapkauskienė, Alfreda, and Ingrida Višinskaitė. "Initial Coin Offerings (ICOs): benefits, risks and success measures." Entrepreneurship and sustainability issues 7.3 (2020): 1472-1483.

[4] Masiak, Christian, Joern H. Block, Tobias Masiak, Matthias Neuenkirch, and Katja N. Pielen. "Initial coin offerings (ICOs): market cycles and relationship with bitcoin and ether." Small Business Economics 55 (2020): 1113-1130.

[5] Moxoto, Ana Claudia De, Paulo Melo, and Elias Soukiazes. "Initial Coin Offering (ICO): a systematic review of the literature." (2021).

[6] CAPE, FILIPPO, MARIA PIERLUIGI, and FILIPPO RIPAMONTI. "An empirical study of the efficiency of initial coin offerings adopting a two-stage DEA model." (2019).

[7] Abraham, Mathew. "Performance analysis of crypto coins and crypto tokens using data envelopment analysis and propensity score matching." International Journal of Electronic Finance 12.3 (2023): 295-314.

[8] Chursook, Anchaya, Ahmad Yahya Dawod, Somsak Chanaim, Nathee Naktnasukanjn, and Nopasit Chakpitak. "Twitter sentiment analysis and expert ratings of initial coin offering fundraising: evidence from australia and singapore markets." TEM Journal 11, no. 1 (2022): 44.

[9] Wang, Jiayue, Runyu Chen, Wei Xu, Yuanyuan Tang, and Yu Qin. "A document analysis deep learning regression model for initial coin offerings success prediction." Expert Systems with Applications 210 (2022): 118367.

[10] Ali, Mohamed Gihan, Ismail Ibrahim Gomaa, and Saad Mohamed Darwish. "An intelligent model for success prediction of initial coin offerings." IEEE Access 10 (2022): 58589-58602.

[11] Karimov, Bedil, and Piotr Wójcik. "Identification of scams in Initial Coin Offerings with machine learning." Frontiers in Artificial Intelligence 4 (2021): 718450.