Predicting ICO Success: A Machine Learning Approach to Analysing Crowdfunding Campaigns

Introduction

Crowdfunding has evolved into a transformative method of securing funding, with Initial Coin Offerings (ICOs) leading the charge in blockchain-based fundraising. Unlike traditional crowdfunding, which typically offers backers products or services, ICOs distribute digital tokens, commonly known as cryptocurrencies. These tokens function as a form of value exchange and are tradable among investors, often possessing utility within the fundraising company's digital ecosystem. ICO campaigns operate on an "all-or-nothing" basis, meaning only those reaching their financial targets receive funding. As ICOs gain traction, comprehending the drivers behind successful fundraising becomes imperative for investors, project teams, and regulators, particularly within the blockchain and tech sectors. Despite their potential, ICOs face challenges stemming from market sentiment, technological viability, and team credibility, compounded by the dynamic nature of the cryptocurrency market. To navigate these complexities, a data-driven approach is necessary to predict campaign success accurately. Thus, this report endeavours to leverage machine learning models to forecast the likelihood of an ICO reaching its fundraising goal. By analysing a comprehensive dataset encompassing attributes like team size, campaign duration, blockchain platform, and cryptocurrency media presence, the study aims to pinpoint significant predictors of ICO success. Such insights will empower investors to make informed decisions and assist campaign organizers in refining their strategies, ultimately fostering market transparency and enhancing investor protection, thereby offering valuable contributions to the evolving landscape of crowdfunding and blockchain technology.

Data Understanding And Preparation:

Overview

The dataset comprises 2767 observations and 16 variables, providing information on various attributes related to Initial Coin Offering (ICO) projects. Each observation represents a distinct fundraising project or company seeking funding through ICOs. The dataset includes both numerical and categorical variables, with 'success' serving as the target variable indicating whether a project achieved its funding goal successfully.

Breakup of features by datatypes is as follows(Appendix 1):

- **Numeric variables:** 'ID', 'hasVideo', 'rating', 'priceUSD', 'teamSize', 'hasGithub', 'hasReddit', 'coinNum', 'minInvestment', and 'distributedPercentage'.
- **Integer variables:** 'ID', 'hasVideo', 'teamSize', 'hasGithub', 'hasReddit', and 'minInvestment'.
- **Character variables:** 'brandSlogan', 'countryRegion', 'startDate', 'endDate', and 'platform'.
- **Factor variables:** 'success' (representing success or failure with levels 'Y' and 'N').

Numerical Variables:

The dataset contains 10 numerical attributes, which were examined utilizing both the str() and summary() functions(Appendix 2).

Additional examination is needed for attributes such as "priceUSD," "teamSize," "distributedPercentage," and "coinNum."

priceUSD:

There are 180 instances where this value is missing, and in 152 observations, the value is “0”.

Histogram of priceUSD:

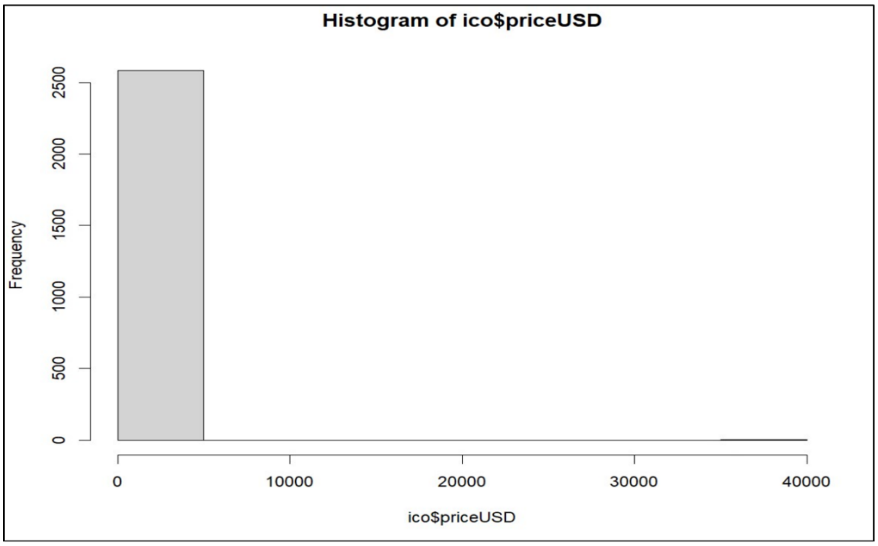

Figure 1 Histogram for “priceUSD” variable.

The histogram for the 'priceUSD' variable reveals insights into its distribution. It comprises eight bins, with the majority of observations falling into the first bin, indicating a prevalence of lower 'priceUSD' values. An outlier is detected, with a 'priceUSD' value exceeding 35000(Appendix 3). The density component shows a skew towards lower values, while the equidistance of bins suggests a uniform width distribution.

Based on the summary data(Appendix 2), it's evident that the maximum observation value is 39384. This indicates that this particular observation significantly deviates from the others.Therefore, this observation will be excluded from the dataset to prevent it from distorting the model.

Data Preparation:

Given the right-skewed distribution(Appendix 3), where the mean exceeds the median, it's evident that the data is skewed by extreme values. To preserve the distribution, 180 missing values will be replaced with the median instead of mean, as the median is less influenced by extreme data points (Appendix 9).

Imputing the values where priceUSD is “0”: there are 152 observations where value of priceUSD is “0” . Since the price of a coin cannot logically be “0”, this suggests incorrect data. Furthermore, considering the right-skewed distribution observed in the preceding paragraph's histogram, the zero values will be replaced with the median. This choice is made because the median is less influenced by extreme values (Appendix 10).

Excluding the outlier in priceUSD with a value of 39,384 (Appendix 11).

teamSize:

There are 154 instances where the value is missing.

Histogram of teamSize

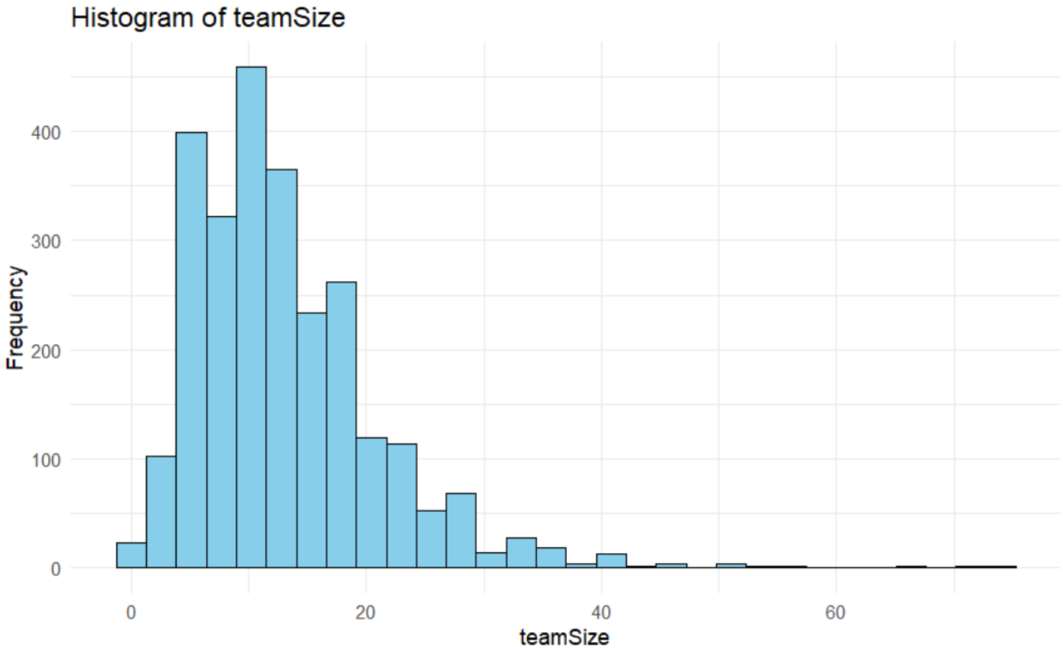

Figure 2 Histogram for "teamSize" variable.

The histogram reveals that most observations cluster within smaller team sizes, with 364 falling into the 0 to 5 range and frequencies gradually decreasing across larger team sizes. For instance, there are 802 observations for team sizes between 5 and 10, 627 for 10 to 15, and 444 for 15 to 20. Frequencies continue to decline for larger teams, with only 3 observations for team sizes between 45 and 50, and just 1 each for sizes between 50 and 55, 55 and 60, and 60 and 65.

Data Preparation:

Addressing missing values in teamSize: there are 154 instances of missing values. As evident from the histogram, the distribution is right-skewed (see Appendix 12). Therefore, to maintain the distribution and account for the skewed nature of the data, the missing values will be imputed with the median(Appendix 13). This choice is made because the median is less sensitive to outliers and helps preserve the overall distribution.

coinNum:

Histogram of coinNum

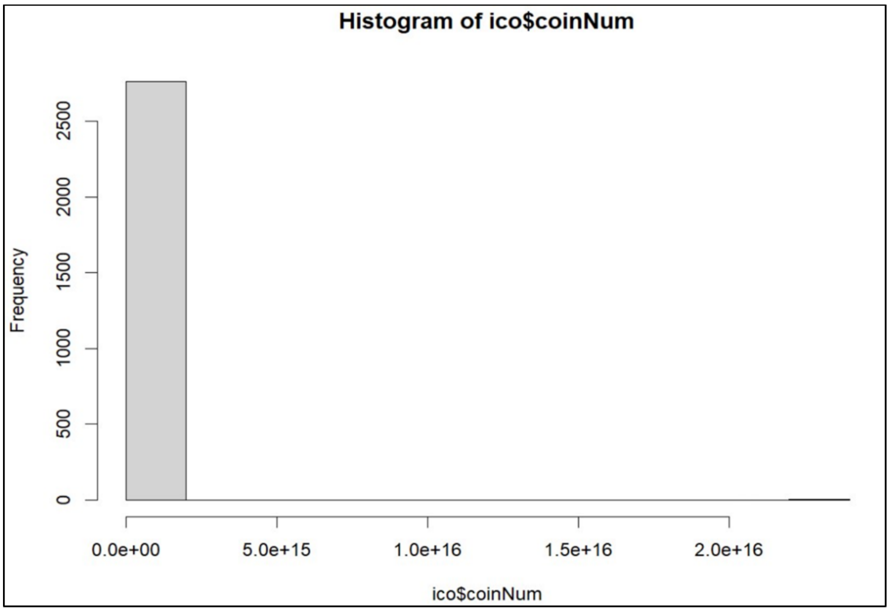

Figure 3 Histogram for "coinNUM" variable.

The hist() function output indicates that there is only a single value falling within the highest bin range, 2.2e+16 to 2.4e+16 (Appendix 4).

The summary details (Appendix 2) reveal that the specific observation corresponds to a coinNum value of 2.262e+16.

It is reasonable to assume that the observation with coinNum = 2.2e+16 to 2.4e+16 is an outlier, and it will be prudent to remove it to prevent any distortion in the prediction results.

Data Preparation:

Removing the outlier in coinNum (Appendix 14)

distributedPercentage:

The percentage of blockchain coin distributed to investors.

Boxplot of distributedPercentage

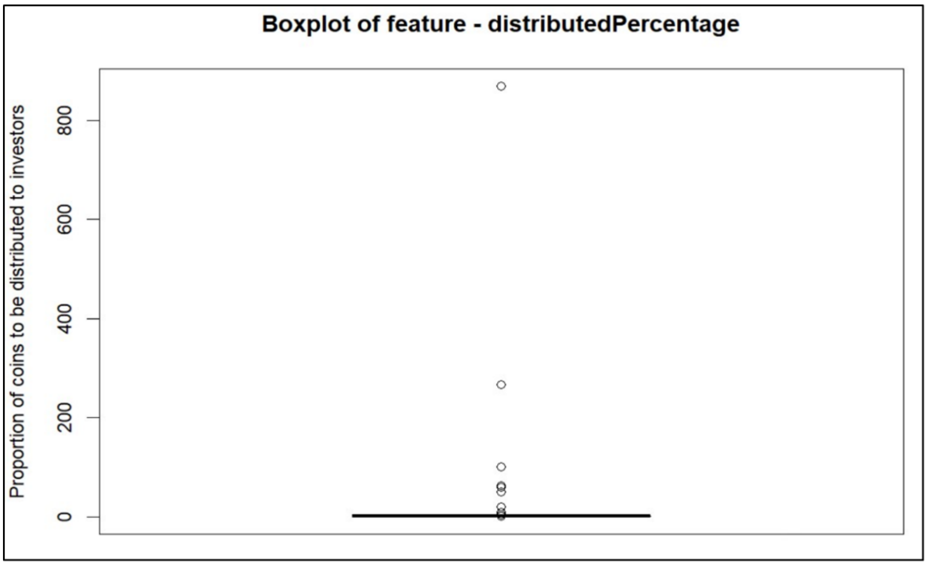

Figure 4 Boxplot for "distributedPercentage" variable.

The output (Appendix 5) indicates that there are 10 observations lying outside the extreme upper whisker, which is set at 1. Since these observations represent percentages, they cannot exceed 1. Therefore, the 10 observations exceeding 1 are incorrect values. Additionally, a value of ‘0’ is unreasonable because no venture would launch an ICO campaign offering no coins. There is one observation where the value is ‘0’, which is also incorrect. Hence, these 11 values will be removed during the data pre-processing stage.

Data preparation:

Addressing incorrect values in the feature "distributedPercentage" (see Appendix 15).

Character features:

countryRegion:

There are 71 observations where the value is blank. Since the country information is missing, these blanks will be replaced with "Unknown."

Upon visual examination of the output generated by the table() function (Appendix 6), it is apparent that some countries are duplicated due to differences in case.. These countries are: India(38) , india(1), Mexico (7), México (1), Singapore (312), SINGAPORE (1), and USA (296) usa (1).

Data Preparation:

Substituting blank values in the "countryRegion" feature with "unknown" (Appendix 16).

Standardizing the case of country names where they appear differently (Appendix 17).

platform:

Name of the blockchain platform.

Output of table() function (Appendix 7) indicates that Ethereum is the most popular platform for the projects. However, there are several values that are similar but written differently due to variations in case, spaces in suffix, or spaces in prefix. These discrepancies need to be corrected.

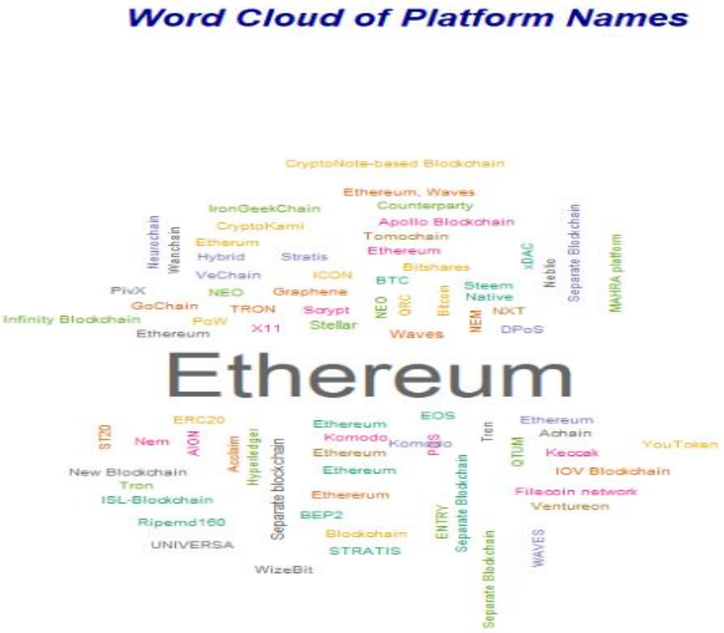

Figure 5 Word Cloud of Platform Names (Appendix 8).

Data Preparation:

Ensuring consistency in the names of platforms: Various inconsistencies are observed in the "platform" feature, primarily involving spaces before and after the name, the use of different cases, and the utilization of alternate platform names.

The approach adopted to ensure uniformity in the names of platforms is as follows:

- The approach adopted involves removing spaces in the prefix and suffix of platform names and converting all names to lowercase (Appendix 18).
- Consolidating similar platform names into one (Appendix 19).
  - Bitcoin and BTC refer to same platform
  - Ethereum and ETH all refer to same platform
  - Stellar and Stellar Protocol refer to same platform

- Blank values in the platform feature will be replaced with "unknown" (Appendix 19).

Success Variable:

An indicator variable which is set as ‘Y’ if the project achieved their funding goal (raised funding successfully) otherwise ‘N’.

Out of a total of 2767 campaigns, 1028 were successful, constituting 37.15% of the dataset, while 1739 campaigns were unsuccessful, accounting for 62.85% of the total.

Feature Engineering:

A new feature "Duration_of_campaign" be created, which represent number of days in between the startDate and endDate of each campaign (Appendix 20).

The features "ID", "startDate", "endDate", and "brandSlogan" are removed from the dataset. "ID" and "brandSlogan" lack predictive power as they solely serve as unique identifiers for each ICO campaign. Additionally, "startDate" and "endDate" are eliminated since a new feature representing the duration of each campaign has been incorporated (Appendix 21).

Modelling:

Five models have been employed, each with its unique characteristics:

Decision Tree (DT):

The decision tree (DT) classification algorithm aims to organize the dataset into a tree structure based on similarities among its elements. It iteratively partitions the dataset until one of three conditions is fulfilled: a) all observations within a leaf node are similar, b) no additional features are available for further division, or c) a specific criterion is satisfied, such as reaching a predefined limit on the number of branches. One of the primary advantages of decision trees is their straightforward interpretability (Appendix 22).

According to the model results, the initial rule is predicated on two conditions: a rating of less than or equal to 3.9 and the country being the USA. This finding aligns with the dataset, as the USA has the highest count of ventures launching ICO campaigns.

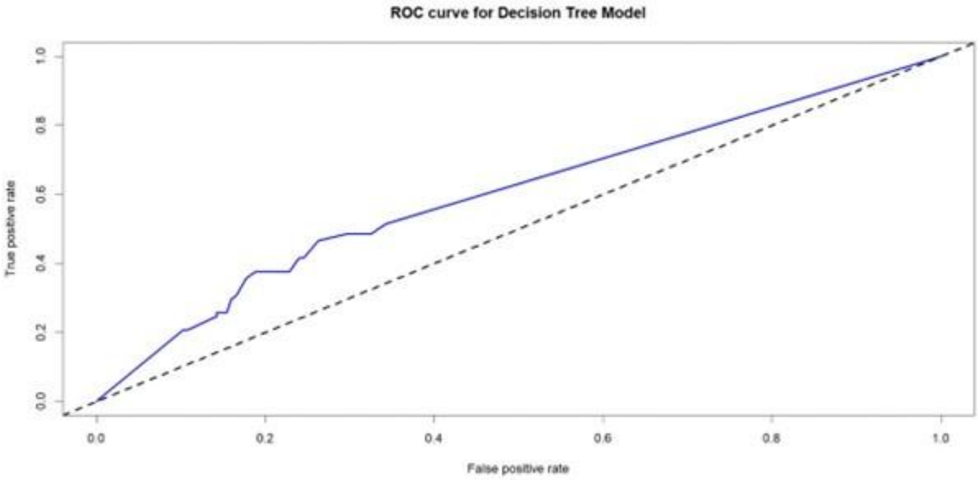

Figure 6 ROC curve for Decision Tree.

Adaboost:

Boosting is a method used to enhance classification accuracy by integrating multiple weak learners, or base algorithms. This process entails iteratively invoking the base algorithm, where each subsequent round assigns higher weights to instances that were previously misclassified. Consequently, in subsequent rounds, the base algorithm is compelled to correctly classify instances with increased weights. Ultimately, the final prediction is determined by a weighted vote based on the outcomes of each round (Appendix 23).

In this scenario, the initial split is based on the condition "rating <= 3." This split effectively divides 1,227 observations, which represents 49.5% of the total observations.

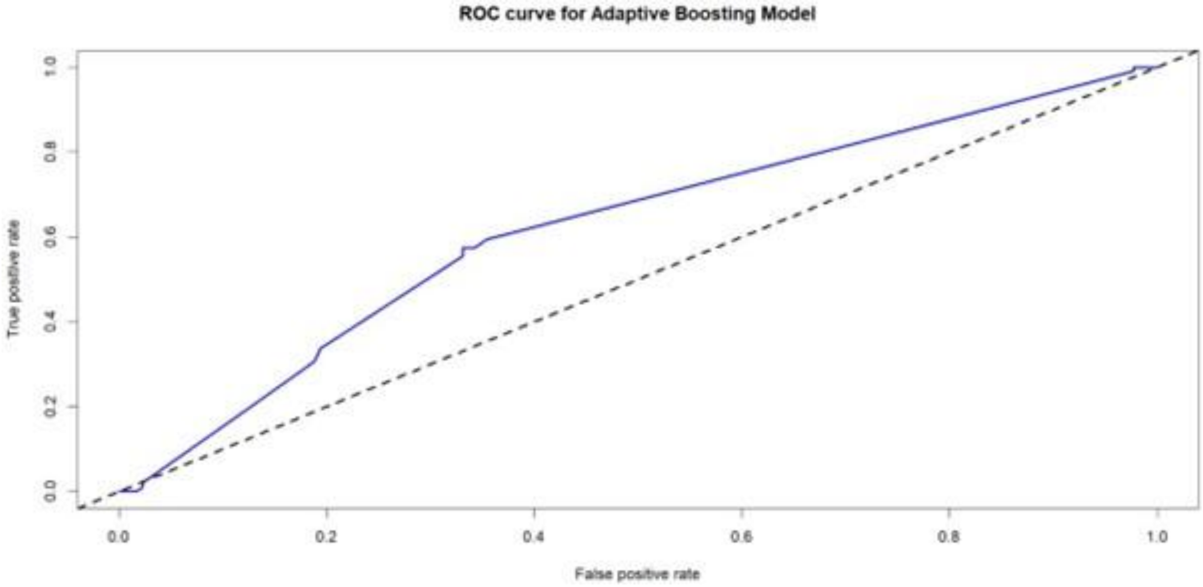

Figure 7 ROC curve for AdaBoost.

Random Forest:

Random forest is a method that aggregates the outputs of several small decision trees into a unified output. It operates on the principles of bagging and random feature selection. One of its significant advantages is its resilience against overfitting, achieved through the averaging of outputs from multiple trees or through voting in classification tasks.

Due to the limitation of random forest in processing features with over 53 categories, we utilized one-hot encoding to generate dummy features for categorical variables with just two categories in our dataset, specifically countryRegion and platform (Appendix 24).

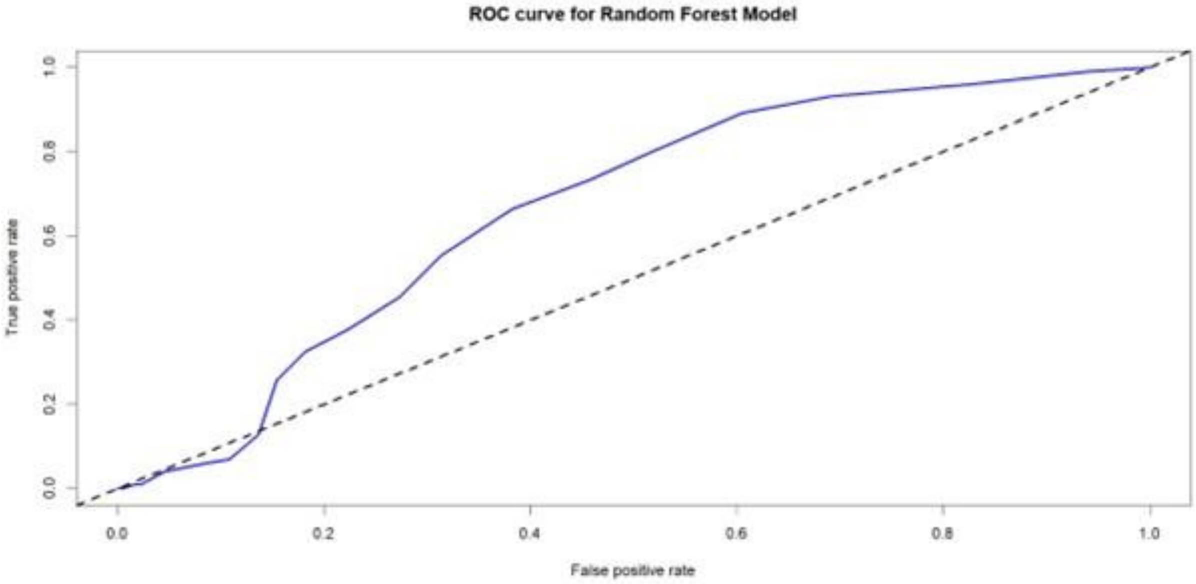

Figure 8 ROC curve for Random Forest.

Support Vector Machines (SVM):

SVM operates by identifying a dividing plane, known as a hyperplane, to separate the data into different classes. Its primary objective is to pinpoint support vectors, which aid in determining the maximum margin hyperplane (MMH). The MMH serves as a boundary line that segregates the data effectively. SVM's efficacy stems from its capability to handle datasets with numerous features and its capacity to identify non-linear hyperplanes (Appendix 25).

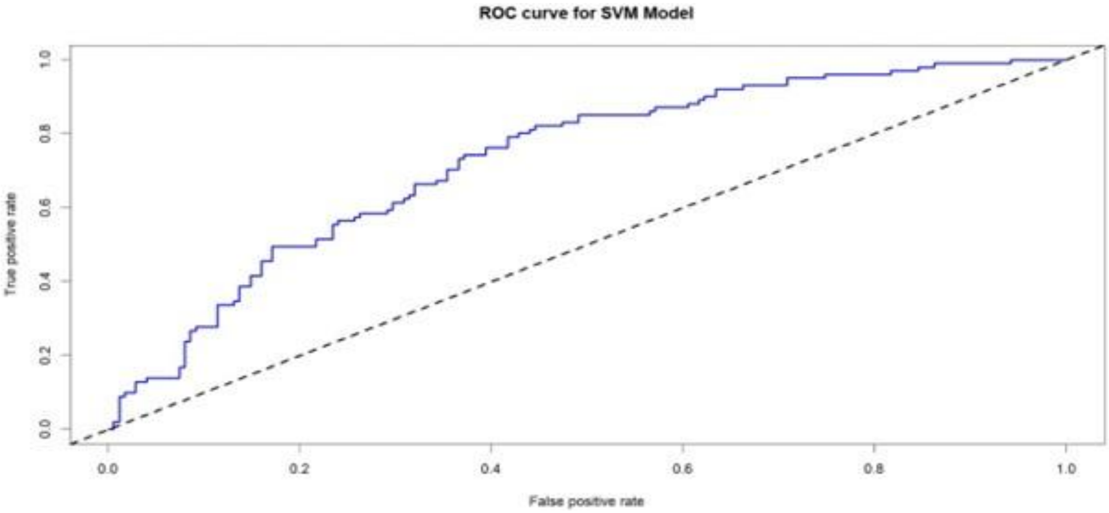

Figure 9 ROC curve for SVM.

Artificial Neural Network (ANN):

ANN, inspired by the functionality of neurons in the human brain, functions by receiving multiple inputs into a node to produce an output. Within ANN, weighted inputs are delivered to nodes, which employ functions to generate outputs. The primary appeal of ANN lies in its ability to yield outstanding results, particularly with datasets containing complex patterns (Appendix 26).

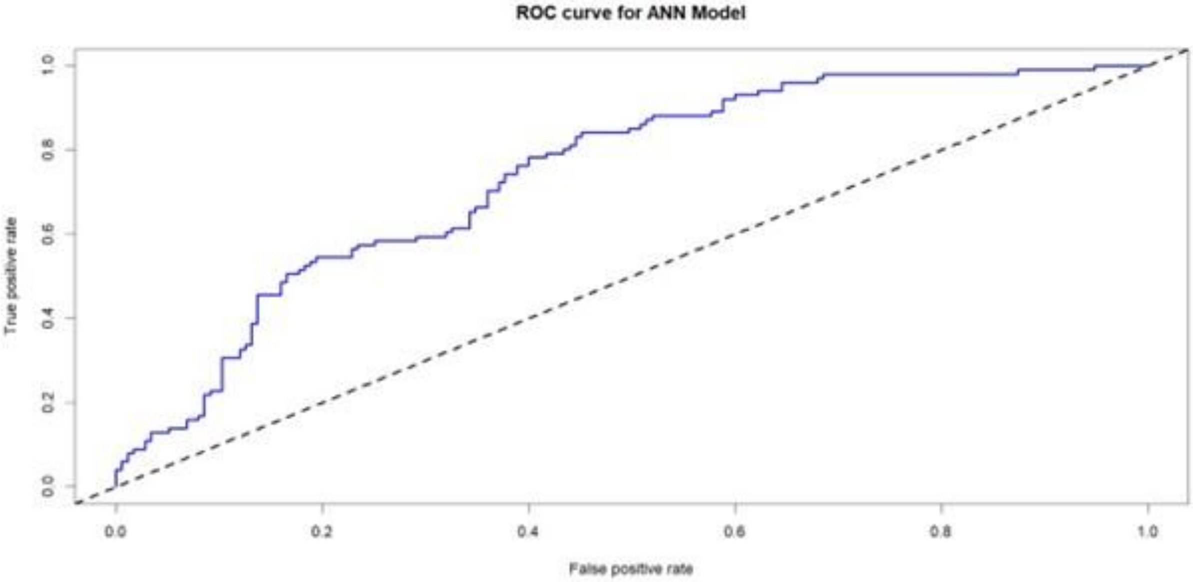

Figure 10 ROC curve for ANN model.

EVALUATION:

Models	Accuracy	AUC	Precision
Decision Tree (DT)	0.6341	0.5989	0.4989

Adaboost	0.6255	0.6144	0.4843
Random Forest	0.6313	0.6637	0.5145
Support Vector Machines(SVM)	0.6748	0.7239	0.5963
Artificial Neural Network(ANN)	0.6965	0.7389	0.6128

Accuracy:

The overall success of model predictions is captured by accuracy. It quantifies the proportion of all correct predictions, regardless of class. Accuracy is particularly valuable when the objective is to precisely detect both the positive and negative classes.

In our case, the ANN model exhibits the highest accuracy of 69.65%. However, considering the imbalance in our dataset where successful ICO campaigns constitute only 37.15% of the data, it's imperative to supplement this metric with another evaluation metric. Accuracy alone may not adequately reflect model performance on imbalanced datasets. For instance, a high accuracy score can still be achieved if the model incorrectly classifies all minority class observations as the majority class. Therefore, additional evaluation metrics are necessary to provide a comprehensive assessment of model performance.

Precision:

Precision quantifies the model's ability to avoid misclassifying negative instances as positive. A high precision indicates that the model has a low false positive rate, meaning that when it predicts a positive outcome, it is likely to be correct. Conversely, a low precision suggests that there is a high likelihood of false positives in the model's predictions.

In our scenario, the artificial neural network (ANN) exhibits the highest precision, reaching 0.6128. This indicates that out of all the instances where the ANN model has predicted an ICO campaign as successful, approximately 61.28% of them were indeed successful.

AUC:

The Area Under the Curve (AUC) is a metric used to evaluate the performance of a classification model. It measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. A higher AUC value indicates better discrimination between positive and negative classes by the model.

In our scenario, the ANN model achieves the highest AUC score of 0.7389, indicating superior predictive capability compared to other models.

Conclusion:

This report analysed factors influencing the success of Initial Coin Offerings (ICOs) and leveraged machine learning techniques to predict whether campaigns would achieve their funding goals. The dataset included 2,767 ICO projects, with the target variable indicating success ('Y' or 'N'). Data quality issues, such as missing values and inconsistencies, were addressed through imputation and standardization. Several numerical attributes like 'priceUSD,' 'teamSize,' and 'coinNum' displayed skewed distributions and outliers, which were treated appropriately.

Feature engineering led to the creation of a new feature, 'Duration_of_campaign,' to capture each project's timeframe. Other attributes, like country names and blockchain platforms, were standardized. Five machine learning models—Decision Tree (DT), Adaboost, Random Forest, Support Vector Machines (SVM), and Artificial Neural Network (ANN)—were tested, with accuracy ranging from 62.55% (Adaboost) to 69.65% (ANN). The Artificial Neural Network (ANN) model outperformed the others in terms of accuracy but still needs careful interpretation due to the class imbalance, as only 37.15% of projects were successful.

Model performance was evaluated beyond accuracy, considering precision, recall, and F1-score to account for class imbalance. Despite its highest accuracy, ANN's results could be further improved by refining features or incorporating ensemble models. Key predictors, such as 'rating' and 'countryRegion,' can assist organizers and investors in better strategizing their projects.

Machine learning approaches offer valuable insights into predicting ICO success, helping investors make informed decisions and guiding campaign organizers in refining their strategies. Future research should integrate more advanced models, new features, and larger datasets to refine predictive capabilities and enhance accuracy. The report demonstrates how machine learning can effectively predict ICO success and provide transparency and investor protection in the dynamic blockchain crowdfunding landscape.