**Data Understanding and Preprocessing**

Author’s Name

Institutional Affiliation

Course Name and Number

Professor’s Name

Assignment Due Date

**Introduction**

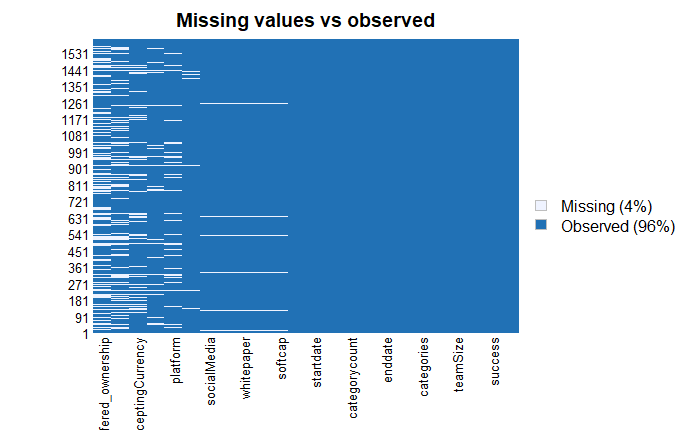
The growth of cryptocurrency in the current financial markets is beyond expectations. It has taken the market by storm and changed the waves of investing and future financial structures. The industry in the last decade has grown to be a novelty product. It has created millionaires overnight and has done tremendous change to the industry. Today, over 4000 different types of cryptos as they are known have been developed. The only common thing about all these is that they work on the same technology of blockchain. However different methods and algorithms are being invented for it. Since blockchain has to create hash files across millions of devices across the planet it is very costly to the planet especially in environmental terms. However this has not slowed down this emerging economy. Different organizations are slowly embracing the idea of purchase with cryptocurrency and are slowly loving the idea. However the volatility cryptos have also poses a danger to its own usage. Simple matters make very big differences to it as no one is sure of its actual market capitalization. But as the future progresses it is getting clear that cryptocurrency is the future currency. It puts society at a level 4 stage of civilization, where physical handling of currency is becoming a thing of the past. Recently a country El Salvador embraced crypto into its government operations and welcomed investors into doing the same.

Similar to IPOs which are basically the act of private corporations going public, ICOs have emerged. It is a new method of raising equity through crowdfunding. Instead of regular currency it uses cryptocurrency. These are basically projects needing funding to grow. They also have different set targets, with minimum and maximum options. The ICOs have durations of raising the funding and can sell their tokens for different prices depending on their assumption of public image. This analysis is a summary of finding the impact of different factors on the success of an ICO. The analysis aims to predict the success of an ICO based on different variables. It uses both logistic regression and deep learning to study these features and how they relate to each other. This report aims not only to summarize this but also too show the prowess of data understanding, presentation and analysis skills. It is divided into different sections each headlining a different concept in the lifecycle of analysis. It culminates with well presented results in tabular and image formats. High frequency of comments is added to the code to ensure that its message is delivered to even persons with a low understanding of statistical inferences. \

**Data Preprocessing**

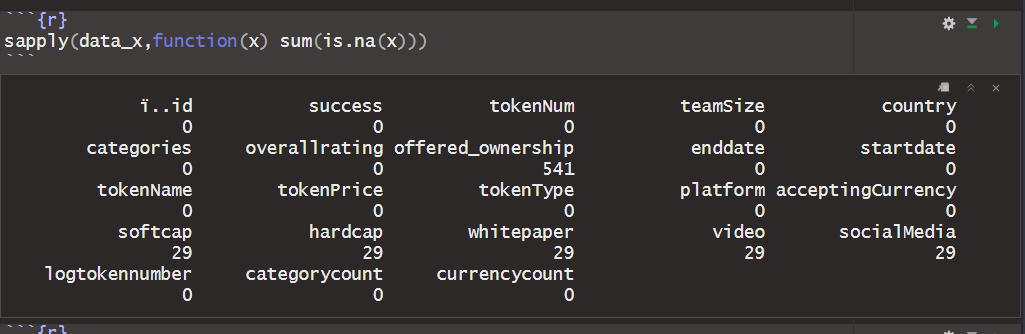
After data collection, the next step is preprocessing. Data preprocessing as highlighted earlier is the gateway to building models and fitting data to models. It is a set of procedures that create, modify, delete, mutate and even change existing datapoints to create the necessary variables under analysis. Similar to a lot of real-life data, ICO data had a lot of mishaps in it. The whole purpose of this particular phase in the lifecycle is to prepare the data for model building. Key structures are enacted that will clean, format, modify, update and nutate the data points in targeted variables into a ready-made dataset. To achieve the most optimal results, our preprocessing was done in two phases. Phase one concerned with data understanding, preprocessing was done in a variable basis manner rather than across the entire dataset. This was to ensure only the most necessary steps were enacted for a certain variable and the maximum information can be retained and learnt from it. Phase two on the other hand, based on the model, required that the data be uniformly preprocessed. This is to ensure that data fed to the model is symmetric in terms of both data shape and length.

The first step of both phases of preprocessing was visualization. It helps give a rough sketch of the data. Questions such as out of the total, how many are unique under each column, pop up. This is done using the sapply(), head() and show() functions in the R environment. The missing variables are then noted and visualization of them is done using the Amelia library. A map is sketched of the total number of missing variables and at what particular stages are they missing. These are the empty slots under each row. The figure below shows the total number under each particular variable. It shows out of the 1606 possible columns, where are the missing located.



**Figure 1.1 Missing Values**

Dealing with the missing values comes after. It shall also be done in two phases. For numerical variables, specifically those with possible variations and distributions, the mean of the entire column will be used to fill the missing values. This is done for variables that are not unique and fall under a certain formation or distribution. The rest of the variables including logistic and binary, factor variables are eliminated by deleting the entire row in the dataset. A smart way of doing this has to be used to ensure th least number of rows are eliminated. As such it will be done in a descending order, whereby the variable with most missing values is used first and this proceeds in subseq1uence. Visualization is done after each elimination procedure to get a hunch of the next elimination procedure. Variables with an extensive number of missing variables will not be used. In this particular analysis, the tokenType is eliminated completely and is not used.



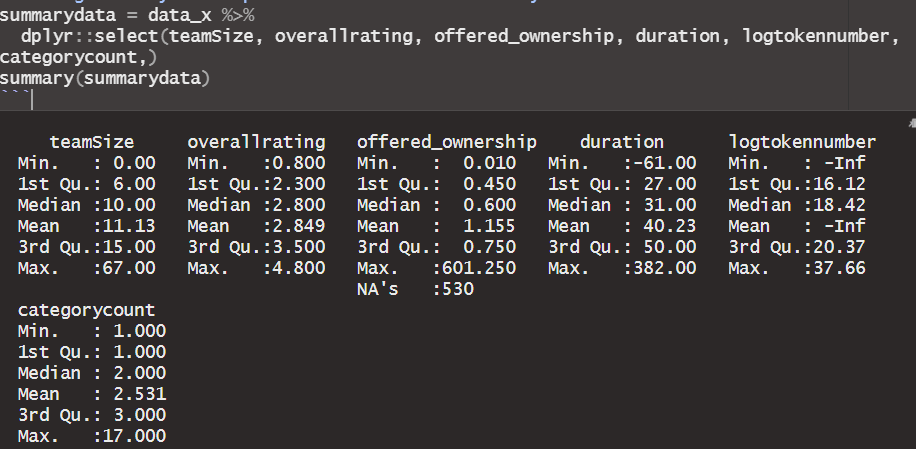
For much more robust analysis, a set number of variables might be necessary for testing their effect on the dependent variable. In this particular project, the variables created are done using the mutate() function in the dplyr package. It creates new columns that will assist in processing. The columns created include duration, which is derived from the subtraction of the end date to the start date, The result is a variable in time format and is converted to a numeric variable. The second variable created is the natural logarithm of the tokenNumber variable. This is due to the huge leaps of the differences between individual ICOs. However, a natural logarithm will ensure the variables are in a somewhat similar trajectory. This is also geared towards understanding the distribution of the variable which would be quite tedious finding using exponential values. The last two new variables created are the categorycount and currency count variables. For variables in list formats such as categories which is a list of the entire goals of the ICO, understanding the number of projects they delve into might assist understanding if it is a motivation to its success. The sapply() function is used together with the length variable to measure the number of elements in that particular list and it is noted down. This is done also for the total number of accepting currencies. The new variables are automatically added to the dataset and a new dataset is copied from the original for further preprocessing.

Some variables require scaling. Scaling is the procedure of fitting variables between a certain range of variables. For our analysis, a minmaxscaler() function will be used on the tokenNumber and logtokennumber variables. The whole reason scaling is done for two main purposes. Scaling is done to ensure the distribution remains the same. The second goal is to ensure the mismatch between different variables is avoided in calculations. Smoothening out the biggest and tiniest numbers will ensure the model does not run into infinites and its log time is short and manageable. It reduces the amount of computing power and time needed to achieve a certain result.

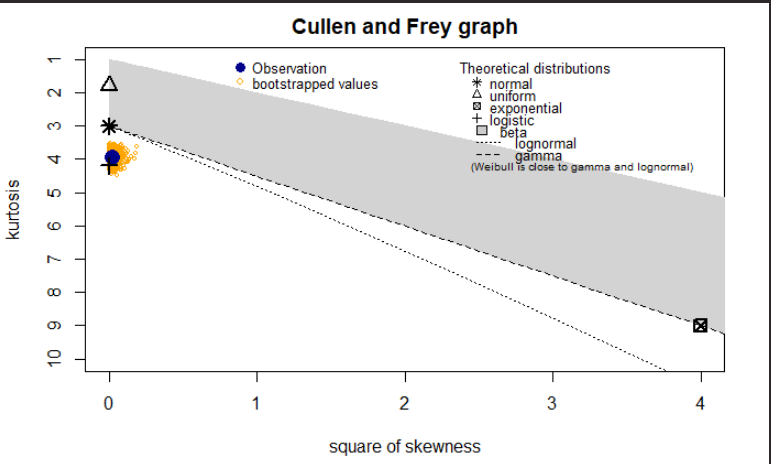
After cleaning, mutating and scaling the next process, is dividing up the dataset into a training set and a testing set, It has to be very unintentional and random to ensure any possible patterns are destroyed to increase the understanding of the model on the data. This will create better generalization patterns that can achieve greater accuracy not only in the training set but also the validation and testing sets. The ratio can be 80% for training and 20% is allocated to the testing dataset. The randomization function in the R package can be used to mix up the rows.

**Data Understanding**

Following the processing done in the previous section, a critical mind map of the data. This mind map however is just a rough sketch of what the data might imply. However, a lot is needed on the statistical aspect of the understanding. This includes distributions of the data, counts of the unique variables, count of each time a unique variable is repeated and graphical representations of the same. For numerical variables, the first step is done using the summary() variable. This creates a brief understanding of the descriptive statistics of selected variables. The image below shows this in real-time. The STAT 5 variables are shown and thus an understanding can be made on possible distribution effect.

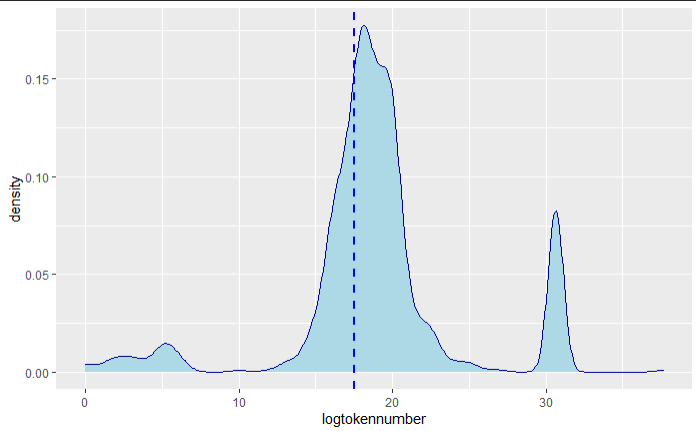


The mean, quartiles and maximum variables show the concentration of the data under question. Better visualization could have been done using boxplots but since their probability distributions shall be plotted using density plots there is no need for doubling the work. Other metrics such as skewness, kurtosis, standard deviation and range will be explicitly shown using the same. The first variable put under question is teamSize. A descriptive function showcasing the possible category of the variable distribution is created using fitdistrplus package and returns a square of skewness to kurtosis plot that shows the possible allocation of the variable. A demo is show in the next diagram. Upon



**Figure 1.1 Descriptive distribution of the logtokennumber variable.**

This shows that the variable is mostly oriented towards the logistic distribution and the lognormal. A density plot of the distribution shows similar results oriented towards the fit that was prescribed using the Cullen and Frey graph. Proper analysis and skimming through the code will show how the results were arrived upon.

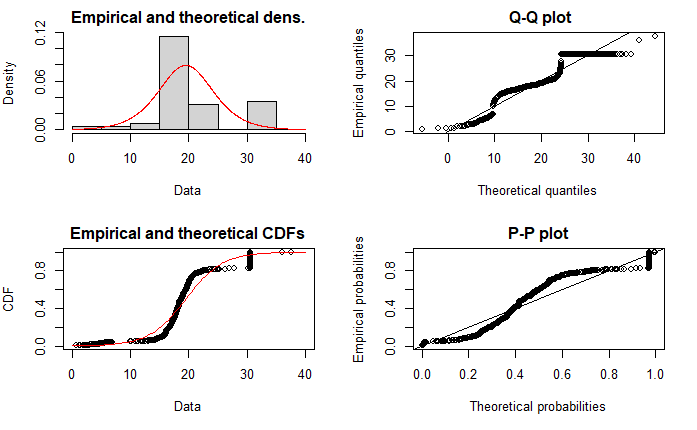


**Figure 1.1 Density plot of the Logtokennumber variable.**

The metrics of the distribution variable are then evaluated using the fitdistribution() function provided by the library.



**Figure 1.1 Estimated variables of the logtokennumber variable.**



**Figure 1.1 Plots showing the fitted distribution in the logtokennumber variable.**

Understanding the distributions perfectly, will allow understanding how the metric makes an impact on the dependent variable. This process is done for the remaining predictor variables using the pre-created functions saved into the R global environment. Images of the density plots are attached in the appendix and most resonate with the predicted distributions. The teamSize variable has a uniform distribution when plotted it resonates. The duration variable has a lognormal distribution and after removing the negative and infinity variables this is clearly shown. The final two variables have normal distributions these being the overallrating variables. Understanding these concepts will go a long way into creating amore creative, robust and concrete structure of the data in the mind.

The final procedure in the data understanding phase is the correlation matrix. Based on the variables sampled. How do the affect and relate to each other? This is an OLS perspective to rethink analysis. It will show either positive or negative relations between the numerical variables and an answer to most questions. It will also act as a precursor analysis to understanding what variables should be fitted into the logistic regression model. The following table shows a snippet of the correlation matrix constructed for the numeric variables to be analysed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| teamSize | 1.00 | 0.55 | 0.02 | -0.04 | 0.16 | 0.07 |
| overallrating | 0.55 | 1.00 | -0.01 | -0.08 | 0.13 | 0.15 |
| offered\_ownership | 0.02 | -0.01 | 1.00 | -0.01 | 0.01 | -0.01 |
| duration | -0.04 | -0.08 | -0.01 | 1.00 | -0.03 | 0.00 |
| logtokennumber | 0.16 | 0.13 | 0.01 | -0.03 | 1.00 | -0.02 |
| categorycount | 0.07 | 0.15 | -0.01 | 0.00 | -0.02 | 1.00 |

The p-values table is also added to show the statistical significance of some of the correlation coefficient of some variables. The resulting with p-values higher than the alpha significant value of a 95% confidence interval are then hypothesized as null. This precursor analysis will resuce the time and effort needed in more preprocessing and computation. .

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| teamSize |  | 0.0000 | 0.5885 | 0.1888 | 0.0000 | 0.0217 |
| overallrating | 0.0000 |  | 0.8765 | 0.0113 | 0.0000 | 0.0000 |
| offered\_ownership | 0.5885 | 0.8765 |  | 0.7545 | 0.8487 | 0.6746 |
| duration | 0.1888 | 0.0113 | 0.7545 |  | 0.2932 | 0.8885 |
| logtokennumber | 0.0000 | 0.0000 | 0.8487 | 0.2932 |  | 0.5050 |
| categorycount | 0.0217 | 0.0000 | 0.6746 | 0.8885 | 0.5050 |  |

**Data Analysis**

Data analysis requires dynamic thinking across all angles. The set of processes needed to understand why a certain phenomenon works as it is requires laying down a roadmap to reach conclusive evidence regarding the same. This section is the data analysis phase of the lifecycle. The model can be seen to have a lot of binary and factor variables and while some might seem strongly related to the idea of ICOs some are not. The analysis however has to predict the variables affecting the success of the model. Thus, the model chosen for this particular venture is the logistic regression model. This model is chosen because of logits background and thus works variably well with binary variables. The model is implemented via the glm function that comes built in R. It is in the family of binomial models. The variables that will not be used in the model are then omitted and visualizing of the remaining dataset is done. The shape parameters of the model must be similar across the entire dataset. The model is then run and it returns a table of results. The table is made up of four columns, which is estimate being the value of the variable coefficient with respect to the dependent variable, the standard error the variable is coupled with. The z statistic which basically is a value of reference to the normal distribution. The p-value is an estimate to the alpha significant level which will basically show the statistical significance of the model.   
## Call:  
## glm(formula = success ~ ., family = binomial(link = "logit"),   
## data = modeldata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.52380 -0.78653 0.01956 0.71573 2.23556   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.656e+00 5.319e-01 -4.993 5.94e-07 \*\*\*  
## id -5.969e-05 1.492e-04 -0.400 0.6891   
## tokenNum 4.386e-13 2.217e-13 1.978 0.0479 \*   
## teamSize 2.074e-02 1.140e-02 1.819 0.0690 .   
## overallrating 1.184e+00 1.336e-01 8.868 < 2e-16 \*\*\*  
## softcap -1.905e-01 1.573e-01 -1.211 0.2260   
## hardcap 9.920e-02 1.842e-01 0.538 0.5903   
## whitepaper -9.378e-01 3.834e-01 -2.446 0.0144 \*   
## video 3.346e-01 1.665e-01 2.009 0.0445 \*   
## socialMedia 4.474e-01 7.515e-02 5.954 2.62e-09 \*\*\*  
## logtokennumber -8.305e-03 1.826e-02 -0.455 0.6492   
## categorycount -7.271e-02 3.910e-02 -1.860 0.0629 .   
## duration -1.851e-03 2.462e-03 -0.752 0.4521   
## currencycount 4.201e-02 4.893e-02 0.859 0.3905   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1848.7 on 1393 degrees of freedom  
## Residual deviance: 1224.2 on 1380 degrees of freedom  
## AIC: 1252.2  
##   
## Number of Fisher Scoring iterations: 14

ANOVA on the model is done to find the difference between the residuals and the required answers are determined. The model is highly commented and thus becomes very intuitive and ready for reading and perusal. All answers are generated in the model and showcased within tables and graphs. The ANOVA is the epitome of the model. It shows the significance of each variable to confirm the hypothesis posed in the last section. Minimal visualization shall be done in this section in terms of graphical results, however a lot hoes in to the tabling to be done. The last section is measuring the variables showing the goodness of the fit. This is done using the” PSCL” library built for the Political Science Class Library. It will measure the r squared value of the model and the adjusted one. This is basically the sum of residuals measured against the sum of results. Other methods such as the F-statistic can also be used to show the viability of the model. The table below showcases these results.

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: success  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 1393 1848.7   
## id 1 0.046 1392 1848.7 0.829324   
## tokenNum 1 265.343 1391 1583.3 < 2.2e-16 \*\*\*  
## teamSize 1 142.216 1390 1441.1 < 2.2e-16 \*\*\*  
## overallrating 1 156.849 1389 1284.3 < 2.2e-16 \*\*\*  
## softcap 1 3.747 1388 1280.5 0.052904 .   
## hardcap 1 0.092 1387 1280.4 0.761288   
## whitepaper 1 6.782 1386 1273.7 0.009207 \*\*   
## video 1 3.493 1385 1270.2 0.061643 .   
## socialMedia 1 41.111 1384 1229.0 1.438e-10 \*\*\*  
## logtokennumber 1 0.127 1383 1228.9 0.721383   
## categorycount 1 3.589 1382 1225.3 0.058154 .   
## duration 1 0.414 1381 1224.9 0.520107   
## currencycount 1 0.737 1380 1224.2 0.390517   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## fitting null model for pseudo-r2

## llh llhNull G2 McFadden r2ML r2CU   
## -612.0890773 -924.3625504 624.5469462 0.3378258 0.3611113 0.4916315

**Future Prospects**

Future prospects of this model are multiple. However, the first and most recommended one is the use of deep learning. Deep learning will increase the accuracy, goodness of fit and the robustness of this model. Multiple regression using an artificial neural network is not only multidimensional it can also incorporate use of tensors and matrices. Models such as Bert can be used to further spike up the model through text analysis. Understanding the individual categories and accepting currencies in terms of semantic embeddings can help generate further knowledge into the fit. Extra data from outside sources such as the economic standards of these countries will also assist further boost the model genius. In summary, model building requires dynamic approaches that are not cutthroat by current technology. However, the maps created by neurons will ensure maximization of generalizations and increase of predictive accuracies. Machine learning is slowly loosing its war to deep learning and going through this route will definitely ensure better, stable and more versatile results.

**References**

Mihas, P. (2019). Qualitative data analysis. In *Oxford research encyclopedia of education*.  
<https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-1195>