Predicting the Success of Initial Coin Offerings (ICOs): A Comparative Analysis of Machine Learning Techniques

1 Introduction

The rise of cryptocurrency has been one of the most significant technological trends in recent years. Cryptocurrencies, built on decentralized networks using blockchain technology, are gaining traction among investors and businesses alike. The success of Bitcoin, advancements in blockchain technology, and increasing adoption of cryptocurrencies like Ethereum have led to the development of many other digital currencies [1]. Before public trading, these new cryptocurrencies often undergo a funding phase known as an Initial Coin Offering (ICO) [2].

An ICO, sometimes called a token sale, is a relatively new fundraising method where cryptocurrency projects sell their tokens to the public in exchange for established cryptocurrencies like Bitcoin or Ethereum [3]. This method allows startups and developers to rapidly raise significant funding with minimal effort, bypassing traditional intermediaries and reducing transaction costs. This gives projects immediate access to global investors [4]. Most ICO crowdfunding campaigns adopt an "all-or-nothing" model, where the fundraising team sets a specific funding goal. If the project reaches this goal within the specified timeline, the ICO is deemed successful, and the funds are released to the project team. Otherwise, the team receives nothing. As a result, predicting ICO success is crucial for both investors and project developers to minimize risk and optimize resource allocation.

The goal of this report is to predict whether an Initial Coin Offering (ICO) project will meet its fundraising goal using machine learning models. By analyzing a dataset that includes features like team size, coin price, rating, platform, marketing presence (e.g., videos, GitHub), and campaign duration, this study aims to identify patterns that contribute to successful campaigns. Understanding the factors that determine success is crucial for boosting investor confidence, assessing market potential, and strategically managing risks. This report will provide valuable insights into the predictors and patterns of ICO success, helping entrepreneurs, developers, and investors make more informed decisions in the rapidly evolving cryptocurrency market.

2 Data Understanding

Our dataset comprises details from 2,767 Initial Coin Offering (ICO) projects, each represented by 16 distinct attributes. These attributes cover a range of both numerical and categorical data that provides a comprehensive view of factors believed to influence the success of ICO fundraising

efforts.

2.0.1 Key Statistics and Data Completeness

- **Total Entries and Features**: The dataset includes 2,767 entries (projects) across 16 features (attributes).
- **Missing Data Insights**:
  - `priceUSD`: 180 entries are missing (6.51% of total data).
  - `countryRegion`: 71 entries are missing (2.57% of total data).
  - `teamSize`: 154 entries are missing (5.57% of total data).
  - `platform`: 6 entries are missing (0.22% of total data).

  Overall, 2,392 projects have complete information across all specified attributes.

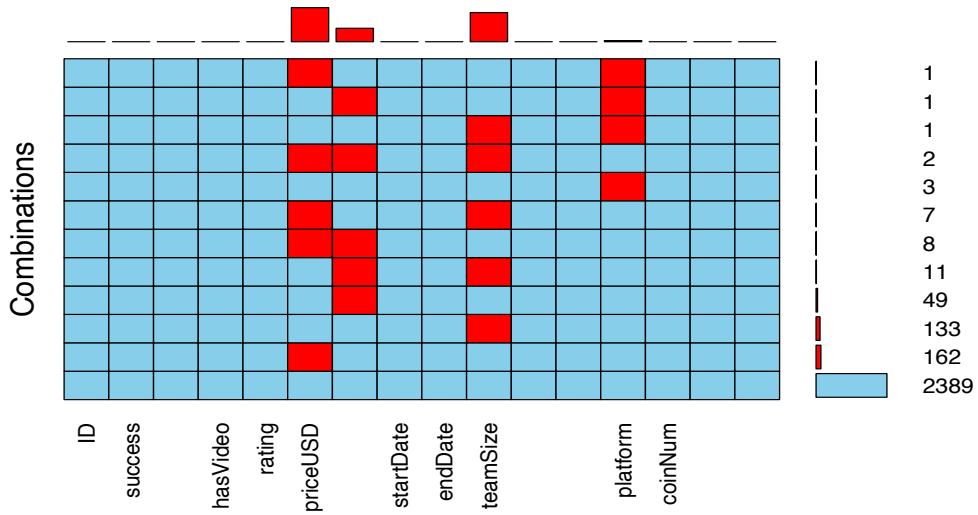


Figure 1: Missing Value Plot highlighting the prevalence of missing data across features.

2.1 Detailed Attribute Analysis

2.1.1 Numerical Attributes

- `priceUSD`: The typical coin price is very low at $0.12, though it can soar up to over $39,000, indicating significant variability.

- `teamSize`: The median team consists of 12 members, with some teams having as many as 75 members.
- `rating`: Projects are generally rated around 3.12 out of 5, suggesting moderate satisfaction.
- `distributedPercentage`: About 55% of coins are typically distributed to investors, with a wide range of distribution percentages observed.

2.1.2 Categorical Attributes

- `success`: Out of the total projects, 1,028 succeeded in reaching their funding goals, whereas 1,739 did not.
- `hasVideo`, `hasGithub`, `hasReddit`: These indicators show whether projects include promotional or informative content such as videos or have active GitHub or Reddit pages.
- `countryRegion`: This indicates the geographical location of the project, which can influence investor interest and regulatory impacts.
- `platform`: Identifies the blockchain technology used, crucial for project compatibility and functionality.

2.2 Visual Analysis and Distributions

The dataset's distributions have been visualized to better understand each feature's behavior and influence (Figure 2).

- **Critical Observations**:
  - `priceUSD` and `coinNum`: Skewed distributions indicate that most projects price their coins low and offer them in large numbers.
  - `rating`: Ratings are well-distributed across the scale from 1 to 5, showing a variety of project qualities.
  - `hasVideo` and `minInvestment`: A majority of projects include videos and specify minimum investment amounts, potentially to increase transparency and investor trust.
  - `teamSize`: Team sizes vary widely, highlighting different operational and developmental scopes.
  - `distributedPercentage`: Distribution percentages mostly fall below 70%, reflecting varying strategies in coin allocation.

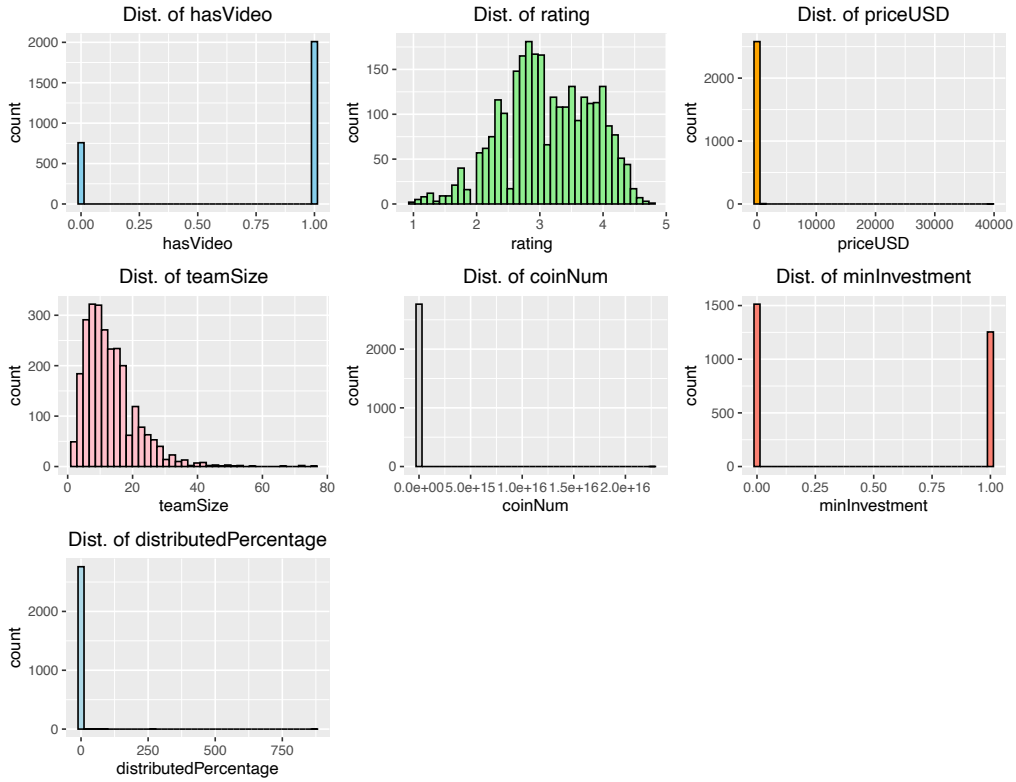


Figure 2: Distribution plots illustrating important features, such as coin price, team size, and others.

These insights from the initial data analysis set the stage for deeper examination and modeling to predict ICO success. The distributions and missing data points noted here will guide the preprocessing steps necessary to ready the dataset for effective machine learning applications.

3 Data Preparation

In this section, we summarize the steps taken to clean and prepare the dataset for machine learning modeling. Key activities included handling missing values, creating new features, and understanding relationships among predictors.

3.1 Data Cleaning

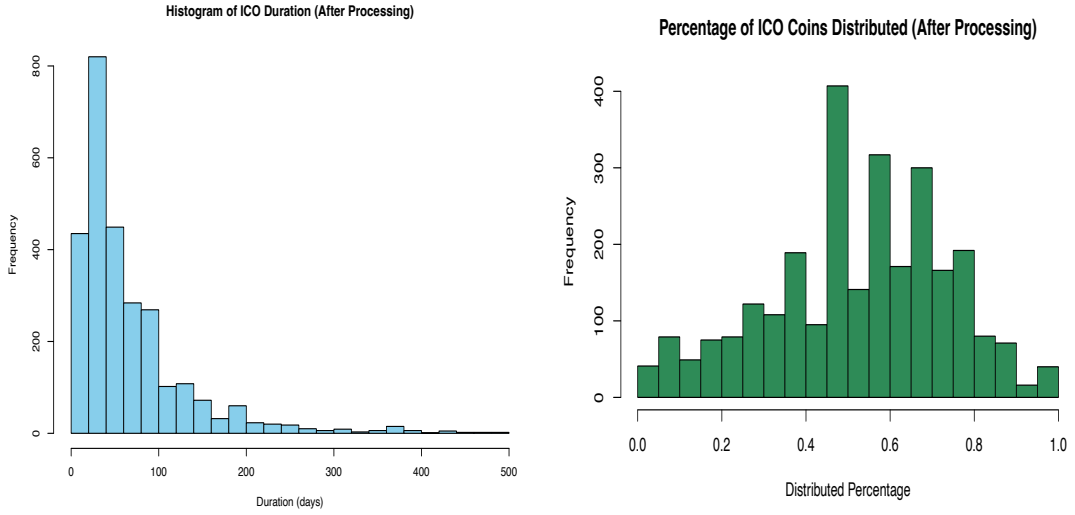
Drop the ID column: The ID feature was dropped since it doesn't add value to predictions.

- **Numerical Variables**: Missing values in `priceUSD` and `teamSize` were filled using their respective median values to maintain data integrity without skewing distributions.
- **Categorical Variables**: Missing values in `countryRegion` were replaced with 'unknown', while those in `platform` were replaced with 'Ethereum', as this was the most frequently used platform, as seen in the word cloud.

3.2 Feature Engineering and Outlier Detection and Handling

- **Date Transformations**: `startDate` and `endDate` were converted from strings to `datetime` objects to facilitate accurate period calculations.
- **Campaign Duration**: `campaignDuration` was created by computing the difference between `endDate` and `startDate`. Entries with negative durations were corrected by swapping the dates, and campaigns lasting over 500 days were removed to exclude outliers. The resulting distribution is similar to an exponential distribution, as seen in Figure 3.
- `distributedPercentage`: Any values above 100% were removed to ensure logical consistency in percentage-based data.

3.3 Visual Analysis and Distributions After Processing



(a) (b)

Figure 3: (a) Histogram of ICO Campaign Duration (b) Percentage of ICO Coins Distributed after Processing, illustrating refined data characteristics.

- `priceUSD`: Extreme values were capped at the 8th and 97th percentile levels to mitigate the influence of outliers.
- `coinNum`: Extreme values were adjusted to the 1st and 99th percentiles. A logarithmic transformation was also applied for normalization, as shown in Figure 4, resembling a normal distribution.

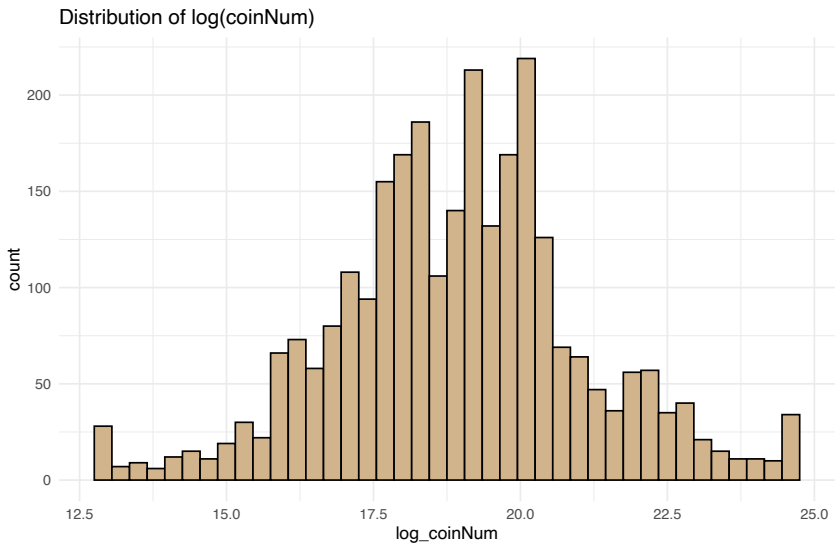


Figure 4: Distribution of Coin Number (Log-Transformed) after Processing, showing normalized data suitable for analysis.

3.4 Relationship Analysis Among Predictors

- **Correlation Matrix**: We used a correlation matrix to identify relationships among numerical predictors. From the correlation plot (Figure 5), we observe that features like `hasGithub`, `hasReddit`, and `hasVideo` are moderately correlated with `rating`, while `teamSize` correlates with `rating`.

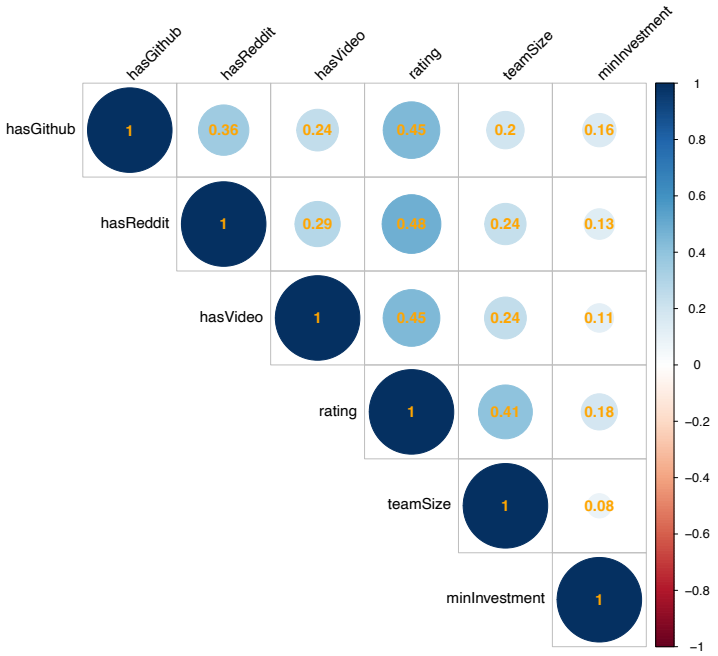


Figure 5: Correlation Plot of Numerical Variables, highlighting key relationships between features.

- **Top 15 Countries by ICO Count**: According to Figure 6, the visualization provides insights into the countries and regions most commonly involved in ICO fundraising.

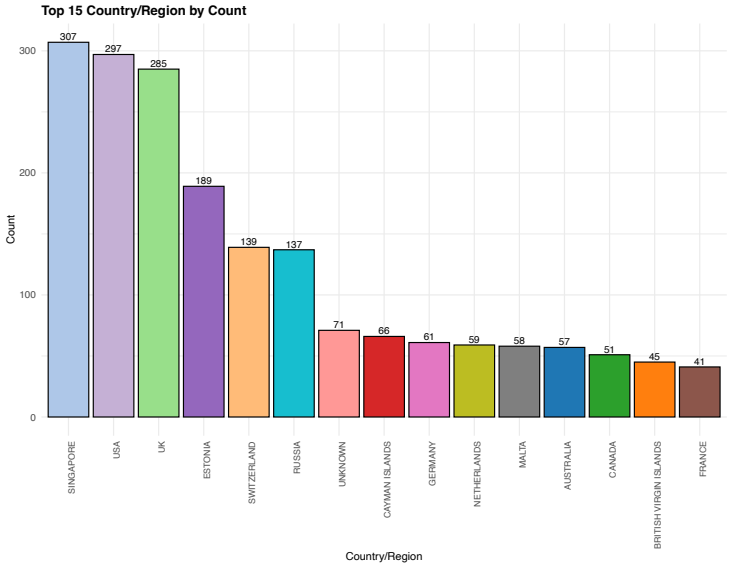


Figure 6: Top 15 Countries by ICO Project Count, showing the prevalence of ICO activities in key financial markets.

- **Word Cloud of Platforms and Regions**: The word cloud visualization provides a snapshot of the most prominent platforms and regions in the dataset. From the word cloud shown in Figure (7), we can infer that most ICOs were launched using the 'Ethereum' platform, with the most common countries being 'Singapore', 'UK', and 'USA'.

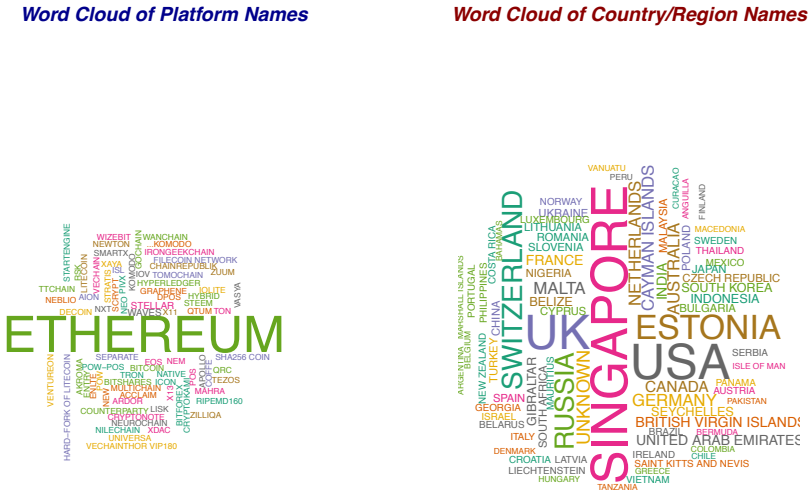


Figure 7: Word Cloud of Platform and Region Names, visually representing the most common platforms and active regions in the dataset.

4 Modeling

In this section, we detail the process of building and evaluating various classification models to predict ICO success. The approach includes selecting suitable machine learning algorithms, training them on the training dataset, and assessing their performance on the test dataset.

4.1 Split Training and Test Dataset

To ensure reliable evaluation of the models, we first split the dataset into training and test subsets. A random seed was set for reproducibility, and 70% of the samples were allocated to the training dataset, leaving the remaining 30% for the test dataset. This resulted in 1,917 samples for training and 821 samples for testing.

4.2 Model Selection

Various classification models were considered, including Decision Trees, Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and Logistic Regression. However, Neural Networks were excluded due to the mixed data types requiring extensive preprocessing. Naive Bayes was excluded because it assumes independent predictor variables, which is not supported by the dataset's correlation matrix.

The selected models were:

- **Decision Tree**: A simple, interpretable model that handles non-linear relationships well, providing a clear understanding of data patterns.
- **Random Forest**: An ensemble method combining multiple decision trees to reduce overfitting and improve predictive accuracy.
- **K-Nearest Neighbors (KNN)**: A non-parametric method that is effective for smaller datasets and provides classification based on majority voting among nearest neighbors.
- **Logistic Regression**: An efficient, interpretable model suitable for binary classification tasks.

4.2.1 Decision Tree Model

The Decision Tree (DT) algorithm splits data into a tree-like structure based on feature similarity. Splits continue until all observations in a leaf node are homogeneous, no features remain for further splitting, or predefined conditions like maximum depth are met. DT models excel in interpretability and pattern recognition. For this dataset, the Decision Tree achieved an accuracy score of 67.1%, indicating it could predict ICO success accurately in most cases.

4.2.2 Random Forest Classification Model

The Random Forest classification model is an ensemble learning method that aggregates predictions from multiple decision trees to improve classification accuracy. Due to its tree-based structure, the model is robust to outliers and requires no data scaling. An additional advantage is that it provides feature importance rankings, helping understand the relative influence of each feature on predictions. The Random Forest model achieved an accuracy rate of 70.6%. Figure 8 highlights the most critical features influencing the model's predictions, where `rating`, `campaign duration`, `teamSize`, and `coinNum` emerged as highly influential.

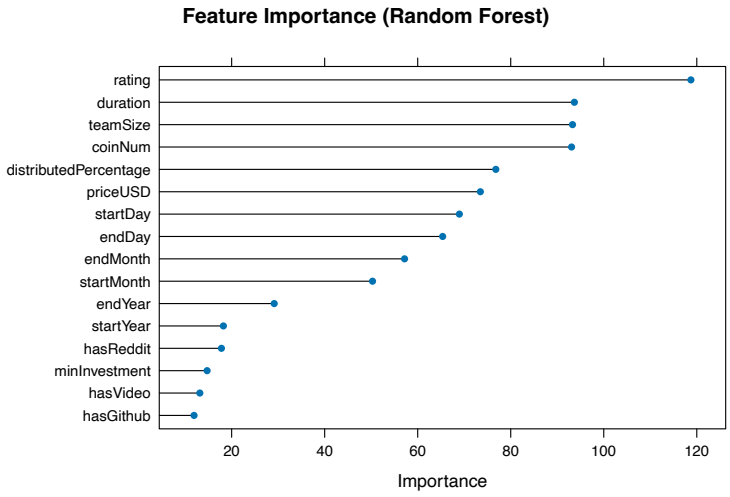


Figure 8: Important features predicted by the Random Forest model.

4.2.3 K-Nearest Neighbors Model

K-Nearest Neighbors (KNN) is a classification method based on distance computation. It classifies each data point by a simple majority vote of its nearest neighbors in feature space. Identifying the optimal K value (number of nearest neighbors considered) is crucial to improving accuracy. The optimal K value was determined by plotting the error rate across various K values (Figure 9) and selecting the value with the lowest error rate. Here, K=27 was found optimal, achieving an accuracy score of 70.03%.

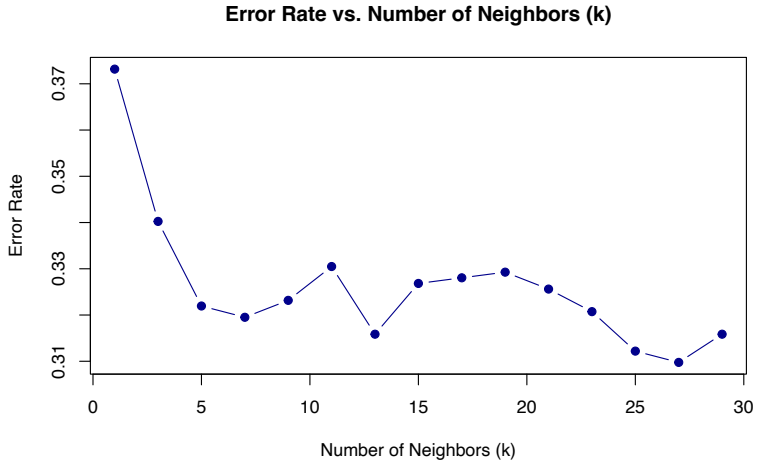


Figure 9: Optimal K value selection for K-Nearest Neighbors model.

4.2.4 Logistic Regression Model

Logistic Regression is a classification algorithm that models the probability of binary outcomes using a logistic function. It is computationally efficient and interpretable, making it suitable for classification tasks. For this dataset, the Logistic Regression model achieved an accuracy of 69.7%. Despite its simplicity, it provided a structured approach to evaluating feature importance and predicting outcomes.

5 Evaluation

To assess each model's performance in predicting ICO success, we compared their accuracy, F1 scores, and AUC values:

Model	Accuracy	F1 Score	AUC
Decision Tree	0.6626	0.7746	0.66
Random Forest	**0.7065**	**0.7935**	**0.73**
K-Nearest Neighbors	0.7004	0.7804	0.65
Logistic Regression	0.6979	0.7825	0.72

Table 1: Performance Metrics for Each Model

Model Interpretation and Insights:

- **Random Forest** achieved the highest accuracy (0.7065), F1 score (0.7935), and AUC (0.73), demonstrating its robustness in predicting successful ICO campaigns. Its ability to combine multiple trees enhances generalizability and improves classification performance.
- **KNN** had an accuracy of 0.7004 and an F1 score of 0.7804. Although this non-parametric method can effectively classify ICOs, it struggled with finding an optimal K value that balances bias and variance.
- **Logistic Regression** closely followed Random Forest with an accuracy of 0.6979, an F1 score of 0.7825, and an AUC of 0.72. Despite its simplicity, logistic regression efficiently handled binary classification, highlighting the importance of factors such as team size, coin price, and campaign duration.
- **Decision Tree** achieved relatively lower scores (accuracy = 0.6626, F1 = 0.7746), indicating difficulties in distinguishing between successful and unsuccessful ICOs due to overfitting.

ROC Curves Analysis: Figure 10 shows the ROC curves for each model. The superior performance of Random Forest and Logistic Regression models is evident, as they maintain higher true positive rates across all false positive rates. This is crucial for ICO projects, as accurately identifying successful campaigns reduces the likelihood of false positives, thus minimizing financial risk for investors and project developers. KNN and Decision Tree exhibit comparatively lower performance, potentially leading to a higher misclassification rate.

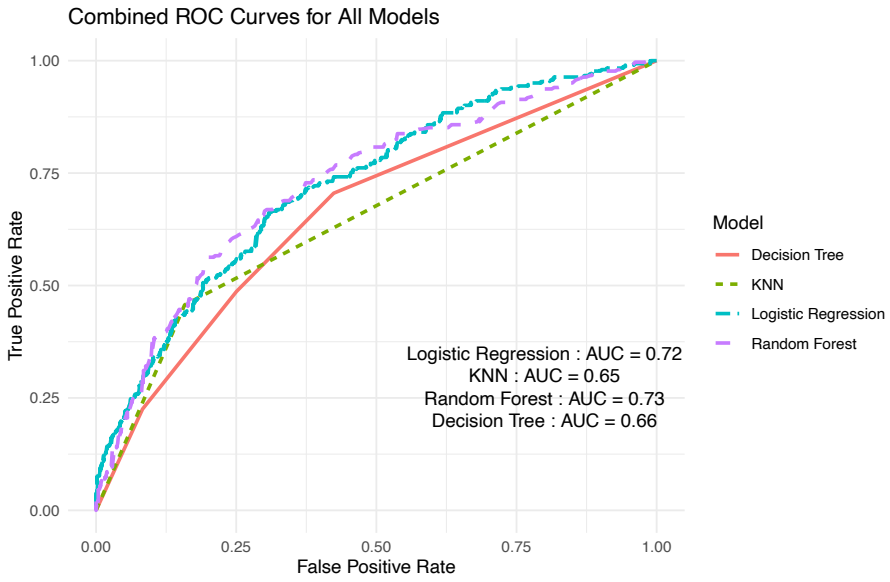


Figure 10: Combined ROC Curves for All Models, showing performance for Decision Tree, Random Forest, KNN, and Logistic Regression.

Confusion Matrices Interpretation: Figure 11 provides the confusion matrices for all models. Both Random Forest and Logistic Regression models have the fewest false predictions, aligning with their higher accuracy scores. This ability to distinguish between successful and unsuccessful campaigns makes them more reliable in this context. For ICO projects, accurate classification ensures that investors can assess projects with confidence, focusing on campaigns that are more likely to meet their funding goals.

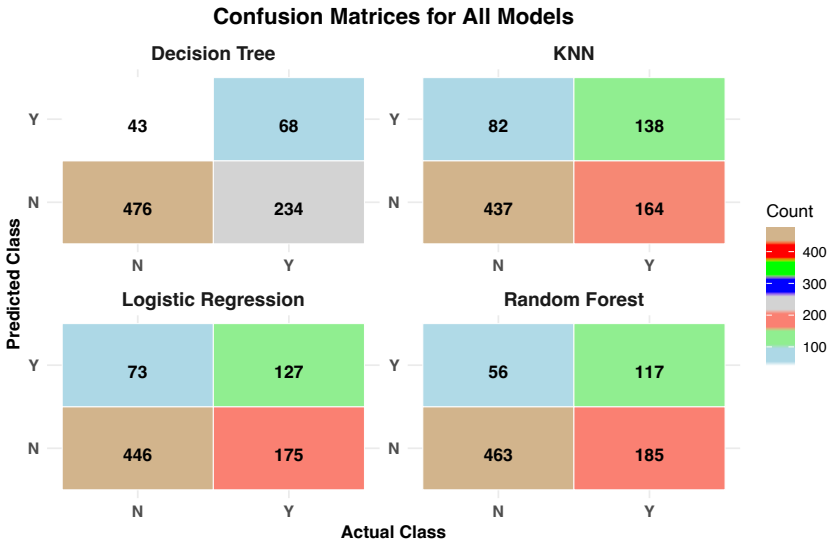


Figure 11: Confusion Matrices for All Models, comparing the classification performance of each model.

6 Conclusion

After evaluating multiple machine learning models to predict ICO project success, the Random Forest and Logistic Regression models emerged as the most reliable predictors, achieving high accuracy, F1 scores, and AUC values. These models demonstrated their effectiveness in distinguishing between successful and unsuccessful ICO campaigns, making them valuable tools for investors and project teams alike.

Interpretation for ICO Project:

- Investor Confidence: The strong performance of Random Forest and Logistic Regression models provides useful guidance for investors by identifying which ICOs have a higher probability of reaching their fundraising goals.
- Project Strategies: Project teams can leverage insights from these models to refine their strategies by emphasizing the influential factors that contribute to successful funding (e.g., campaign duration, rating, team size).
- Resource Allocation: By accurately identifying potential risks and market opportunities, investors and project developers can strategically allocate resources, improving the success rate of their projects.

7 Limitations of Work and Future Work

7.1 Limitations of Work

1. **Imputation Bias**: The approach of imputing missing values in `priceUSD` and `teamSize` with the median may introduce bias into the results. This can alter the dataset's natural distribution and potentially skew model outcomes. Obtaining accurate values directly from the data owners would enhance the reliability and validity of the models.

2. **Feature Dependency**: Some models, such as Naive Bayes, assume that predictor variables are independent of each other. This assumption may not hold true in this dataset, where features might be interdependent, potentially leading to suboptimal model performance.

7.2 Future Work

1. **Sentiment Analysis**: Incorporating the feature `brandSlogan` using text mining or sentiment analysis techniques could provide valuable insights into how a project's messaging influences its success. Analyzing the sentiment and content of these slogans may reveal patterns that correlate with fundraising success, offering a novel angle for future model enhancements.

2. **Additional Features**: Future research should consider exploring additional features, such as the composition of the project team or prevailing market trends at the time of the ICO. These factors could further refine the predictive accuracy of the models and provide a more nuanced understanding of the dynamics influencing ICO outcomes.

References

[1] S. Arsi, S. Ben Khelifa, Y. Ghabri, and H. Mzoughi, "Cryptocurrencies: Key risks and challenges," in *Cryptofinance: A new currency for a new economy*, pp. 121–145, World Scientific, 2022.

[2] C. Fisch, "Initial coin offerings (icos) to finance new ventures," *Journal of Business Venturing*, vol. 34, no. 1, pp. 1–22, 2019.

[3] P. P. Momtaz, "Initial coin offerings," *Plos one*, vol. 15, no. 5, p. e0233018, 2020.

[4] W. Xu, T. Wang, R. Chen, and J. L. Zhao, "Prediction of initial coin offering success based on team knowledge and expert evaluation," *Decision Support Systems*, vol. 147, p. 113574, 2021.