Success Indicators of Initial Coin Offerings

by

Meghana Kolasani

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Computer Science

(Honors Associate)

Presented June 5, 2019

Commencement June 2019

AN ABSTRACT OF THE THESIS OF

Meghana Kolasani for the degree of Honors Baccalaureate of Science in Computer Science presented on June 5, 2019. Title: Success Indicators of Initial Coin Offerings

Abstract approved:____

Bin Zhu

Abstract:

In a big data day and age, there is an abundance of information and we now have the tools to understand it with data mining and machine learning algorithms. With the rise and fall and rise again of Bitcoin, the finance industry and society itself is caught in between. There are many factors that an investor considers when choosing to invest in a particular coin and understanding those factors can allow the investor to make a more informed decision. Data on how an initial coin offering is presented on the rating platform, ICOBench, was collected and analyzed using the random forest classifier. This resulted in a prediction model regarding the success of a coin in terms of average rating and number of ratings to determine what factors hold the most influence.

Key Words: data mining, random forest, bitcoin, cryptocurrency

Corresponding e-mail address: kolasanm@oregonstate.edu

Success Indicators of Initial Coin Offerings

by

Meghana Kolasani

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Computer Science

(Honors Associate)

Presented June 5, 2019

Commencement June 2019

Honors Baccalaureate of Science in Computer Science project of Meghana Kolasani presented on June 5, 2017.

APPROVED:

Bin Zhu, Mentor, College of Business

Shaokun Fan, Committee Member, College of Business

Byron Marshall, Committee Member, College of Business

Toni Doolen, Dean, Oregon State University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, Honors College. My signature below authorizes release of my project to any reader upon request.

Meghana Kolasani, Author

Table of Contents

Chapter 1: Introduction

1.1 Summary

Initial coin offerings (ICOs) are a way to fundraise projects by exchanging crypto tokens for cryptocurrencies like Bitcoin or Ethereum. ICOs are presently the most prominent trend in cryptocurrency and have the potential to become the securities and shares of tomorrow, revolutionizing the financial system. Blockchain projects are budding with opportunity and have evolved significantly over the past few years, experiencing significant highs and lows.

We believe that specific factors about how an ICO is presented to an audience can be analyzed from an aggregate data set to create a prediction model that can generate the success likelihood of a particular coin and provide a data driven approach for selecting ICO investments.

Background research was conducted to select variables to be tracked for a given ICO. These variables were used to extract data from the ICObench website, a well-known ICO rating platform. The data was then analyzed using a random forest classifier that resulted in a prediction model that can be used to change how ICOs are marketed and/or developed. Understanding the elements that influence investors to invest in some coins over others will help anticipate the success of ICOs.

1.2 Overview of Cryptocurrency

Out of the 180 currencies that exist in the world that are used across 195 countries, cryptocurrencies have increased significantly due to growth in technology with over 4000 digital currencies available in the market (World Atlas, 2018). One of the most valued and diverse alternative stores of value is bitcoin, and its popularity and the future of digital currency continue to grow as time goes on.

The concept itself was based on allowing a seamless system to exchange money, which can be rather cumbersome with the current financial system. The exchange during financial transactions is time consuming, even with newer forms of payment systems like PayPal or Venmo. However, this digital currency results in instantaneous value exchange, making it a strong asset. Figure 1 presents current currency, block, mining and network data about the bitcoin blockchain.

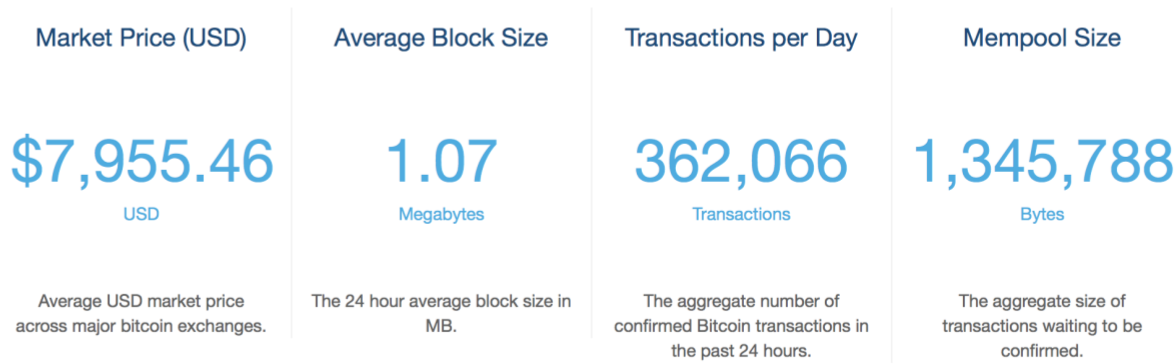

Figure 1: Popular Blockchain Statistics (Blockchain, 2019)

1.3 Bitcoin Transaction Process

The bitcoin transaction process is relatively straightforward, as shown in Figure 2. To start, a user begins with a wallet-like software and initiates a bitcoin payment. All initiated and pending transactions are broadcasted globally on the bitcoin network using blockchain. Blockchain is a breakthrough technology that is essentially a shared public ledger serving as the backbone for the entire bitcoin network without needing a bank or any other intermediaries. This eliminates mistakes and delays from different ledgers trying to communicate with each other. Transactions are handled with private keys that sign a transaction to provide proof of the transaction while also denying any alterations of the transaction. Miners confirm pending transactions by packing a large number of them in a block that is processed and verified by the network using strict cryptographic rules that enforce a chronological order in the blockchain (Bitcoin). The updated blockchain with the new block is disseminated to the entire network and new transactions will be recorded in the next block. The user on the other end of the transaction can check the arrival of the bitcoin using the same wallet-like software (The Economist, 2015).

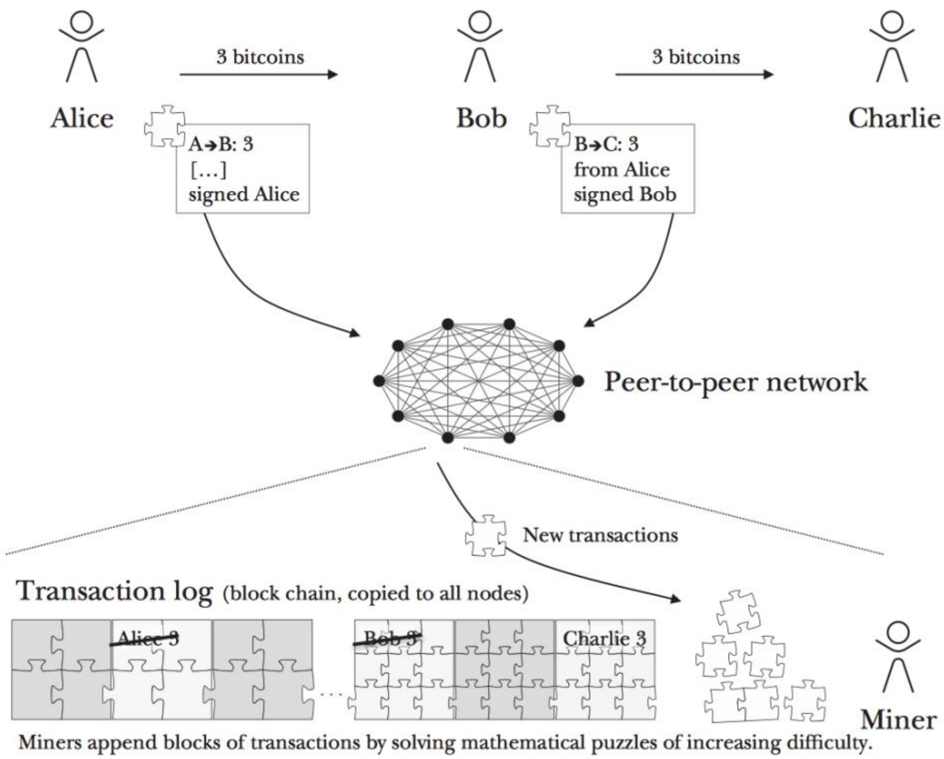

Figure 2: Bitcoin's Approach to Transaction Flow and Validation (Bohme, Christin and Moore, 2015)

1.4 Bitcoin in Industries

The traditional centralized financial services model that industries currently use is being disrupted by this digital currency. Established financial organizations like JPMorgan Chase, Citigroup and Credit Suisse are realizing that adapting to this powerful technology is essential to stay viable and reduce economic fraud and crime. Even companies that dominate the IPO business such as Intercontinental Exchange, NASDAQ Inc., and Goldman Sachs are becoming some of the largest investors in blockchain ventures (Harvard Business Review, 2017). Bitcoin still hasn't been integrated into many smaller emerging markets due to setbacks since its initial introduction in 2009. Financial markets tend to vary and market crashes have dire consequences for both market participants and society as a whole. This is why many smaller companies are hesitant about the volatility of cryptocurrency values, since the fluctuation results in losses that are hard to absorb. Figure 3 is a graph that depicts the total USD value of trading volume on major bitcoin exchanges of all time. There was clearly a peak in 2018 that has since settled down, but the market for bitcoin is still relatively active and liquid compared to previous years.

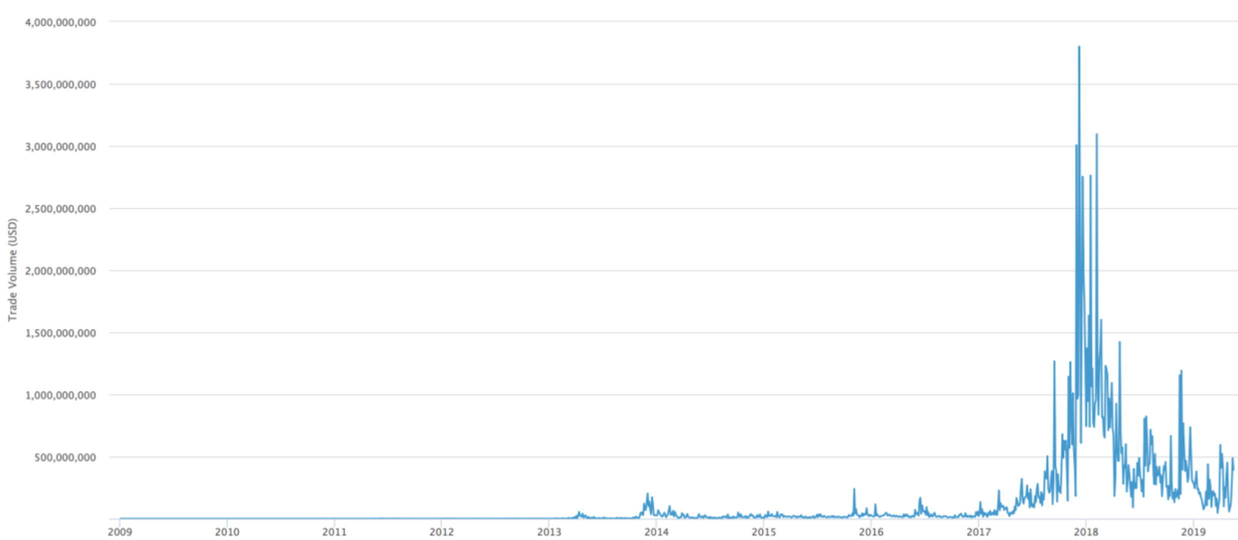

Figure 3: USD Exchange Trade Volume (Blockchain, 2019)

At the same time, there are industries that have embraced cryptocurrency and have had success in allowing this method of payment, and Figure 4 shows a recent climb in average USD market price across major bitcoin exchanges in the past year. Shopping companies with online stores such as Target, Nike, Amazon and Walmart, travel companies like Expedia, charity organizations like the Human Rights Foundation, Save the Children and the Water Project, certain hotels, restaurants and online casinos are embracing this payment method (Nasdaq, 2018).

Figure 4: Average USD Market Price Across Major Bitcoin Exchanges (Blockchain, 2019)

Chapter 2: Data

2.1 Initial Coin Offerings

Entrepreneurial finance is intrigued by the innovative ventures using distributed ledger technology such as blockchain. Initial coin offerings (ICOs) are essentially a smart contract for blockchain technologies, similar to an initial public offering (IPO) in the mainstream investment world with tokens instead of shares (Investopedia, 2018). ICOs use mechanisms to fundraise for business objectives and operations by conducting peer-to-peer crowd sale of digital tokens in exchange for Bitcoin or Ethereum. The token itself holds the same value as native currency, but can also give a user stakes in the project.

By providing just a 'white paper' that contains the details of the project, creators can draw in potential investors without the need for unnecessary paperwork (Fisch, 2019). There is a sense of community, credibility, and exposure with this successful crowdfunding technique that often drives an incentive for developers to design innovative projects. At the same time, these characteristics can attract scammers and are based on speculation. However, the ability to access new tokens early when they have the potential to become significantly valuable gives investors more opportunity for a high return on investment, and it is at their discretion to choose the right ICO to invest in.

2.2 ICObench

ICObench is a free informational platform that provides ICO ratings from experts that can provide legal, analytical and technical insights to potential investors. It is one of the more popular online ICO analysis websites, as it has a team of analysts with diverse backgrounds that have ranked thousands of ICOs and provide consultancy services for all stages of the project as well as detailed write-ups of the ICOs presented in an organized manner (Bitcoin Exchange, 2018). Their website provides stats on the 5521 ICOs listed on the platform that look at the highest number of ICOs and highest funds raised by country, by industry and by platform which can be useful when analyzing the ICO market world.

Each ICO page may include a description of the coin, platform tags, a video, information about the project, timeline, team, financial data, milestones, user ratings, blog and white paper. The white paper summarizes details about the project background, its benefits and originality, technological features, the team, and future plans. There are also more specific ratings regarding profile, team, vision and product. ICObench created a bot, Benchy, to execute an analytical assessment. Users can also communicate with Benchy to receive more information about the ICOs or ICObench itself. Ratings are dynamic and calculated at least once per day using a combination of the experts' ratings, Benchy, the ICO profile analysis, and an algorithm that uses more than 20 different criteria (ICObench).

The ICO analyzer is an objective tool to report the availability of important data to allow teams to improve their ICO profile to appeal to investors. Active members of ICObench can apply to become an expert, giving them the right to rate ICOs as long as they conform to the rules. Team members can also be given an ICO success score (ISS), which is a combined evaluation of every ICO the member is a part of. Having a high number of successful ICOs results in a higher ISS. Although ICObench provides a paid feature for premium listing, this does not affect the ICO's ratings. The ICO will just be prioritized for publication before other pending projects and advertised more by being placed higher up on the browse list in relevant categories. Experts are also not paid and users cannot pay for quantity or quality of reviews.

2.3 Data Description

ICOBench was the platform used to extract data from. A random set of the HTML pages for over 2000 ICO profiles were selected to analyze and several variables were chosen to review regarding how an ICO profile is presented on this platform. The focus was on business characteristics that can influence investors. Data about these characteristics was collected for each ICO profile in the dataset.

- **Description Length**: The number of words in the short description of the project
- **Flesch Reading Ease:** Score that determines complexity of a text
- **SMOG Index:** Grade level that determines readability of a text
- **Number of Tags:** Number of categories the ICO falls under
- **Video:** Whether the ICO has a video or not
- **Average Rating:** Average of the given ratings
- **Number of Ratings:** Number of users who have rated the ICO
- **Number of Tokens:** Number of tokens for sale
- **Team Presence:** Whether the ICO has a team with more than one member or not
- **Number of LinkedIn:** Number of team members who provide their LinkedIn profile
- **Number of ISS:** Number of team members who have an ICO success score
- **Milestones:** Number of milestones the team has achieved up to date

2.4 Data Acquisition

Web scraping was used to acquire this data, by extracting the desired information from data downloaded from the website. The web inspector developer tool was used to determine what elements of the HTML page contain the data relevant to this project. The library

BeautifulSoup was utilized to parse the HTML and store the retrieved data, as its methods are easy to use and it provides a variety of functionality for flexible and fast web scraping.

During the first phase of data extraction, more data items were extracted such as name of token, price, profile rating, team rating, vision rating, and time remaining. However, there was an issue of missing, null or inconsistent values. The price column was dropped, as many profiles would price the ICO in different quantities of tokens or in different currencies. The name column was dropped, as the analysis is dependent on the information about the ICO but not the specific ICO itself. The time column was dropped, as profiles would vary between start and end date and time remaining, and many profiles did not have a value at all. The specific ratings such as profile, team and vision were dropped, as they are based on an aggregate of user ratings and could be easily biased by one skewed rating if the quantity of ratings is low.

The remaining data was cleaned up to replace invalid characters with null values and maintain consistencies between the rows, resulting in a data set with relevant information stored in a csv file and ready to be used for analysis.

Chapter 3: Research Methodology

3.1 Data Mining

Data mining is the process of exploring large data to find meaningful patterns, correlations and anomalies using a broad range of techniques (SAS). The volume of data society is generating is increasing exponentially, but an overwhelming amount of information does not necessarily result in automatic insights. With data mining, repetitive and chaotic noise in data can be filtered out to better understand relevant information that can be used to assess outcomes and make informed decisions (SAS).

Data mining can conduct supervised or unsupervised learning. Unsupervised learning uses models such as association analysis, clustering and principal component analysis to describe and understand data by revealing its underlying patterns. Supervised learning uses models such as linear regressions, time series, logistic regressions and neural networks to classify or predict an observation (MicroStrategy). Supervised learning with classification will be the data mining process used for analysis, since we are trying to classify ICOs into categories of success based on the factors presented on its profile.

3.2 Random Forest Model

Random forest is the supervised learning algorithm used to classify the data in this project. It is flexible and relatively easy to use for both classification and regression tasks. The algorithm works by taking the dataset and selecting some random samples, creating a decision tree for every sample, building a prediction for every tree, voting on each prediction, and selecting final prediction based on the number of votes, as shown in Figure 5.

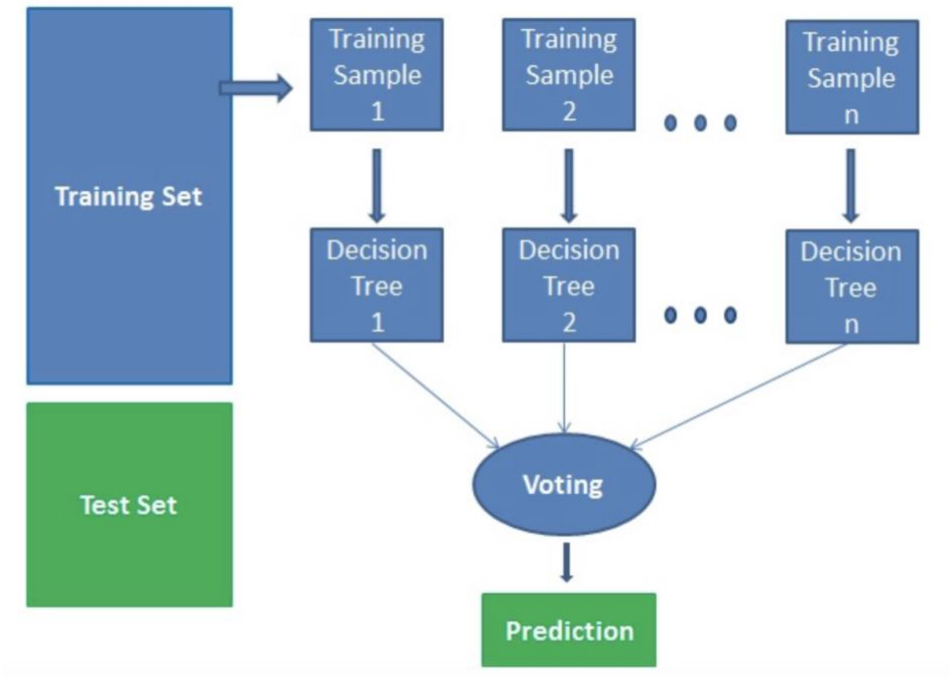

Figure 5: Random Forest Algorithm Process (DataCamp, 2018)

The Random Forest model often outperforms the Decision Tree model and is typically more accurate, since the 'forest' is comprised of a number of unpruned and diverse trees. A single decision tree is a top-down tree structure where the root node carries the target variable and the remaining data is partitioned into branches where a decision is made at every branch until a result or classification is made at the leaf nodes. While decision trees are faster and easier to interpret, they can suffer from overfitting (when a tree essentially memorizes a training set), a problem the random forest prevents by creating trees on random subsets and assembling these trees together (DataCamp, 2018). The random forest algorithm also assists in evaluating the relative importance on individual features influencing the prediction, resulting in a more thorough analysis.

3.3 Software Selection

Several Python libraries were utilized to construct a prediction model, including NumPy, Matplot and Pandas to load, manipulate and summarize the data, and Scikit-learn for the machine learning algorithms that model the data. Scikit-learn is an open source software that is simple, efficient and provides an array of tools for data mining and analysis. The ensemble method RandomForestClassifier was used to classify the data.

Scikit-learn was chosen over TensorFlow, another library used for machine learning operations. TensorFlow has more capabilities for deep learning, but that is beyond the scope of this project. Jupyter Notebook was the online notebook used to write and execute the software. This interface enables a user to combine computational information such as the code and data with multimedia and graphs to establish a narrative when presenting the data.

Chapter 4: Data Analysis

4.1 Data Processing

Additional preprocessing of the data needed to occur before a prediction model could be applied. The sum of null values in each category was calculated to ensure that most columns had sufficient data. The data set is large enough, so any rows with remaining null values were dropped.

The values in the features Average Rating and Number of Ratings were organized into three categories of 'low', 'average' and 'high' for simplicity, since the values were so specific in a broad range. The distribution was first graphed for each of these features as shown in Figure 6 and Figure 7, respectively.

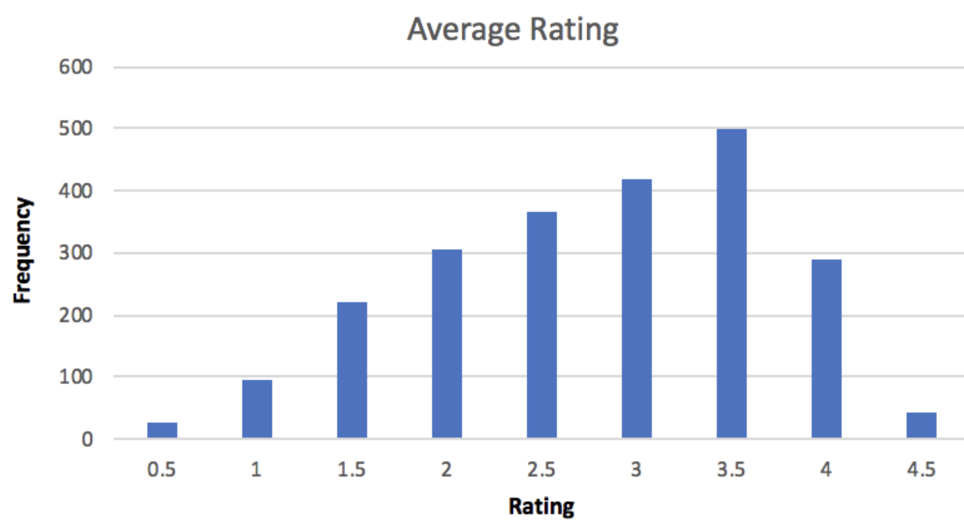

Figure 6: Distribution of Average Ratings

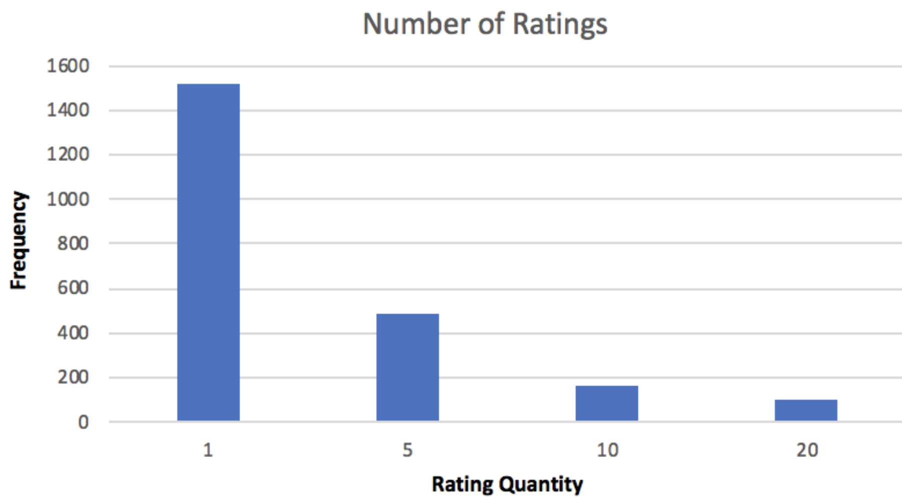

Figure 7: Distribution of Number of Ratings

Based on these distributions, Average Rating was split up where 0-2.5 is a 'low' rating, 2.5-3.5 is an 'average' rating, and 3.5-5 is a 'high' rating. Number of Ratings was split up where 1-4 is a 'low' quantity, 4-10 is an 'average' quantity, and 10+ is a 'high' quantity. The resulting distribution graphs are shown in Figure 8 and Figure 9 where 0 represents average, 1 represents high, and 2 represents low.

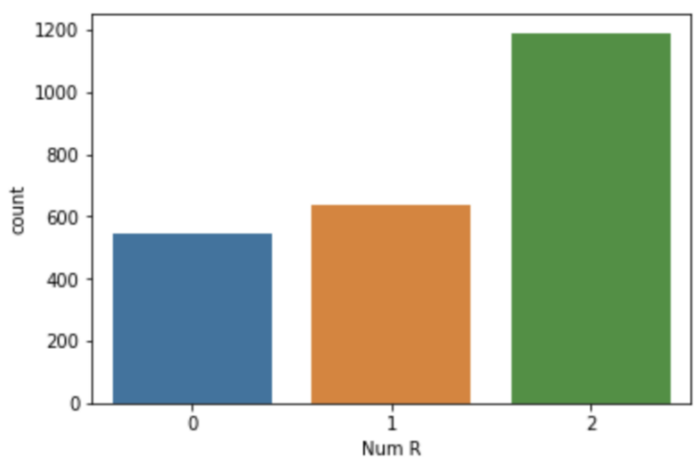

Figure 8: Categorized Distribution of Number of Ratings

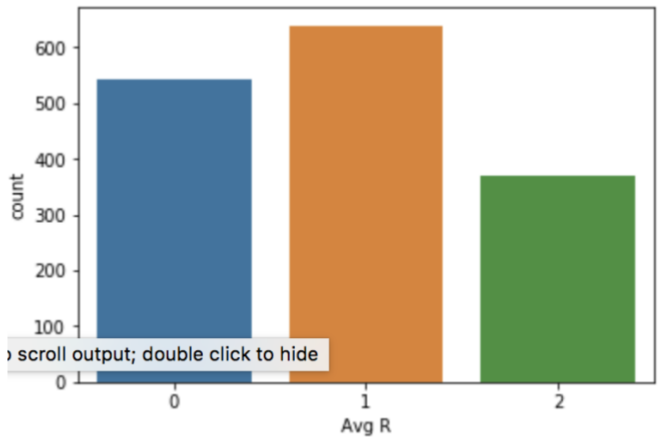

Figure 9: Categorized Distribution of Average Ratings

With the distributions in place, all categorical values were converted to numerical values for the purpose of simplifying the model. The fit transform method from the sklearn library was used on the Average Rating, Number of Ratings, Video, and Team columns.

4.2 Process

The data had to be separated into response variables and feature variables. Average Rating and Number of Ratings were chosen to be the target features, as they can both be indicators of the success of an ICO. ICOs with a higher average rating and higher number of expert ratings tend to be higher on ICObench's ranking list.

Then the data was split into a training set and test set, with a test size of 20% and a random state of 42, used to seed the random number generator to make the model's output replicable. Standard scaling was used to get an optimized result. Scikit-learn has a StandardScaler package that normalizes the features so that the distribution will have a mean of 0 and standard deviation of 1.

Finally, the RandomForestClassifier was applied with n_estimators of 200. This hyperparameter represents the number of trees the forest will build before making a prediction. Having a larger number of trees calls for slower computation, but can stabilize the predictions and increase performance. The model is trained using the training data and a prediction is made.

4.3 Results

Classification reports containing the main classification metrics for the target values Average Rating and Number of Ratings were generated after running the classifier shown in Figure 10 and Figure 11, where 0 is average, 1 is high, and 2 is low. There was a 77% accuracy for predicting the Number of Ratings and an 82% accuracy for predicting the Average Rating. This is based on the micro-average, which is the average of the total true positives, false negatives and false positives.

Looking closer into the classification reports, there are four metrics. Precision is the ratio of true positives to total predicted positives, used to determine the ability of the model to not to label a negative sample as positive. Recall is the ratio of true positives to all predictions, used to determine the ability of the model to find all positive samples. F1 Score is the weighted average of precision and recall, taking both false positives and negatives into consideration. Support is the number of truly predicted samples in the class.

	precision	recall	f1-score	support
0	0.17	0.02	0.04	41
1	0.62	0.45	0.52	40
2	0.80	0.97	0.88	230
micro avg	0.77	0.77	0.77	311
macro avg	0.53	0.48	0.48	311
weighted avg	0.70	0.77	0.72	311

Figure 10: Classification Report for Number of Ratings

	precision	recall	f1-score	support
0	0.83	0.63	0.72	112
1	0.85	0.99	0.92	124
2	0.75	0.81	0.78	75
micro avg	0.82	0.82	0.82	311
macro avg	0.81	0.81	0.81	311
weighted avg	0.82	0.82	0.81	311

Figure 11: Classification Report for Average Ratings

A confusion matrix was also generated for each of the target values after running the classifier, shown in Figure 12 and Figure 13. The confusion matrix describes the performance of the model given the test data. For each class, the number of true and false predictions are summarized.

$$\begin{bmatrix} 71 & 21 & 20 \\ 1 & 123 & 0 \\ 14 & 0 & 61 \end{bmatrix}$$

Figure 12: Confusion Matrix for Average Rating

$$\begin{bmatrix} 1 & 6 & 34 \\ 2 & 18 & 20 \\ 3 & 5 & 222 \end{bmatrix}$$

Figure 13: Confusion Matrix for Number of Ratings

For the Number of Ratings, the model was able to correctly predict 222 out of 230 samples as ‘low’, 18 out of 40 samples as ‘high’, and 1 out of 41 samples as ‘average’. There is high recall for predicting a low number of ratings, with minimal false negatives, but the precision was brought down slightly, due to the false positives. For the Average Rating, the model was able to correctly predict 61 out of 75 samples as ‘low’, 123 out of 124 samples as ‘high’ and 71 out of 112 samples as ‘average’. There is a high recall for predicting a high average rating, with only one false positive bringing down the precision.

Feature importance was also generated to determine the significance of the different estimators on the prediction model. The graph shown in Figure 14 shows that the three most important features (in order), were having a video, number of ICO success scores and number of LinkedIn profiles, followed by description length, reading ease, number of ratings, number of tags and milestones.

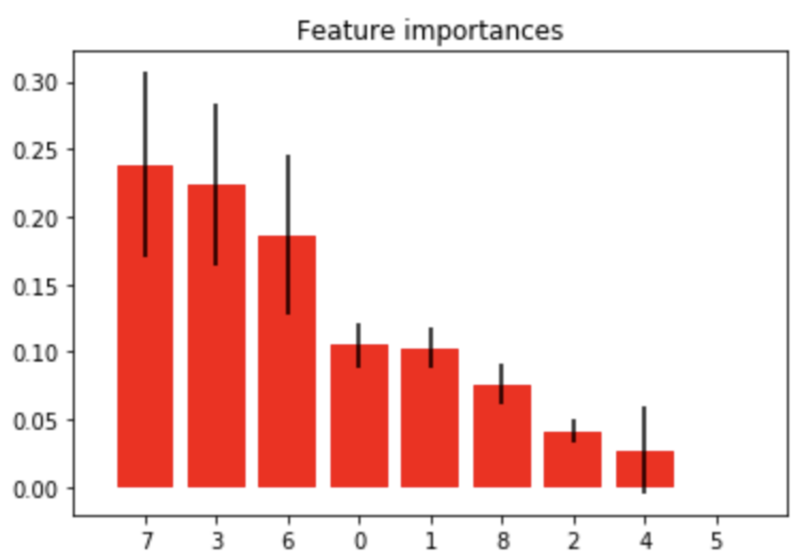

Figure 14: Importance of Individual Features

Chapter 5: Conclusion

5.1 General Discussion

The predictions the Random Forest model came up with were fairly accurate, meaning that there are indeed certain factors that can indicate the success of a particular ICO. Based on the results, having a video, number of ICO success scores, and number of LinkedIn profiles were the primary features to influence an ICO's success in terms of average rating and number of ratings. In particular, these features were very relevant when predicting whether an ICO will have a high average rating or low number of ratings.

The features of high importance are all related to media presence. Having a video on an ICO profile demonstrates that the team put in the resources and effort to really showcase their ICO on the platform as well as shared on YouTube. Teams with multiple members established on LinkedIn illustrates a sense of credibility, as users can look into team members' professional experience. As for ICO success scores (ISS), having a score alone displays that the team member is experienced and has multiple successful ICOs rated on the platform and having a high ISS quantifies a member's success.

Users put in the time to research and look through ICO rating platforms like ICObench in order to discover ICOs and be better informed about a particular ICO before deciding to invest. By recognizing these high importance features, a user can be more aware of what to look for and make more informed decisions.

5.2 Business Contribution

There was a lot of research conducted about Bitcoin when it initially started rising in popularity regarding how and in what ways it disrupted the market. However, after the peak in late 2017, the market for Bitcoin died down significantly. This research aims to inform current potential investors about features of promising coins. The unprecedented cryptocurrency market crash in 2018 has caused investors to be much more speculative when it comes to investing in this digital currency.

In reality, the bitcoin market is maturing, not failing, with a price increase of 150% from its lowest point. The market has made an impressive recovery in the past few months, with an increasing daily trading volume highly due to technical milestones, more widespread adoption, shifting sentiments and gold investment wavering (Visual Capitalist, 2019). The public's perception of Bitcoin is changing, and research needs to keep track of these changes.

5.3 Directions for Future Research

Bitcoin technology is still relatively new and has immense room for growth. As this currency emerges into more industries and drives financial innovation, the number of investors will grow and the number of ICO platforms will grow along with it. Further research can be done by comparing different platforms and exploring how predictive factors may vary or have different levels of importance. Beyond just the marketing features of how an ICO is presented, research can be done to look into evaluating the specifics of the white paper provided and technical documentation to provide a more detailed analysis for a more experienced investor.

Bibliography:

"Bitcoin Charts & Graphs - Blockchain." *Blockchain.com*, 2019, www.blockchain.com/en/charts.

"Data Mining Explained." *MicroStrategy*, www.microstrategy.com/us/resources/introductory-guides/data-mining-explained#database.com.

"7 Major Companies That Accept Cryptocurrency." *NASDAQ.com*, Nasdaq, 31 Jan. 2018, www.nasdaq.com/article/7-major-companies-that-accept-cryptocurrency-cm913745.

Fisch, Christian. "Initial Coin Offerings (ICOs) to Finance New Ventures: An Exploratory Study." *SSRN Electronic Journal*, 2018, doi:10.2139/ssrn.3147521.

Frankenfield, Jake. "Initial Coin Offering (ICO)." *Investopedia*, Investopedia, 12 Mar. 2019, www.investopedia.com/terms/i/initial-coin-offering-ico.asp.

"How Does Bitcoin Work?" *Bitcoin*, bitcoin.org/en/how-it-works.

"ICOBench.com Review: Best Cryptocurrency ICO Token Data & Details?" *BitcoinExchangeGuide*, 11 Apr. 2018, bitcoinexchangeguide.com/icobench/.

Momtaz, Paul P. "Token Sales and Initial Coin Offerings: Introduction." *SSRN Electronic Journal*, 2018, doi:10.2139/ssrn.3277707.

"Random Forests Classifiers in Python." *DataCamp Community*, www.datacamp.com/community/tutorials/random-forests-classifier-python.

Sawe, Benjamin Elisha. "How Many Currencies Exist in the World?" *WorldAtlas*, 28 June 2018, www.worldatlas.com/articles/how-many-currencies-are-in-the-world.html.

Tapscott, Alex TapscottDon. “How Blockchain Is Changing Finance.” *Harvard Business Review*, 1 Mar. 2017, hbr.org/2017/03/how-blockchain-is-changing-finance.

Viens, Ashley. “The Beginning of a Bitcoin Bull Run?” *Visual Capitalist*, 17 May 2019, www.visualcapitalist.com/bitcoin-bull-run/.

“What Is Data Mining?” *SAS*, www.sas.com/en_us/insights/analytics/data-mining.html.