Do signals interact?

A machine learning approach to study the multi-signal environment of Initial Coin Offerings.

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Management of Technology**

Faculty of Technology, Policy and Management

by

Job Wever

Student number: 5174171

To be defended in public on December 22nd 2022

Graduation committee

Chairperson, Second Supervisor	: Dr. ir., S., Van Cranenburgh, Section T&L
First Supervisor	: Dr. ir., Z., Roosenboom-Kwee, Section ETI
Advisor	: A.A., Ralcheva, Section ETI

Do signals interact?

A machine learning approach to study the multi-signal environment of Initial Coin Offerings.

by

J. Wever

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday December 22, 2022 at 10:00 AM.

Student number:	5174171	
Thesis committee:	Dr. ir. S. Van Cranenburgh,	TU Delft, chair, second supervisor
	Dr. ir. Z. Roosenboom-Kwee,	TU Delft, first supervisor
	A. A. Ralcheva,	TU Delft, advisor

TU Delft

Preface

Dear reader,

I would like to thank my supervisor, Aleksandrina Ralcheva, whose insight and knowledge into the subject matter steered me through this research. I am sure you are the one to lay the basis for my scientific eagerness in the field of Entrepreneurial Finance, and I hope that our fruitful cooperation in the scientific field will continue. I am sincerely grateful to my first supervisor, Zenlin, for her invaluable advice on the research subject and for establishing the graduation committee. As a thoughtful lecturer, she was the first to introduce me to the field of Finance. Finally, I would like to express my gratitude to the committee chairman, Sander van Cranenburgh, for his feedback and contributions to the methodological part of this thesis. Without you, the hyperparameters would still be lost somewhere in a dark random forest.

Executive Summary

Many firms seek external financial resources to sustain and expand their businesses (Choi and Wang, 2009). Initial Coin Offerings (ICOs) represent a novel way for ventures to raise capital for a wide variety of projects. To obtain these resources, ICOs need to provide information on the quality of their economic activities and their capabilities to potential contributors. However, in the absence of official prospectus and mandatory disclosure of information documents, it is challenging to assess the quality of an ICO. Because their underlying quality is not readily observable, ICOs send multiple signals which proxy the potential of their tokens. Signaling theory (Spence, 1978) can be used to explain how ventures convey relevant information to potential external contributors to acquire financial resources. This thesis offers some novel insights into how the ICO signaling process works and has evolved.

While the traditional signaling framework has proven to be an effective mechanism for understanding how external parties assess the quality of new ICOs, much of the literature has explored the impact of signals in isolation (Connelly et al., 2011) and in relatively low noise environments. Overlooked has been the complementary, overlapping, and potentially competing effect that signals may have on heterogeneous decision-makers in more chaotic environments such as ICOs. Recent developments on the theoretic foundation of signaling literature emphasize the need for a more sophisticated approach. For example, Drover et al. (2018) argue that some of the assumptions underlying the signaling framework underestimate the complexity inherent to the signaling process, as it takes place in a high-noise environment, and it is subject to cognitive processes and sets of decision-making criteria being time and context dependent.

To better understand the ICO signaling process, this thesis departs from studying signals in isolation and focuses instead on the complex interactions between signals. The present thesis is concerned with two specific themes – namely, the temporal and context dependency of signal effectiveness – and studied it in two particular ways: logistic regression and a data-driven approach. The former largely being the tool of the traditional signaling framework, howbeit extended with signal interaction terms. The latter refers to machine learning methods that are being expected to learn such relationships from the data without explicitly having to program them a priori.

In light of ICOs, where relatively little is known or understood about the interactions between signals, specifying interaction terms a priori may be a challenging and time consuming process. It is along these lines that we may begin to situate the significance and benefits of studying signals with data-driven methods – not so much in terms of generating the highest prediction performance but rather as a potential tool to understand and explore the ICO signaling dynamics in more detail. Nevertheless, it is still unclear if signaling research could benefit from a data-driven approach. Therefore, the main goal and research question central to this thesis are:

Research goal: To examine the suitability of a data-driven approach to effectively study signal interactions in the context of ICOs.

Main RQ: Can a data-driven approach effectively be applied to study signal interactions in the ICO environment?

The research started with an extensive literature survey aimed at identifying signals that were previously found to be associated with ICO success, also known as success determinants. Following this, relevant data was scraped from multiple sources (i.e., ICOmarks, TORD, Twitter, Telegram, Facebook, and GitHub) resulting in a data set of 957 ICOs launched between 2017 and 2019. To the best of the author's knowledge, this is thus far one of the largest and most comprehensive data sets considered for ICO signaling research. In the first part of this thesis, the analysis was performed with the aid of the logistic regression. The first step was to identify the success determinants within the full data set. The results from the logistic regression were largely consistent with previous studies but also provided some additional insights. It was found that ICOs with larger Twitter and Telegram networks are more likely to succeed. Furthermore, specifying a higher number of milestones was found to be negatively associated with success.

Next, this process was repeated for different subsamples to get a first impression of the time and context dependency of signals. First, the pooled sample was subdivided into time windows (i.e., 2017, 2018, and 2019) and a separate analysis was performed within each window. Second, ICOs were grouped based on categories and a separate analysis was performed within each of the following categories: Data and AI, Financial Services, and Entertainment and Gaming. Consistent with expectations (Bellavitis et al., 2021), the subsample analyses revealed large swings in the statistical significance of signals between time samples (i.e., 2017, 2018, and 2019) and industry sector samples (i.e., Data and AI, Financial Services, and Entertainment and Gaming).

In the analysis that followed, the time and context dependency of signals was further investigated by extending the logistic regression with interaction terms. Specifically, the time dependency was investigated by creating interaction terms between signals and time (i.e., the years 2017 and 2018). The context dependency was explored by creating interaction terms between signals and categories (i.e., Financial Services and Entertainment and Gaming). Here, initial evidence was provided that effective signals differ over time and are context dependent. Overall, this supports the notion that signals interact, and, therefore, should not be studied in isolation but rather in portfolios of signals (Drover et al., 2018).

The second part of the thesis was devoted to evaluate the suitability of a data-driven method. For this purpose, four data-driven models were selected: Ridge Classifier, Random Forest Classifier, Extra Trees Classifier, and Support Vector Machine. A short explanation is provided below:

- **Ridge Classifier:** Addresses the problem of overfitting and multicollinearity in regression methods by means of L2 regularization - i.e., by adding a penalty term to their cost function.
- **Random Forest Classifier:** Employs the bootstrapping technique to create multiple decision trees on randomly selected data samples, known as bags, gets a prediction from each tree, and selects the best solution by utilizing voting.
- **Extra Trees Classifier:** A decision tree based classifier that aggregates the results of multiple de-correlated decision trees collected in a "forest." In concept, it is very similar to a Random Forest Classifier and differs from it only in the manner of construction of the decision trees in the forest.
- **Support Vector Machine (linear kernel):** Its objective is to separate the two classes of data points using one hyperplane among all the ones available. The algorithm searches for the hyperplane that has the maximum margin (distance) between data points of both classes.

The first step was to choose, among the selected models, the one that best accomplished the underlying objective of distinguishing successful ICOs from unsuccessful ICOs. Here, classification performance was considered a good proxy of the ability of an evaluation method to capture the complexities underlying the ICO signaling process. The stratified cross-validation method (Schaffer, 1993) was utilized to evaluate the performance of the models in terms of the AUC-ROC, Accuracy and Precision metrics.

Proceeding with the model that performed best, Random Forests, the last part of the thesis was devoted to test the suitability of this model to study the ICO signaling process. First, the most important signals were identified by evaluating their feature importance scores. Second, the SHapley Additive exPlanations (SHAP) approach was adopted to study the effect of signals in more detail. Unlike the feature importance scores, SHAP differentiates between the negative effects and positive effects of signals on ICO success. This approach resulted in several interesting insights. The SHAP analysis indicated that the size of an ICO's Twitter, Telegram, and Facebook network is exponentially related to the probability of funding success. Moreover, small network sizes were found to have a negative impact on funding success, whereas larger network sizes were found have a positive effect on funding outcomes. Finally, the impact of interaction terms on the SHAP values was determined to assess whether the data-driven was capable of capturing the pre-identified signal interactions - i.e., the signal interactions identified by extending the logistic regression with interaction terms. The results demonstrated that the Random Forest effectively captured the signal interactions.

Although the results should not be treated uncritically, they suggest that a data-driven approach, such as Random Forests, can help us understand the effect of signals and signal interactions on ICO funding success. Unlike established methods (e.g., logistic regression), the main advantage of this data-driven approach is that it does not require the analyst to specify interaction terms a priori but effectively learns them from the data. While the results show potential uses and interpretations of the outcomes of Random Forest models, it is essential to recall that such outcomes should not be treated as equivalent to the outcomes of a statistical model like logistic regression. In other words, generalizations from machine learning may lead to misjudgments if correlations are taken for causalities. Machine learning, therefore, is not just a new toolbox for the same problems. It should rather be seen as a different way of exploring signaling issues which is adequate in cases where data is complex and theoretical expectations are missing or are drawn into question. This thesis emphasizes that a data-driven approach can be used as an effective tool to explore potentially relevant signal interactions, whereas established tools such as logistic regression can be used as a confirmatory tool.

From a managerial perspective, the present challenges of ICOs require legislation that acknowledges and understands a global digital market's economic and technical peculiarities (Zetzsche et al., 2019). Failure to do so will foster fraudulent behavior and result in regulatory arbitrage with fuzzy implications (De Andrés et al., 2022). The interplay between a large number of signals broadcasted simultaneously, and the ability of regulators to evaluate rich information in noisy environments likely plays a key role in designing effective regulations. The proposed data-driven method offers a novel and efficient approach that facilitates learning about such high-noise signaling environments. Therefore, the proposed method could complement current practices to improve investor protection, make the ICO market more efficient, and enhance success outcomes of ICOs.

For investors and entrepreneurs, this thesis shows that having a larger Twitter and Telegram network size, a lower number of milestones, a shorter planned duration of the ICO campaign, no specified hard cap, a higher expert rating, and a larger team size were positively related to funding success. Depending on the sample specification, higher Environmental, Social, and Governance (ESG) scores, offering bonuses, the Know Your Customer program, having a Minimal Viable Product (MVP), having Reddit,

having Slack, and having Bitcointalk positively impact ICO funding.

As the problem of information asymmetry is widely diffused throughout management practices, the applicability of the proposed data-driven approach may also be extended to other areas of inquiry. Future research could explore to what extent the proposed approach can help to alleviate information asymmetry in other signaling environments. Developing models that provide a more accurate approximation of the signaling process are generally desirable. The data-driven approach provides a new way for mining potentially relevant signal interactions.

Contents

1 Introduction **1**
- 1.1 Research Objective . . . 2
- 1.2 Scientific and managerial implications . . . 3

2 Initial Coin Offerings **5**
- 2.1 Definition of ICOs . . . 5
- 2.2 Signaling in ICOs . . . 7

3 Signaling: a new paradigm **11**
- 3.1 Machine Learning in Entrepreneurial Finance . . . 12
- 3.2 Preliminaries . . . 13

4 Methodology **15**
- 4.1 Sample and data description . . . 15
- 4.2 Modeling approach . . . 18
  - 4.2.1 Benchmark model . . . 18
  - 4.2.2 Signal interaction terms . . . 18
  - 4.2.3 Machine learning . . . 19
  - 4.2.4 Evaluation . . . 24
  - 4.2.5 Interpretation and validation . . . 24

5 Results **25**
- 5.1 Logistic regression: Signal interactions . . . 25
- 5.2 Machine Learning approach . . . 33
- 5.3 Comparison: Logistic Regression and Random Forest . . . 42

6 Discussion and Conclusion **45**
- 6.1 Research conclusion . . . 45
- 6.2 Scientific relevance . . . 48
- 6.3 Managerial implications . . . 49
- 6.4 Limitations . . . 49
- 6.5 Future recommendations . . . 50

A Grid search **59**

B Parsimonious model supplement **63**

C Correlation matrix **65**

D AUC-ROC **69**

E Regression models **73**

1

Introduction

Initial Coin Offerings (ICOs) emerged as a disruptive tool in capital formation and are contributing to disintermediate financial markets. ICOs are smart contracts based on blockchain technology that allow fundraisers to solicit money from the crowd by issuing digital currencies. Being a digital, decentralized, disintermediated, global, and unregulated market, ICOs present novel challenges but also some innovative solutions to existing fundraising methods (Ackermann et al., 2020). Since its peak in 2018, several developments, such as the first year-on-year decrease in 2019, raised important questions about the future of the ICO market (Fromberger and Haffke, 2019; de Andrés et al., 2022).

One of the biggest challenges in the ICO market is to distinguish successful ICOs from unsuccessful ICOs. Bypassing any regulation that normally applies to businesses placing securities to investors, dozens of ICOs raise money without official prospectuses, with no particular protection for contributors and disclosing only a limited set of information. As a result, contributors find it challenging to assess the prospects of ICOs and suffer from high information asymmetries (Momtaz, 2021a). Existing research relies on signaling theory to describe the market and to differentiate between successful and unsuccessful tokens. In the context of ICOs, this framework argues that because token quality cannot be directly observed, decision-makers must rely on information signals thought to correlate with quality. This means that evaluators (e.g., regulators or investors) search for signals that can be used to make inferences about the underlying quality of an ICO.

Thus far in the context of ICOs, the vast majority of signaling research tends to investigate the ways in which a positive signal – in isolation – correlates to the quality of an ICO, assuming that signals are temporally stable, their effect is symmetric over the range of the signal, the audience is relatively homogeneous, and receivers interpret the signal in the same way over time (Connelly et al., 2011; Fisch, 2019; Momtaz, 2021b). These assumptions, however, may underestimate the complexity underlying the ICO signaling process, which resembles a high-noise environment where multiple ventures send multiple signals simultaneously. This means that multiple signals compete for the receivers' attention, and these signals may be drowned out, amplified, or distorted along the way. Currently, large swings in ICO signal interpretations make it difficult to establish consistent empirical relationships (Bellavitis et al., 2021). As such, the traditional signaling framework building on assumptions of stable and independent relationships may be less suitable, at least for the early-stage analyses of the ICO market.

The issues put forward that may be problematic for the traditional signaling framework may be summarized in three main points. First, fundraising initiatives tend to be pursued after considering the in-

terdependencies between and among the signals (Douglas et al., 2020). Yet, the traditional framework explains the signaling process as the linear additive impact of the signals considered in isolation – i.e., independently of the effect of other signals (Drover et al., 2018). Second, entrepreneurial phenomena are often characterized by relationship asymmetry, meaning that a signal may be both positively and negatively associated with funding success, depending on the configurations (Edelman et al., 2021). Third, the heterogeneity of fundraising is reflective of inter-configurational differences between ICOs, whereas the traditional framework focuses on the commonalities across all the configurations and thus suppresses inter-configurational differences that may be causal for the observed heterogeneity (Douglas et al., 2020).

Most of these issues may be addressed by extending the traditional framework with interaction terms. While clearly useful, specifying models that capture the complexities inherent to the ICO signaling process is challenging. The specification process of these models usually relies on prior knowledge from the analyst concerning, for example, how signals interact and which interactions are relevant to include. Furthermore, finding a proper model specification usually involves a trial-and-error procedure, in which several candidate specifications are tested and the most parsimonious or plausible model is chosen. This process is already cumbersome for the traditional signaling framework (Spence, 1978; Fisch, 2019) but will be even more complicated when incorporating interaction effects (Drover et al., 2018). The presence of considerably more signals, possible combinations of chosen alternatives, and potential interactions effects between alternatives will increase the complexity of the model specification process.

Instead of testing prior problem formulations, machine learning algorithms learn from the data to discover patterns, which can then be used for prediction and generating new information. It sacrifices possible intuitive insights in favor of a method which does not require any a priori assumptions either of interactions between signals or of context dependent relationships. This may improve the efficiency of the feature and model selection practices while making it less vulnerable for biases and interpretations guided by the interest of the research. Additionally, adopting such a data-driven approach might reveal relationships that are nonintuitive.

1.1. Research Objective

Although literature suggests many benefits over traditional signaling methods, it is still unclear if signaling research could benefit from a data-driven approach. The main goal of this research is to examine the suitability of a data-driven approach to effectively study signal interactions in the context of ICOs. In addition, a systematic understanding of how signal interactions influence ICOs is still lacking. As an exploratory empirical study, the first step of this thesis is to develop initial evidence about the role of signal interactions in the context of ICO funding. In doing so, I depart from the tradition of studying singular signals working in isolation and focus instead on the complex interactions within a portfolio of signals. Building on recent theoretical development (Drover et al., 2018), I aim to explore the time dependent, complementary, overlapping, and potentially competing effects that signals may have. I expect that signals are time and context dependent, that signals moderate each other through complex interactions, and that different configurations can lead to the same outcome. Taken together, I seek to answer the following research question:

Main RQ: Can a data-driven approach effectively be applied to study signal interactions in the ICO environment?

The present thesis is concerned with two specific interaction effects: the temporal and context depen-

dency of signal effectiveness. These effects are studied using two methods: a logistic regression and a data-driven approach. The former largely being the tool of the traditional signaling framework, howbeit extended with signal interaction terms. The latter being expected to learn such relationships from the data without explicitly having to program them a priori.

The first step is to identify the signals that have previously been associated with ICO success[1]. Here, I address the following question:

RQ1: What are the determinants of ICO success?

Next, I examine the time and context dependency of these signals by extending the logistic regression with interaction effects. Specifically, I ask:

RQ2: What is the influence of signal interactions on ICO success?

The second part of this thesis is concerned with testing the suitability of the a data-driven approach to study signal interactions. Four data-driven models have been considered in the present thesis: Random Forest (RF), Extra Trees (ET), Ridge Regression (Ridge), and Support Vector Machine (SVM). The first objective is to choose, among the selected set of machine learning models, the one that best accomplishes the underlying objective to distinguish successful from unsuccessful ICOs.

RQ3: To what extent do the respective data-driven models (i.e., RF, ET, Ridge, and SVM) improve the classification performance (if at all) as compared to the logistic regression?

Once the best performing model is selected, the remainder of the analysis focuses on evaluating the suitability of the respective model to study the ICO signaling environment.

RQ4: To what extent can a data-driven approach be used to extract relevant information concerning the ICO signaling process and signal interactions?

1.2. Scientific and managerial implications

The proposed methods to delineating the evaluation of signals arises from the need to better understand the multi-signal ICOs environment. This research makes several contributions to the literature. First, I extend the literature on ICO signaling by departing from the tradition of studying singular signals working in isolation, and focused instead on the interactions between signals. In doing so, I show that signals interact and jointly affect ICO funding success. Specifically, this thesis provides initial evidence that signals in the high noise ICO environment are time and context dependent. Second, the present study tested the suitability of a data-driven approach to study the high-noise ICO signaling environment.This data-driven approach results in a modest improvement of classification performance in comparison to the baseline approach and was shown to be an effective tool to explore signal interactions. This research also presents how to leverage machine learning methodologies in signaling research and discusses their potential for complementing current practices.

This thesis further shows that having a larger Twitter and Telegram network size, a lower number of milestones, a shorter planned duration of the ICO campaign, no specified hard cap, a higher expert rating, and a larger team size were positively related to funding success. Depending on the sample specification, higher Environmental, Social, and Governance (ESG) scores, offering bonuses, the

[1] In this work, success is defined as a binary variable indicating whether an ICO achieved their minimum required funding target, also known as a ”softcap”.

Know Your Customer program, having a Minimal Viable Product (MVP), having Reddit, having Slack, and having Bitcointalk positively impact ICO funding.

The findings of this thesis are especially informative for two parties involved in the ICO process, namely regulators and investors. In absence of official legal jurisdictions and the mandatory disclosure of documents, regulators and investors are relying purely on the signals to evaluate ICOs and understand the market. As such, a better understanding of the underlying signaling process is crucial. The present challenges of ICOs require legislation that acknowledges a global digital market's economic and technical peculiarities (Zetzsche et al., 2019). The interplay between a large number of signals broadcasted simultaneously, and the ability of regulators and policy makers to evaluate rich information in noisy environments likely plays a key role in designing effective regulations. Furthermore, the continuous development of the market is also likely to further evolve ICOs in ways that may change the dynamics between investors and entrepreneurs. The extensions presented in this research provide a new way that facilitates learning about and accounting for such developments. The proposed data-driven approach could complement current practices to improve investor protection, make the ICO market more efficient, and enhance success outcomes of ICOs.

Although the focus of this thesis has primarily centred on the evaluation of ICO funding success, the applicability of the proposed methods can also extend to other areas of inquiry. For example, the ideas advanced in this thesis may provide new insights in the area of crowdfunding and angel investing. The applicability of signaling theory and the present extensions, though, can also extend to areas beyond the boundaries of entrepreneurial finance. For instance, how does the model help inform the interacting attributes of individual assessments, such as job applicants? Here, researchers might explore how certain signals jointly relate to the quality of new hires. Furthermore, the information asymmetry concept is widely diffused throughout management research. Hence, developing models that provide a more accurate approximation of the signaling process are generally desirable. Given that the specification process of interaction terms can be cumbersome for established methods, coupled with theoretical underdevelopment, the data-driven approach provides a new way to mining potentially relevant signal interactions.

The remainder of this work proceeds as follows. Firstly, I will offer a definition of Initial Coin Offerings, contrast ICOs to conventional fundraising methods, and review related signaling literature. Next, I will provide a brief description of machine learning, its applications in entrepreneurial finance, and a short theoretical explanation of the models selected in this thesis. I then present the models and discuss their results, extensions, and implications. Finally, the thesis is concluded with suggested topics for further research.

2

Initial Coin Offerings

2.1. Definition of ICOs

Initial Coin Offerings, also known as Token offerings, are smart contracts on a blockchain used to raise external funding by issuing tokens or coins. A blockchain is a decentralized, immutable, public ledger that is used to record transactions – referred to as ”blocks” – across many computers such that the record cannot be modified without the permission of all subsequent block owners (Iansiti and Lakhani 2017). Smart contracts are computer protocols on existing blockchains that automate value-exchange transactions. In essence, smart contracts replace intermediaries, reducing the transaction costs in the exchange process (Momtaz 2018c). ICOs draw inspiration from concepts like crowdfunding and initial public offerings (IPOs), but represent their own unique category of fundraising (Momtaz, 2019). One aspect that distinguishes ICOs from traditional entrepreneurial financing is the concept of selling tokens. A token corresponds to a unit of value issued by a venture and typically takes two forms: ”utility token” or ”equity-based token” (Amsden and Schweizer, 2018).

Utility tokens, the most prevalent type of token, refer to a digital property that enables the exchange of utility. These tokens, often called ICOs, serve as a currency in the venture's digital environment and can be redeemed for a product or service once developed. Although tokens usually have no financial or tangible value at the start of an ICO (Kaal and Dell'Erba, 2017), tokens can be bought in exchange for cryptocurrencies such as Bitcoin and Ether. In some cases, however, they may also be bought in exchange for fiat money (Fenu et al., 2018). Moreover, while utility tokens do not possess ownership rights, they are popular because of the lack of regulation in most jurisdictions (Momtaz, 2020). At the same time, the lack of regulation accentuates the risk of this investment vehicle and highlights the need for more accurate screening and due diligence practices. Therefore, the scope of the present thesis is limited to utility tokens.

In contrast, the equity-based tokens, known as Security Token Offerings (STOs), are subject to securities laws. The value of an STO is derived from a tradeable asset, and they are typically used as investment vehicles (Fisch, 2019). They come in the form of equity tokens, which are similar to conventional stocks, or akin to all sorts of securities, offering the possessor a form of ownership, voting rights, dividends, or other monetary benefits (Fisch, 2019). Because STOs are regulated securities and do not share the same risks as ICOs, these type of tokens are considered outside the scope of the present thesis.

Since ICOs are a relatively new financing method, the related literature is still nascent. To get a better understanding of this distinct form of financing, I provide parallels with other sources of entrepreneurial finance. More specifically, I compare ICOs with equity-based crowdfunding, reward-based crowdfunding, and IPOs along four dimensions: (i) funding characteristics, (ii) investor characteristics, (iii) deal characteristics, and (iv) post-deal characteristics. The discussion is summarized in table 2.1.

Table 2.1: Brief comparison between ICOs and conventional funding mechanisms.

	ICOs	**Equity Crowdfunding**	**Reward Crowdfunding**	**IPOs**
Funding				
Stage of funding:	*All stages*	*Early stage*	*Early seed stage*	*After funding rounds*
Format:	*Utility tokens, security tokens*	*Equity instrument*	*Product*	*Equity shares*
Investors				
Type of investor:	*Early adopters and public*	*Early adopters*	*Early adopters*	*Public*
Motive of investors:	*Financial and non-financial*	*Financial and non-financial*	*Financial and non-financial*	*Financial*
Deal				
Amounts (in USD $):	*>1k*	*100k – 2m*	1k - 150k	*>10m*
Transaction costs:	*Lowest*	*Low*	*Low*	*High*
Regulation:	*Low*	*Low*	*Low*	*High*
Information source:	*Whitepaper*	*Campaign description*	*Campaign description*	*IPO prospectus*
Post-deal				
Liquidity:	*High*	*Low*	*Low*	*High*
Voting power:	*Security tokens: yes; Utility tokens: No*	*No*	*No*	*Yes*
Exit options:	*Secondary market*	*IPO, acquisition*	*IPO, acquisition*	*Tradeable on market*

While ICOs are not limited to a specific funding stage, they typically provide access to capital for early stage ventures whose products or platforms are still in the development stage (Momtaz, 2020). In that sense, they are comparable to crowdfunding, which is generally used as a fundraising tool in early stages. In contrast, IPOs are reserved for established start-ups to acquire a high volume of growth capital (Huang et al., 2020). Similarly as with tokens, there is a distinction between equity-based crowdfunding and reward-based crowdfunding. Here, STOs are more comparable to equity based crowdfunding as their value is derived from a share of ownership. ICOs are most similar to reward based crowdfunding as both financing tools raise funds in exchange for a form of utility (e.g., product or digital service). However, an important distinction between ICOs and reward-based crowdfunding is that ICOs are tradeable, whereas the product obtained in reward-based crowdfunding is essentially illiquid. In terms of liquidity, ICOs can best be compared to IPOs as investors obtain tradeable stocks in exchange for funding capital. On the question of the investors' motives, investors in ICOs and crowdfunding investors are driven by financial and non-financial motives and typically attract early adopters (Lipusch, 2018; Fisch et al., 2018). On the contrary, IPOs investors are generally driven by financial motives only (Momtaz, 2020). Another significant reason for investors to invest in ICOs is the after market liquidity. Most tokens are listed on an exchange platform on which they can be traded on a 24/7 basis. Neither crowdfunding nor IPOs can offer the same amount of liquidity (Momtaz, 2020). Provided that they are listed on a trading exchange, ICOs provide the opportunity for investors to exit at any time. Exits in other early stage financing methods are often not feasible until a specific maturity level is reached. Similar to crowdfunding campaigns, ICOs have close-to-zero transaction costs and low levels of regulation in comparison to IPOs. With regards to the funding amounts, the largest ICO ranked among the three largest IPOs globally in 2018 (Howell et al., 2020). According to Fisch (2019), the most successful ICO campaign exceeded the aggregated funding raised on Kickstarter – one of the largest crowdfunding platforms – since its launch in 2009. For example, previous studies report successful ICO campaigns ranging from USD $1,000 up to USD $4.2 billion (Momtaz, 2019).

Having defined what ICOs are, in general, and how they complement other funding methods, in particular, I move on to study the ICO environment in more detail. Before proceeding to discuss the factors that drive successful fundraising initiatives in ICOs, I first introduce the theoretical foundation of the

present thesis: signaling theorem.

2.2. Signaling in ICOs

Literature on the theoretical foundations of the ICO market is sparse (Fridgen et al., 2018; Li and Mann, 2018; Liebau and Schueffel, 2019; Ofir and Sadeh, 2020; Chitsazan et al., 2022). Only a few studies explored theoretical motivations for the factors that shape ICO success (Catalini and Gans, 2018). Others have employed organizational legitimacy theory (Chanson et al., 2018); behavioral economics in information systems theory (Albrecht et al., 2020); and agency theory (Liebau and Schueffel, 2019; Momtaz, 2021a) to study the ICO market.

Most studies adopted signaling theory to investigate success determinants of ICO campaigns (Amsden and Schweizer, 2018; Fisch, 2019; Momtaz, 2019). Signaling theory is a suitable basis for ICO research since the market suffers from a significant degree of uncertainty and information asymmetry between the primary participants (i.e., issuers and investors) in the ICO process. As a result, issuers must entice investors by providing information that signals the quality of their business (Spence, 1978). In this theoretical framework, the primary components of analysis are issuers, signals, and investors in the ICO market (Spence, 1978; Connelly et al., 2011).

Several signals have been associated with ICO success in previous studies. These signals may be subdivided into three categories: (i) token characteristics, (ii) campaign characteristics, and (iii) social capital.

Token characteristics

One of the signals that has been found to correlate to an ICO's success is the white paper (Zetzsche et al., 2018). A white paper is a disclosure document that functions as a company's prospectus and provides information on the token characteristics (e.g., location, utility, blockchain technology), campaign characteristics (e.g., pricing, milestones, duration) and team characteristics (e.g., team size, advisors, experience). As the main source of information, whitepapers provide value by minimizing information asymmetry between ventures and investors (Amsden and Schweizer, 2018; Bourveau et al., 2018; Fisch, 2019; Ofir and Sadeh, 2020).

The geographic location also plays an important role in the success of a token (Huang et al., 2020; Ackermann et al., 2020; Fisch, 2019; Davies and Giovannetti, 2018). This can be attributed to the presence of specific conditions in markets that enable the emergence of ICOs. For example, the existence of developed financial markets, blockchain technologies, and the development of regulations were found to increase the interest in ICOs significantly (Huang et al., 2020). Especially projects launched in Europe or America are more successful (Fisch, 2019), along with ICOs from China and Israel (Fenu et al., 2018). Crowdfunding literature also demonstrates that campaigns located in bigger cities are more likely to succeed (Ralcheva and Roosenboom, 2020).

Similarly, the industry sector to which an ICO belongs is directly linked to its success. Previous research has shown that the outcome of a campaign is dependent on the specific sector or industry a product or service belongs to (Davies and Giovannetti, 2018; Fisch, 2019). One potential explanation is that that higher technological capabilities are correlated to innovative potential thus are crucial for a token's success, particularly in industry sectors where innovation is the basis of competition (Fisch, 2019).

Several attempts have been made to imbue some kind of regulation into the ICO market, two of which

have been linked to success in prior studies: a Know Your Customer requirement and a whitelist. The Know Your Customer (KYC) program requires investors to provide information to confirm their identities, thereby mitigating some of the risks and information asymmetry between investors and issuers (Shrestha et al., 2021). Likewise, a whitelist is a list of approved investors that are granted exclusive access to invest in ICOs. According to Belitski and Boreiko (2021), ICOs that use a whitelist are more likely to succeed.

Previous studies have also examined the impact of ICOs' sustainability scores on funding success. According to Mansouri and Momtaz (2022), ICOs with higher Environment, Society, and Governance (ESG) scores are more successful on average. The ESG scores were calculated using a machine learning model that evaluated the content of ICOs' whitepapers.

Campaign characteristics

The time preceding an ICO is considered critical and mainly devoted to publicity efforts and the development of the campaign (Momtaz, 2019). In the development stage, issuers can solicit feedback from early investors through the pre-sale of tokens. The pre-sale of tokens – known as the pre-ICO phase – is an optional stage preceding the launch of an ICO where early investors are offered discounts and bonuses in compensation for the risk (Liu and Wang, 2019b). Literature reports both positive (Roosenboom et al., 2020; Lyandres et al., 2019; Adhami et al., 2018) and negative (Amsden and Schweizer, 2018; Momtaz, 2021b) correlations between pre-sale campaigns and project success. Possible explanations for this discrepancy might be the interaction between signals or the instability of the signal (i.e., the effect of the signal differs over time).

Previously published studies on the effect of financing thresholds on project outcomes are also not consistent. At the start of an ICO campaign, entrepeneurs have the option to specify their fundraising targets in terms of a soft-cap and/or hard-cap limit. According to Amsden and Schweizer (2018), hard-cap limits positively affect a project's success, as they help investors predict the value of the tokens more accurately (Amsden and Schweizer, 2018). However, high hard-cap limits, which seem unattainable, negatively affect the project's success (Lyandres et al., 2019). Studies on the effect of soft-cap limits are less consistent and provide evidence for a positive effect (Amsden and Schweizer, 2018) as well as a negative effect (Bourveau et al., 2018), thus requires additional research.

Another signal that has been identified in literature is the duration of a campaign. Previous studies indicate that the time span of a campaign is negatively correlated with funding success (Roosenboom et al., 2020; Ackermann et al., 2020; Fisch, 2019; Davies and Giovannetti, 2018). This means that shorter campaigns have a higher probability of success, whereas longer campaigns are negatively correlated to funding success.

Previous studies also suggest that using bonus schemes negatively correlates to funding success (Roosenboom et al., 2020). One potential explanation is that a bonus scheme may indicate that an ICO is struggling to attract sufficient interest in the ICO by itself, thus resorting to bonuses and price discounts to gather the attention of investors. An additional form of discount, the bounty program, constitutes an offer and sale of tokens where tokens can be earned in exchange for services designed to advance or develop the product or service related to an ICO. According to Fisch and Momtaz (2020), offering a bounty program has a negative effect on funding success, possibly because it signals a lack of competency and independency of the issuers.

Social capital

Social capital is a critical ingredient for the development of innovative digital opportunities (Al-Omoush et al., 2020). Literature has linked several social capital characteristics to the success of an ICO project, including team size, team experience, and social media presence.

Multiple studies examined the influence of team size on ICO project performance, and there is consensus that larger teams are more likely to succeed (Roosenboom et al., 2020; Liu and Wang, 2019b; Amsden and Schweizer, 2018; Adhami et al., 2018; Ante et al., 2018). Regarding team characteristics, only past managerial experience tends to be relevant for the success of an ICO, while education, professional, and entrepreneurial experience are irrelevant (Adhami et al., 2018). Similar conclusions were found in traditional crowdfunding literature (Allison et al., 2017).

Social networks have been linked to the performance of ICO projects as they foster innovation by promoting collaborative work (Al-Omoush et al., 2020; Fenu et al., 2018). They provide legitimacy to projects, as ICO issuers do not entirely control the project's content in the social networks' environment. Research shows that media-provided content outperforms firm-provided content in influencing investors' behaviour (Chanson et al., 2018). Nevertheless, well-managed social media platforms can encourage early investors, and regular updates about the project contribute to a project's success (Bourveau et al., 2018).

Next to having an active social media network, the availability of a GitHub platform as a deposit of public code boosts a project's success (Albrecht et al., 2020), particularly during the pre-sales of tokens (Roosenboom et al., 2020). While social media platforms add value in general, GitHub deposits signal quality in particular because it demonstrates technological capabilities.

Another way of reducing information asymmetries and identifying quality projects is by using experts' ratings. Despite some concerns, expert ratings are widely used as a substitute for traditional third-party involvement and were found to determine the success of an ICO with considerable precision (Liu and Wang, 2019a). Moreover, multiple studies have linked experts' ratings to project success and found that higher expert ratings are positively correlated with funding success (Fenu et al., 2018; Fisch and Momtaz, 2020; Momtaz, 2021b; Lee et al., 2022).

3

Signaling: a new paradigm

According to signaling theory (Spence, 1978), token issuers intentionally provide investors with specific information (i.e., signals) about their capabilities and business quality to alleviate information asymmetry and motivate investors to fund their ICO project. In turn, investors receive the signals and evaluate whether to participate in a specific ICO project. Hence, the information transmission between issuers and investors influences investors' decisions to participate in ICO initiatives, thus determining funding outcomes. Similarly, funding outcomes also provide feedback for issuers, which they can use to optimize the signaling process and encourage investors to participate. This signaling process affects issuers, investors, and ICO campaigns, consequently influencing ICO success.

Thus far in the context of ICOs, signaling research focused on singular positive signals, assuming that signals are temporally stable, their effect is symmetric over the range of the signal, the audience is relatively homogeneous, and receivers interpret the signal in the same way over time. Recent developments on the theoretic foundation of signaling literature emphasize the need for a more sophisticated approach. For example, Drover et al. (2018) argue that some of the assumptions underlying the signaling framework underestimate the complexity inherent to the signaling process, as it takes place in a high-noise environment, and it is subject to cognitive processes and sets of decision-making criteria being time and context dependent. Unlike Spence (1978), Drover et al. (2018) claim that signals should not be studied in isolation but rather in pools or configurations. They extend the traditional signaling framework by focusing on multi-signal interpretations and incongruent signals.

While the extension proposed by Drover et al. (2018) is relatively new, there are a variety of ways to study multi-signal configurations. For example, Edelman et al. (2021) adopted this extension in the form of a crisp-set qualitative comparative analysis (cs/QCA) to study signal configurations in angel investments. Because they are specifically designed to study causal complexity and equifinality[1], set-theoretic approaches such as cs/QCA are appropriate for analyzing complicated phenomena such as multi-signal configurations in fundraising. Critics, however, question the efficiency and suitability of cs/QCA methods when dealing with large sample sizes (Bail, 2015).

In the realm of crowdfunding, the effectiveness of the multi-signal approach has been exemplified in a report by Steigenberger and Wilhelm (2018). In contrast to Edelman et al. (2021), Steigenberger and Wilhelm (2018) followed an econometric methodology in the form of the generalized method of

[1] Here, causal complexity and equifinality refer to the situation in which a specific outcome may result from multiple possible configurations of signals.

moments (GMM). One potential drawback of this approach is that it may suffer from equifinality and functional biases - i.e., it assumes the relations between variables to be linear.

Although it has not been applied to the multi-signal framework yet, one way to tackle these issues while capturing causal complexity inside large data sets is the notion of machine learning (Bail, 2015). For example, random forest models combine traditional regression tree models with bootstrapping techniques to classify a data set into multiple branches — i.e., variable configurations — with sufficient sample size (Breiman, 2001). These novel tools show considerable promise for analyzing multi-signal environments. This is not solely because they detect causal complexity – or equifinality – but also because they could identify patterns within data that humans would not be capable of recognizing. In the sections that follow, I provide a more detailed account on the synthesis and evaluation of machine learning practices in the context of entrepreneurial finance.

3.1. Machine Learning in Entrepreneurial Finance

The rise of data and machine learning has led to digital innovations and technology-enabled business models in the financial sector, especially in the ”fintech” industry (Dixon et al., 2020). Among other innovations, fintech includes equity crowdfunding, mobile payment systems, trading systems, cryptocurrencies and other blockchain applications. Behavioural prediction is often a critical aspect of product design and risk management needed for fintech-based business models; consumers and investors are presented with well-defined choices but have ambiguous economic needs and limitations, thus do not necessarily behave in a strictly economically rational fashion. Hence, it is required to treat parts of the system as a ”black box” by using methods capable of capturing relations that cannot be known in advance.

The fundraising landscape has changed alongside the emergence of technology-driven entrepreneurship (Ferrati et al., 2021). Subsequently, the number of early-stage proposals submitted to Venture Capitalists has increased significantly (Block et al., 2021). Furthermore, because data mining — specifically, machine learning — has radically transformed numerous operations in the financial sector (e.g., stock trading, insurance and risk management, and wealth management), a similar impact is expected in the field of entrepreneurial finance (Giudici et al., 2021).

Recent developments in the field of crowdfunding highlight the potential of advanced statistical models in entrepreneurial finance. In these advanced statistical models – also referred to as machine learning – algorithms are able to learn from data to discover patterns and support decisions, as opposed to being explicitly programmed by the researcher based on prior expectations or knowledge. I review the burgeoning literature that bring models, prediction algorithms, and practices from machine learning to the entrepreneurial finance field (Yeh and Chen, 2020; Wang et al., 2020; Ren et al., 2021; Wei et al., 2022; Greenberg et al., 2013).

Scholars have put forward several motives for applying machine learning practices in the analysis of crowdfunding campaigns. I discuss the four main motifs that may be of interest to the ICO literature as well. Firstly, machine learning models can address some challenges associated with traditional statistical models (e.g., linear regression and logistic regression), such as the optimization of model specifications and the adverse effects caused by model misspecification. In traditional signaling models, the structure of the problem is imposed by the researcher based on prior statistical inferences or existing empirical evidence. Subsequently, in the time-consuming and often ad hoc process that follows, the final model is determined based on a set of performance comparisons (Vulkan et al., 2016; Walthoff-Borm et al., 2018). If the resulting model turns out to be a poor estimate of the true underlying

relationship between signals, model inferences and predictions can be erroneous. Instead of testing prior problem formulations, machine learning algorithms learn from the data to discover patterns, which can then be used for prediction and generating new information. This may improve the efficiency of the feature and model selection practices while making it less vulnerable for biases and interpretations guided by the interest of the research. Additionally, adopting a data driven approach might reveal relations that are nonintuitive. Secondly, machine learning algorithms often outperform traditional approaches in accuracy and provide better goodness-of-fit, especially in prediction practices (Yeh and Chen, 2020). While goodness-of-fit is rarely the main objective of studies in entrepreneurial finance, a better performance is generally desirable as it signals that the model resembles the underlying signaling process and market dynamics more accurately. Thirdly, another advantage of ML methods over traditional models is that ML algorithms can manage a large amount of structured and unstructured data and generally make reliable decisions or forecasts (Yeh and Chen, 2020). As regulators and evaluators are commonly exposed to a sheer amount of data, machine learning provides new ways for mining meaningful, statistically robust, and potentially hidden insights. Finally, machine learning methods (e.g., Random Forest) can work with types of data that can be problematic for traditional statistical methods, such as multicollinearity and nonlinearity. Introducing machine learning to the analysis of crowdfunding market dynamics thus provided an opportunity to tackle these issues and make existing analysis and classification methods more robust.

In summary, crowdfunding literature demonstrates a clear potential for machine learning models to enrich analysis and evaluation practices in ICO research. Yet, scant evidence is available about the incremental value of using such data-driven methods to analyze the ICO market. Before proceeding to discuss the suitability of specific models, it is necessary to explain some concepts related to machine learning.

3.2. Preliminaries

Machine learning is a branch of computer science concerned with developing computer algorithms that learn and improve automatically from analyzing and interpreting patterns and structures in data without being explicitly programmed (Bishop and Nasrabadi, 2006). Machine learning uses the input and output data to train an algorithm, automatically tries to find patterns within the data, and constructs a model from these patterns. Once the algorithm has identified patterns and learned relationships between the input and output data, the resulting model can be used to predict the output, starting from new input.

Typically, machine learning algorithms require a data set that consist of a set of instances that all have particular features with corresponding feature values. For example, a data set may comprise of records of ICO listings containing the expert ratings and team size of the ICO projects. In this example, each instance represents a particular ICO, the features are expert rating and team size, and the feature values are, for example, 4/5 stars and 5 team members.

One of the subcategories of machine learning is supervised learning. Supervised learning involves using labelled instances to train computer algorithms for predicting a particular target variable of an unseen instance. For example, one might want to train a model that is able to predict whether an ICO (consisting of several features) is successful (target). Supervised learning models can be distinguished based on the target variable type. The target variable can be continuous or categorical, referring to a regression or classification problem, respectively. This work focuses on a particular type of classification problem, the binary classification problem, in which the target variable takes two values (i.e., 0 or 1).

Other subcategories of machine learning are unsupervised learning and reinforcement learning. Unsupervised learning involves training a computer algorithm by exploring patterns in a set of unlabelled instances. Reinforcement learning focuses on maximizing a particular reward by learning which actions to perform.

4

Methodology

4.1. Sample and data description

Most of the information on ICOs is retrieved from ICOmarks, a database that contains information about more than 7000 ICOs[1]. ICOmarks is the most comprehensive and high-quality database of ICO campaigns available, and therefore serves as the foundation of the sample. Data is collected with the aid of an Application Programming Interface (API). To account for missing information, this database has been supplemented with data from additional sources (i.e., TORD, Twitter, Telegram, Facebook, and GitHub). This process resulted in a list of 957 ICOs. The resulting data set is cleaned using the following preprocessing steps:

- *Data cleaning*, where all redundant and irrelevant information is removed from the database, including duplicates, missing values, and outliers;
- *Data scaling and transformation*, where the goal is to rescale the values of numeric columns in the dataset without distorting the differences in the ranges of values, while reducing the impact of magnitude in the variance;
- *Data selection*, where the objective of the analysis (e.g., subsample analysis) determines which data will be taken into account in the final data set;
- *Oversampling and undersampling*, in case the training dataset has an unequal distribution of the target class, it will be fixed using the Synthetic Minority Over-sampling Technique (SMOTE);

Several measures of success have been proposed in literature. Four measures were identified that have been subsumed under the term 'success': (a) the tradability of tokens, (b) the total amount of capital raised, (c) reaching or exceeding a percentage of the hard cap, and (d) reaching or exceeding the softcap. In the present thesis, success is defined as a binary variable indicating whether the soft cap (i.e., the minimum required funding target) has been reached. On average, 34% of the ICOs in the sample reach their predefined soft cap.

In the analysis, multiple signals are included that, based on the literature review in Chapter 2.2, are expected to impact funding. An overview of the variables and their descriptions can be found in Table 4.1. Table 4.2 presents the descriptive statistics after the data scaling and transformation preprocessing step. Next to the variables identified in literature, several novel variables are introduced. First of all,

[1] See https://icomarks.com (last accessed November 2022).

instead of evaluating the ESG scores based on a whitepaper text analysis tool (Mansouri and Momtaz, 2022), it is explored if the same effect can be derived by performing a similar analysis on the ICO description[2]. The reason for performing the analysis on the ICO description is twofold: i) whitepapers are not always easily accessible, and ii) whitepaper analysis is a time-consuming process. Hence, this thesis tests a more accessible and efficient way to assess ESG signals. Furthermore, as campaign videos were found to have a positive influence on funding success in crowdfunding (Mollick, 2014), the present thesis tests if campaign videos have a similar effect in the context of ICOs. Here, the effect of a campaign video is modeled by a dummy variable indicating the presence of a campaign video.

In addition, the effect of social media is modeled by the size of an ICO's social media networks (i.e., Twitter, Telegram, and Facebook) instead of modeling the impact of these networks by means of a dummy variable (Perez et al., 2020). The size of a project's social media network was found to be a significant predictor of crowdfunding success (Lu et al., 2014; Thies et al., 2014). The online attention, the reach of a network, and the corresponding "Electronic Word of Mouth" were found to be valuable assets for marketing performances. Similarly to crowdfunding, ICOs have official Twitter, Telegram, and Facebook accounts. Therefore, this factor is expected to have a similar impact on ICO funding. The size of their social network accounts is measured by their respective number of followers. Lastly, this thesis considers additional communication platforms that are frequently used as communication tools related to ICOs: Reddit, Slack, Discord, and Bitcointalk. In the absence of mandatory disclosure of information, voluntary disclosure and associated discussions on various social media platforms are expected to play an important role in due diligence practices for ICOs (Boreiko and Risteski, 2021).

Table 4.2: Descriptive statistics

	Count	Mean	std	Min	Median	Max
Success	957.0	0.34	0.48	0.00	1.00	1.00
Whitepaper	957.0	0.98	0.15	1.00	1.00	1.00
Whitelist	957.0	0.28	0.45	0.00	1.00	1.00
KYC	957.0	0.43	0.50	0.00	1.00	1.00
ESG	957.0	0.18	0.16	0.07	0.27	1.00
Bounty	957.0	0.32	0.46	0.00	1.00	1.00
Bonus	957.0	0.15	0.36	0.00	0.00	1.00
Hard cap	957.0	0.66	0.47	0.00	1.00	1.00
Duration	957.0	55.44	73.63	21.00	66.00	1096.00
Number of milestones	957.0	8.10	4.72	5.00	11.00	29.00
Presale	957.0	0.38	0.49	0.00	1.00	1.00
MVP	957.0	0.23	0.42	0.00	0.00	1.00
Video	957.0	0.77	0.42	1.00	1.00	1.00
Twitter (log)	957.0	7.39	2.64	6.55	9.06	14.04
Telegram (log)	957.0	6.02	3.35	4.37	8.46	12.10
Facebook (log)	957.0	5.82	4.04	0.00	8.89	15.38
GitHub	957.0	0.60	0.49	0.00	1.00	1.00
Reddit	957.0	0.66	0.47	0.00	1.00	1.00
Slack	957.0	0.18	0.39	0.00	0.00	1.00
Discord	957.0	0.08	0.28	0.00	0.00	1.00
Bitcointalk	957.0	0.75	0.44	0.00	1.00	1.00
Team size	957.0	9.96	6.28	6.00	13.00	50.00
Number of advisors	957.0	4.20	4.24	0.00	7.00	22.00
Expert rating	957.0	1.69	1.23	0.00	3.00	3.00

[2]The ICO description is a short summary about the ICO, containing the most important aspects of the "business plan".

Table 4.1: Variable definitions

	Variable description	Source
Dependent variable		
Success	Binary variable indicating whether the soft cap has been reached: 1 if reached or exceeded, 0 otherwise	ICOmarks, TORD
ICO characteristics		
Whitepaper	Dummy variable indicating the presence of a whitepaper by the time of launch.	ICOmarks
Continent	Geographical location in which the ICO team is located in terms of the following regions: Asia, North America, Europe, and Other. Here, "Other" refers to the minority regions that contained only a few ICOs.	ICOmarks, TORD
Year	Year in which the ICO is launched	ICOmarks
Whitelist	Availability of the Whitelist	ICOmarks
KYC	Availability of the Know Your Customer procedure	ICOmarks
ESG	Environmental, Social, and Governance (ESG) scores, calculated based on an ICO's description using the algorithm provided by Mansouri and Momtaz (2022)	ICOmarks
Category	Industry sector of ICO	ICOmarks
Campaign characteristics		
Bounty	Dummy variable indicating the availability of a bounty program. In a bounty program people can participate in promotion and marketing activities in exchange for tokens.	ICOmarks
Bonus	Dummy variable indicating discounts during the ICO	ICOmarks
Duration	Duration of ICO campaign in number of days as specified in the whitepaper	ICOmarks
Number of milestones	Number of milestones specified in the whitepaper	ICOmarks
Presale	Dummy variable indicating whether the ICO launched a presale	ICOmarks
MVP	Dummy indicating the presence of a Minimum Viable Product	ICOmarks
Video	Dummy indicating the existence of a campaign video	ICOmarks
Social media		
Twitter	Size of twitter network measured in number of twitter followers	Twitter
Telegram	Size of Telegram network in numbers of followers	Telegram
Facebook	Size of Facebook network in numbers of followers	Facebook
GitHub	Dummy indicating the availability of a GitHub repository	GitHub
Reddit	Dummy indicating the presence of an ICO network on Reddit	ICOmarks
Slack	Dummy indicating the presence of an ICO network on Slack	ICOmarks
Discord	Dummy indicating the presence of an ICO network on Discord	ICOmarks
Bitcointalk	Dummy indicating the presence of an ICO network on Bitcointalk	ICOmarks
Social capital		
Team size	Number of team members measured by LinkedIn	ICOmarks
Number of advisors	Number of advisors measured by LinkedIn	ICOmarks
Expert Rating	Categorical variable indicating the expected quality of an ICO: Good (>75%) Medium (55%-75%) Bad (<55%) None (No rating available)	ICOmarks, TORD

4.2. Modeling approach

Complementing previous research on traditional signaling theory (Spence, 1978) by providing a dynamic overview and adopting the configuration based extension proposed by Drover et al. (2018), we developed a predictive model based on supervised machine learning algorithms to predict the funding success of ICO campaigns. Here, success is defined as a binary outcome, which takes the value of one if the predefined soft cap threshold has been reached. In detail, the entire approach can be divided into six phases: (1) Benchmark model (traditional approach), (2) Extending traditional approach with interaction terms, (3) Machine learning approach, (4) Performance evaluation, and (5) Interpretation and validation.

4.2.1. Benchmark model

To evaluate the respective models' performances in classifying an ICO, a benchmark model is developed based on the traditional signaling framework. In addition, the benchmark model is used as a supplement to answer the first research question. Specifically, the developed model is used to confirm which of the identified success determinants in Chapter 2.2 are associated with ICO success. Because the success determinants in Chapter 2.2 are drawn from multiple different studies each using different samples, and large swings in signal significance are anticipated (Bellavitis et al., 2021), this process is repeated for different subsamples. First, the pooled sample is subdivided into time windows (i.e., 2017, 2018, 2019) and a separate analysis is performed within each window. Second, ICOs are grouped based on categories and a separate analysis is performed within each of the following categories: Data and AI, Financial Services, and Entertainment and Gaming.

In this thesis, the logistic regression is used as a benchmark model because it is one of the most frequently employed statistical models to study the signaling environment. The main advantage of this model is that it can be used for both statistical inferences and classification and prediction of the binary dependent variable. One major drawback of logistic regression is that it assumes linearity between the predicted (dependent) variable and the predictor (independent) variables. Unless explicitly programmed in terms of interaction variables, the logistic regression can only be used to study signals in isolation and ignores the complementary, overlapping, and potentially competing effects that signals may have.

Logistic regression is a type of classification algorithm where the dependent variable is binary (or binomial). Logistic regression is essentially a logit transformation of the linear regression method. The resulting equation includes each variable's impact on the log-odds ratio of the observed event of interest. The ordinary logistic regression with binary response y_i is given by the probability P of the response success:

$$P\left(y_i = 1\right) = \frac{1}{1 + e^{-x_i\beta}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}} \tag{4.1}$$

where x_i is the i-th row of an matrix of n observations with p variables and a column of ones to accommodate the intercept, and β is the column vector of the regression coefficients β_p. The parameters estimates are obtained by maximizing the log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log\left(1 + e^{x\beta}\right)\right] \tag{4.2}$$

4.2.2. Signal interaction terms

To explore the potentially interactive effects that signals may have, the benchmark model is extended with interaction terms. The present thesis is particularly concerned with two interactions: the time and

context dependency of signals. Here, time is modeled in terms of years and context refers to the domain and industry sector a particular ICO belongs to. To examine the time dependency of signals, three models are developed: (i) interaction effects between signals and 2017, (ii) interaction effects between signals and 2018, and (iii) combination of (i) and (ii).

Similarly, to investigate the context dependency of signals the benchmark model is extended with interaction terms among two categories: Financial Services and Entertainment and Gaming. Again, three models are created: (i) interaction effects between signals and Financial Services, (ii) interaction effects between signals and Entertainment and Gaming, and (iii) combination of (i) and (ii).

4.2.3. Machine learning

The objective of the second part of this thesis is to test the suitability of a data-driven approach to study the ICO signaling environment, in general, and the interaction effects, in particular. For this purpose, four data-driven models have been adopted: Ridge Classifier, Random Forest Classifier, Extra Trees Classifier, and Support Vector Machine. An overview of the general advantages and disadvantages of using each method can be found in Table 4.6. All of the models are developed using the Python software package and scikit-learn library (Pedregosa et al., 2011). The models are discussed and specified below.

Ridge Classifier

Despite its suitability for making statistical inferences and explaining causal relationships between signals, logistic regression classification models may suffer from overfitting and multicollinearity. In case the data suffers from multicollinearity, using Logistic regression results in determinate regression coefficients with high standard errors. Ridge regression accounts for multicollinearity by adding a penalty term to the cost function of the logistic regression, also referred to as L2 regularization. This method is generally suited for high dimensional problems where the amount of features is high. The main reason for adopting this method in this thesis is because of its ability to reduce overfitting and deal with a relatively low amount of observations with respect to the amount of independent variables.

Starting from the cost function of the logistic regression (i.e., equation 4.2). The ridge regression estimator depends on the choice of a tuning parameter $\lambda \geq 0$, to be determined separately. The coefficients estimates are the values that optimize the following cost function:

$$l_{\lambda}^{R}(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log\left(1 + e^{x_i \beta}\right) \right] - \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.3}$$

As the shrinkage penalty λ increases, the ridge coefficient estimates will decrease to reduce overfitting. This thesis uses the optimal value of λ which was determined using a grid search[3] on the pooled sample and resulted in a value of $\lambda = 5.27$.

Random Forest

The Random Forest model (Breiman, 2001) is a classification algorithm that consists of multiple, uncorrelated Decision Trees (DTs). The uncorrelated forest of trees is created by means of bagging - i.e., the application of bootstrapping techniques to decision trees. Here, each individual tree in Random Forest is trained on a random subset of instances known as bags. The uncorrelated forest of trees is constructed by training multiple trees on different bags. The prediction performance of the Random Forest – by the wisdom of crowds [4] – is more accurate than that of any individual tree. Unlike the decision tree

[3] The grid search was performed with the aid of the Python package scikit-learn and the corresponding GridSearchCV function.

[4] According to the wisdom of crowds, a large number of relatively uncorrelated models (e.g., trees) operating as a whole will outperform any of the individual constituent models.

classifier, where the prediction is based on a single tree, Random Forest aggregates the prediction of multiple, uncorrelated trees and picks the most frequent result as an output to the instance. The main reason for adopting this algorithm is because of its effectiveness of dealing with variable interactions. The design of the algorithm allows for configurational dependent outcomes and implicitly deals with moderated variables by means of decision trees. A schematic example of a Random Forest model can be found in Figure 4.1. The full details of the Random Forest model and, in particular, the mathematical formulation are considered outside the scope of this thesis; the interested reader is referred to more technical descriptions such as that of Breiman (2001).

One important aspect related to the performance of the Random Forest Classifiers is the hyperparameter selection. In this thesis, a grid search was performed to determine the optimal hyperparameters (i.e., the number of DTs, the maximum depth of each tree, and the number of variables used per split) for the present analysis. The optimal hyperparameters can be found in Table 4.3. The procedure is described in detail in Appendix A.

Table 4.3: Random Forest hyperparameters (tuned).

Parameter	Values
Number of trees	100
Depth	Max
Maximum variables per split	$\sqrt{\text{Number of features}}$

Extra Trees

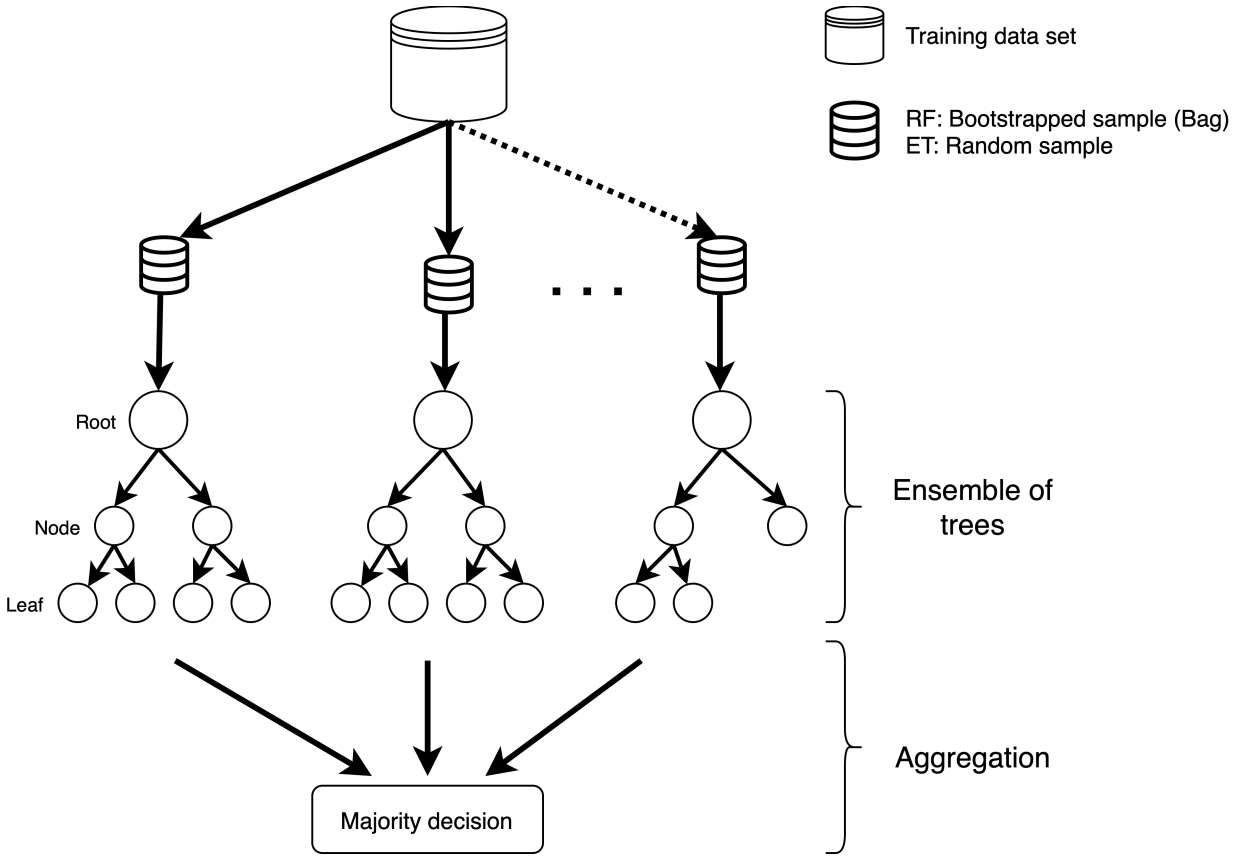
A similar decision tree algorithm is the extra tree classifier. Extra Trees – short for extremely randomized trees – was proposed by Geurts et al. (2006), with the primary objective of further randomizing tree building in the context of numerical input features, where the choice of the optimal cut-point is responsible for a large proportion of the variance of the induced tree. In contrast to Random Forests, the Extra Trees method does not employ bootstrapping techniques that optimize the cut-point for each randomly chosen feature at each node. Instead, Extra Trees selects a cut-point randomly. This method enhances performance in problems characterized by a large amount of numerical features varying more or less continuously. In the context of the ICO market, where multiple firms transmit multiple signals simultaneously, and the amount of noise variables may be high, this method may perform better than random forest. From a statistical perspective, dropping the bootstrapping idea reduces bias, whereas cut-point randomization often has an excellent variance reduction effect. This method has shown promising results in several high-dimensional complex problems, thus is expected to perform well in the multi-signal ICO environment. Next to its effectiveness of dealing with variable interactions, the main reason for adopting this algorithm is that it reduces overfitting with respect to Random Forest. The full details of the Extra Trees classifier and, in particular, the mathematical formulation are considered outside the scope of this thesis; the interested reader is referred to more technical descriptions such as that of Geurts et al. (2006).

Similarly as for the Random Forest Classifiers, a grid search was performed to determine the optimal hyperparameters (i.e., the number of DTs, the maximum depth of each tree, and the number of variables used per split) for the Extra Trees classifier. The resulting hyperparameters can be found in Table 4.4.

Table 4.4: Extra Trees hyperparameters (tuned).

Parameter	Values
Number of trees	100
Depth	Max
Maximum variables per split	$\sqrt{\text{Number of features}}$

Figure 4.1: Simplified example of the Random Forest (RF) and Extra Trees (ET) models.



Support Vector Machine

The SVM classifier derives boundaries between data points that belong to different classes. Points within certain boundaries are normally part of a common class. In ideal circumstances, the data points belonging to different classes are separable via a linear boundary. However, this is often not possible in real world data sets. SVM addresses this issue by casting the data points to a higher dimensional space in which the data becomes linearly separable with a hyperplane by using a specific kernel function. This technique is based on the Cover's Theorem stating that non-linearly separable data points would highly likely be separated by a hyperplane if projected to a higher-dimensional space via a non-linear transformation. The boundary hyperplane will be realized by referencing the borderline data points, which are called the support vectors. The identified support vectors should be away from the boundary by a given margin. One advantage of SVM models is that it is capable of dealing with non-linear relationships. SVM models work comparably well when there is an understandable margin of dissociation between classes but perform poorly otherwise. The main reason for adopting this algorithm is twofold: (i) it performs well when the data is separable, and (ii) it provides an indication about the level of separability of signals (i.e., how much signals overlap). The full details of the Support Vector Machine classifier and, in particular, the mathematical formulation are considered outside the scope of this thesis; the interested reader is referred to more detailed descriptions such as that of Hearst et al. (1998).

For the Support Vector Machine algorithm, the Python package scikit-learn and the corresponding GridSearchCV function were utilized as an automated hyper-parameter tuning approach. The results can be found in Table 4.5.

Table 4.5: Support Vector Machine hyperparameters (tuned).

Kernel	Python package	Alpha	Penalty	Maximum iterations
Linear	SGDClassifier (scikit-learn)	0.005	L2 regularization	1000

Table 4.6: Overveiw of general advantages and disadvantages of using each of the five algorithms.

	Logistic Regression	Random Forest	Extra Trees	Ridge Classifier	Support Vector Machine
Advantages	Captures well linear relationships	Capable of handling high-dimensional datasets	Capable of handling high-dimensional datasets	Captures well linear relationships	Effective in high-dimensional datasets
	Simple	Non-linear modeling	Non-linear modeling	Reduces overfitting	Reduces overfitting
	Directly interpretable	Captures well interaction terms	Captures well interaction terms	Works well with relatively low number of observations	Works well when data is separable
	Causal identification	Reduces outlier influence	Reduces outlier influence	Reduces complexity	
		Finds optimum ”split”	Reduces overfitting	Suppresses multicollinearity	
Disadvantages	Prone to overfitting	Not directly interpretable	Not directly interpretable	Shrinks coefficients to zero	Not directly interpretable
	Assumes linear relationships between response and predictor variables	Less accurate for highly linear relationships	Less accurate for highly linear relationships	Assumes linear relationships between response and predictor variables	Hyperparameter tuning is complicated
	Interactions need to be specified a priori	Numeric solution	Numeric solution		Numeric solution
	Assumes linear relationships between response and predictor variables		Chooses ”split” random		Not suitable for large datasets

4.2.4. Evaluation

The objective of this phase is to choose, among the selected ML models, the one that better accomplishes the underlying objective of distinguishing successful ICOs from unsuccessful ICOs. Here, classification performance is considered a good proxy of the ability of an evaluation method to capture the complexities underlying the ICO signaling process. The stratified cross-validation method is utilized to evaluate the performance of the ML models (Schaffer, 1993). The stratified cross-validation method splits the data set into k folds while ensuring that each fold has the same proportion of observations with a given categorical value (i.e., preserving the success ratio in each fold). Subsequently, one fold is separated out as a test sample and the remaining folds are used as training samples. In the present thesis, the evaluation is performed using 10 folds and the entire process is repeated ten times.

The process can be subdivided into four steps: (i) data sampling, (ii) splitting the data set into folds using the stratified cross-validation method, (iii) testing and training the models, and (iv) evaluating their performances. During the training phase, the ML models are trained using the training folds as input data. The instances in the test data set remain unseen and will be used only to evaluate the model's prediction performance. In doing so, each model will be evaluated in terms of three performance metrics: AUC-ROC, Accuracy, and Precision.

The AUC is the area under the Receiver Operating Curve (ROC). The ROC curve represents the relationship between the false-positive rate (FPR) and the true positive rate (TPR) for different probability thresholds of model predictions.[5] Here, an AUC of 1 corresponds to a perfect classifier, whereas an AUC of 0.5 means that the model cannot distinguish between the positive and negative classes. Given the goal of the present thesis - namely, to distinguish between successful and unsuccessful ICOs - this is considered the most important metric because it measures the ability of a model to distinguish between successful and unsuccessful ICOs.

In addition to the AUC-ROC, Accuracy and Precision are adopted. Accuracy is the most intuitive and straightforward performance metric and is the ratio of correctly predicted observations to the total number of predictions. In this case, Accuracy is a useful metric to measure the amount of correctly predicted instances because the target class is balanced.[6] Lastly, because the Accuracy metric can be misleading, it is complemented with the Precision metric. The Precision metric quantifies the number of positive class predictions that truly belong to the positive class.

4.2.5. Interpretation and validation

The analysis proceeds with the best performing model identified in the previous step. First, the feature importance scores of all signals are determined on a collective and individual basis. The feature importance score measures how much a single variable contributed to the performance of the ML model. Next, the impact of the most important signals on funding success is examined in more detail. For this purpose, the SHapley Additive exPlanations (SHAP) approach is adopted as a unified measure of signal impact on funding success. Unlike the feature importance scores, SHAP differentiates between the negative effects and positive effects of signals on ICO success. The interpretation of SHAP values is as follows: positive SHAP values indicate positive effects of a specific signal on funding success, and negative SHAP values resemble negative effects of a specific feature on funding success. Finally, the impact of interaction terms on the SHAP values is determined to assess whether the data-driven approach is a suitable alternative for studying signal interactions. The resulting values are compared with the results obtained in the second step (see Section 4.2.2).

[5] The FPR is also referred to as Recall or Sensitivity.
[6] Note: SMOTE sampling is used to ensure an equal distribution of the target class (i.e., success).

5

Results

5.1. Logistic regression: Signal interactions

To examine the success determinants identified in ICO literature, the analysis is conducted on the pooled sample of ICOs using the benchmark model. Additionally, the pooled sample is split by time and industry sector in terms of years (i.e., 2017-2019) and categories (i.e., Data and AI, Financial services, and entertainment and gaming) as an initial step to explore differences among them. The results of the logistic regressions can be found in Table 5.1 and Table 5.2 for the time samples and category samples, respectively. The correlation matrix (see Appendix, Table C.1) and Variance Inflation Factors (see Appendix, Table C.2) report no sign of multicollinearity. In light of the omitted-variable bias[1] (Mood, 2010), the variables that are insignificant are retained, even though there is no sign of collinearity.

The results for the pooled sample (i.e., Model 1) are largely consistent with previous studies but also offer some additional insights. Instead of modeling the effect of social media activity as a dummy variable denoting an ICO's official presence on Twitter, Facebook, and Telegram, social media activity was modeled by the size of their social networks (i.e., number of followers). It was found that the size of an ICO's Twitter and Telegram networks both have a positive impact on ICO success, as indicated by the significant positive coefficients (b=0.11, p<0.01) and (b=0.05, p<0.10), respectively. However, no significant relationship was found between the size of an ICO's Facebook network and the funding outcome. Furthermore, additional social media platforms (i.e., Reddit, Bitcointalk, Slack, and Discord) were included as dummy variables, but none of these signals were statistically significant in the pooled sample. In contrast to crowdfunding literature (Mollick, 2014), no significant relation was found between the availability of a campaign video and funding success. In previous studies, the duration of a project was found to be negatively associated with the likelihood of success (Ackermann et al., 2020; Roosenboom et al., 2020). Model (1) reports a significant and weak negative correlation between the duration of a campaign and funding success (b=-0.003, p<0.10). The availability of a bounty program is negatively correlated (b=-0.46, p<0.01) to the probability of funding success. A possible explanation is that a bounty program could signal a lack of competence to independently develop the product or service. The number of milestones was found to be negatively associated with success (b=-0.05, p<0.01), possibly because a higher number of milestones was found to reduce the informative content of the whitepaper (Florysiak and Schandlbauer, 2022). Furthermore, it was found that the hard

[1] As counter intuitive as it may sound, removing variables potentially related to the outcome, even if insignificant, is tricky in logistic regression, given its inherent omitted-variable bias. Removing variables related to outcome can lead to bias in the estimates of the coefficients of the retained variables, even if the retained variables are not correlated with the removed variable.

cap was negatively associated (b=-0.34, p<0.05) with funding outcomes, potentially because high hard cap limits can be perceived as unattainable (Lyandres et al., 2019). In accordance with Amsden and Schweizer (2018), the results indicate a negative correlation (b=-0.34, p<0.10) between a presale and the likelihood of funding success. Similarly, larger teams were found to have a higher probability of success as demonstrated by a positive significant coefficient (b=0.03, p<0.05), confirming previous research (Amsden and Schweizer, 2018; Ante et al., 2018; Roosenboom et al., 2020).

Table 5.1: The determinants of ICO success per time sample.

	Model (1): Full Sample		Model (2): 2017		Model (3): 2018		Model (4): 2019	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Whitepaper	0.40	(0.52)	-1.84**	(0.93)	1.06	(1.08)	1.48	(1.32)
Whitelist	0.19	(0.21)	-1.76	(1.88)	0.14	(0.25)	0.15	(0.72)
KYC	0.21	(0.21)	2.41*	(1.43)	0.23	(0.25)	0.62	(0.78)
ESG	-0.18	(0.50)	0.63	(1.42)	-0.81	(0.68)	2.88*	(1.65)
Bounty	-0.46**	(0.20)	-0.41	(1.02)	-0.59**	(0.24)	1.21	(0.76)
Bonus	0.17	(0.22)	1.54**	(0.67)	-0.20	(0.29)	0.77	(0.74)
Hard cap	-0.34**	(0.17)	0.25	(0.34)	-0.27	(0.22)	-2.18***	(0.78)
Duration	-0.003*	(0.001)	-0.003	(0.002)	-0.003**	(0.002)	-0.01	(0.00)
Number of milestones	-0.05***	(0.02)	-0.06	(0.04)	-0.07***	(0.02)	-0.10	(0.08)
Presale	-0.34*	(0.18)	-1.40	(0.92)	-0.20	(0.21)	-1.11*	(0.66)
MVP	-0.10	(0.22)	1.32	(1.24)	-0.13	(0.28)	1.17*	(0.67)
Video	-0.06	(0.19)	0.35	(0.42)	-0.02	(0.27)	-0.62	(0.72)
Twitter	0.11***	(0.03)	0.29***	(0.08)	0.05	(0.04)	0.25*	(0.15)
Telegram	0.05*	(0.03)	0.04	(0.05)	0.09**	(0.04)	-0.17	(0.11)
Facebook	-0.01	(0.02)	-0.03	(0.05)	0.02	(0.03)	-0.01	(0.08)
GitHub	-0.08	(0.16)	-0.43	(0.37)	-0.15	(0.22)	-0.04	(0.64)
Reddit	0.23	(0.18)	0.90**	(0.39)	0.23	(0.24)	0.49	(0.77)
Slack	0.16	(0.21)	-0.26	(0.37)	0.53*	(0.31)	0.38	(1.11)
Discord	-0.12	(0.29)	0.01	(1.63)	-0.61	(0.39)	0.18	(0.80)
Bitcointalk	0.10	(0.20)	-0.77	(0.47)	0.24	(0.28)	-0.18	(0.74)
Team size	0.03**	(0.01)	0.04	(0.03)	0.03*	(0.02)	-0.00	(0.05)
Number of advisors	-0.03	(0.02)	-0.09*	(0.05)	-0.02	(0.03)	-0.15*	(0.09)
Expert rating	0.34***	(0.07)	0.07	(0.15)	0.40***	(0.08)	0.50	(0.44)
Asia	-0.17	(0.76)	0.31	(1.09)	-0.55	(1.16)	0.05	(1.09)
Europe	-0.71	(0.76)	0.21	(1.09)	-1.15	(1.17)	-1.33	(1.11)
North America	-0.24	(0.77)	0.19	(1.07)	-0.85	(1.17)	0.00	(1.22)
Other continents	-0.62	(0.75)	-0.82	(1.05)	-1.12	(1.16)	-0.50	(1.14)
Data and AI	-0.20	(0.26)	-0.66	(0.73)	-0.15	(0.33)	0.41	(1.05)
Entertainment and Gaming	-0.42	(0.27)	-0.65	(0.50)	-0.21	(0.40)	-1.83	(1.47)
Financial Services	-0.50***	(0.17)	-1.04***	(0.40)	-0.40*	(0.24)	0.20	(0.74)
2017	-0.38	(0.52)						
2018	-0.16	(0.47)						
2019	-0.28	(0.49)						
N	957		228		573		156	
Log-Likelihood	-540.44		-121.00		-311.99		-47.71	
Pseudo R-Sq	0.1224		0.2066		0.1509		0.3505	
Wald χ	-615.84***		-152.51***		-367.43***		-73.46***	
AUC	66.0%		63.6%		67.7%		56.5%	
Accuracy	64.1%		58.5%		61.6%		60.6%	
Precision	48.4%		48.9%		45.8%		39.0%	

***,**, and * denote statistical significance of p<0.01, p<0.05, and p<0.10, respectively.

Next, a subsample analysis is performed. Starting with the time samples (i.e., 2017, 2018, and 2019), the first step is to explore whether the determinants of ICO success differ per time window. The results can be found in Models 2, 3, and 4 of Table 5.1. It was found that some determinants differ significantly from year to year. For example, the results suggest a significant positive (b=2.41, p<0.10) influence of imposing the Know Your Customer (KYC) program in 2017, which was found to be insignificant in 2018 and 2019. Likewise, higher ESG scores were positively associated (b=2.88, p<0.10) with success in 2019 but were not found to affect the funding outcomes in other time windows. Furthermore, a whitepaper reduced the chances of success (b=-1.84, p<0.05) in the 2017 sample, but no evidence was found in the 2018 and 2019 samples. Model 2 demonstrates that official presence on Reddit positively (b=0.90, p<0.05) affected funding success in 2017. Similarly, model 3 suggests that official presence on Slack positively (b=0.53, p<0.10) influenced funding success in 2018. Interestingly, some features that were found to be significant in the subsamples were not significant in the pooled sample (i.e., KYC, ESG, Bonus, MVP, Reddit, Slack, and Number of advisors). While the samples cannot be compared on a one-to-one basis, this may be a first indication of the time dependency of signals. Another potential explanation is that some effects cancel out when pooling them together and require a time-dependent modeling approach to be discovered.

An additional subsample analysis was performed to explore whether effective signals differ per industry sector. The results can be found in models 5, 6, and 7 of Table 5.2. Although caution must be taken given the sample size, success determinants differed per subsample. For ICOs in the Data and AI, Financial Services, and Entertainment and Gaming sector, the findings show that the significant determinants differ per category sample. More specifically, it was found that ICOs with higher Expert ratings and presence on Reddit are more likely to succeed in the Data and AI sector, as indicated by significant positive coefficients (b=0.86, p<0.01) and (b=1.81, p<0.05). A different set of significant determinants was identified in the Financial Services sample. Here, higher expert ratings (b=0.18, p<0.10), larger team sizes (b=0.04, p<0.05), and larger Twitter networks (b=0.11, p<0.05) are positively associated with the success of a campaign, whereas the availability of an MVP (b=-0.73, p<0.05) and a hard cap (b=-0.087, p<0.01) negatively impact funding outcomes. Regarding the Entertainment and Gaming sector, a significant negative correlation coefficient (b=-0.41, p<0.01) suggests that a higher number of specified milestones reduced funding performance within this sector. In line with previous studies, higher expert ratings (b=0.89, p<0.10) and the disclosure of a whitepaper (b=4.93, p<0.05) lead to a higher probability of funding success. Interestingly, the predictions of the logistic regression were found to be insignificant in this subsample, as demonstrated by the insignificant Wald χ square test. In other words, the logistic model could not produce statistically meaningful predictions from the current set of signals in the Entertainment and Gaming sector. A possible explanation is that the logistic model suffers from a relatively high number of noise variables compared to the low number of instances. This might also explain why the geographical locations (i.e., Europe, North America, and minority regions) were found to be negatively correlated with ICO success, which differs from the findings presented by Fisch (2019) and Huang et al. (2020).

Table 5.2: The determinants of ICO success per category sample.

	Model (1): Full Sample		Model (5): Data and AI		Model (6): Financial Serv.		Model (7): Enter. and gaming	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
2017	-0.38	(0.52)	-1.43	(2.35)	-0.48	(0.78)	-2.51	(1.90)
2018	-0.16	(0.47)	-0.80	(2.02)	-0.51	(0.72)	0.05	(1.76)
2019	-0.28	(0.49)	-0.16	(2.15)	-0.34	(0.75)	-3.73	(2.65)
Whitepaper	0.40	(0.52)	-2.13	(2.25)	1.00	(1.17)	4.93**	(2.29)
Whitelist	0.19	(0.21)	0.22	(0.77)	0.26	(0.30)	-0.52	(1.23)
KYC	0.21	(0.21)	-0.15	(0.89)	0.51	(0.32)	-0.14	(1.09)
ESG	-0.18	(0.50)	-0.67	(1.93)	-0.10	(0.75)	-2.88	(3.84)
Bounty	-0.46**	(0.20)	-0.66	(0.80)	-0.16	(0.29)	-0.44	(1.32)
Bonus	0.17	(0.22)	1.13	(0.78)	-0.41	(0.35)	-0.37	(1.46)
Hard cap	-0.34**	(0.17)	-0.25	(0.79)	-0.87***	(0.26)	1.76**	(0.90)
Duration	-0.00*	(0.00)	-0.00	(0.00)	-0.00	(0.00)	-0.01	(0.01)
Number of milestones	-0.05***	(0.02)	-0.08	(0.07)	-0.06*	(0.03)	-0.41***	(0.15)
Presale	-0.34*	(0.18)	-0.20	(0.75)	-0.02	(0.28)	-1.51	(1.04)
MVP	-0.10	(0.22)	0.47	(0.78)	-0.73**	(0.32)	0.94	(1.45)
Video	-0.06	(0.19)	0.46	(0.92)	-0.07	(0.30)	0.05	(0.99)
Twitter	0.11***	(0.03)	0.25	(0.19)	0.11**	(0.05)	0.11	(0.15)
Telegram	0.05*	(0.03)	-0.07	(0.11)	0.04	(0.04)	0.04	(0.13)
Facebook	-0.01	(0.02)	-0.02	(0.09)	-0.02	(0.04)	0.10	(0.12)
GitHub	-0.08	(0.16)	1.00	(0.71)	-0.01	(0.26)	0.80	(0.90)
Reddit	0.23	(0.18)	1.81**	(0.85)	0.20	(0.28)	-1.30	(0.97)
Slack	0.16	(0.21)	0.49	(0.91)	-0.07	(0.33)	-0.66	(0.91)
Discord	-0.12	(0.29)	0.32	(1.34)	-0.27	(0.42)	1.79	(1.72)
Bitcointalk	0.10	(0.20)	0.28	(0.91)	-0.18	(0.32)	1.93*	(1.14)
Team size	0.03**	(0.01)	-0.01	(0.06)	0.04**	(0.02)	0.08	(0.09)
Number of advisors	-0.03	(0.02)	-0.07	(0.08)	-0.02	(0.03)	-0.13	(0.12)
Expert rating	0.34***	(0.07)	0.86***	(0.30)	0.18*	(0.10)	0.89**	(0.38)
Asia	-0.17	(0.76)	1.38	(3.42)	-0.94	(1.39)	-0.37	(1.65)
Europe	-0.71	(0.76)	0.14	(3.34)	-1.37	(1.39)	-4.31***	(1.57)
North America	-0.24	(0.77)	2.22	(3.45)	-1.01	(1.40)	-3.52**	(1.61)
Other continents	-0.62	(0.75)	1.79	(3.26)	-1.47	(1.35)	-3.44**	(1.61)
Data and AI	-0.20	(0.26)						
Entertainment and Gaming	-0.42	(0.27)						
Financial Services	-0.50***	(0.17)						
N	957		282		463		212	
Log-Likelihood	-540.44		-44.034		-243.35		-34.098	
Pseudo R-Sq	0.1224		0.3854		0.1292		0.4200	
Wald χ	-615.84***		-71.648***		-279.45***		-58.788	
AUC	66.0%		63.9%		70.9%		67.3%	
Accuracy	64.1%		63.6%		66.7%		63.8%	
Precision	48.4%		47.8%		57.1%		48.1%	

***,**, and * denote statistical significance of p<0.01, p<0.05, and p<0.10, respectively.

To further investigate the time and context dependency of signals, the benchmark model is extended with signal interaction terms concerning time (i.e., 2017 and 2018) and categories (i.e., Financial Services and Entertainment and Gaming). Table 5.3 shows the results of our time interaction analysis. Model 1 represents the baseline model, excluding interaction effects. Model 8 presents estimates for the signals interacting with the time window: the product between 2017 and the respective signal. Model 9 tests these interactions for the year 2018: the product between 2018 and the respective signal. Model 10 includes both sets of interaction effects simultaneously. Table 5.3 reports only the interaction effects found to be significant in 2017 and 2018. A full overview can be found in the Appendix.

Model 10 provides evidence supporting our expectations, demonstrating that signals are time-dependent, as indicated by the moderating effects of 2017 and 2018. Here, it was found that the year 2017 negatively moderates the relationship between the whitepaper (b=-3.32, p<0.01) and funding success and positively moderates the effect of bonuses (b=1.30, p<0.10), hard cap (b=1.62, p<0.01), Twitter (b=0.11, p<0.05), and expert rating (b=0.36, p<0.10) on funding success. In contrast, the year 2018 positively moderates (b=0.20, p<0.01) the effect between the size of the Telegram network and funding success. These results further support the idea that some signals are time-dependent. It also explains why some signals (e.g., bonus and whitepaper) were found to be significant in our subsample analysis while they were found to be insignificant in the pooled sample. Interestingly, the size of an ICO's Twitter network was found to be significant in 2017 but not in 2018. The impact of the Telegram network showed an opposite trend: significant under the moderation of 2018, insignificant under the moderation of 2017.

Regarding the context dependency of signals, we investigated the interaction effects between signals and the industry sector (i.e., Financial services and Entertainment and Gaming). The results can be found in Table 5.4. Again, Model 1 presents the baseline model, excluding interaction effects. Model 11 presents estimates for the signals interacting with the Financial Services sector: the product between the Financial Services dummy and the respective signal. Model 12 examines the interactions with the Entertainment and Gaming sector: the product between Entertainment and Gaming dummy and the respective signals. Model 13 includes both sets of interaction effects simultaneously. Similarly as before, Table 5.4 reports only the interaction effects that were found to be significant. A full overview is provided in the Appendix.

Model 13 provides support for the context dependency of signals as it reveals a different set of significant interaction terms within the Financial services sector (i.e., KYC, bonus, hard cap, MVP, and expert rating) as opposed to the Entertainment and Gaming sector (i.e., hard cap). The results posit that within the Financial Services sector, the KYC program (b=0.87, p<0.05) and Expert rating (b=0.36, p<0.05) signals have a positive and significant impact on funding success. In contrast, the signals bonus (b=-0.92, p<0.05), hard cap target (b=-0.56, p<0.10), and MVP (b=-1.00, p<0.05) significantly negatively influence funding success.

Table 5.3: Interaction effects: time dependency of signals

	Model (1)		Model (8):		Model (9)		Model (10)	
	No interaction		Signals x 2017		Signals x 2018		(2) and (3)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Whitepaper	0.40	(0.52)	0.99*	(0.51)	-0.29	(0.49)	1.77**	(0.79)
2017	-0.38	(0.52)	-0.44	(0.91)	-0.49	(0.44)	-0.46	(0.91)
2018	-0.16	(0.47)	-0.63	(0.42)	-0.90	(1.11)	-0.83	(1.13)
2019	-0.28	(0.49)	-0.67	(0.44)	-0.41	(0.43)	-0.43	(0.52)
Whitelist	0.19	(0.21)	0.17	(0.21)	0.17	(0.21)	0.16	(0.21)
KYC	0.21	(0.21)	0.23	(0.21)	0.19	(0.21)	0.21	(0.21)
ESG	-0.18	(0.50)	-0.31	(0.51)	-0.28	(0.50)	-0.29	(0.52)
Bounty	-0.46**	(0.20)	-0.40*	(0.20)	-0.44**	(0.20)	-0.39*	(0.20)
Bonus	0.17	(0.22)	-0.09	(0.25)	0.61*	(0.35)	0.26	(0.54)
Hard cap	-0.34**	(0.17)	-0.51***	(0.20)	-0.41	(0.25)	-1.50***	(0.48)
Duration	-0.00*	(0.00)	-0.00*	(0.00)	-0.00*	(0.00)	-0.00*	(0.00)
Number of milestones	-0.05***	(0.02)	-0.06***	(0.02)	-0.06***	(0.02)	-0.06***	(0.02)
Presale	-0.34*	(0.18)	-0.30*	(0.18)	-0.36**	(0.18)	-0.29	(0.18)
MVP	-0.10	(0.22)	-0.05	(0.22)	-0.12	(0.22)	-0.04	(0.22)
Video	-0.06	(0.19)	-0.04	(0.19)	-0.05	(0.19)	-0.07	(0.20)
Twitter	0.11***	(0.03)	0.08**	(0.04)	0.16***	(0.05)	0.17**	(0.09)
Telegram	0.05*	(0.03)	0.04	(0.03)	-0.00	(0.04)	-0.12*	(0.07)
Facebook	-0.01	(0.02)	-0.00	(0.02)	-0.00	(0.02)	-0.00	(0.02)
GitHub	-0.08	(0.16)	-0.07	(0.17)	-0.08	(0.16)	-0.09	(0.17)
Reddit	0.23	(0.18)	0.24	(0.18)	0.23	(0.18)	0.26	(0.18)
Slack	0.16	(0.21)	0.20	(0.21)	0.14	(0.21)	0.25	(0.21)
Discord	-0.12	(0.29)	-0.22	(0.29)	-0.20	(0.29)	-0.28	(0.30)
Bitcointalk	0.10	(0.20)	0.09	(0.21)	0.10	(0.20)	0.09	(0.21)
Team size	0.03**	(0.01)	0.03***	(0.01)	0.03**	(0.01)	0.03**	(0.01)
Number of advisors	-0.03	(0.02)	-0.03	(0.02)	-0.03	(0.02)	-0.03*	(0.02)
Expert rating	0.34***	(0.07)	0.42***	(0.08)	0.23**	(0.11)	0.43	(0.29)
Asia	-0.17	(0.76)	0.31	(0.31)	0.41	(0.31)	0.34	(0.32)
Europe	-0.71	(0.76)	-0.25	(0.29)	-0.17	(0.29)	-0.22	(0.30)
North America	-0.24	(0.77)	0.19	(0.32)	0.28	(0.31)	0.17	(0.33)
Other continents	-0.62	(0.75)	-0.20	(0.33)	-0.10	(0.32)	-0.23	(0.34)
Data and AI	-0.20	(0.26)	-0.20	(0.27)	-0.18	(0.26)	-0.18	(0.27)
Entert. and gaming	-0.42	(0.27)	-0.45	(0.27)	-0.47*	(0.27)	-0.38	(0.28)
Financial Services	-0.50***	(0.17)	-0.53***	(0.17)	-0.50***	(0.17)	-0.49***	(0.18)
2017 x Whitepaper			-2.53***	(0.98)			-3.32***	(1.15)
2017 x Bonus			1.63***	(0.63)			1.30*	(0.79)
2017 x Hard cap			0.64*	(0.36)			1.62***	(0.57)
2017 x Twitter			0.15**	(0.07)			0.11**	(0.05)
2017 x Telegram			0.00	(0.05)			0.16	(0.08)
2017 x Expert rating			0.36**	(0.16)			0.36*	(0.32)
2018 x Bonus					-0.74	(0.45)	-0.42	(0.60)
2018 x Expert rating					0.19	(0.14)	0.01	(0.30)
2018 x Hard cap					0.11	(0.33)	1.17	(0.52)
2018 x Telegram					0.09*	(0.05)	0.20***	(0.08)
2018 x Twitter					-0.10	(0.06)	-0.12	(0.10)
2018 x Whitepaper					1.21	(1.12)	-0.82	(1.28)
N	957		957		957		957	
Log-Likelihood	-540.44		-529.45		-536.32		-501.11	
Pseudo R-Sq	0.1224		0.1403		0.1291		0.1863	
Wald χ	-615.84***		-615.84***		-615.84***		-615.84***	
AUC	66.0%		66.8%		66.0%		67.8%	
Accuracy	64.1%		65.2%		64.3%		65.4%	
Precision	48.4%		59.6%		58.3%		60.0%	

***,**, and * denote statistical significance of p<0.01, p<0.05, and p<0.10, respectively.

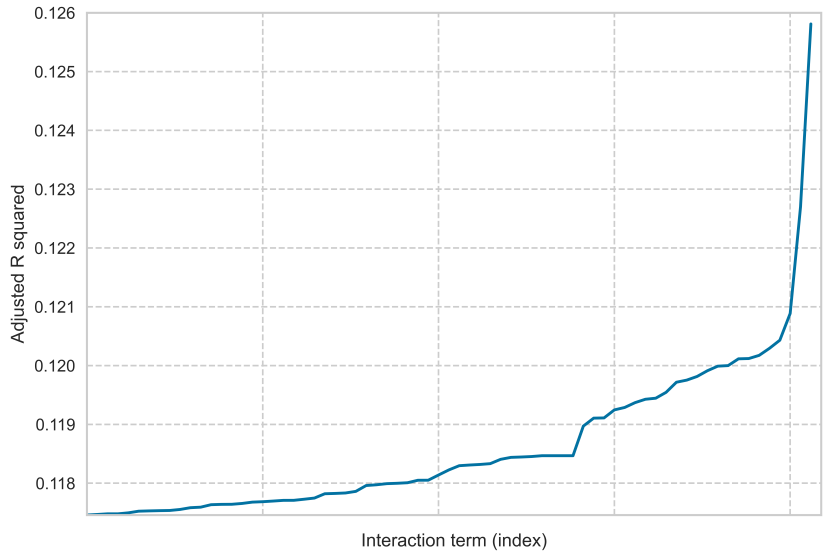
Table 5.4: Interaction effects of signals on Financial Services and Entertainment and Gaming categories.

	Model 1		Model 11		Model 12		Model 13	
	No interaction		x Fin Services		x Entert Gaming		(2) and (3)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Whitepaper	0.40	(0.52)	0.51	(0.54)	0.44	(0.53)	0.52	(0.55)
2017	-0.38	(0.52)	-0.43	(0.53)	-0.35	(0.52)	-0.41	(0.54)
2018	-0.16	(0.47)	-0.26	(0.48)	-0.13	(0.48)	-0.25	(0.49)
2019	-0.28	(0.49)	-0.28	(0.50)	-0.26	(0.50)	-0.28	(0.51)
Whitelist	0.19	(0.21)	0.20	(0.21)	0.22	(0.21)	0.22	(0.21)
KYC	0.21	(0.21)	-0.24	(0.26)	0.21	(0.21)	-0.25	(0.28)
ESG	-0.18	(0.50)	-0.08	(0.51)	-0.12	(0.50)	-0.04	(0.51)
Bounty	-0.46**	(0.20)	-0.43**	(0.20)	-0.42**	(0.20)	-0.40*	(0.20)
Bonus	0.17	(0.22)	0.49	(0.30)	0.17	(0.23)	0.57*	(0.32)
Hard cap	-0.34**	(0.17)	-0.02	(0.23)	-0.46***	(0.17)	-0.21	(0.25)
Duration	-0.00*	(0.00)	-0.00*	(0.00)	-0.00*	(0.00)	-0.00*	(0.00)
Number of milestones	-0.05***	(0.02)	-0.05***	(0.02)	-0.05***	(0.02)	-0.05***	(0.02)
Presale	-0.34*	(0.18)	-0.34*	(0.18)	-0.33*	(0.18)	-0.35*	(0.18)
MVP	-0.10	(0.22)	0.40	(0.30)	-0.14	(0.23)	0.43	(0.33)
Video	-0.06	(0.19)	-0.05	(0.19)	-0.08	(0.19)	-0.05	(0.19)
Twitter	0.11***	(0.03)	0.11***	(0.03)	0.11***	(0.03)	0.10***	(0.03)
Telegram	0.05*	(0.03)	0.05*	(0.03)	0.05*	(0.03)	0.04*	(0.03)
Facebook	-0.01	(0.02)	-0.01	(0.02)	-0.01	(0.02)	-0.01	(0.02)
GitHub	-0.08	(0.16)	-0.10	(0.17)	-0.07	(0.16)	-0.09	(0.17)
Reddit	0.23	(0.18)	0.25	(0.18)	0.25	(0.18)	0.27	(0.18)
Slack	0.16	(0.21)	0.05	(0.21)	0.13	(0.21)	0.04	(0.22)
Discord	-0.12	(0.29)	-0.10	(0.29)	-0.17	(0.29)	-0.10	(0.29)
Bitcointalk	0.10	(0.20)	0.07	(0.21)	0.08	(0.21)	0.07	(0.21)
Team size	0.03**	(0.01)	0.03**	(0.01)	0.03**	(0.01)	0.03**	(0.01)
Number of advisors	-0.03	(0.02)	-0.03	(0.02)	-0.02	(0.02)	-0.03	(0.02)
Expert rating	0.34***	(0.07)	0.53***	(0.09)	0.32***	(0.07)	0.52***	(0.10)
Asia	-0.17	(0.76)	-0.08	(0.79)	-0.18	(0.77)	0.00	(0.79)
Europe	-0.71	(0.76)	-0.66	(0.78)	-0.70	(0.77)	-0.56	(0.79)
North America	-0.24	(0.77)	-0.13	(0.79)	-0.25	(0.77)	-0.05	(0.80)
Other continents	-0.62	(0.75)	-0.49	(0.78)	-0.62	(0.76)	-0.40	(0.78)
UK	-0.68	(0.81)	-0.56	(0.83)	-0.63	(0.81)	-0.44	(0.84)
Data and AI	-0.20	(0.26)	-0.20	(0.27)	-0.19	(0.26)	-0.20	(0.27)
Entertainment and Gaming	-0.42	(0.27)	-0.48*	(0.28)	-0.82	(0.52)	-1.00*	(0.54)
Financial Services	-0.50***	(0.17)	-0.67**	(0.31)	-0.49***	(0.17)	-0.76**	(0.32)
Fin Services x KYC			0.87***	(0.34)			0.87**	(0.35)
Fin Services x Bonus			-0.84*	(0.45)			-0.92**	(0.46)
Fin Services x Hard cap			-0.74**	(0.33)			-0.56*	(0.35)
Fin Services x MVP			-0.97**	(0.42)			-1.00**	(0.44)
Fin Services x Expert rating			0.36***	(0.13)			0.36**	(0.14)
Entert Gaming x MVP					0.50	(0.81)	-0.10	(0.85)
Entert Gaming x KYC					-0.45	(0.61)	0.03	(0.64)
Entert Gaming x Bonus					-0.01	(0.87)	-0.42	(0.90)
Entert Gaming x Expert rating					0.24	(0.22)	0.05	(0.23)
Entert Gaming x Hard cap					1.28**	(0.57)	1.05*	(0.60)
N	957		957		957		957	
Log-Likelihood	-540.44		-526.92		-537.14		-525.27	
Pseudo R-Sq	0.1224		0.1444		0.1278		0.1471	
Wald χ	-615.84***		-615.84***		-615.84***		-615.84***	
AUC	66.0%		66.7%		66.1%		66.9%	
Accuracy	64.1%		64.8%		64.2%		64.9%	
Precision	48.4%		56.6%		52.3%		57.3%	

***,**, and * denote statistical significance of p<0.01, p<0.05, and p<0.10, respectively.

Lastly, a parsimonious evaluation tool is developed to explore the impact of interaction terms beyond the effects of time and context. Starting from the baseline model that includes all the initial determinants (i.e., Model 1), the model adds only a single interaction term to the baseline model, fits an Ordinary Least Squares (OLS) model, and records the adjusted R-squared of the model. This process is repeated for every possible first-order interaction term until all terms are evaluated. The adjusted R-squared measure is a useful indicator because it increases only if the newly added interaction term improves the OLS model. Next, the interaction terms that were found to improve the adjusted R-squared most were plotted against the corresponding adjusted R-squared values. The resulting plot is shown in Figure 5.1. Here, the x-axis represents the interaction terms, and the y-axis denotes the Adjusted R-squared for adding a single interaction term. The x-axis is intentionally left blank for presentation purposes because listing the corresponding interaction terms made the plot unreadable. The selected set of interaction terms can be found in Appendix B.

Figure 5.1: The individual effect of the interaction terms on the adjusted R-squared.



The graph shows that the some interaction terms improve the adjusted R-squared of the baseline model by nearly 10% (i.e., from 0.117 to 0.125). Even though these results indicate that some interaction terms are expected to improve the goodness-of-fit significantly, investigating and evaluating the effect of every single interaction term in detail is considered outside the scope of the present thesis.

Overall, using subsample analysis and interaction terms it was identified that success determinants (i.e., effective signals) are context and time dependent. Additionally, for smaller sample sizes (e.g., subsample Entertainment and Gaming) in which the amount of noise variables was relatively high in comparison to the amount of observations, the benchmark model was not capable of making statistically significant and meaningful predictions as indicated by an insignificant Wald χ squared score. Besides, studying signal interactions with the traditional method was a time-consuming process. Taken all of this together, this highlights the need for more efficient and robust evaluation methods to study the complexities inherent to the ICO signaling process. The remainder of this thesis moves on by evaluating the suitability of a data driven approach to study signal interactions in the high-noise ICO environment.

5.2. Machine Learning approach

Following the steps in Chapter 4, the first objective is to compare the performance and stability of the selected machine learning models to test their suitability for analyzing the high-dimensional, heterogeneous, and noisy signaling environment by predicting the success of ICOs. The experiments are conducted based on the same samples used to evaluate the potential success determinants in the previous section. The baseline model (i.e., Model 1) is included here as a benchmark against which to compare the other models. The analysis deviates from the traditional forecasting methodology, where only the subset of explanatory variables with significant explanatory power is used to develop the models for three reasons. Firstly, it is assumed that more advanced modeling techniques may find nonintuitive relationships and patterns. Secondly, it was shown that effective signals are context-dependent, meaning they interact with other signals to become effective. Thirdly, in the real ICO setting, it may be hard to conduct feature selection a priori and thus it is important to find methods that are capable of dealing with high-noise environments.

The results of the experiments are given in Table 5.5, in terms of the evaluation metrics that show the mean value of the prediction performance across 10 repeats of 10-fold cross-validation for each combination of machine learning algorithm (rows) and evaluation metric (columns). Here, relative performance is defined as the performance difference of the given model with respect to the benchmark model (i.e., logistic regression).

The results show that the Random Forest model outperformed the other models in each of the (sub)samples evaluated. In the pooled sample, both the Random Forest and Extra Trees classifiers performed well and outperformed the benchmark model. As expected, Random forest outperforms the Extra Tree model, possibly because it chooses the optimum split while Extra Trees chooses it randomly. In most cases SVM shows similar or worse performance than the logistic regression, except for the sub-sample 2017. One potential reason why SVM does not perform well relates to the similarity between the signals, meaning that the features show similar or overlapping properties and behavior. As a result of this, signals are hard to separate, even by higher order mechanisms like SVMs.

Even though ridge regression generally performs well when the amount of features is high, the penalized logistic regression model (i.e., ridge) did not improve prediction performance with respect to the benchmark model, as demonstrated by a decrease in AUC (Δ=-0.7%), and nearly similar Accuracy (Δ=+0.3%) and Precision (Δ=+0.2%). One potential explanation relates to the high-noise signaling environment. When the variance of a given set of signals (i.e., the variance in coefficient estimates) is high, ridge regression explicitly restricts the model from overfitting by increasing the penalty term. Consequently, the resulting model may become too simple to explain the complexity inherent to the signaling process. In terms of the high-noise ICO environment, many ventures send multiple signals simultaneously and effective signals are context dependent. Therefore, the variance of the effect of a certain signal may be too high, resulting in an oversimplified model.

Table 5.5 also shows which sample has the best prediction scores and which is most difficult to predict. Out of all the samples investigated, the best prediction performance was obtained for the year 2019. Interestingly, the predictability of ICO success improved between 2017 and 2019, with prediction accuracies of 69.8% and 82.5%, respectively. This may imply that investors learn over time, making more rational and consistent decisions and thus become more predictable. Another possible explanation might be that the ICO market matures, constituting a more stable signaling environment and reducing noise. In the same period, the relative performance of the Random Forest model also increased from

ΔAUC=+9.7% to ΔAUC=+17.0% and from ΔAcc.=+11.3% to ΔAcc.=+21.9% in 2017 and 2019, respectively. Regarding the categories, the best prediction performance is obtained for the Random Forest model in the Entertainment and Gaming industry (i.e., AUC = 69.4%, Accuracy = 72.4%, and Precision = 64%). Taken together, these results suggest that the Random Forest better captures the underlying signaling dynamics, potentially because the nature of the algorithm effectively handles feature interactions.

Table 5.5: Performance evaluation Machine Learning models

		Overall Performance			Relative Performance		
Sample	Model	AUC	Acc.	Prec.	Δ AUC	Δ Acc.	Δ Prec.
Full Sample	Random Forest Classifier	70.2%	70.4%	60.5%	4.2%	6.3%	12.1%
	Extra Trees Classifier	68.0%	69.1%	57.4%	2.0%	4.9%	9.0%
	SVM - Linear Kernel	66.8%	63.8%	60.4%	0.8%	-0.3%	12.1%
	Logistic Regression	66.0%	64.1%	48.4%			
	Ridge Classifier	65.3%	64.4%	48.6%	-0.7%	0.3%	0.2%
2017	Random Forest Classifier	73.4%	69.8%	65.2%	9.7%	11.3%	16.3%
	SVM - Linear Kernel	66.7%	56.7%	53.3%	3.1%	-1.9%	4.4%
	Logistic Regression	63.6%	58.5%	48.9%			
	Extra Trees Classifier	62.2%	64.2%	53.7%	-1.4%	5.6%	4.8%
	Ridge Classifier	62.0%	56.7%	46.0%	-1.6%	-1.9%	-3.0%
2018	Random Forest Classifier	68.9%	70.8%	60.1%	1.2%	9.2%	14.2%
	Ridge Classifier	68.2%	62.1%	46.6%	0.5%	0.5%	0.7%
	Logistic Regression	67.7%	61.6%	45.8%			
	Extra Trees Classifier	66.8%	69.1%	58.5%	-0.9%	7.5%	12.7%
	SVM - Linear Kernel	63.3%	61.1%	51.2%	-4.4%	-0.5%	5.3%
2019	Random Forest Classifier	73.5%	82.5%	68.3%	17.0%	21.9%	29.3%
	Extra Trees Classifier	64.4%	72.1%	35.0%	7.9%	11.5%	-4.0%
	Logistic Regression	56.5%	60.6%	39.0%			
	SVM - Linear Kernel	52.8%	70.8%	54.5%	-4.4%	10.3%	15.5%
	Ridge Classifier	52.1%	57.1%	25.0%	-4.4%	-3.5%	-14.0%
Data and AI	Random Forest Classifier	68.3%	70.8%	61.6%	3.3%	7.2%	13.8%
	Logistic Regression	65.0%	63.6%	47.8%			
	Extra Trees Classifier	63.9%	66.9%	52.1%	-1.1%	3.4%	4.3%
	SVM - Linear Kernel	62.2%	55.9%	58.0%	-2.8%	-7.7%	10.2%
	Ridge Classifier	61.2%	62.2%	46.5%	-3.8%	-1.4%	-1.3%
Financial Services	Random Forest Classifier	73.6%	71.3%	67.1%	2.7%	4.7%	10.0%
	Extra Trees Classifier	72.5%	69.3%	65.4%	1.6%	2.6%	8.3%
	Logistic Regression	70.9%	66.7%	57.1%			
	SVM - Linear Kernel	67.6%	64.0%	59.0%	-3.3%	-2.6%	1.9%
	Ridge Classifier	67.3%	63.2%	53.1%	-3.6%	-3.5%	-4.0%
Entertainment and Gaming	Random Forest Classifier	69.4%	72.4%	64.0%	2.1%	8.6%	15.9%
	Logistic Regression	67.3%	63.8%	48.1%			
	Ridge Classifier	66.7%	63.8%	47.9%	-0.6%	0.0%	-0.2%
	Extra Trees Classifier	66.1%	68.6%	56.5%	-1.2%	4.8%	8.4%
	SVM - Linear Kernel	65.2%	63.0%	51.1%	-2.1%	-0.8%	3.0%

Relative performance is measured with respect to the baseline model: Logistic Regression (i.e., Model 1).

Another important aspect is the consistency of an evaluation method. Following 10 repeats of 10-fold cross-validations the stability of the methods was evaluated within each sample. The results have been summarized in Table 5.6. Again, Random Forest and Extra Tree classifiers perform best showing structurally lower standard deviations in comparison to the other methods. In contrast, logistic regression (baseline) shows huge variation in prediction performance, meaning that it performs extremely well in some folds, and extremely poor in other folds. The data reported here appear to support the assumption that the ICO signaling process is context dependent and can not be captured by a single dominant net-effects evaluation that excludes the effect of signal interactions. Furthermore, when interested in a specific subsample only, Random Forest is the most reliable analysis method.

Table 5.6: Standard deviation of prediction accuracy per sample in %.

	Full Sample	2017	2018	2019	Data and AI	Financial Services	Entertainment and Gaming
Logistic regression	4.57%	16.06%	7.50%	10.00%	20.08%	6.98%	18.41%
Random Forest	2.93%	5.41%	4.70%	4.08%	5.77%	5.13%	5.24%
Extra Trees	3.09%	8.41%	5.03%	10.58%	8.23%	5.15%	10.87%
Ridge	4.68%	13.46%	8.46%	13.78%	17.91%	5.90%	19.99%
SVM	9.36%	9.68%	5.40%	19.66%	21.28%	16.58%	19.87%

Evaluation and interpretation

The remainder of the analysis is performed using the Random Forest model. First, the feature importance scores are computed. The feature importance scores are the contribution of each variable to the correct classification of successful ICOs. Table 5.7 summarizes the importance of each feature in identifying successful ICOs for each of the seven samples. In addition, the factors have been grouped by similarity to examine their collective impact.

Table 5.7: Feature Importance per sample

	Sample (1): Full sample	Sample (2): 2017	Sample (3): 2018	Sample (4): 2019	Sample (5): Data and AI	Sample (6): Fin. Serv.	Sample (7): Ent. Gaming
ICO Characteristics	28.0%	21.5%	24.6%	24.0%	24.1%	21.4%	20.7%
Whitepaper	0.1%	0.5%	0.1%	0.2%	0.2%	0.4%	0.2%
Continent	8.0%	8.2%	7.7%	13.0%	8.9%	8.8%	7.6%
Whitelist	1.9%	0.1%	2.4%	1.4%	2.7%	1.7%	1.6%
KYC	2.1%	0.1%	2.1%	1.3%	2.3%	1.6%	2.0%
ESG	4.5%	6.0%	4.7%	3.5%	5.4%	5.4%	4.9%
Category	7.4%	6.6%	7.7%	4.5%			
Year	3.9%				4.6%	3.6%	4.3%
Campaign characteristics	32.3%	25.6%	31.5%	36.8%	30.2%	30.5%	33.1%
Bounty	1.8%	0.7%	4.1%	1.5%	3.3%	2.0%	3.7%
Bonus	1.1%	4.0%	1.2%	1.1%	1.1%	1.6%	1.1%
Hard cap	6.0%	2.2%	4.2%	6.1%	3.9%	2.7%	4.2%
Duration	12.1%	9.2%	9.6%	13.7%	10.1%	12.3%	11.9%
Number of milestones	5.0%	6.3%	7.3%	8.0%	5.6%	6.6%	6.4%
Presale	2.5%	0.2%	1.6%	3.3%	1.8%	1.9%	1.9%
MVP	1.9%	0.6%	2.2%	1.4%	2.4%	1.5%	2.3%
Video	1.9%	2.3%	1.4%	1.7%	2.0%	1.9%	1.7%
Social media	25.8%	35.3%	27.7%	27.7%	27.5%	29.6%	28.8%
Twitter	7.3%	10.9%	7.0%	8.7%	7.2%	8.5%	8.3%
Telegram	6.3%	9.0%	6.7%	4.8%	5.9%	7.4%	6.4%
Facebook	5.0%	6.8%	6.3%	3.7%	6.1%	6.4%	6.2%
GitHub	1.8%	2.1%	1.9%	4.3%	2.3%	1.8%	2.3%
Reddit	1.7%	2.0%	2.0%	2.3%	1.9%	1.3%	2.2%
Slack	1.4%	2.1%	0.9%	0.8%	1.7%	1.4%	1.2%
Discord	0.8%	0.2%	1.1%	1.1%	0.8%	1.1%	0.6%
Bitcointalk	1.5%	2.2%	1.8%	2.0%	1.5%	1.7%	1.5%
Social capital	13.9%	17.6%	16.2%	11.4%	18.2%	18.5%	17.3%
Team size	5.0%	6.0%	5.6%	3.7%	6.5%	5.3%	5.7%
Number of advisors	3.8%	5.0%	4.6%	5.2%	4.9%	4.2%	4.7%
Expert rating	5.1%	6.5%	6.0%	2.5%	6.8%	9.0%	6.9%

The majority of the findings in the pooled sample conform to the factors identified in prior research. The ML model identifies ICO characteristics (whitepaper, continent, quarter, whitelist, KYC, ESG ratings, and category) to have a collective feature importance of 28.0%, agreeing with previous research (Fisch, 2019). Another substantial collection of features concerns the role of social media accounting for 32.3% of the prediction performance, which confirm and expand earlier findings (Fenu et al., 2018; Bourveau et al., 2018). The size of the Twitter, Telegram, and Facebook networks, in particular, are the most important features in this group, supporting the crowdfunding literature (Clauss et al., 2020). The remaining social media platforms appear to have a relatively low influence on ICO success prediction, possibly biased due to the limited information provided by the analysis method (i.e., dummy variables indicating official presence on a platform). Another explanation may be that these signals are context dependent, meaning that their importance is moderated by other signals. If the Random Forest models captures such dependencies, these features will automatically end up lower in the decision trees resulting in a lower feature importance score. It is interesting to see that the Random Forest algorithm identified the duration of an ICO campaign as the most important feature, whereas a weak effect was identified using the logistic regression.

Overall, significant differences are observed between the feature importances over the years, both at the grouped and at the individual level of features. The time samples (Samples 2-4) suggest that the collective importance of Campaign characteristics increases by 11.2% between 2017 and 2019. This effect can largely be attributed to the growing importance of the hard cap and duration variables (+3.9% and +4.5%, respectively). This suggest that ICO with a more accurate description of the fundraising campaign, in terms of a funding target and a road map, have a higher probability of success in 2019.Interestingly, the importance of social media suddenly decreases from 2017 to 2018, from 35.3% in 2017 to 27.7% in 2018. A large fraction of this change is caused by the diminishing impact of the size of an ICO's Twitter and Telegram networks, reducing by –3.9% and –2.3%, respectively. This aligns with the findings in the previous section, where Twitter was found to be statistically significant in 2017 but statistically insignificant in 2018 and 2019 (see Table 5.1). Furthermore, the importance of a GitHub repository remains relatively low except for the 2019 subsample (4.3%). This complements the findings of Roosenboom et al. (2020), who found no significant effect of GitHub repositories for ICOs launched until December 2017.

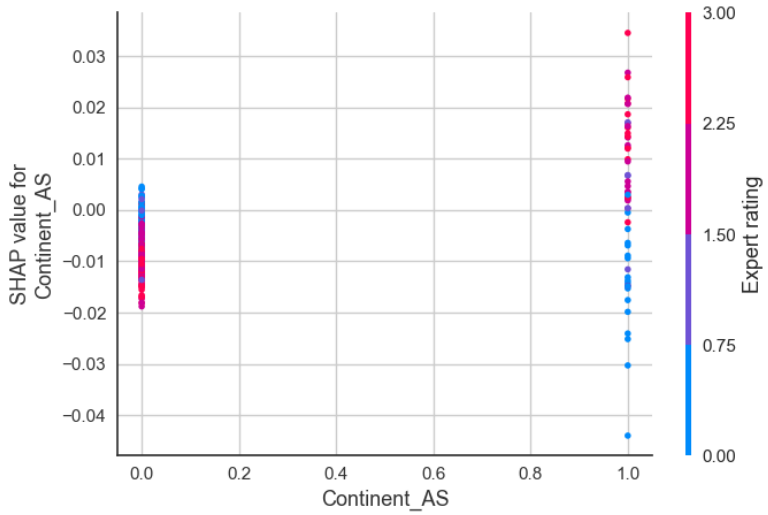
On the other hand, the individual and collective feature importances remain considerably stable between categories. Except for the Expert Ratings, which is slightly more important in the Financial services sample in comparison to the other sectors as indicated by a difference of approximately 3%. On a collective level, no substantial differences were observed in the importance of social capital between the categories.

Taken together, it is worth noting that the signal importances of the pooled sample represent the configuration-based samples (Samples 2-7) reasonably well. To evaluate and validate the analysis results in more detail, the SHapley Additive exPlanations (SHAP) method is adopted to interpret the outputs of the RF algorithm for the full sample. The interpretation of SHAP values is as follows: positive SHAP values indicate positive effects of a specific signal on funding success, and negative SHAP values resemble negative effects of a specific feature on funding success. Since identifying the most influential signals and exploring their influence patterns are the main objectives of this study, the remainder of the analysis is performed for the most important signals as identified in Table 5.7. Figure 5.2 shows the distribution of all instances' SHAP values per signal.

Consistent with the results in the first part of this thesis, it can be seen that longer campaign durations (red) have a negative influence on funding success, whereas shorter campaign durations (blue) have a positive effect on funding success. Regarding the influence of Expert ratings on funding outcomes, we see that higher ratings (red) increase the probability of funding success, whereas lower values (blue) reduce the likelihood of success. This agrees well with literature (Roosenboom et al., 2020) and our previous analysis (see Table 5.1). Also in line with our previous analysis, we find that ICOs in the Financial Services sector are less likely to succeed than other categories. Concerning the geographical locations, ICOs launched in Europe have a lower probability of reaching their fundraising targets, while ICOs launched in North America are more likely to reach their funding targets. Moreover, Figure 5.2 displays conflicting differences in SHAP values for ICOs launched in Asia. However, a more detailed analysis reveals that the effect of this signal is moderated by the expert rating, as shown in Figure 5.3.

Figure 5.2: SHAP value scatter plot per signal

Figure 5.3: Interaction effects between the geographical location (Asia) and expert ratings.



Next, the impact of social media network size on funding success is explored in more detail by plotting each instance against its corresponding SHAP value for a given platform. The results can be found in Figures 5.4-5.6. The results demonstrate that ICOs with a larger social media network have a higher probability of succeeding (Mollick, 2014), as indicated by increasing SHAP values for a larger number of Twitter followers, Telegram, and Facebook. Furthermore, the graphs show exponential relationships between the network size (log) and the SHAP values. What is also interesting in these figures is that small network sizes are shown to have a negative impact on funding success.

Figure 5.4: SHAP dependence plot: Size of an ICO's Twitter network size (log)

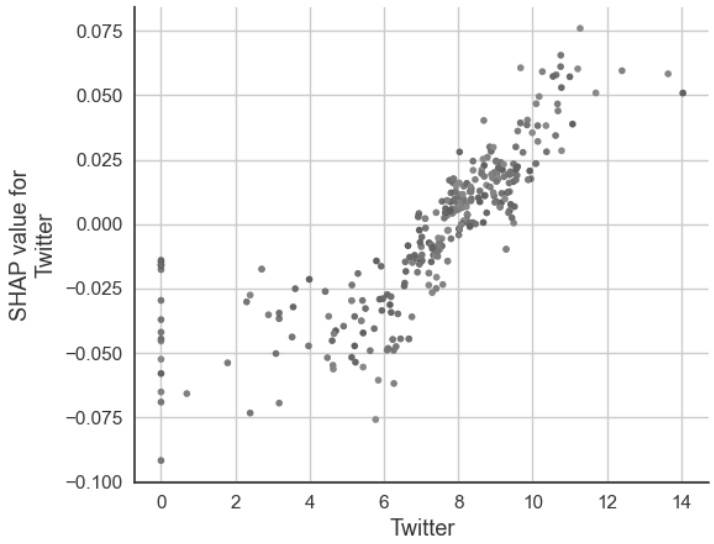


Figure 5.5: SHAP dependence plot: Size of an ICO's Telegram network size (log)

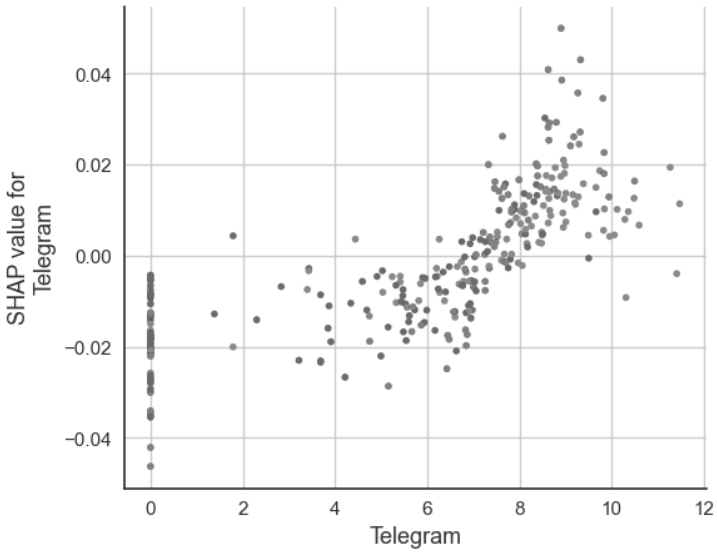
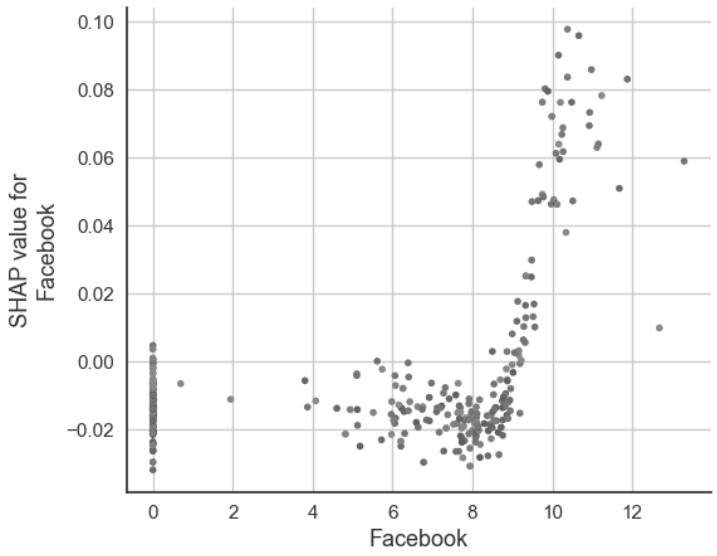


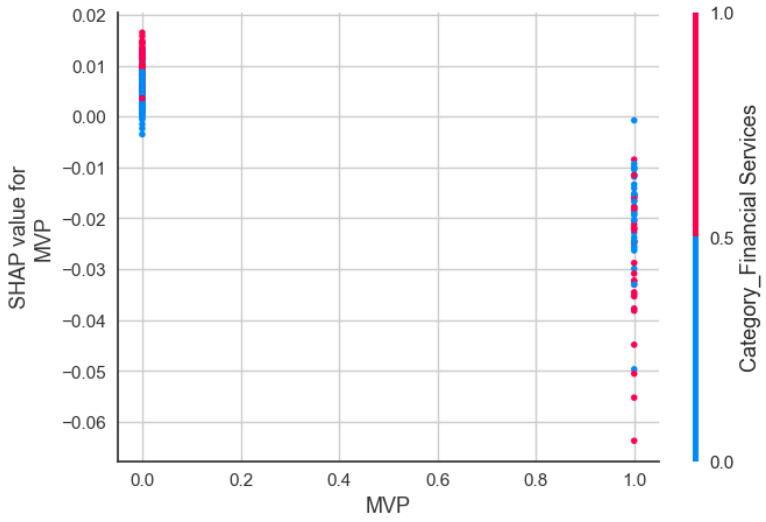
Figure 5.6: SHAP dependence plot: Size of an ICO's Facebook network size (log)



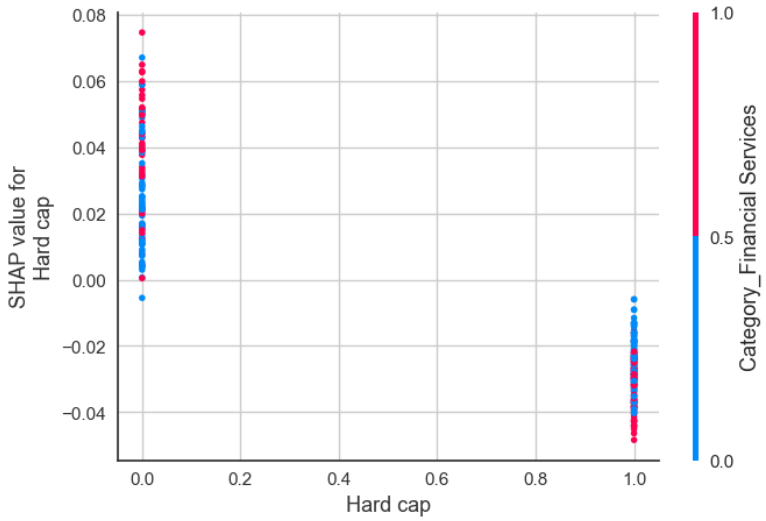
The analysis is concluded by evaluating whether the Random Forest model is a suitable alternative to study signal interactions and provides statistically meaningful results. To do so, the influence of the Financial services dummy on the SHAP values of the MVP, Hard cap, and KYC signals is determined. The respective plots can be found in Figure 5.7. In the previous section (Table 5.3), it was shown that the Financial Services category positively moderates the effect of the KYC signal on funding success and negatively moderates the influence of the MVP and Hard cap signals on funding success. Figure 5.7a provides evidence that the RF model identified this interaction, demonstrating that the Financial Services sector positively moderates the relationship between KYC and funding success as indicated by the SHAP values of the blue dots (financial services dummy=0, negative effect) and red dots (financial services dummy =1, positive effect). Similarly, Figure 5.7b and Figure 5.7c align with earlier findings as they reveal that the SHAP values of the MVP and Hard cap signals depend on the Financial Services dummy. In other words, the RF model correctly identifies that the Financial Services dummy moderates the impact of the MVP and Hard cap signals on the funding success. These results provide initial evidence that the Random Forest is capable of detecting statistically meaningful signal interactions.

Figure 5.7: SHAP dependence plots to validate interaction effects

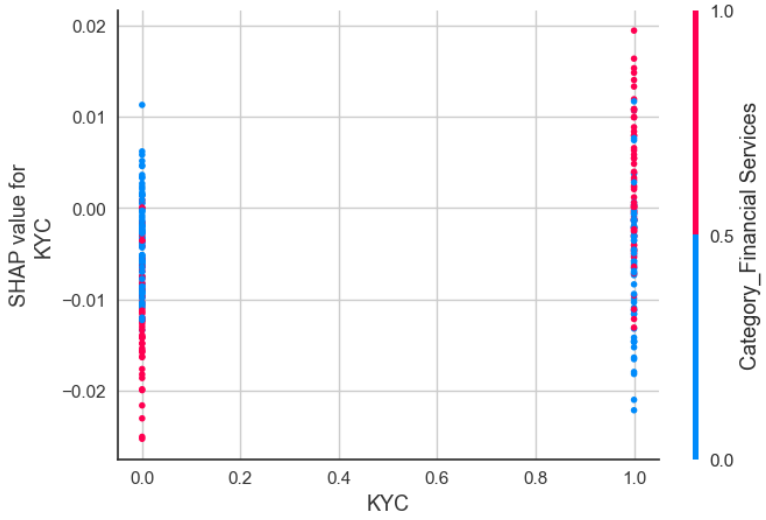
(a) Interaction effect: Financial Services x MVP



(b) Interaction effect: Financial Services x Hard cap



(c) Interaction effect: Financial Services x KYC



5.3. Comparison: Logistic Regression and Random Forest

Overall, both the statistical approach and the data-driven approach offer strengths and weaknesses, depending on their application. Here, I differentiate between prediction and explanation. The former entails the ”effects of causes”, while the latter focuses on the ”causes of effects” (Gelman and Imbens, 2013). An overview of the comparison is provided in Table 5.8.

Table 5.8: Overview of general advantages of using logistic regression and Random Forest.

	Traditional signaling method: Logistic regression	Data-driven approach: Random Forest
Advantages	Simple	Automatically identifies trends and patterns
	Directly interpretable	Captures effectively signal interactions
	Causal inference	Handles well high-dimensional data
	Captures well linear trends	Captures non-linear trends
		Prediction performance (moderate improvements)
Disadvantages	Interaction terms need to be specified a priori	Not directly interpretable
	Lower prediction performance*	Cannot be used for causal inference
	Assumes linear relationship between response and predictor variables	Requires additional machine learning knowledge
	Prone to overfitting on high-dimensional datasets	

*Though the relative performance difference with respect to the Random Forest model is small when the logistic regression is extended with interaction terms.

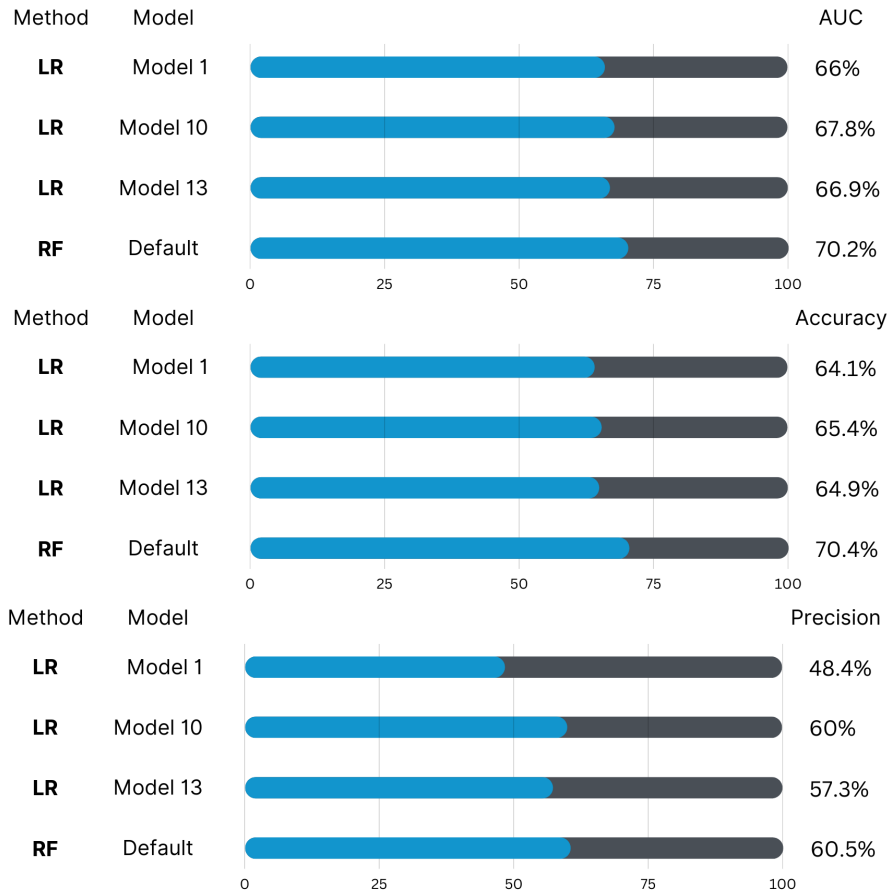
In the case of ICO funding, understanding how signals contribute to the success of ICOs is critical. One of the main advantages of logistic regression is that it provides interpretable coefficients. These coefficients explain precisely how the signals relate to the success of an ICO. Therefore, logistic regression is preferred when the goal is to explain the relationship between ICO success and one or more signals. However, logistic regression models do not account for interactions natively. Instead, signal interactions need to be specified a priori. Failing to specify key interactions could lead to erroneous results. If the logistic regression model is used for inference, this might postulate spurious relationships. Besides, rigorous causal identification may be infeasible for practical reasons. For example, large multi-signal data sets, coupled with theoretical underdevelopment, may undermine the credibility of causal modeling assumptions.

In light of ICOs, where relatively little is known or understood about the interactions between signals, data-driven methods may be a preferred initial step to explore relevant relationships. Although the results should not be treated uncritically, they suggest that a data-driven approach such as Random Forests can help us understand the effect of signals and signal interactions on ICO funding success. Therefore, I am cautiously optimistic about using data-driven approaches to explain and explore relevant signals and signal interactions related to successful funding outcomes. One potential application of data-driven models is to assist in the specification of logistic regression models. For example, data-driven models can help identify relevant signal interactions, which can then be evaluated and explained

in more detail using the logistic model.

While identifying causal relationships is one essential scientific goal, prediction is another important objective for many substantive issues in ICOs. For example, information asymmetry is incredibly destructive, and accurately predicting its onset is a critical issue for regulators who must try to improve market efficiency and find ways to reduce it. The summary presented in Figure 5.8 shows that Random Forests offers modest improvement in predictive power compared to the logistic regression in distinguishing successful ICOs from unsuccessful ICOs. AUC scores, Accuracy, and Precision all demonstrate modest improvement in prediction performance of Random Forests over logistic regression. It can also be seen that the performance of the logistic regression can be enhanced by incorporating interaction terms. For example, by incorporating time effects (i.e., Model 10) the performance of the logistic regression with respect to the baseline model (i.e., Model 1) increases by +1.9%, +1.3%, and +11.6% in terms of AUC, Accuracy, and Precision, respectively. These results suggest that the performance difference between the Random Forest and logistic regression can be minimized by incorporating relevant interaction terms. If scientists are solely interested in distinguishing successful from non-successful tokens, a data-driven approach is a useful place to start. However, the logistic regression reaches similar levels of performance if extended with relevant interaction terms.

Figure 5.8: Performance comparison Logistic Regression (LR) models vs. Random Forest (RF)



Lastly, generalizations from machine learning may lead to misjudgments if correlations are taken for causalities. Machine learning, therefore, is not just a new toolbox for the same problems. It should rather be seen as a different way of exploring signaling issues which is adequate in cases where data is complex and theoretical expectations are missing or are drawn into question. On the other hand, the logistic shows similar levels of classification performance when extended with relevant interaction terms and offers the additional benefit of causal inference. Given their complementary traits, using a combination of both methods is the desired convergence to take advantage of the best of both approaches in advancing ICO signaling literature. Here, the data-driven approach can be used as an exploratory tool, while the logistic regression can be used as a confirmatory tool.

Table 5.9: Purpose Logistic Regression vs. Random Forest.

Method	Inference	Interaction terms	Purpose	Preferred when
Logistic Regression	✓	Specified a priori	Explanation	Causal identification / Confirmatory
Random Forest	X	Derived from data	Prediction / Exploration	Data is complex / Theoretical underdevelopment

6

Discussion and Conclusion

ICOs represent a novel way for ventures to raise capital for a wide variety of projects. Given its rapid rise, the dynamics of ICOs have been largely unstudied. This thesis offers some exploratory insights into how the ICO signaling process works and has evolved. While the traditional signaling framework has proven to be an effective mechanism for understanding how external parties assess the quality of new ICOs, much of the literature has explored the impact of signals in isolation (Drover et al., 2018) and in relatively low noise environments. However, in high-noise environments, such as ICOs, signals may have complementary, overlapping or potentially competing effects. It is along these lines that we may begin to situate the significance and benefits of studying signals in portfolios rather than in isolation – not so much in terms of generating the highest prediction performance but rather as a critical ingredient to understand the ICO signaling dynamics in more detail. In the present thesis, however, I have not been concerned with laying out the full range of ways in which the interaction and configurational effects of signals have contributed or is contemporaneously bound up with the success of an ICO. Rather this thesis addressed two specific themes – namely, the temporal and context dependency of signal effectiveness – and studied it in two particular ways: extending the traditional approach and introducing a data-driven approach. These methods have been tested in the noisy setting of ICOs, where signals of quality from multiple transmitters compete for the investors' attention.

6.1. Research conclusion

As an exploratory empirical study, the main goal was to assess the suitability of a data-driven approach to effectively study signal interactions. Following the subquestions posed in Chapter 1.1, this section concludes by answering the main research question.

RQ1: What are the determinants of ICO success?

With the aid of an extensive literature survey, the first step was to identify an extended set of signals that were potentially related to ICO success. The results from the logistic regression were largely consistent with previous studies but also provided some additional insights. It was found that ICOs with larger Twitter and Telegram networks were more likely to succeed. Furthermore, the amount of milestones was found to be negatively associated with success, possibly because a higher number of milestones was found to reduce the informative content of the whitepaper (Florysiak and Schandlbauer, 2022). In addition, the analysis was performed on different subsamples. Depending on the sample specification, higher Environmental, Social, and Governance (ESG) scores, offering bonuses, the Know

Your Customer program, having a Minimal Viable Product (MVP), having Reddit, having Slack, and having Bitcointalk positively impact ICO funding. Consistent with Bellavitis et al. (2021), the subsample analysis revealed large swings in the statistical significance of signals between time samples (i.e., 2017, 2018, and 2019) and industry sector samples (i.e., Data and AI, Financial Services, and Entertainment and Gaming).

RQ2: What is the influence of signal interactions on ICO success?

Overall, the results confirm the notion that signals interact, and, therefore, should not be studied in isolation but rather in portfolios of signals (Drover et al., 2018). This has been shown by providing initial evidence on the time and context dependency of signals. The time-dependency of signals was first explored using subsample analysis, which showed that the significance of signals differed per subsample (i.e., 2017, 2018, and 2019). Interaction term analysis further supported this result, demonstrating that several signals (i.e., whitepaper, bonus, hard cap, Twitter, Telegram, and Expert Rating) are dependent on time. The analysis also revealed that in the absence of signal interaction terms, the effect of some signals (e.g., bonus and number of advisors) might be left unnoticed.

The context dependency of signals was explored in a similar fashion. Here, context dependency refers to the industry sector an ICO belongs to. Subsample analysis provided initial insights in the context dependency of signals demonstrating that the significance of signals differed per subsample: Data and AI, Entertainment and Gaming, and Financial Services. Interaction term analysis provided further evidence that signals are context-dependent. Specifically, signal interaction terms with the Financial Services and Entertainment and Gaming sectors demonstrated that the effect of several signals (i.e., KYC, bonus, hard cap, MVP, and expert rating) is dependent on the industry sector in which they are transmitted. Moreover, this thesis demonstrated that a context-based approach might provide the signal receiver with credible information on, in this case, industry-specific and hidden effects. One clear example is the Know Your Customer signal, whose effect is significantly moderated by Financial Services and would otherwise not be identified. At the same time, a more holistic approach ensures that the effect of certain signals is attributed to a specific context instead of only the singular signal.

RQ3: To what extent do the respective data-driven models (i.e., RF, ET, Ridge, and SVM) improve the classification performance (if at all) as compared to the logistic regression?

While prediction performance is rarely an aim in and of itself, it is considered a good proxy to evaluate whether the respective models capture the underlying signaling process and market dynamics more accurately as opposed to already established signaling methods. It was found that on each of the samples evaluated, the Random Forest model moderately outperformed the other methods in terms of AUC, Accuracy, Precision, and consistency. One possible explanation of this is that, by the nature of the algorithm, Random Forest allows for context-dependent outcomes and is better suited to capture conditional dependencies between signals without explicitly having to program them. In addition, the analysis revealed that the predictability of ICO success improved over time. Although interpretations should be treated with caution, this might indicate that investors learn over time, make more informed and rational decisions, and thus become more predictable.

RQ4: To what extent can a data-driven approach be used to extract relevant information concerning the ICO signaling process and signal interactions?

Based on the results of the previous question, the remainder of the analysis was performed using the

Random Forest model. First, the individual and collective importance of signals was evaluated. In line with the results obtained in the first part of the thesis, significant differences were observed between the time samples in the collective importance of campaign and social media related signals. These changes were mostly driven by changes on the individual level (i.e., hard cap, duration, Twitter, and Telegram). Overall, the collective feature importance scores of the pooled sample resembled the subsamples reasonably well. However, feature importance scores only provide information about the relevance of a specific signal, and they do not inform whether a signal has a positive or negative effect on ICO funding success. Moreover, feature importance scores provide little insights in the interaction between signals. A more detailed view on the effect of signals and signal interactions was obtained with the aid of the SHAP approach. Confirmed using the findings from the logistic regression, it was validated that the Random Forest approach effectively captured signal interactions with the additional benefit of not having to program them a priori. In addition, the analysis performed with the data-driven approach provided further insights into the impact of social media network sizes on funding success. It was found that the social media networks Twitter, Telegram, and Facebook can have a positive and negative effect, depending on the size of the respective network. More specifically, larger network sizes were positively related to funding success, whereas smaller network sizes were negatively related to funding success.

Main RQ: Can a data-driven approach effectively be applied to study signal interactions in the ICO environment?

Although the results should not be treated uncritically, they suggest that a data-driven approach, such as Random Forests, can help us understand the effect of signals and signal interactions on ICO funding success. Unlike the logistic regression, the main advantage of this data-driven approach is that it does not require the analyst to specify interaction terms a priori but effectively learns them from the data. As an additional advantage, the time dedicated to training the Random Forest model and identifying relevant signal interactions was generally shorter than the time lost in the trial-and-error process needed to identify and specify relevant signal interactions using the logistic regression model. In light of ICOs, where relatively little is known or understood about the interactions between signals, data-driven methods may be a preferred initial step to explore relevant relationships.

Nevertheless, these advantages come with a price not limited to the extra effort necessary to learn new and complex methods. Most importantly, there is a trade-off between prediction performance and model interpretability. While this research shows potential uses and interpretations of the outcomes of Random Forest models, it is essential to recall that such outcomes should not be treated as equivalent to the outcomes of a statistical model like logistic regression. In other words, generalizations from machine learning may lead to misjudgments if correlations are taken for causalities. Machine learning, therefore, is not just a new toolbox for the same problems. It should rather be seen as a different way of exploring signaling issues which is adequate in cases where data is complex and theoretical expectations are missing or are drawn into question. For instance, finding that a signal has a high feature importance does not necessarily mean that the respective signal specified in the traditional signaling model will be statistically significant. This research shows that signals with the highest feature importance scores do not necessarily lead to large coefficients in the logistic regression model (e.g., ICO duration). This thesis emphasizes that a data-driven approach such as Random Forests can be used as an effective tool to explore potentially relevant signal interactions, whereas logistic regression can be used as a confirmatory tool.

6.2. Scientific relevance

The proposed methods to delineating the evaluation of signals arised from the need to better understand the multi-signal ICOs environment. This research makes several contributions to the literature. First, a limited amount of studies have considered signal interactions and empirical evidence on the recent developments of the signaling framework (Drover et al., 2018) is still nascent. In particular, this thesis is the first to examine the interdependence of signals in the context of ICOs. In doing so, the literature on ICO signaling was revised by departing from the tradition of studying singular signals working in isolation, and focus instead on the interactions between signals. It was shown that signals interact and jointly affect ICO funding success. Specifically, this thesis provided initial evidence that signals in the high noise ICO environment are time and context-dependent.

This thesis further shows that having a larger Twitter and Telegram network size, a lower number of milestones, a shorter planned duration of the ICO campaign, no specified hard cap, a higher expert rating, and a larger team size were positively related to funding success. Depending on the sample specification, higher ESG scores, offering bonuses, having a KYC program, having a MVP, having Reddit, having Slack, and having Bitcointalk positively impact ICO funding.

Furthermore, incorporating time dependency has important implications because it relaxes a key assumption of traditional signaling theory, namely, that signals are stable over time. The results highlight the standing issue that traditional signaling research places little emphasis on time, which may affect whether or not signals are effective. Because the influence of a signal may vary over time, accounting for time effects may be crucial to obtain consistent results across different signaling environments. Considering that previous research reported large swings in signal interpretations (Bellavitis et al., 2021), the results of the present thesis suggest that this outcome may result from overlooking the importance of time dependency.

In addition, this thesis examined the suitability of a data-driven approach to study the ICO signaling environment. This data-driven approach resulted in a modest improvement of classification performance in comparison to the baseline approach and was shown to be an effective tool to explore signal interactions. This novel approach may be particularly useful in cases where data is complex and theoretical expectations are missing or are drawn into question.

This research also presents that in the ICO setting the recent extension on the signaling framework (Drover et al., 2018) provides new and meaningful success determinants. It was demonstrated that the effect of many success determinants can only be identified when focusing on signal interactions. Moreover, including signal interactions enhances the classification performance and model-fit of the logistic regression method. Taken all together, these findings emphasize the importance of signal interactions to better understand the ICO signaling environment.

Machine learning plays a role in many deployed decision systems, often in ways that are difficult or impossible to understand by human stakeholders. Explaining, in a human-understandable way, the relationship between the input and output of machine learning models is essential to the development of trustworthy machine learning based systems. This thesis adds to the growing body of literature that seeks to improve the explainability in machine learning practices. More specifically, this thesis presents how to interpret and leverage the outcome of machine learning methods from the perspective of signaling theory.

6.3. Managerial implications

Overall, the findings of this thesis are especially informative for two parties involved in the ICO process, namely regulators and investors. In absence of official legal jurisdictions and the mandatory disclosure of documents, regulators and investors are relying purely on the signals to evaluate ICOs and understand the market. As such, a better understanding of the underlying signaling process is crucial. This is particularly relevant for regulators or future policy-makers. It is evident that the sandbox approach and mere warnings and recommendations[1] are not effective. Zetzsche et al. (2019) claims that by advertising the ICO market's unregulated nature, warnings may adversely attract fraudulent participants. Instead, the present challenges of ICOs require legislation that acknowledges a global digital market's economic and technical peculiarities. Failure to do so will foster fraudulent behavior and result in regulatory arbitrage with fuzzy implications (De Andrés et al., 2022). One clear example is the Know Your Customer (KYC) initiative[2], whose effect can hardly be measured without taking into account signal interactions. Hence, the interplay between a large number of signals broadcasted simultaneously, and the ability of regulators and policy makers to evaluate rich information in noisy environments likely plays a key role in designing effective regulations. Furthermore, the continuous development of the market is also likely to further evolve ICOs in ways that may change the dynamics between investors and entrepreneurs. The extensions presented provide a new way that facilitates learning about and accounting for such developments.

As the problem of information asymmetry reaches far beyond the boundaries of the current research setting, the applicability of signaling theory and the present extensions may also advance other areas of inquiry. For example, many organizations and individuals today seek resources in high-noise environments. Firms communicate information about a product to customers in a crowded market, and job seekers compete with hundreds of other applicants for one open position (e.g., see Bangerter et al. (2012)). Because signal receivers in high-noise environments may focus on singular signals working in isolation, important signal interactions may be overlooked. Consequently, applying signaling theory may be ineffective or result in adverse effects. The present thesis suggests that such outcomes may result from overlooking the importance of signal interactions, specifically concerning their time and context-dependency.

6.4. Limitations

While the results confirm the importance of a context based approach (Drover et al., 2018), they represent only a first foray into the phenomena of studying the complementing and interactive effect of signals, and they have a number of limitations. Firstly, the analysis presented in this thesis is scoped on the application of Random Forest as a suitable candidate for assisting signaling practices. While this thesis has clearly shown that Random Forest can be successfully applied for specific purposes, many other data-driven methods might yield similar or even better results. Another downside related to the data-driven approach is that it produces a numerical solution that cannot be converted to an analytical solution. That is, framing the problem in a well-understood form and calculating the exact solution without the aid of a computer. The analysis of the present thesis was based on the assumption that funding success is an accurate proxy for the underlying quality of an ICO. While clearly important, several other dimensions of token quality, such as the post-ICO performance, were not evaluated.

Similarly, while a large number of signals were considered , many of these signals were modeled as dummy variables, thus contained limited information. This limitation is likely to affect the interpretation

[1] E.g., see SEC (2013)

[2] As mentioned in Chapter 2, the Know Your Customer initiative was an attempt to imbue some initial form of regulation into the ICO system.

of a specific signal significantly. For example, it was shown that the social media network sizes of Twitter, Telegram, and Facebook can have a positive and negative effect on ICO funding. If modeled as a dummy variable, this would only indicate the positive effects of these networks (e.g., see Fisch (2019)). However, the findings indicate that a small Twitter, Telegram, and Facebook network size has a negative impact on ICO funding, whereas large network sizes have a positive influence on ICO funding. Hence, it is likely that many of the dummy variables considered in this research provide an incomplete view on the effect of the respective signals.

Although the signaling framework may reduce information asymmetry, it primarily addresses how ICOs signal ”quality” to those outside their boundaries. This approach provides limited insights about the signal receiver – how and why he or she might attend to signals and what transpires when multiple signals are considered simultaneously. From this perspective, the signaling framework is incomplete because it treats the cognitions of the receiver as a ’black box’. In other words, this view assumes that signal receivers perceive and attend to all signals once available. Given that signal receivers are bounded by rationality and cannot attend to all available signals (Shepherd et al., 2017), this assumption is imperfect. While studying portfolios of signals instead of singular signals in isolation already provides a step in the right direction, the cognitive limits to receive, process, and attend to all signals is left unanswered.

Furthermore, the present work did not account for market conditions. While such effects may be partially absorbed by the time variable, future studies should control for cryptocurrency market conditions such as the price and volatility of Bitcoin and Ethereum. Additionally, ICOs may be a phenomenon of short-lived importance. Currently, the year-on-year decrease since 2019 raised important questions about the future of the ICO market (Fromberger and Haffke, 2019). Given that the present study only considered ICOs launched between 2017 and 2019, the results offer limited insights into the developments of the market after this period. However, even if ICOs continue to make up a small proportion of new venture funding, the results of this exploratory study suggest that a number of findings could be of interest for future research in different signaling environments.

6.5. Future recommendations

Given the initial evidence that the importance of signals change over time, it is likely that the effect of many other signals in ICO literature, such as the whitepaper content, are also time-dependent and dynamic. This suggests that making temporal considerations explicit and modeling time directly should be part of future empirical models. Future research on ICOs could perform a longitudinal study over time to examine if signals became more stable in recent years.

Furthermore, future studies could investigate the ways in which signals are working together and connect this to the specific role of the respective signals. For example, the role of social media networks may be to grasp the attention of investors whereas the role of the Know Your Customer program may be to signal reliability.

Future research could also adopt the proposed data-driven approach to explore the effect of signals on moral hazard. According to Connelly et al. (2011), there are two types of signals: signals of quality and signals of intent. The former is used to reduce information asymmetry and was the focal point of this thesis. The latter is used to overcome the problem of moral hazard.

The present thesis demonstrates that including signal interactions in the ICO setting provides meaningful and new insights about the signaling process. Accounting for signal interactions may be particularly

relevant in cases where signaling has been found to be ineffective (e.g., see Park et al. (2016)). Specifically, the results of thesis suggests that this outcome may result from overlooking the importance of time and context dependent signals. Future studies could revise these cases. Especially when prior expectations or knowledge about signal interactions are missing or nonintuitive, adopting the proposed data-driven approach may help to obtain new insights.

Environments with signaling effects may also have learning effects, which points to a potential explanation for the time-dependency of signals that was observed. In particular, signal receivers can learn from past ICO experience, including past success and failure that may enable them to make better inferences about the underlying quality of an ICO. This study was unable to separate learning effects from the signaling effects. Future research along these lines can make significant contributions to signaling and learning theories.

Lastly, future research could also link signal interactions to cognition, to explain how and why signals are context dependent, and to provide a better understanding of the investors' decision-making rationale in the context of ICOs. This may also reveal new information explaining how and why signals are attended to, and which kind of signals transpire when the receiver is exposed to multiple signals at the same time.

Bibliography

Ackermann, E., Bock, C., and Bürger, R. (2020). Democratising entrepreneurial finance: the impact of crowdfunding and initial coin offerings (icos). In *Contemporary developments in entrepreneurial finance*, pages 277–308. Springer.

Adhami, S., Giudici, G., and Martinazzi, S. (2018). Why do businesses go crypto? an empirical analysis of initial coin offerings. *Journal of Economics and Business*, 100:64–75.

Al-Omoush, K. S., Simón-Moya, V., and Sendra-García, J. (2020). The impact of social capital and collaborative knowledge creation on e-business proactiveness and organizational agility in responding to the covid-19 crisis. *Journal of Innovation & Knowledge*, 5(4):279–288.

Albrecht, S., Lutz, B., and Neumann, D. (2020). The behavior of blockchain ventures on twitter as a determinant for funding success. *Electronic Markets*, 30(2):241–257.

Allison, T. H., Davis, B. C., Webb, J. W., and Short, J. C. (2017). Persuasion in crowdfunding: An elaboration likelihood model of crowdfunding performance. *Journal of Business Venturing*, 32(6):707–725.

Amsden, R. and Schweizer, D. (2018). Are blockchain crowdsales the new'gold rush'? success determinants of initial coin offerings. *Success determinants of initial coin offerings (April 16, 2018)*.

Ante, L., Sandner, P., and Fiedler, I. (2018). Blockchain-based icos: pure hype or the dawn of a new era of startup financing? *Journal of Risk and Financial Management*, 11(4):80.

Bail, C. A. (2015). Lost in a random forest: Using big data to study rare events. *Big Data & Society*, 2(2):2053951715604333.

Bangerter, A., Roulin, N., and König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology*, 97(4):719.

Belitski, M. and Boreiko, D. (2021). Success factors of initial coin offerings. *The Journal of Technology Transfer*, pages 1–17.

Bellavitis, C., Fisch, C., and Wiklund, J. (2021). A comprehensive review of the global development of initial coin offerings (icos) and their regulation. *Journal of Business Venturing Insights*, 15:e00213.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Block, J. H., Groh, A., Hornuf, L., Vanacker, T., and Vismara, S. (2021). The entrepreneurial finance markets of the future: a comparison of crowdfunding and initial coin offerings. *Small Business Economics*, 57(2):865–882.

Boreiko, D. and Risteski, D. (2021). Serial and large investors in initial coin offerings. *Small Business Economics*, 57(2):1053–1071.

Bourveau, T., De George, E. T., Ellahie, A., and Macciocchi, D. (2018). Initial coin offerings: Early evidence on the role of disclosure in the unregulated crypto market. *Available at SSRN*, 3193392.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Catalini, C. and Gans, J. S. (2018). Initial coin offerings and the value of crypto tokens. Technical report, National Bureau of Economic Research.

Chanson, M., Gjoen, J., Risius, M., and Wortmann, F. (2018). Initial coin offerings (icos): The role of social media for organizational legitimacy and underpricing.

Chitsazan, H., Bagheri, A., and Tajeddin, M. (2022). Initial coin offerings (icos) success: Conceptualization, theories and systematic analysis of empirical studies. *Technological Forecasting and Social Change*, 180:121729.

Choi, J. and Wang, H. (2009). Stakeholder relations and the persistence of corporate financial performance. *Strategic management journal*, 30(8):895–907.

Clauss, T., Niemand, T., Kraus, S., Schnetzer, P., and Brem, A. (2020). Increasing crowdfunding success through social media: The importance of reach and utilisation in reward-based crowdfunding. *International Journal of Innovation Management*, 24(03):2050026.

Connelly, B. L., Certo, S. T., Ireland, R. D., and Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of management*, 37(1):39–67.

Davies, W. E. and Giovannetti, E. (2018). Signalling experience & reciprocity to temper asymmetric information in crowdfunding evidence from 10,000 projects. *Technological Forecasting and Social Change*, 133:118–131.

De Andrés, P., Arroyo, D., Correia, R., and Rezola, A. (2022). Challenges of the market for initial coin offerings. *International Review of Financial Analysis*, 79:101966.

de Andrés, P., Arroyo, D., Correia, R., and Rezola, A. (2022). Challenges of the market for initial coin offerings. *International Review of Financial Analysis*, 79:101966.

Dixon, M. F., Halperin, I., and Bilokon, P. (2020). *Machine learning in Finance*, volume 1406. Springer.

Douglas, E. J., Shepherd, D. A., and Prentice, C. (2020). Using fuzzy-set qualitative comparative analysis for a finer-grained understanding of entrepreneurship. *Journal of Business Venturing*, 35(1):105970.

Drover, W., Wood, M. S., and Corbett, A. C. (2018). Toward a cognitive view of signalling theory: individual attention and signal set interpretation. *Journal of Management Studies*, 55(2):209–231.

Edelman, L. F., Manolova, T. S., Brush, C. G., and Chow, C. M. (2021). Signal configurations: Exploring set-theoretic relationships in angel investing. *Journal of Business Venturing*, 36(2):106086.

Fenu, G., Marchesi, L., Marchesi, M., and Tonelli, R. (2018). The ICO phenomenon and its relationships with ethereum smart contract environment. In *2018 International Workshop on Blockchain Oriented Software Engineering (IWBOSE)*, pages 26–32. IEEE.

Ferrati, F., Muffatto, M., et al. (2021). Entrepreneurial finance: emerging approaches using machine learning and big data. *Foundations and Trends® in Entrepreneurship*, 17(3):232–329.

Fisch, C. (2019). Initial coin offerings (ICOs) to finance new ventures. *Journal of Business Venturing*, 34(1):1–22.

Fisch, C., Masiak, C., Vismara, S., and Block, J. H. (2018). Motives to invest in initial coin offerings (icos). *Available at SSRN 3287046*.

Fisch, C. and Momtaz, P. P. (2020). Institutional investors and post-ico performance: an empirical analysis of investor returns in initial coin offerings (icos). *Journal of Corporate Finance*, 64:101679.

Florysiak, D. and Schandlbauer, A. (2022). Experts or charlatans? ico analysts and white paper informativeness. *Journal of Banking & Finance*, 139:106476.

Fridgen, G., Regner, F., Schweizer, A., and Urbach, N. (2018). Don't slip on the initial coin offering (ico): A taxonomy for a blockchain-enabled form of crowdfunding. In *26th European Conference on Information Systems (ECIS)*.

Fromberger, M. and Haffke, L. (2019). Ico market report 2018/2019–performance analysis of 2018's initial coin offerings. *Available at SSRN 3512125*.

Gelman, A. and Imbens, G. (2013). Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Giudici, G., Vismara, S., et al. (2021). Ipos and entrepreneurial firms. *Foundations and Trends® in Entrepreneurship*, 17(8):766–852.

Greenberg, M. D., Pardo, B., Hariharan, K., and Gerber, E. (2013). Crowdfunding support tools: predicting success & failure. In *CHI'13 extended abstracts on human factors in computing systems*, pages 1815–1820.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Howell, S. T., Niessner, M., and Yermack, D. (2020). Initial coin offerings: Financing growth with cryptocurrency token sales. *The Review of Financial Studies*, 33(9):3925–3974.

Huang, W., Meoli, M., and Vismara, S. (2020). The geography of initial coin offerings. *Small Business Economics*, 55(1):77–102.

Kaal, W. A. and Dell'Erba, M. (2017). Initial coin offerings: emerging practices, risk factors, and red flags. *Verlag CH Beck (2018), U of St. Thomas (Minnesota) Legal Studies Research Paper*, (17-18).

Lee, J., Li, T., and Shin, D. (2022). The wisdom of crowds in fintech: Evidence from initial coin offerings. *The Review of Corporate Finance Studies*, 11(1):1–46.

Li, J. and Mann, W. (2018). Initial coin offering and platform building. *SSRN Electronic Journal*, pages 1–56.

Liebau, D. and Schueffel, P. (2019). Cryptocurrencies & initial coin offerings: Are they scams?-an empirical study. *The Journal of The British Blockchain Association*, 2(1):7749.

Lipusch, N. (2018). Initial coin offerings–a paradigm shift in funding disruptive innovation. *Available at SSRN 3148181*.

Liu, C. and Wang, H. (2019a). Crypto tokens and token offerings: an introduction. *Cryptofinance and mechanisms of exchange*, pages 125–144.

Liu, C. and Wang, H. (2019b). Initial coin offerings: What do we know and what are the success factors? *Cryptofinance and Mechanisms of Exchange*, pages 145–164.

Lu, C.-T., Xie, S., Kong, X., and Yu, P. S. (2014). Inferring the impacts of social media on crowdfunding. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 573–582.

Lyandres, E., Palazzo, B., and Rabetti, D. (2019). Do tokens behave like securities? an anatomy of initial coin offerings. *SSRN Electronic Journal*.

Mansouri, S. and Momtaz, P. P. (2022). Financing sustainable entrepreneurship: Esg measurement, valuation, and performance. *Journal of Business Venturing*, 37(6):106258.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, 29(1):1–16.

Momtaz, P. P. (2019). Token sales and initial coin offerings: introduction. *The Journal of Alternative Investments*, 21(4):7–12.

Momtaz, P. P. (2020). Initial coin offerings. *Plos one*, 15(5):e0233018.

Momtaz, P. P. (2021a). Entrepreneurial finance and moral hazard: evidence from token offerings. *Journal of Business Venturing*, 36(5):106001.

Momtaz, P. P. (2021b). Initial coin offerings, asymmetric information, and loyal ceos. *Small business economics*, 57(2):975–997.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, 26(1):67–82.

Ofir, M. and Sadeh, I. (2020). Ico vs. ipo: Empirical findings, information asymmetry, and the appropriate regulatory framework. *Vand. J. Transnat'l L.*, 53:525.

Park, U. D., Borah, A., and Kotha, S. (2016). Signaling revisited: The use of signals in the market for ipo s. *Strategic Management Journal*, 37(11):2362–2377.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Perez, C., Sokolova, K., and Konate, M. (2020). Digital social capital and performance of initial coin offerings. *Technological forecasting and social change*, 152:119888.

Ralcheva, A. and Roosenboom, P. (2020). Forecasting success in equity crowdfunding. *Small Business Economics*, 55(1):39–56.

Ren, J., Raghupathi, V., and Raghupathi, W. (2021). Exploring the subjective nature of crowdfunding decisions. *Journal of Business Venturing Insights*, 15:e00233.

Roosenboom, P., van der Kolk, T., and de Jong, A. (2020). What determines success in initial coin offerings? *Venture Capital*, 22(2):161–183.

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine learning*, 13(1):135–143.

SEC (2013). Ponzi schemes using virtual currencies. *SEC Pub. No. 153 (7/13)*.

Shepherd, D. A., Mcmullen, J. S., and Ocasio, W. (2017). Is that an opportunity? an attention model of top managers' opportunity beliefs for strategic action. *Strategic Management Journal*, 38(3):626–644.

Shrestha, P., Arslan-Ayaydin, Ö., Thewissen, J., and Torsin, W. (2021). Institutions, regulations and initial coin offerings: An international perspective. *International Review of Economics & Finance*, 72:102–120.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics*, pages 281–306. Elsevier.

Steigenberger, N. and Wilhelm, H. (2018). Extending signaling theory to rhetorical signals: Evidence from crowdfunding. *Organization Science*, 29(3):529–546.

Thies, F., Wessel, M., and Benlian, A. (2014). Understanding the dynamic interplay of social buzz and contribution behavior within and between online platforms–evidence from crowdfunding.

Vulkan, N., Åstebro, T., and Sierra, M. F. (2016). Equity crowdfunding: A new phenomena. *Journal of Business Venturing Insights*, 5:37–49.

Walthoff-Borm, X., Schwienbacher, A., and Vanacker, T. (2018). Equity crowdfunding: First resort or last resort? *Journal of Business Venturing*, 33(4):513–533.

Wang, W., Zheng, H., and Wu, Y. J. (2020). Prediction of fundraising outcomes for crowdfunding projects based on deep learning: a multimodel comparative study. *Soft Computing*, 24(11):8323–8341.

Wei, Y. M., Hong, J., and Tellis, G. J. (2022). Machine learning for creativity: Using similarity networks to design better crowdfunding projects. *Journal of Marketing*, 86(2):87–104.

Yeh, J.-Y. and Chen, C.-H. (2020). A machine learning approach to predict the success of crowdfunding fintech project. *Journal of Enterprise Information Management*.

Zetzsche, D. A., Buckley, R. P., Arner, D. W., and Föhr, L. (2018). The ico gold rush: It's a scam, it's a bubble, it's a super challenge for regulators. *CGN: Investment in R&D & Innovation*.

Zetzsche, D. A., Buckley, R. P., Arner, D. W., and Fohr, L. (2019). The ico gold rush: It's a scam, it's a bubble, it's a super challenge for regulators. *Harv. Int'l LJ*, 60:267.

A

Grid search

A grid search was performed to find the best combination of hyperparameters for the Random Forest model. The combination of hyperparameters that reports the highest test (out-of-sample) AUC, Accuracy, and Precision is used for the analysis. Table A.1 presents the range of hyperparameters considered during the tuning process.

Table A.1: Hyperparameter tuning values for the Random Forest model

Parameter	Values
Number of trees	From 10 to 1,000, in multiples of 10
Depth	3, 5, 10, max (Default)
Maximum variables per split	$4, 8, 16, \sqrt{\text{Number of features}}$

To identify the best combination of hyperparameters is obtained in two steps. First, each possible RF model is trained under different combinations of hyperparameters. In the second step, either the tree depth or the maximum number of variables per split is fixed, and the AUC, Accuracy, and Precision are plotted for different specifications of the other parameter as a function of the maximum number of trees. The combination of hyperparameters that reports the best performance in terms of AUC, Accuracy, and Precision is used for the analysis in this thesis.

Figures A.1-A.3 show the AUC, Accuracy, and Precision values for different number of variables per split and different number of trees, for a tree depth fixed at the maximum number of layers (default), trained on the full sample. It was found that the optimal hyperparameters for the Random Forest model contains the maximum depth (default), the sqrt of the number of variables per split and 100 decision trees lead to the best overall performance in terms of AUC, Accuracy, and Precision. In addition, the same analysis is repeated using the Python package scikit-learn and the corresponding GridSearchCV function, which resulted in the same hyperparameter values.

Figure A.1: AUC of Random Forest models for different hyperparameter specifications.

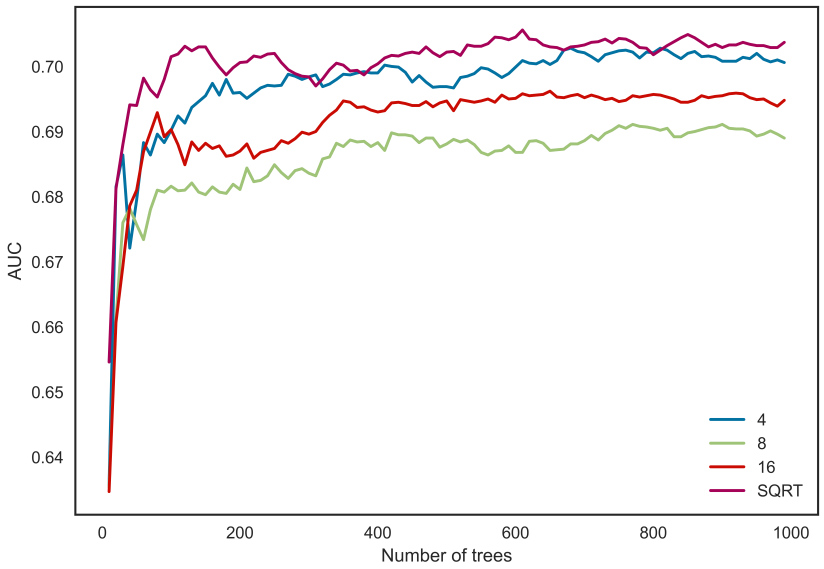


Figure A.2: Accuracy of Random Forest models for different hyperparameter specifications.

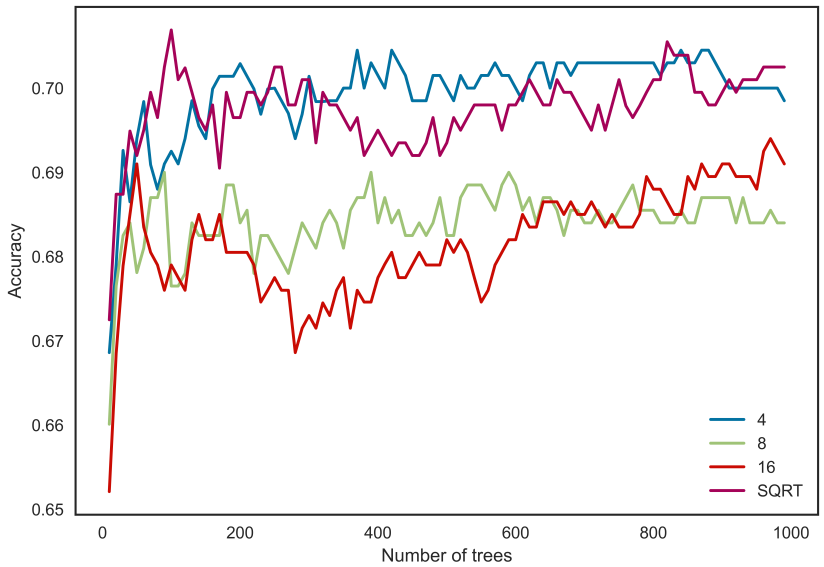
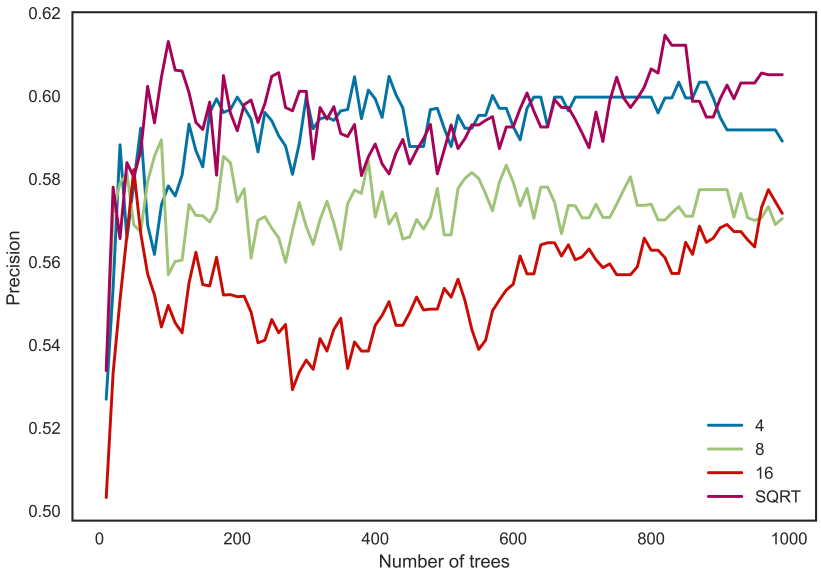


Figure A.3: Precision of Random Forest models for different hyperparameter specifications.



B

Parsimonious model supplement

Table B.1: Interaction terms and corresponding R-squared (parsimonious).

Interaction terms	Adjusted R-squared
Hard cap x GitHub	0.125681
KYC x Continent Europe	0.124572
Whitelist Bitcointalk	0.123655
2017 x Hard cap	0.123233
Bonus x Asia	0.123081
Hard cap x Video	0.122990
Bonus x GitHub	0.122773
Whitelist x Slack	0.122552
ESG x Bounty	0.122550
Telegram x Asia	0.122469
Whitepaper x 2018	0.121814
Whitelist x Telegram	0.121812
2019 x Bounty	0.121681
Hard cap x Duration	0.121381
Hard cap x Financial Services	0.121245
Video x Europe	0.120788
Duration x Number of milestones	0.120659
2019 x Duration	0.120400
Hard cap x Reddit	0.120312
ESG x Duration	0.120214
2019 x Bonus	0.120206
Facebook x Team size	0.120176
Bonus x Reddit	0.120152
2017 x Number of milestones	0.120150
2017 x Duration	0.120135
2019 x Telegram	0.120116
2019 x Bitcointalk	0.120066
2017 x Asia	0.120005
Hard cap x Data and AI	0.120000
2018 x Number of milestones	0.119969

C

Correlation matrix

Table C.1: Correlation matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
(1) Success	1.0***																													
(2) Whitepaper	0.01	1.0***																												
(3) Whitelist	0.03	0.07*	1.0***																											
(4) KYC	-0.01	0.06	0.54***	1.0***																										
(5) ESG	-0.03	-0.0	0.03	0.12***	1.0***																									
(6) Bounty	-0.13***	0.07*	0.22***	0.27***	0.06	1.0***																								
(7) Bonus	-0.0	0.05	0.07*	0.07*	0.03	0.2***	1.0***																							
(8) Hard cap	-0.15***	0.08*	0.07*	0.13***	-0.01	0.23***	0.16***	1.0***																						
(9) Duration	-0.14***	0.02	0.06	0.07*	0.03	0.19***	0.02	0.17***	1.0***																					
(10) Number of milestones	-0.13***	0.04	0.1**	0.11***	0.04	0.12***	0.03	0.13***	0.05	1.0***																				
(11) Presale	-0.11***	0.05	0.2***	0.24***	0.16***	0.28***	0.16***	0.18***	0.08*	0.13***	1.0***																			
(12) MVP	-0.11***	0.0	0.18***	0.3***	0.08*	0.38***	0.11***	0.2***	0.13***	0.1**	0.27***	1.0***																		
(13) Video	-0.03	0.13***	0.05	0.09**	0.08*	0.09**	0.09**	0.08*	0.05	0.14***	0.19***	0.11**	1.0***																	
(14) Twitter	0.15***	0.05	0.11***	0.07*	0.01	0.13***	0.08*	-0.02	-0.01	-0.03	0.01	0.05	0.09**	1.0***																
(15) Telegram	0.05	0.05	0.24***	0.22***	0.05	0.2***	0.14***	0.07*	0.01	0.08**	0.25***	0.19***	0.11***	0.26***	1.0***															
(16) Facebook	0.03	0.09**	0.13***	0.1**	-0.01	0.16***	0.04	0.02	-0.04	0.12***	0.06	0.03	0.16***	0.31***	0.21***	1.0***														
(17) GitHub	-0.01	0.09**	0.11***	0.09**	0.01	0.15***	0.08*	0.1**	0.02	0.07*	0.1**	0.15***	0.17***	0.11***	0.09**	0.14***	1.0***													
(18) Reddit	0.01	0.1**	0.13***	0.16***	0.01	0.23***	0.07*	0.1**	0.03	0.16***	0.16***	0.16***	0.22***	0.17***	0.12***	0.17***	0.3***	1.0***												
(19) Slack	0.06	-0.02	-0.1**	-0.19***	-0.03	-0.13***	-0.05	-0.07*	-0.06	-0.0	-0.1**	-0.14***	0.07*	0.06	-0.2***	0.13***	0.1**	0.07*	1.0***											
(20) Discord	-0.02	0.02	0.1**	0.08*	0.05	0.1**	0.04	0.01	0.06	-0.03	0.12***	0.12***	0.01	0.05	0.03	-0.03	0.07*	0.08**	0.04	1.0***										
(21) Bitcointalk	-0.03	0.13***	0.05	0.03	-0.03	0.24***	0.08*	0.11***	-0.0	0.2***	0.13***	0.07*	0.21***	0.1**	0.06	0.27***	0.21***	0.3***	0.12***	-0.03	1.0***									
(22) Team size	0.06	0.02	0.16***	0.17***	0.07*	0.12***	0.04	0.1**	0.01	0.18***	0.12***	0.14***	0.2***	0.11***	0.16***	0.21***	0.14***	0.16***	0.01	0.0	0.12***	1.0***								
(23) Number of advisors	-0.01	0.06*	0.19***	0.24***	0.08*	0.1**	0.08*	0.04	0.01	0.1**	0.12***	0.14***	0.14***	0.14***	0.25***	0.26***	0.19***	0.2***	-0.03	-0.02	0.14***	0.17***	1.0***							
(24) Expert rating	-0.25***	0.06*	0.13***	0.27***	0.09**	0.27***	0.05	0.25***	0.19***	0.11***	0.22***	0.3***	0.07*	-0.01	0.16***	-0.0	0.04	0.08**	-0.15***	0.06	0.06	0.02	0.06	1.0***					
(25) 2017	0.05	-0.06	-0.33***	-0.46***	-0.15***	-0.33***	-0.11***	-0.22***	-0.06	-0.09**	-0.36***	-0.26***	-0.03	0.02	-0.25***	0.08*	-0.05	-0.13***	0.32***	-0.14***	0.07*	-0.07*	-0.11***	-0.3***	1.0***					
(26) 2018	-0.01	0.1**	0.26***	0.27***	0.02	0.22***	0.01	0.15***	0.01	0.1**	0.22***	0.05	0.08*	-0.05	0.12***	0.07*	0.06	0.09**	-0.19***	-0.0	0.12***	0.12***	0.16***	0.06*	-0.68***	1.0***				
(27) 2019	-0.05	-0.07*	0.06	0.2***	0.12***	0.09**	0.02	0.05	0.03	0.0	0.1**	0.21***	-0.03	0.02	0.11***	-0.07*	0.01	0.07*	-0.1**	0.1**	-0.14***	-0.01	-0.03	0.23***	-0.21***	-0.47***	1.0***			
(28) Data and AI	0.02	-0.01	0.09**	0.07*	0.08**	0.05	0.09**	0.03	0.01	0.03	0.11**	0.01	0.02	0.04	0.04	-0.02	-0.0	0.04	-0.02	-0.01	0.01	-0.01	-0.02	0.03	-0.07*	0.06	-0.01	1.0***		
(29) Entertainment and gaming	-0.0	-0.04	-0.04	-0.07*	-0.09**	-0.09**	-0.0	-0.02	-0.01	-0.05	-0.07*	-0.03	-0.02	0.04	-0.03	-0.01	-0.08*	-0.06	0.05	0.0	-0.07*	-0.05	-0.02	-0.03	0.1**	-0.09**	-0.04	-0.12***	1.0***	
(30) Financial Services	-0.11***	0.05	0.06	0.09**	0.03	0.12***	0.01	0.07*	0.08*	0.06	0.05	0.11***	0.03	-0.04	0.01	0.03	0.05	0.08*	-0.04	0.04	0.09**	0.07*	-0.01	0.07*	-0.06	0.05	-0.01	-0.35***	-0.32***	1.0***

Table C.2: Variance Inflation Factor matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
(1) Success	-																													
(2) Whitepaper	1.0	-																												
(3) Whitelist	1.0	1.0	-																											
(4) KYC	1.0	1.0	1.41	-																										
(5) ESG	1.0	1.0	1.0	1.02	-																									
(6) Bounty	1.02	1.01	1.05	1.08	1.0	-																								
(7) Bonus	1.0	1.0	1.01	1.0	1.0	1.04	-																							
(8) Hard cap	1.02	1.01	1.01	1.02	1.0	1.06	1.03	-																						
(9) Duration	1.02	1.0	1.0	1.0	1.0	1.04	1.0	1.03	-																					
(10) Number of milestones	1.02	1.0	1.01	1.01	1.0	1.01	1.0	1.02	1.0	-																				
(11) Presale	1.01	1.0	1.04	1.06	1.03	1.08	1.03	1.03	1.01	1.02	-																			
(12) MVP	1.01	1.0	1.03	1.1	1.01	1.17	1.01	1.04	1.02	1.01	1.08	-																		
(13) Video	1.0	1.02	1.0	1.01	1.01	1.01	1.01	1.01	1.0	1.02	1.04	1.01	-																	
(14) Twitter	1.02	1.0	1.01	1.01	1.0	1.02	1.01	1.0	1.0	1.0	1.0	1.0	1.01	-																
(15) Telegram	1.0	1.0	1.06	1.05	1.0	1.04	1.02	1.0	1.0	1.01	1.06	1.04	1.01	1.07	-															
(16) Facebook	1.0	1.01	1.02	1.01	1.0	1.03	1.0	1.0	1.0	1.01	1.0	1.0	1.03	1.11	1.05	-														
(17) GitHub	1.0	1.01	1.01	1.01	1.0	1.02	1.01	1.01	1.0	1.0	1.01	1.02	1.03	1.01	1.01	1.02	-													
(18) Reddit	1.0	1.01	1.02	1.03	1.0	1.05	1.01	1.01	1.0	1.03	1.03	1.03	1.05	1.03	1.02	1.03	1.1	-												
(19) Slack	1.0	1.0	1.01	1.04	1.0	1.02	1.0	1.0	1.0	1.0	1.01	1.02	1.0	1.0	1.04	1.02	1.01	1.01	-											
(20) Discord	1.0	1.0	1.01	1.01	1.0	1.01	1.0	1.0	1.0	1.0	1.01	1.01	1.0	1.0	1.0	1.0	1.01	1.01	1.0	-										
(21) Bitcointalk	1.0	1.02	1.0	1.0	1.0	1.06	1.01	1.01	1.0	1.04	1.02	1.0	1.05	1.01	1.0	1.08	1.04	1.1	1.02	1.0	-									
(22) Team size	1.0	1.0	1.03	1.03	1.01	1.02	1.0	1.01	1.0	1.03	1.01	1.02	1.04	1.01	1.03	1.04	1.02	1.03	1.0	1.0	1.01	-								
(23) Number of advisors	1.0	1.0	1.04	1.06	1.01	1.01	1.01	1.0	1.0	1.01	1.02	1.02	1.02	1.02	1.06	1.07	1.04	1.04	1.0	1.0	1.02	1.03	-							
(24) Expert rating	1.07	1.0	1.02	1.08	1.01	1.08	1.0	1.07	1.04	1.01	1.05	1.1	1.0	1.0	1.03	1.0	1.0	1.01	1.02	1.0	1.0	1.0	1.0	-						
(25) 2017	1.0	1.0	1.12	1.27	1.02	1.12	1.01	1.05	1.0	1.01	1.15	1.08	1.0	1.0	1.07	1.01	1.0	1.02	1.11	1.02	1.01	1.01	1.01	1.1	-					
(26) 2018	1.0	1.01	1.07	1.08	1.0	1.05	1.0	1.02	1.0	1.01	1.05	1.0	1.01	1.0	1.01	1.01	1.0	1.01	1.04	1.0	1.01	1.01	1.03	1.0	1.88	-				
(27) 2019	1.0	1.0	1.0	1.04	1.02	1.01	1.0	1.0	1.0	1.0	1.01	1.04	1.0	1.0	1.01	1.01	1.0	1.01	1.01	1.01	1.02	1.0	1.0	1.06	1.05	1.28	-			
(28) Data and AI	1.0	1.0	1.01	1.0	1.01	1.0	1.01	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	-		
(29) Entertainment and gaming	1.0	1.0	1.0	1.0	1.01	1.01	1.0	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.0	1.0	1.01	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.01	1.01	1.0	1.01	-	
(30) Financial Services	1.01	1.0	1.0	1.01	1.0	1.01	1.0	1.0	1.01	1.0	1.0	1.01	1.0	1.0	1.0	1.0	1.0	1.01	1.0	1.0	1.01	1.01	1.0	1.01	1.0	1.0	1.0	1.14	1.11	-

D

AUC-ROC

Figure D.1: ROC Curve logistic regression (model 1)

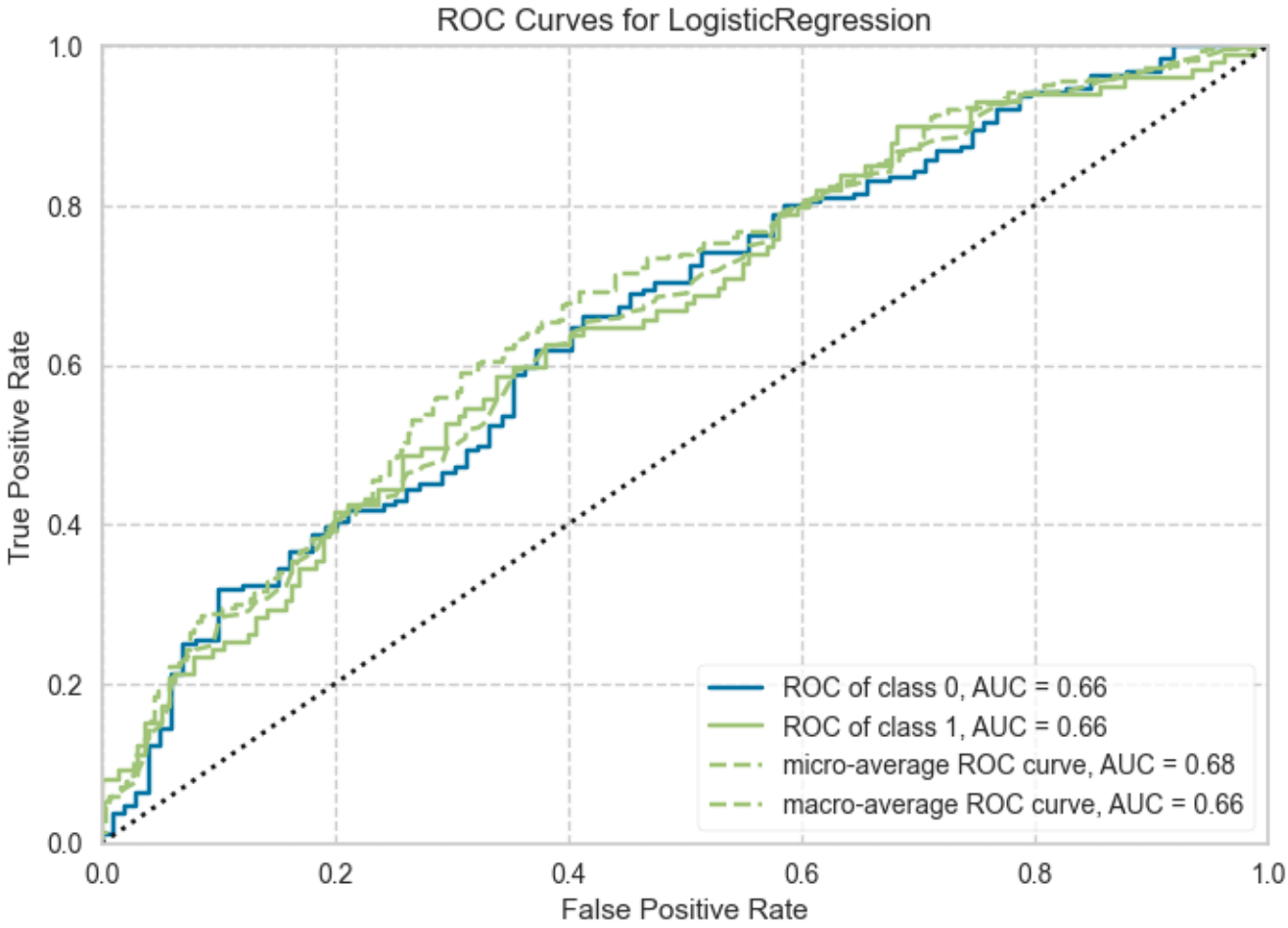
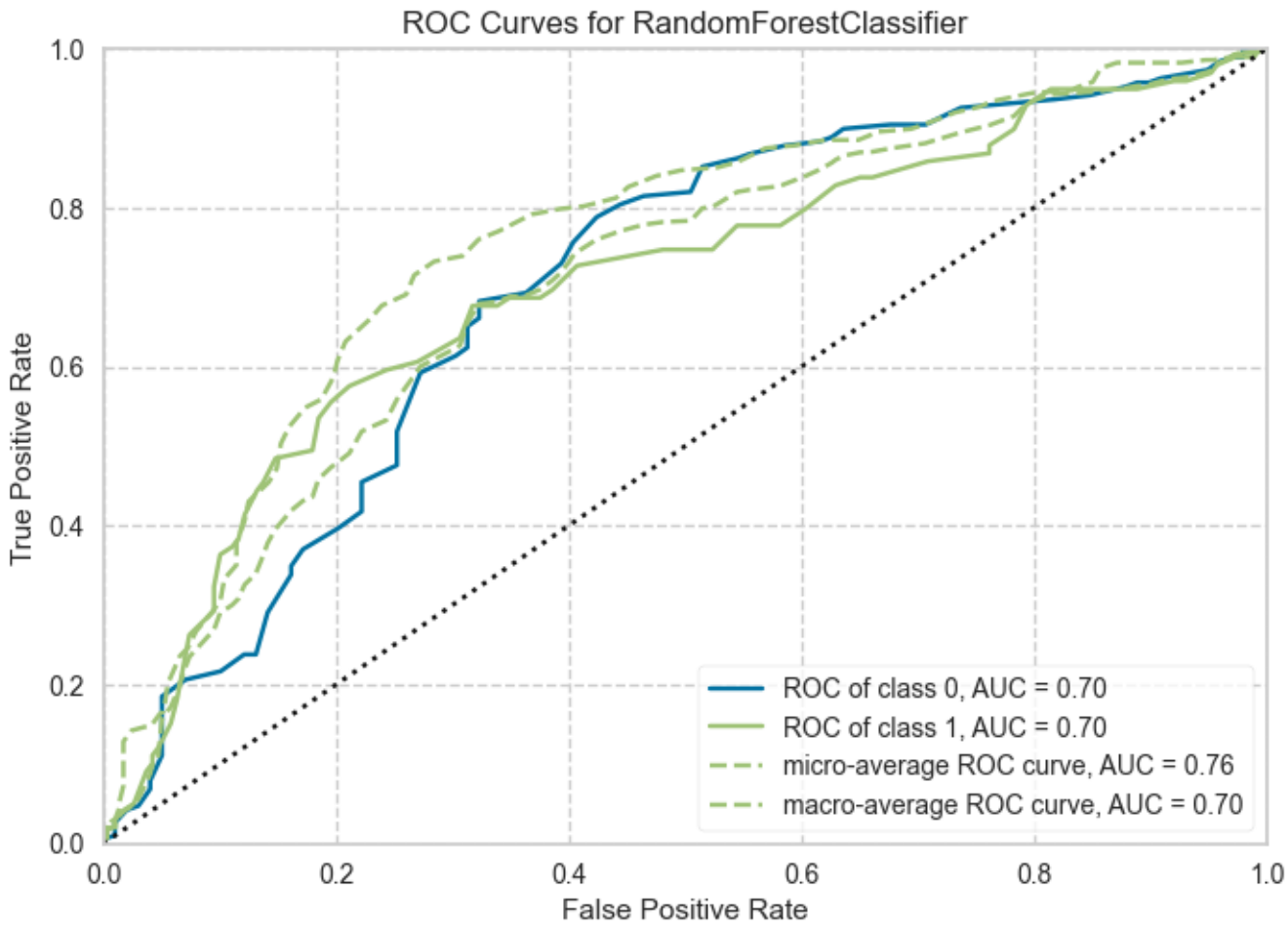


Figure D.2: ROC Curve Random Forests



E

Regression models

Table E.1: Logit regression results: Interaction effects of Financial Services and Entrepreneurship and Gaming on features

	Model 1	Model 2	Model 3	Model 4
Whitepaper	0.40	0.33	0.44	0.35
	(0.52)	(0.63)	(0.53)	(0.64)
2017	-0.38	-0.40	-0.35	-0.34
	(0.52)	(0.73)	(0.52)	(0.75)
2018	-0.16	-0.07	-0.13	-0.04
	(0.47)	(0.67)	(0.48)	(0.69)
2019	-0.28	-0.26	-0.26	-0.24
	(0.49)	(0.69)	(0.50)	(0.70)
Whitelist	0.19	0.15	0.22	0.20
	(0.21)	(0.30)	(0.21)	(0.30)
KYC	0.21	-0.18	0.21	-0.20
	(0.21)	(0.29)	(0.21)	(0.31)
ESG	-0.18	-0.19	-0.12	-0.10
	(0.50)	(0.72)	(0.50)	(0.73)
Bounty	-0.46**	-0.70**	-0.42**	-0.65**
	(0.20)	(0.29)	(0.20)	(0.30)
Bonus	0.17	0.57*	0.17	0.65*
	(0.22)	(0.31)	(0.23)	(0.33)
Hard cap	-0.34**	-0.04	-0.46***	-0.22
	(0.17)	(0.23)	(0.17)	(0.26)
Duration	-0.00*	-0.00	-0.00*	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Number of milestones	-0.05***	-0.05**	-0.05***	-0.05**
	(0.02)	(0.02)	(0.02)	(0.02)
Presale	-0.34*	-0.38**	-0.33*	-0.38**
	(0.18)	(0.19)	(0.18)	(0.19)
MVP	-0.10	0.51	-0.14	0.55
	(0.22)	(0.32)	(0.23)	(0.35)
Video	-0.06	-0.06	-0.08	-0.07

Table E.1 continued from previous page

	(0.19)	(0.26)	(0.19)	(0.26)
Twitter	0.11***	0.11**	0.11***	0.11**
	(0.03)	(0.04)	(0.03)	(0.04)
Telegram	0.05*	0.04	0.05*	0.04
	(0.03)	(0.04)	(0.03)	(0.04)
Facebook	-0.01	0.00	-0.01	0.00
	(0.02)	(0.03)	(0.02)	(0.03)
GitHub	-0.08	-0.15	-0.07	-0.14
	(0.16)	(0.23)	(0.16)	(0.23)
Reddit	0.23	0.30	0.25	0.33
	(0.18)	(0.25)	(0.18)	(0.25)
Slack	0.16	0.10	0.13	0.08
	(0.21)	(0.29)	(0.21)	(0.29)
Discord	-0.12	-0.01	-0.17	-0.02
	(0.29)	(0.41)	(0.29)	(0.42)
Bitcointalk	0.10	0.21	0.08	0.19
	(0.20)	(0.27)	(0.21)	(0.28)
Team size	0.03**	0.02	0.03**	0.02
	(0.01)	(0.02)	(0.01)	(0.02)
Number of advisors	-0.03	-0.04	-0.02	-0.03
	(0.02)	(0.03)	(0.02)	(0.03)
Expert rating	0.34***	0.51***	0.32***	0.51***
	(0.07)	(0.10)	(0.07)	(0.11)
Continent_AS	-0.17	0.06	-0.18	0.11
	(0.76)	(1.02)	(0.77)	(1.03)
Continent_EU	-0.71	-0.66	-0.70	-0.59
	(0.76)	(1.02)	(0.77)	(1.03)
Continent_NorthAm	-0.24	-0.02	-0.25	0.03
	(0.77)	(1.02)	(0.77)	(1.03)
Continent_Other	-0.62	-0.41	-0.62	-0.34
	(0.75)	(1.01)	(0.76)	(1.02)
Continent_UK	-0.68	-0.46	-0.63	-0.36
	(0.81)	(1.05)	(0.81)	(1.07)
Category_Data and AI	-0.20	-0.22	-0.19	-0.22
	(0.26)	(0.28)	(0.26)	(0.28)
Category_Entertainment and gaming	-0.42	-0.50*	-0.82	-0.99*
	(0.27)	(0.29)	(0.52)	(0.54)
Category_Financial Services	-0.50***	-1.07	-0.49***	-1.13
	(0.17)	(1.69)	(0.17)	(1.69)
Fin Services x Whitepaper		0.72		0.70
		(1.33)		(1.34)
Fin Services x 2017		-0.14		-0.20
		(1.07)		(1.08)
Fin Services x 2018		-0.41		-0.45
		(0.98)		(0.99)
Fin Services x 2019		-0.06		-0.07
		(1.01)		(1.02)
Fin Services x Whitelist		0.11		0.06

Table E.1 continued from previous page

	(0.43)		(0.43)
Fin Services x KYC	0.74*		0.76*
	(0.43)		(0.44)
Fin Services x ESG	0.21		0.13
	(1.04)		(1.04)
Fin Services x Bounty	0.56		0.51
	(0.41)		(0.42)
Fin Services x Bonus	-0.97**		-1.05**
	(0.46)		(0.48)
Fin Services x Hard cap	-0.79**		-0.61*
	(0.34)		(0.36)
Fin Services x Duration	0.00		0.00
	(0.00)		(0.00)
Fin Services x Milestones	-0.00		0.00
	(0.04)		(0.04)
Fin Services x MVP	-1.19***		-1.22***
	(0.45)		(0.47)
Fin Services x Video	0.03		0.04
	(0.40)		(0.40)
Fin Services x Twitter	0.00		0.01
	(0.07)		(0.07)
Fin Services x Telegram	0.01		0.01
	(0.05)		(0.05)
Fin Services x Facebook	-0.02		-0.02
	(0.05)		(0.05)
Fin Services x GitHub	0.13		0.13
	(0.34)		(0.34)
Fin Services x Reddit	-0.09		-0.13
	(0.37)		(0.37)
Fin Services x Slack	-0.12		-0.10
	(0.44)		(0.44)
Fin Services x Discord	-0.19		-0.18
	(0.59)		(0.60)
Fin Services x Bitcointalk	-0.35		-0.33
	(0.42)		(0.42)
Fin Services x Team size	0.02		0.02
	(0.03)		(0.03)
Fin Services x Number of advisors	0.02		0.02
	(0.04)		(0.04)
Fin Services x Expert rating	0.34**		0.33**
	(0.14)		(0.15)
Fin Services x Asia	-0.15		-0.12
	(0.47)		(0.47)
Fin Services x Europe	0.14		0.15
	(0.43)		(0.43)
Fin Services x North America	-0.12		-0.09
	(0.49)		(0.49)
Entert Gaming x Bonus		-0.01	-0.33

Table E.1 continued from previous page

			(0.87)	(0.93)
Entert Gaming x Expert rating			-0.24	-0.03
			(0.22)	(0.24)
Entert Gaming x Hard cap			1.28**	1.00*
			(0.57)	(0.60)
Entert Gaming x KYC			-0.45	-0.01
			(0.61)	(0.67)
Entert Gaming x MVP			0.50	-0.17
			(0.81)	(0.88)
N	957	957	957	957
Log-Likelihood	-540.44	-524.36	-537.14	-522.90
Pseudo R-Sq	0.1224	0.1485	0.1278	0.1509
Wald χ	-615.84***	-615.84***	-615.84***	-615.84***

Table E.2: Time dependency of signals: 2017 and 2018

	Model (1):	Model (2):	Model (3):	Model (4):
Whitepaper	0.40	1.08**	-0.29	1.87*
	(0.52)	(0.53)	(0.52)	(0.97)
2017	-0.38	-0.25	0.04	-0.07
	(0.52)	(1.04)	(0.50)	(1.05)
2018	-0.16	-0.75*	-1.06	-0.85
	(0.47)	(0.44)	(1.19)	(1.20)
2019	-0.28	-0.70	-0.45	-0.50
	(0.49)	(0.46)	(0.47)	(0.69)
Whitelist	0.19	0.17	0.14	0.45
	(0.21)	(0.22)	(0.44)	(0.61)
KYC	0.21	0.18	0.39	0.57
	(0.21)	(0.22)	(0.41)	(0.62)
ESG	-0.18	-0.30	0.56	2.27*
	(0.50)	(0.57)	(0.78)	(1.36)
Bounty	-0.46**	-0.41*	-0.00	0.75
	(0.20)	(0.21)	(0.41)	(0.62)
Bonus	0.17	-0.13	0.48	-0.12
	(0.22)	(0.25)	(0.37)	(0.59)
Hard cap	-0.34**	-0.48**	-0.40	-1.46**
	(0.17)	(0.20)	(0.26)	(0.60)
Duration	-0.00*	-0.00**	-0.00	-0.01
	(0.00)	(0.00)	(0.00)	(0.00)
Number of milestones	-0.05***	-0.07***	-0.05**	-0.09
	(0.02)	(0.02)	(0.03)	(0.06)
Presale	-0.34*	-0.33*	-0.38**	-0.32*
	(0.18)	(0.18)	(0.18)	(0.19)
MVP	-0.10	-0.06	-0.02	0.53
	(0.22)	(0.23)	(0.38)	(0.54)
Video	-0.06	-0.12	-0.15	-0.69
	(0.19)	(0.23)	(0.28)	(0.56)
Twitter	0.11***	0.07*	0.15***	0.21*
	(0.03)	(0.04)	(0.05)	(0.11)
Telegram	0.05*	0.05	-0.00	-0.12
	(0.03)	(0.03)	(0.04)	(0.08)
Facebook	-0.01	0.01	-0.02	-0.02
	(0.02)	(0.03)	(0.04)	(0.08)
GitHub	-0.08	-0.04	-0.04	0.20
	(0.16)	(0.20)	(0.25)	(0.53)
Reddit	0.23	0.13	0.37	-0.09
	(0.18)	(0.21)	(0.28)	(0.59)
Slack	0.16	0.45	-0.15	0.39
	(0.21)	(0.28)	(0.31)	(0.94)
Discord	-0.12	-0.21	0.19	0.24
	(0.29)	(0.30)	(0.46)	(0.62)
Bitcointalk	0.10	0.27	-0.02	0.19
	(0.20)	(0.24)	(0.32)	(0.64)

Table E.2 continued from previous page

	Model (1):	Model (2):	Model (3):	Model (4):
Team size	0.03**	0.03**	0.04	0.01
	(0.01)	(0.01)	(0.02)	(0.04)
Number of advisors	-0.03	-0.02	-0.06*	-0.08
	(0.02)	(0.02)	(0.03)	(0.07)
Expert rating	0.34***	0.41***	0.27**	0.56
	(0.07)	(0.08)	(0.12)	(0.38)
Continent Asia	-0.17	0.27	0.18	0.03
	(0.76)	(0.34)	(0.42)	(0.71)
Continent Europe	-0.71	-0.38	-0.40	-1.94**
	(0.76)	(0.33)	(0.38)	(0.78)
Continent NorthAm	-0.24	0.22	0.29	0.20
	(0.77)	(0.36)	(0.41)	(0.73)
Continent Other	-0.62	-0.21	-0.18	-0.42
	(0.75)	(0.34)	(0.33)	(0.37)
Category Data and AI	-0.20	-0.03	-0.38	-0.10
	(0.26)	(0.30)	(0.45)	(0.83)
Category Entertainment and gaming	-0.42	-0.34	-0.80**	-0.57
	(0.27)	(0.35)	(0.40)	(0.90)
Category Financial Services	-0.50***	-0.41**	-0.65**	-0.30
	(0.17)	(0.20)	(0.28)	(0.58)
2017 x Whitepaper		-2.64**		-3.46***
		(1.04)		(1.33)
2017 x Whitelist		-2.03		-2.30
		(1.88)		(1.96)
2017 x KYC		2.11		1.71
		(1.43)		(1.55)
2017 x ESG		0.62		-1.96
		(1.48)		(1.93)
2017 x Fin. Serv.		-0.60		-0.71
		(0.44)		(0.70)
2017 x Data and AI		-0.64		-0.57
		(0.76)		(1.08)
2017 x Entertainment and Gaming		-0.30		-0.07
		(0.60)		(1.02)
2017 x Bounty		-0.04		-1.19
		(0.99)		(1.15)
2017 x Bonus		1.64**		1.63*
		(0.70)		(0.88)
2017 x Hard cap		0.69*		1.66**
		(0.39)		(0.69)
2017 x Duration		0.00		0.00
		(0.00)		(0.00)
2017 x Milestones		0.02		0.04
		(0.04)		(0.07)
2017 x MVP		0.97		0.37
		(1.08)		(1.19)
2017 x Video		0.35		0.92

Table E.2 continued from previous page

	Model (1):	Model (2):	Model (3):	Model (4):
		(0.47)		(0.70)
2017 x Twitter		0.19**		0.13**
		(0.08)		(0.05)
2017 x Telegram		-0.02		0.15
		(0.06)		(0.10)
2017 x Facebook		-0.04		-0.01
		(0.06)		(0.09)
2017 x GitHub		-0.36		-0.60
		(0.41)		(0.64)
2017 x Reddit		0.73*		0.95
		(0.44)		(0.70)
2017 x Slack		-0.64		-0.58
		(0.46)		(1.00)
2017 x Discord		0.33		-0.17
		(1.68)		(1.75)
2017 x Bitcointalk		-0.85*		-0.80
		(0.51)		(0.78)
2017 x Team size		0.02		0.04
		(0.04)		(0.05)
2017 x Number of advisors		-0.07		-0.01
		(0.05)		(0.09)
2017 x Expert rating		0.35**		0.50*
		(0.17)		(0.41)
2017 x Asia		0.11		0.22
		(0.59)		(0.85)
2017 x Europe		0.60		2.03**
		(0.50)		(0.86)
2017 x North America		0.05		-0.08
		(0.58)		(0.85)
2018 x Whitepaper			1.35	-0.83
			(1.19)	(1.45)
2018 x Whitelist			-0.01	-0.33
			(0.50)	(0.66)
2018 x Video			0.15	0.66
			(0.39)	(0.62)
2018 x Twitter			-0.11	-0.16
			(0.07)	(0.12)
2018 x Telegram			0.09*	0.20**
			(0.05)	(0.09)
2018 x Team size			-0.00	0.02
			(0.03)	(0.05)
2018 x Slack			0.68	0.12
			(0.44)	(0.99)
2018 x Reddit			-0.14	0.33
			(0.38)	(0.64)
2018 x Number of advisors			0.04	0.06
			(0.04)	(0.08)

Table E.2 continued from previous page

	Model (1):	Model (2):	Model (3):	Model (4):
2018 x North America			0.01	-0.07
			(0.49)	(0.75)
2018 x Milestones			-0.02	0.02
			(0.04)	(0.07)
2018 x MVP			-0.07	-0.62
			(0.47)	(0.61)
2018 x KYC			-0.18	-0.36
			(0.49)	(0.67)
2018 x Hard cap			0.13	1.17
			(0.34)	(0.64)
2018 x GitHub			-0.10	-0.34
			(0.34)	(0.57)
2018 x Data and AI			0.23	-0.07
			(0.56)	(0.89)
2018 x Fin. Serv.			0.22	-0.13
			(0.36)	(0.62)
2018 x Facebook			0.04	0.04
			(0.05)	(0.08)
2018 x Expert rating			0.13	0.16
			(0.15)	(0.39)
2018 x Europe			0.40	1.78**
			(0.43)	(0.79)
2018 x Entertainment and Gaming			0.57	0.34
			(0.56)	(0.98)
2018 x ESG			-1.37	-3.09**
			(1.04)	(1.52)
2018 x Duration			-0.00	0.00
			(0.00)	(0.00)
2018 x Discord			-0.82	-0.88
			(0.60)	(0.73)
2018 x Bounty			-0.57	-1.33**
			(0.47)	(0.67)
2018 x Bonus			-0.67	-0.09
			(0.47)	(0.66)
2018 x Bitcointalk			0.24	0.01
			(0.42)	(0.70)
2018 x Asia			0.40	0.40
			(0.48)	(0.73)
N	957	957	957	957
Log-Likelihood	-540.44	-519.25	-526.84	-498.59
Pseudo R-Sq	0.1224	0.1569	0.1445	0.1904
Wald χ	-615.84***	-615.84***	-615.84***	-615.84***