# Business Understanding

Crowdfunding, a popular strategy for businesses and startups, has evolved with the advent of Initial Coin Offerings (ICOs). This has gained major attention owing to its distinct method of fundraising via issuing digital coins or tokens. ICOs use blockchain technology — as opposed to conventional forms of fundraising — to provide investors the opportunity to partake in early-stage ventures by buying digital assets.

While crowdfunding is not a new concept, ICOs have revolutionized fundraising by introducing digital tokens that serve as utility or ownership rights, facilitated by blockchain technology. The ICO process involves various stakeholders, including investors, fundraising teams, and listing websites. Investors analyse the project or venture’s potential to determine if they should contribute funds against digital tokens. The fundraising team develops and presents the concept or project to investors, explaining its goals, roadmap, and the advantages of taking part in the ICO.

Whether an ICO campaign will be successful depends on factors like project viability, market demand, business credibility, and investor sentiment. Due to the high risks involved and the dynamics of the cryptocurrency market, estimating the result of an ICO fundraising campaign accurately, is key for both fundraisers and investors.

In this technical report, we explore how to predict the success of ICO fundraising by leveraging machine learning techniques. Using a dataset that contains attributes of different ICO projects, our focus is to create predictive models that can assess if a fundraising venture will successfully attain its funding goal. Using rigorous data analysis, evaluation, and modelling, we aim to offer actionable insights for fundraising teams and investors, helping them optimize their strategies and make informed decisions in the ever-evolving landscape of ICOs.

# Data Understanding

The dataset provided for this machine learning assignment contains information on Initial Coin Offering (ICO) projects from various fundraising teams or companies. Before delving into the modelling and prediction task, it's essential to thoroughly understand the data to make informed decisions throughout the analysis process.

**How big is the data?**

The dataset consists of 2767 observations (rows) and 16 variables (columns), providing a substantial amount of data to analyse and model.

**How does the data look like?**

By examining a random sample of 5 rows from the dataset, we can observe the structure and content of the data. As shown in Figure 1, each row represents a different ICO project, with information such as the project's ID, success indicator, start and end dates of the campaign, number of coins issued, price per coin, team size, country, brand slogan, rating, minimum investment requirement, percentage of distributed coins, platform, and indicators for the presence of video, Github page, and Reddit page.

A screen shot of a computer

Description automatically generated

*Figure 1: Random 5 rows of dataset*

**What are the data types of all variables?**

The data types of the columns include integers, numeric, and characters. These data types represent numerical variables, such as ID, priceUSD, rating, distributedPercentage, coinNum, teamSize, as well as categorical variables, such as countryRegion, and platform as shown in Figure 2. We also notice that brandSlogan is a long text string whereas startDate and endDate are in character format.

A screenshot of a computer program

Description automatically generated

*Figure 2: Structure of dataset*

**Are there any missing values?**

Upon inspection, it was found that some columns contain missing values (including NA, blanks, and whitespaces) such as priceUSD, teamSize, countryRegion, and platform as shown in Figure 3.

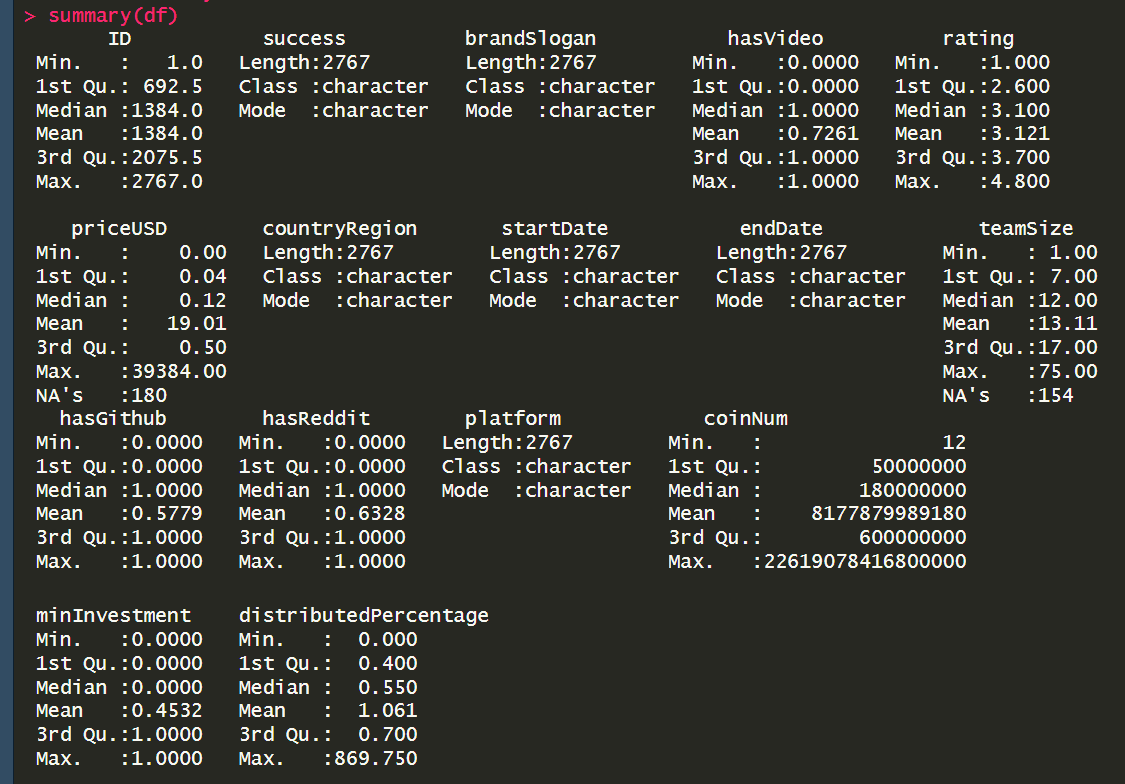
A black screen with white text

Description automatically generated

*Figure 3: Missing values in dataset*

**How does the data look mathematically?**

We also looked into the descriptive statistics for all variables which revealed the distribution and variability of the data. For instance, the mean and standard deviation of variables like rating, priceUSD, teamSize, and coinNum provide a sense of the typical values and their spread within the dataset as shown in Figure 4.



*Figure 4: Summary of dataset*

**Are there any duplicate values?**

There are no duplicate rows present in the dataset as we see in Figure 5, indicating that each observation represents a unique ICO project.

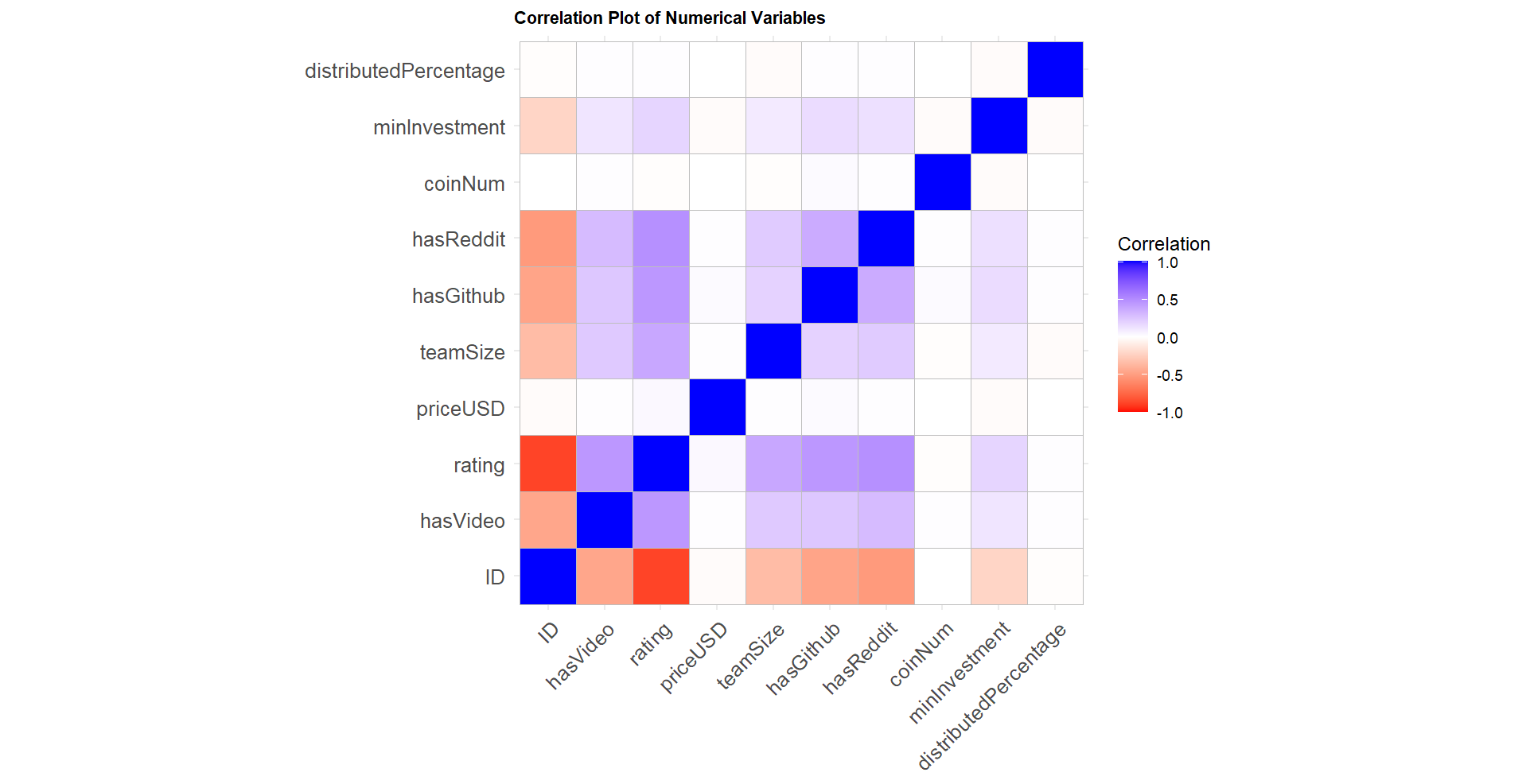
A black screen with white text

Description automatically generated

*Figure 5: Duplicates in dataset*

**Is there any correlation between the variables?**

By computing the correlation plot between numerical variables, we can assess the relationships between different relevant features. We can see from Figure 6 that there is a high positive correlation between rating to variables like hasReddit, hasGithub, hasVideo, and teamSize implying that projects who have either mentioned their official Reddit & GitHub pages or have provided a video on their campaign page along with projects with bigger teams tend to have a better rating than others.



*Figure 6: Correlation plot of numerical variables*

Understanding these aspects of the dataset lays the foundation for subsequent data preparation, modelling, and evaluation steps in the machine learning analysis process.

# Data Cleaning

The first step in the data cleaning process was to drop the ID column, as it served no meaningful purpose in our analysis. Subsequently, attention was turned to understanding and preprocessing the categorical variable 'success,' which indicated whether a campaign achieved its fundraising goal. This involved factorizing the success column from 'Y' and 'N' to 'Yes' and 'No,' respectively, to enhance readability and interpretation. Additionally, we notice from Figure 7 that 37.2% (1028 out of 2767) of the ICO projects were successful in reaching their campaign goal.

A black screen with pink text

Description automatically generated

*Figure 7: Distribution of target variable*

We now look into standardizing and cleaning the 'countryRegion' variable. This entailed converting all values to lowercase, removing irrelevant spaces, and checking for unique values. Based on this assessment, minor corrections were made, such as updating 'curaçao' to 'curacao' and 'méxico' to 'mexico,' ensuring consistency and accuracy in the representation of countries. Furthermore, the start and end dates of each crowdfunding campaign were transformed into the Date format to facilitate temporal analysis. The duration of each ICO campaign was then calculated using the start and end dates as shown in Figure 8, and records with negative durations were removed as they seemed wrong.

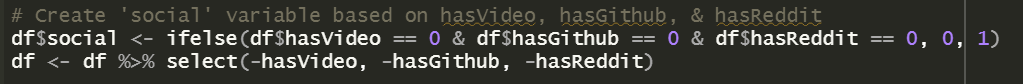
A screenshot of a computer program

Description automatically generated

*Figure 8: Calculating ICO duration*

Similar to the 'countryRegion' variable, the ‘platform’ variable was cleaned, and certain values were updated to ensure consistency and eliminate redundancy. For instance, 'btc' was updated to 'bitcoin,' 'eth’ ‘ethererum’ and ‘etherum' were updated to 'ethereum,' and so forth. Additionally, values greater than 1 in the 'distributedPercentage' variable and zero values in priceUSD were filtered out, as percentages cannot exceed 1 and price of each coin cannot be zero.

During the data understanding phase, we noticed that the variables hasVideo, hasReddit, and hasGithub have a strong positive correlation amongst each other. Hence, we create a new dummy variable ‘social’ as shown in Figure 9, which has the value of 0 if all these 3 variables are 0, else it will have the value of 1. So ‘social’ having value of 1 means that the company has either put up a video or have provided their reddit and/or github page. The original 3 variables were then dropped as those were already factored in with the newly created variable.



*Figure 9: Creating social variable*

Incorporating external data sources was a key aspect of enhancing the richness and depth of the dataset. This involved adding data for GDP, Inflation, Unemployment, and Bitcoin prices (in USD) from external sources such as the World Bank Open Data and CoinMarketCap . The bitcoin prices data was downloaded from the CoinMarketCap website, which is available to download as an open license, whereas the GDP, Inflation, and Unemployment data was fetched directly from the WDI package in R. To facilitate the integration of this external data with the existing dataset, new columns were created to store relevant information such as the year which was extracted from startDate and the ISO3 country code obtained from countryRegion using the countrycode package in R. Finally, the external data was merged with the existing dataset, and a final check for missing values was conducted before saving the cleaned dataset for further analysis. At the end of the data cleaning process, we now have a total of 2594 observations (rows) and 18 variables (columns) in our data set as we can see in Figure 10.

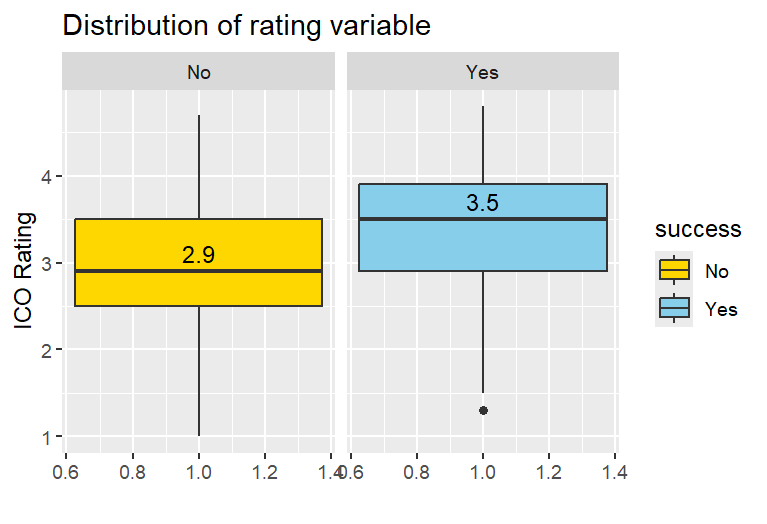
A computer screen shot of a black and white screen

Description automatically generated

*Figure 10: Missing values & dimension of cleaned dataset*

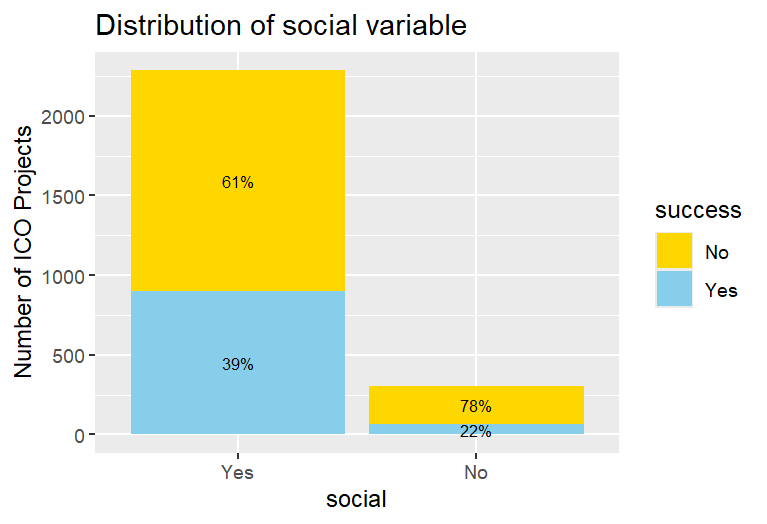
# Data Exploration

The primary objective of the data exploration phase is to gain a deeper understanding of the distributions and relationships of the predictors with regards to the target variable. We began by examining the distribution of each variable to identify potential insights into the factors influencing crowdfunding campaign success. For variables such as ‘rating’ we explored the distribution, which provided insights into the quality and credibility of crowdfunding projects. The box plot analysis in Figure 11 revealed that successful projects typically had a higher median rating of 3.5 compared to unsuccessful ones which have a median rating of 2.9, highlighting the significance of project quality in attracting investor support.



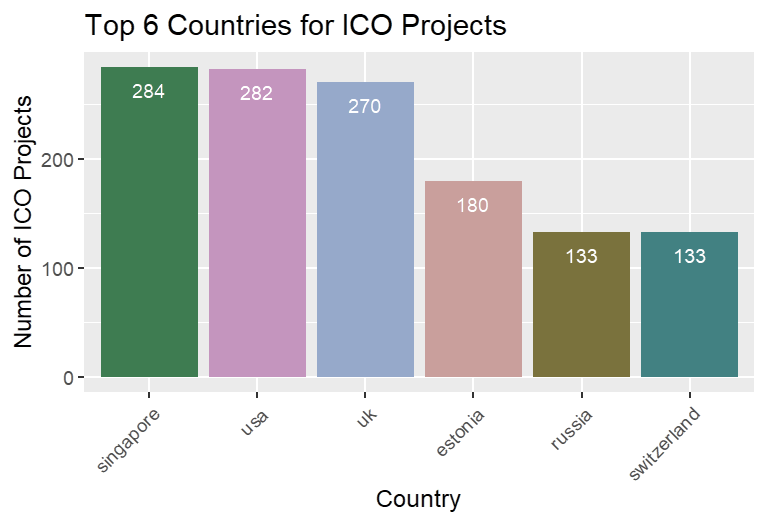
*Figure 11: Distribution of rating*

Upon delving on the ‘social’ variable, we can easily see from Figure 12 that teams or projects that do not have any videos or have no presence on reddit/github have a higher chance (78%) of being unsuccessful when it comes to reaching the crowdfunding target.

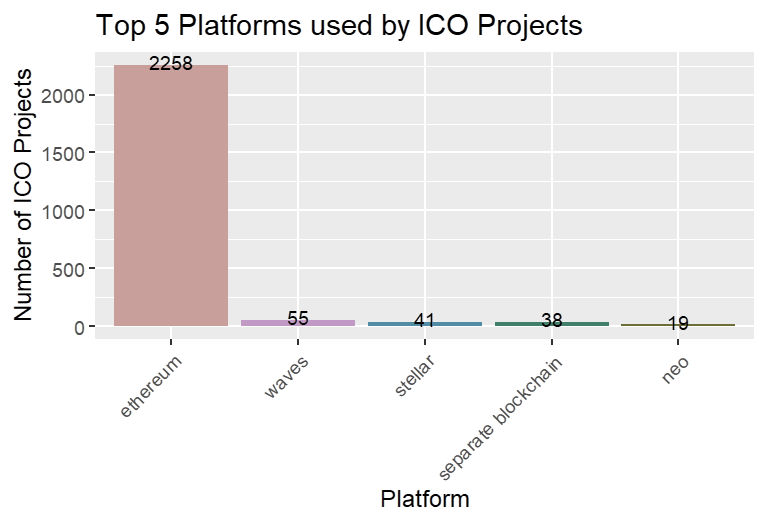


*Figure 12: Distribution of social*

Furthermore, we analysed the distribution of 'countryRegion' and 'platform' variables to identify geographical and technological trends in crowdfunding activity. Notably, we observed from Figure 13 and Figure 14 that Singapore, USA, and the UK were among the top contributors to ICO projects, while Ethereum emerged as the dominant platform choice for most projects.

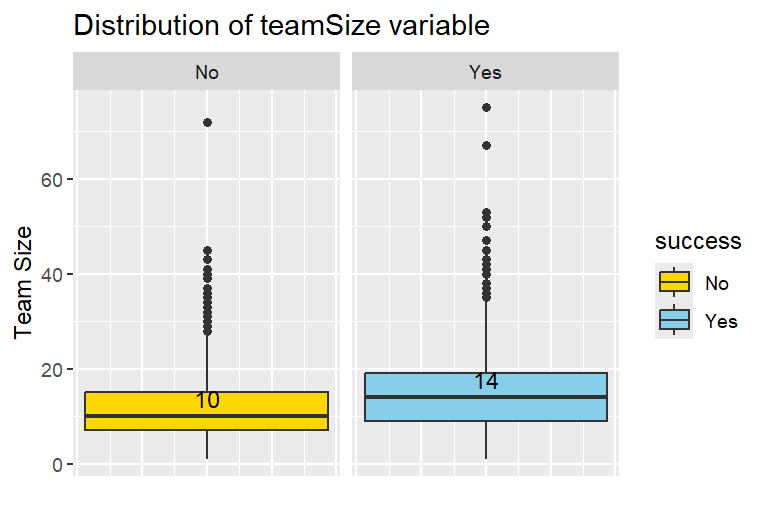


*Figure 13: Top 6 countries for ICO projects*



*Figure 14: Top 5 platforms used by ICO projects*

We also examined the distribution of 'teamSize' variable in Figure 15 to understand its impact on project success. Interestingly, successful projects tended to have larger team size of median 14 compared to unsuccessful ones which has a median team size of 10. This suggests that project team dynamics may play a role in shaping investor perceptions and fundraising outcomes.



*Figure 15: Distribution of teamSize*

Lastly, we conducted text analysis on the 'brandSlogan' variable to gain insights into the messaging and content strategies employed by crowdfunding campaigns. By preprocessing and analysing the text data, we identified recurring themes and keywords such as ‘blockchain’, 'platform', and 'cryptocurrency' which can be seen in Figure 16.



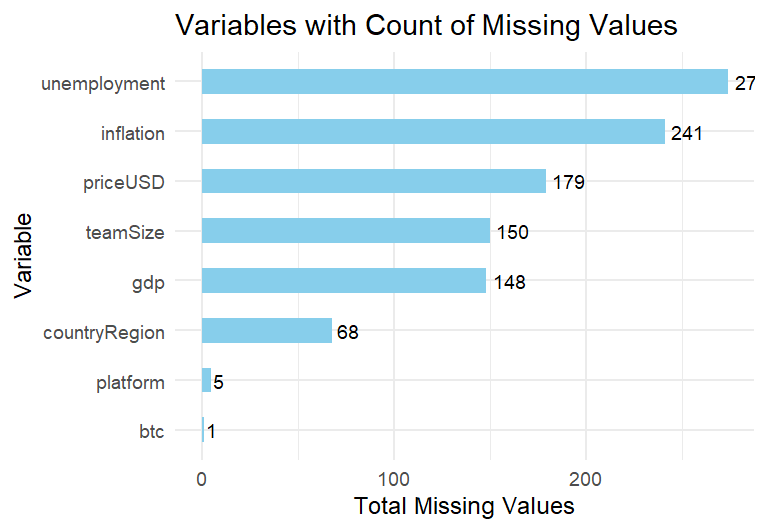
*Figure 16: brandSlogan wordcloud*

# Data Pre-processing

In this phase, we aim to transform the raw dataset into a format suitable for modelling and analysis. This involved several steps, including handling missing values, addressing outliers, transforming data formats, feature engineering, and investigating relationships among predictors.

## Handling Missing Values

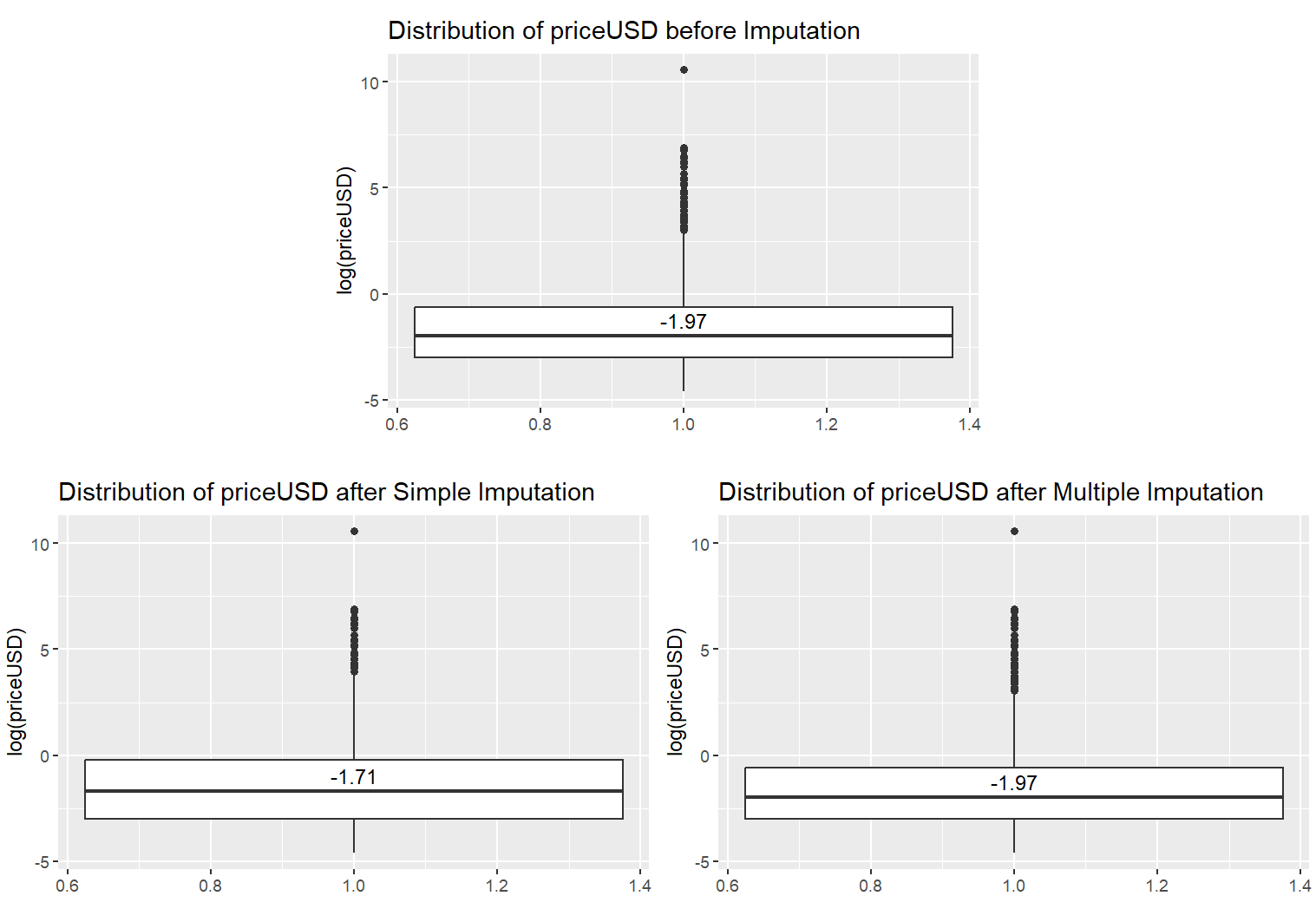
We begin by identifying variables with missing values and visualizing the distribution of missing observations across different variables using a horizontal bar plot as shown in Figure 17.



*Figure 17: Variables with missing values*

For categorical variables such as 'countryRegion' and 'platform,' we replaced missing values with 'unknown'. However, for numerical variables like 'priceUSD' and 'teamSize,' simple imputation using the mean value was performed. Additionally, we recognized that imputing missing values for 'gdp' and 'btc' based solely on mean values would not be logical, as these values depend on country and year. Therefore, we applied multiple imputation, which employs a series of regression models where each missing data is modelled conditional upon other variables in the dataset.

Upon obtaining imputed datasets, we compared the distribution of variables such as 'priceUSD' and 'teamSize' between different imputation methods through boxplots shown in Figure 18. The results indicated that multiple imputation produced data distributions similar to the original dataset, validating our choice of imputation method.



*Figure 18: Distribution of priceUSD before and after simple & multiple imputation*

## Dealing with Outliers

We now proceed to identify any outliers present in the dataset. To remove outliers, we employed the Winsorization method for ‘priceUSD’, ‘coinNum’, ‘inflation’, ‘unemployment’, and ‘ico\_duration’, which involves capping extreme values at the 1st and 99th percentiles of each variable's distribution as shown for the ‘ico\_duration’ variable in Figure 19.

A screenshot of a computer program

Description automatically generated

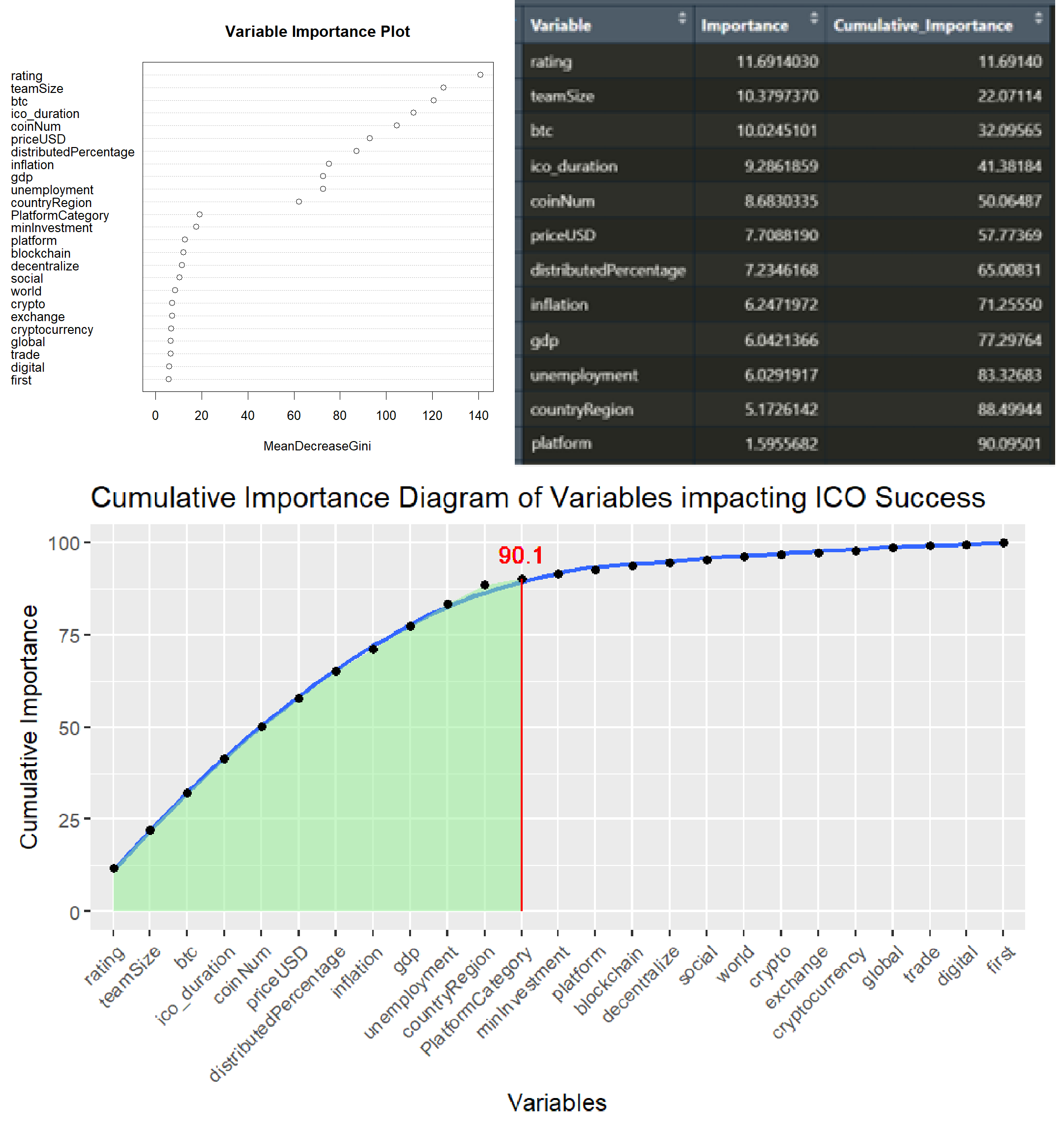
*Figure 19: Removing outliers of ico\_duration*

## Data Encoding & Feature Selection

Once the data is processed, we focused on data encoding and feature selection to prepare the dataset for machine learning modelling. This involved dropping irrelevant columns such as 'startDate' and 'endDate' and factoring categorical variables like 'success,' 'countryRegion,' and 'platform' to convert them into a format suitable for modelling. We performed Z-scaling on numerical variables to ensure consistency in their ranges. It was also crosschecked that there are no attributes which are highly correlated as that might result in multicollinearity. For categorical variables like 'countryRegion,' and 'platform', we implemented frequency encoding to match each category with the frequency of its occurrence in the dataset.

In addition, we encoded text data from the 'brandSlogan' column using preprocessing techniques such as transforming text to lowercase, removing punctuation, stopwords, URLs, replacing accented characters with ascii encoding, and creating a document term matrix with term frequency weighting to capture the importance of terms in brandSlogan. The top terms occurring more than 100 times were selected for the final encoded dataset.

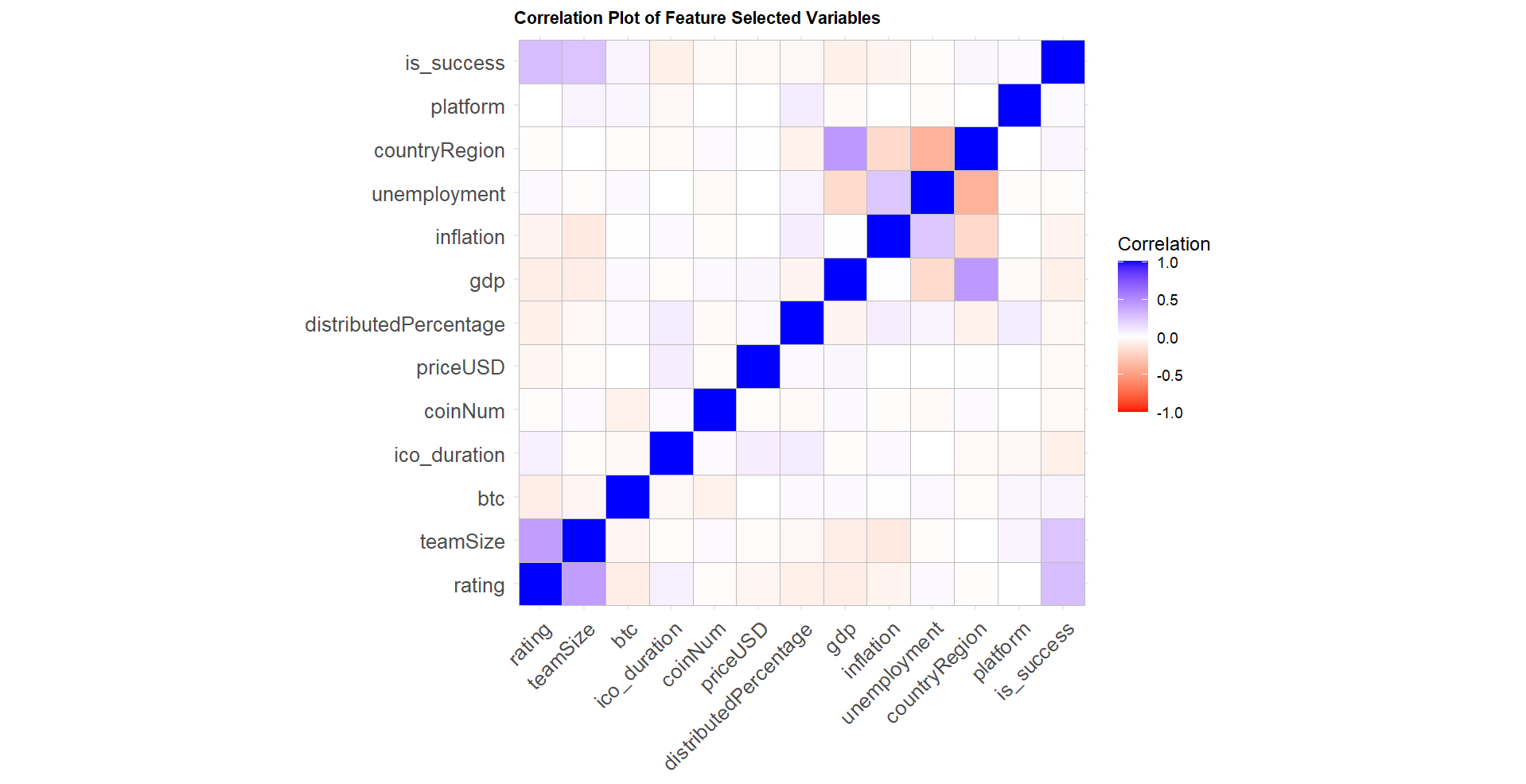
Finally, we applied a Random Forest model to identify the top features contributing to the prediction of project success. This feature selection process involved calculating the importance factor ‘MeanDecreaseGini’ for each variable which indicates how much each variable contributes to decreasing the impurity in the decision tree nodes. We selected the top features that contributed to around 90% of the overall importance as shown in Figure 20. These selected features were then retained for further analysis and modelling in our next steps.



*Figure 20: Feature selection*

# Investigating Relationships Among Predictors:

Lastly, we conducted exploratory analysis to investigate relationships among predictors and identify potential correlations. For instance, we see from Figure 21 the relationship of 'is\_success' with variables like 'teamSize' and 'rating' and noticed that there is a positive correlation which means that a project with higher rating or a larger team size might result in success. Furthermore, we can see that ‘countryRegion’ is positively correlated with ‘gdp’ but negatively correlated with ‘unemployment’ and ‘inflation’. So, we can interpret that a country with higher GDP produces not only more ICO projects, but also has a lower inflation and unemployment index.



*Figure 21: Correlation plot of feature selected variables*

# Modelling

Now that we have the dataset ready to be fit into different models, we split the dataset into 2 sections – 80% of the data was used for training the model whereas the remaining 20% of the data was used for testing.

Based on the problem definition of predicting project success, it's essential to consider the nature of the dataset, the characteristics of the problem, and the requirements of the task. After careful consideration, the four models that seemed most logical for this problem are Logistic Regression, Decision Tree, Random Forest, and Support Vector Machines (SVM).

* Logistic Regression is a standard statistical method for binary classification which represents the likelihood of a binary result hinging on one or several predictor variables. It’s simpler than other models, which makes it easier to understand and implement. Even though simple, the method provides interpretable outcomes and is fit for both linear and nonlinear relationships between the target variable and predictors.
* Decision Trees are resourceful algorithms that recursively divide the feature space into areas depending on the values of predictor variables. Like logistic regression, they provide interpretable outcomes and are fit for both categorical and numerical data. They can also handle nonlinear relationships between the target variable and predictors, that may be present in the dataset. This workability facilitates more accurate forecasts, particulary when linear models like Logistic Regression may not be adequate.
* Random Forest is a wholesome method that incorporates several Decision Trees to optimise predictive performance. By integrating the forecasts of individual trees, Random Forest tends to decrease overfitting and attain greater accuracy than a singular Decision Tree. Random Forest is popular for its strength and multifacetedness, needing minimal hyperparameter tuning to provide competitive performance. Its capability to handle complex interactions and nonlinear relationships in the data makes it a great choice for different regression and classification tasks.
* Support Vector Machines (SVM) is a robust supervised learning algorithm that develops hyperplanes in a high-dimensional space to divide classes. It handles high-dimensional data well, making it appropriate for datasets with a large number of predictors. While forecasting project success, there might be several factors affecting the result, and SVM can effectively capture such complexity. With suitable kernel functions, SVM can handle nonlinear relationships between the target variable and predictors. This workability allows SVM to model complex decision boundaries, boosting its forecasting ability.

We applied these machine learning algorithms and employed appropriate hyperparameter tuning techniques using cross-validation to optimize performance of each model. For each of the models, we first fit the model on the training dataset, then use the model to make predictions on the test dataset followed by extracting measures like accuracy, precision, sensitivity, and specificity which helped assess the effectiveness of each algorithm in predicting project success.

# Evaluation

After training and optimizing each model, we evaluated their performance using various metrics such as accuracy, precision, sensitivity, and specificity to assess their effectiveness in predicting project success.

Through systematically fine-tuning the parameters of each model, we were able to improve the accuracy and reliability of our predictions, thereby maximizing the effectiveness of our models in identifying successful crowdfunding campaigns.

Accuracy measures the overall correctness of the model predictions, representing the proportion of correctly classified outcomes out of all outcomes. Precision measures the proportion of true positive predictions out of all positive predictions, indicating the model's ability to avoid false positives. Sensitivity (also known as recall) measures the proportion of true positive predictions out of all actual positive instances, highlighting the model's ability to detect positive cases. Specificity measures the proportion of true negative predictions out of all actual negative instances, indicating the model's ability to avoid false negatives. We also noticed that 62.7% of the target variable is ‘No’ whereas only 37.3% are ‘Yes’. This might lead to class imbalance, and hence we calculated the F1-Score for each of the models which can be seen in Table 1 along with all measures.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODEL** | **Accuracy** | **Precision** | **Sensitivity** | **Specificity** | **F1-Score** |
| Logistic Regression | **67.823%** | **68.354%** | **86.859%** | 39.614% | **0.7650** |
| Decision Tree | 66.281% | 66.998% | 86.538% | 35.749% | 0.7552 |
| Random Forest | 65.511% | 38.647% | 60.606% | 67.183% | 0.4720 |
| SVM | **67.823%** | 35.749% | 68.519% | **67.640%** | 0.4698 |

*Table 1: Measures of models*

Among these models, Logistic Regression and SVM have the highest accuracy of 67.823%. However, Logistic Regression outperforms SVM in terms of precision, sensitivity, specificity, and F1-Score. It achieves the highest precision (68.354%) and sensitivity (86.859%), indicating its ability to correctly identify positive instances (successful projects) and minimize false negatives. While Decision Tree and Random Forest models show competitive accuracy rates, they demonstrate lower precision and sensitivity compared to Logistic Regression. Moreover, Random Forest notably lags in precision and sensitivity. SVM performs the worst among the models, with the lowest precision and F1-Score.

Therefore, based on these metrics and the goal of accurately predicting project success, Logistic Regression emerges as the best-performing model. It strikes a balance between accuracy, precision, sensitivity, and specificity, making it a suitable choice for this classification task.

# Conclusion

Our analysis uncovered multiple crucial predictors that significantly affect the success of crowdfunding campaigns. Factors like team size, bitcoin prices, campaign duration, the number of coins issued, and the price of blockchain coins emerged as key determiners of fundraising success. Furthermore, variables connected to project quality perception, including overall rating and percentage of coins distributed to investors played a notable role in garnering investor interest and attaining fundraising goals. Additionally, macroeconomic factors like GDP, inflation, and unemployment also showed a significant impact on project success.

The insights obtained from our analysis have realistic implications for businesses and fundraising teams looking to launch successful crowdfunding campaigns. Companies can adapt their campaign strategies to increase the likelihood of achieving their funding goals by understanding the main factors that drive fundraising success.

By understanding the key determinants identified, investors can make data-driven decisions when evaluating investment opportunities in ICOs. The machine learning model will offer investors a framework for assessing the potential success of a crowdfunding campaign, allowing them to allocate their investment capital more strategically and mitigate risks.

By incorporating the identified predictors of fundraising success into their evaluation criteria, listing websites can also enhance their platforms by providing more informed recommendations to investors, thereby increasing investor confidence and engagement on their platforms.

Overall, these insights underscore the significance of utilising data-driven approaches and machine learning techniques to improve decision-making and optimise campaign results.