Cryptocurrency Exchanges: Predicting which Markets will Remain Active

George Milunovich[1]

Department of Actuarial Studies and
Business Analytics
Macquarie University, Sydney

Seung Ah Lee

Department of Actuarial Studies and
Business Analytics
Macquarie University, Sydney

March 8, 2021

Abstract

About 99 percent of cryptocurrency trades occur on organised exchanges and many investors subsequently keep their digital assets in accounts with cryptocurrency markets. This generates exposure to the risk of exchange closures. We construct a database containing eight key characteristics on 238 cryptocurrency exchanges and employ machine learning techniques to predict whether a cryptocurrency market will remain active or whether it will go out of business. Both in-sample and out-of-sample measures of forecasting performance are computed and ranked for four popular machine learning algorithms. While all four models produce satisfactory classification accuracy, our best model is a random forest classifier. It reaches accuracy of 90.4 percent on training data and 86.1 percent on test data. From the list of predictors we find that exchange lifetime, transacted volume and cyber security measures such as security audit, cold storage and bug bounty programs rank high in terms of feature importance across multiple algorithms. On the other hand, whether an exchange has previously experienced a security breach does not rank highly according to its contribution to classification accuracy.

Key Words: Cryptocurrency Exchange, Remain Active, Closure, Classification, Machine Learning, Forecasting

[1] Corresponding Author: Department of Actuarial Studies and Business Analytics, Macquarie Business School, Macquarie University, NSW, 2109, Australia. Email: george.milunovich@mq.edu.au; Tel: +61 2 9850 8543.

1 Introduction

Cryptocurrencies are digital assets which can be transferred and used without an intermediary, such as a bank, and whose issuance is not under the control of any central authority. They are created via a process called *mining* and are managed by decentralised open source code. Cryptocurrencies transact on peer-to-peer (P2P) networks that enable any two parties to interact directly. Bitcoin – one of the first digital currencies – was introduced in Nakamoto (2008) and started trading in 2010. While economists argue both in favour of and against digital currencies[2], the global cryptocurrency market capitalisation has reached $1.59 trillion and currently transacts in excess of $135 billion daily[3].

Despite advocating for anonymity and decentralization the cryptocurrency market is in fact highly centralized. According to a U.S. Senate hearing about 99 percent of all cryptocurrency trades occur on centralized exchanges that are used to convert funds between national currencies and digital assets (Roubini, 2018). In addition, the top five exchanges account for more than 50 percent of all cryptocurrency transactions. Many investors also choose to keep their cryptocurrencies in accounts with cryptoexchanges which retain control of private keys which are required to transfer clients' digital assets (Hileman and Rauchs, 2017). Such practices create and amplify exposure to the risk of exchange closures. Indeed, several studies have now documented a high number of digital exchange closures during which investor funds are either fully or partially lost, see e.g. Oosthoek and Doerr (2020) and Moore et al. (2018).

We address the risk investors face from closures of digital exchanges by attempting to forecast such closures using publicly available data. If it is possible to accurately predict which markets will remain open and which ones will go out of business then investors can take this information into account and avoid exchanges that are likely to face closure. For this purpose we compile a database containing eight publicly available characteristics on 238 cryptocurrency markets and employ four popular machine learning (ML) techniques to train predictive algorithms. Our ML methods comprise the following classifiers: decision tree, random forest, logistic regression and support vector machine. Amongst few studies which investigate risk factors in the context of digital exchange closures, our study is the first attempt to assess classification performance on a test dataset. However, such out-of-sample analysis is crucial for understanding how well the models generalize to new data, and gauges how effectively investors can protect themselves by relying on generated forecasts. Further to measuring predictive ability we also investigate feature importance in order to identify key predictors that contribute to classification accuracy.

In order to build a database of useful predictors we draw on several papers from the current literature. For instance, Moore and Christin (2013) investigate determinants of digital exchange closures using a Cox proportional hazard model. They report that transaction volume is a statistically significant predictor that is inversely related to exchange closures. Experiencing a security breach, on the other hand, appears to have no statistically significant impact on the

[2] For example, IMF's He lists several advantages of crypto assets over fiat currencies He (2018). On the other hand, an article titled "Stiglitz, Roubini and Rogoff lead joint attack on bitcoin" argues against bitcoin (Newlands, 2018).

[3] As of February 19, 2021. Source: https://coinmarketcap.com.

probability of closures, although it is positively correlated with it. More recent results reported in Oosthoek and Doerr (2020) similarly suggest that there is a decreasing trend in exchange closures resulting from security breaches over the 2011 – 2019 time frame. Moore et al. (2018) employ longitudinal analysis to investigate 80 bitcoin exchanges established over the 2010 - 2015 period. They find that higher-volume exchanges are less likely to close and that experiencing a security breach in the same quarter increases the odds of closure. Their analysis includes several other predictors such as two-factor authentication and an anti-money laundering index for the countries in which exchanges operate. Interestingly they find that these additional features have no impact on the probability of exchange closures. Another study of interest is Johnson et al. (2018) who develop an economic model to capture the short-term incentives of cryptocurrency exchanges. In their model security investment plays a crucial role in determining profitability and risk levels of digital exchanges. We take these result into account by adding multiple cyber-security features to our list of predictor variables.

After compiling our database and splitting the sample into training and test datasets, we optimize machine learning algorithms and assess their classification performance. Our results may be summarised as follows. All four classifiers (decision tree, random forest, logistic regression and support vector machine) perform satisfactorily, and achieve classification accuracy ranging between 78.3 and 90.4 percent. While the random forest classifier ranks first according to three out of four performance criteria we employ, its advantage over competing algorithms is diminished when measured out-of-sample. This suggests that the recorded accuracy is driven by simple patterns that may be exploited by alternative algorithms. In terms of feature importance transacted volume, exchange lifetime and security audit rank high across all algorithms for which we are able to readily compute importances. Moreover, both random forest and logistic regression identify two additional cyber-security features – bug bounty and cold storage – as important predictors of which exchanges will remain active. On the other hand two-factor authentication and experiencing a previous security breach does not seem to significantly impact the probability of an exchange remaining active according to logistic regression and decision tree classifiers.

In the rest of the paper Section 2 describes our empirical methodology and Section 3 discusses our data on cryptocurrency exchanges. Section 4 provides empirical results and Section 5 concludes.

2 Methodology

Our empirical method consists of three steps: i) training and optimising ML algorithms, ii) evaluating in-sample (training dataset) and out-of-sample (test dataset) classification performance and ranking the algorithms according to their predictive ability, and iii) examining feature importance and determining which predictors contribute most to forecasting ability. We start by discussing the problem of predicting which digital exchanges will remain active and which ones will face closure.

2.1 The Prediction Problem

Our aim is to predict which cryptocurrency exchange will remain active and which will go out of business, given relevant predictor variables. This results in a classification problem where the target variable is defined as follows

$$y_i = \begin{cases} 1 & \text{if cryptocurrency exchange } i \text{ remains active} \\ 0 & \text{if cryptocurrency exchange } i \text{ has closed down.} \end{cases} \tag{1}$$

Forecasts of y_i are denoted $\hat{y}_i^a$ and are generated on the basis of eight available predictor variables $(x_{1i}, x_{2i}, \ldots, x_{8i})$ which are discussed in detail in Section 3. Thus, the forecasts are constructed according to the following equation

$$\hat{y}_i^a = \phi^a(x_{1i}, x_{2i}, \ldots, x_{8i}), \tag{2}$$

where ϕ^a is a function describing the relationship between the forecast and predictor variables that depends on which forecasting algorithm a is used.

While there is a plethora of prediction models one could employ, we decide to limit our investigation to four popular and commonly used ML algorithms. These are as follows: i) logistic regression, ii) decision tree, iii) random forest and iv) support vector machine. These algorithms are flexible and capable of capturing complicated relationships between the target and relevant features. A brief description of each classifier is provided next.

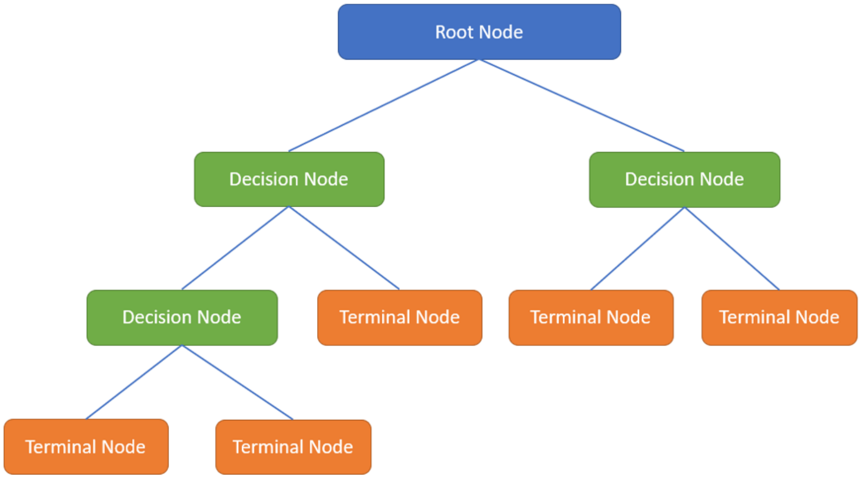
Logistic regression is one of the first, see e.g. Wilson and Worcester (1943), and most widely used methods to model binary dependent variables. It specifies the conditional probability of success given the vector of predictors x_i, as a sigmoid function of the following form $P(y_i = 1|x_i) = \frac{1}{1+e^{-w'x_i}}$ where w refers to the vector of weights, including the intercept. Logistic regression prediction are then generated as follows

$$\hat{y}_i^{LR} = \begin{cases} 1 & \text{if } P(y_i = 1|x_{1i}, x_{2i}, \ldots, x_{8i}) \geq 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Since we do not implement any regularization in the logistic regression we are also able to estimate standard errors, and thus gauge the statistical significance of the estimated coefficients.

Another popular classification algorithm is the decision tree classifier. It can build complex decision boundaries by dividing the feature space into rectangles. A decision tree consists of a root node, decision nodes and terminal nodes as illustrated in Figure 1. These nodes are formed by starting at the tree root and splitting the data on the feature that results in the largest information gain (IG). The splitting procedure is repeated at each decision node until the leaves either contain elements from only one class or by setting a limit for the maximal depth of the tree (which avoids overfitting). In our application we employ grid search cross-validation to optimise two hyperparameters i) tree depth, and ii) criterion used to compute IG.

Figure 1: A Decision Tree Classifier



A random forest classifier is an ensemble of decision trees. Random forests combine predictions of multiple decision trees by averaging across their estimated probabilities, thus reducing the degree of overfitting. A random forest of k trees may be constructed via the following algorithm:

i) Draw a random bootstrap sample of size n from the training dataset (with replacement);

ii) Grow a decision tree from the bootstrap sample:

  (a) Randomly select d features without replacement;

  (b) Build a tree using these d features;

iii) Repeat i) - ii) k times;

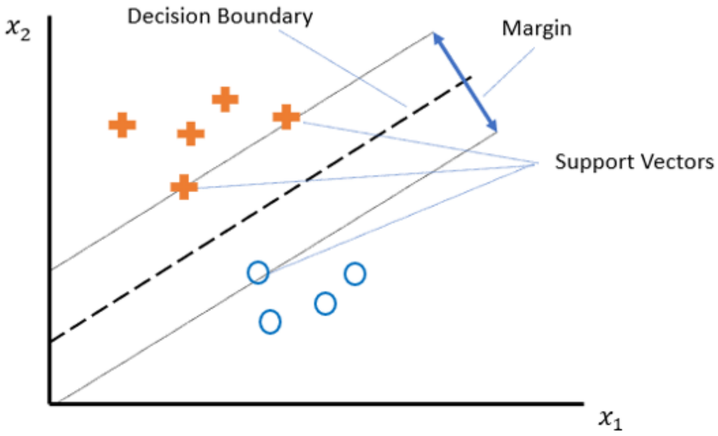
iv) Combine classifiers by averaging their probabilistic predictions. Assign class label according to greatest probability.

In our application we optimize three hyperparameters: i) maximum tree depth, ii) number of trees k, and iii) criterion used to compute IG.

Lastly, support vector machine classifier is an algorithm designed to be robust to outliers. It works by maximizing the margin which is the distance between the decision boundary and the training examples that are closest to the boundary, i.e. support vectors, as illustrated in Figure 2. We implement the algorithms with L2 regularization and optimize the regularization strength parameter. In addition, we cross validate the kernel function as a hyperparameter across the following values {linear, rbf, polynomial, sigmoid}.

In order to improve convergence properties of ML algorithms we normalize all continuous predictors as zero mean and unit variance variables. Binary indicator features are left unchanged. All algorithms are implemented in Python using scikit-learn libraries.

Figure 2: A Support Vector Machine Classifier



2.2 Training Algorithms and Performance Evaluation

In order to be able to assess forecasting performance on an independent dataset we divide our data into training and test subsamples. The training dataset contains 70 percent of the data (166 observations), while the test dataset consists of the remaining 30 percent (72 observations). When splitting the data we preserve the proportions of examples in each class by stratifying data according to the target variable. Moreover, alternative 80:20 and 60:40 splits between the training and test datasets result in similar classification accuracy.

The models are first trained and their hyperparameters optimized using the training dataset and K-fold cross-validation where $K = 10$. In the second step, we compute both in-sample (training dataset) and out-of-sample (test dataset) forecasting performance according to the following measures: i) classification accuracy, ii) precision, and iii) recall, and iv) F1 score. These metrics gauge somewhat different aspects of forecasting ability that cryptocurrency investors may care about. Denoting true positives as TP, true negatives as TN, false positives as FP and false negatives as FN we explain the four performance metrics as follows.

i) Accuracy $= \frac{\text{number of correctly classified examples}}{\text{sample size}} = \frac{TP+TN}{TP+TN+FP+FN}$. Classification accuracy is defined as the ratio of correctly predicted examples to the total sample size and is probably the most commonly used measure of classification performance. Nevertheless, it has a disadvantage that in situations where there is a class imbalance the model can predict the value of the majority class for all samples and still achieve a high classification accuracy.

ii) Precision $= \frac{TP}{TP+FP}$. Precision computes the ratio of true positives to all positively labelled (predicted) examples. It answers the question of how many exchanges actually remain active out of all the exchanges which are predicted to survive.

iii) Recall $= \frac{TP}{TP+FN}$. Recall is the ratio of correctly predicted positive example to all positive samples. It tells us what portion of all exchanges which remain active we classify as remaining active.

iv) F1 Score $= 2 \times \frac{\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$. F1 score is computed as the weighted average of precision and recall and aims to balance these two metrics.

3 Data and Descriptive Statistics

Our dataset comprises cross-sectional data on 238 exchanges collected for the June 2010 – January 2021 time period. We construct this database by collecting information from publicly available sources, as well as incorporating some of the data published in previous studies such as Moore et al. (2018) and Oosthoek and Doerr (2020). In particular, we obtain information on security breaches from online lists compiled by Hackernews (2019), Selfkey (2019) and Slowmist (2021) and from other various media reporting. Information on transaction volumes and exchange lifetimes is collected from online information portals and news websites such as coinmarketcap.com, coingecko.com, cryptowisser.com and coinpaprika.com. Lastly, each exchange's website is inspected for information on cyber-security programs and any additional relevant information. To view the websites of closed exchanges we rely on the Wayback Machine which is described as a digital archive of the World Wide Web (archive.org).

The target *active* is a binary variable signifying if an exchange remains active or has closed down, as defined in (1). Our list of predictors comprises eight features including i) *volume* – average daily traded volume in USD, ii) *lifetime* – exchange lifetime in days and iii) *breach* – a variable tracking whether there has been a security breach or not. Amongst other predictors are binary variables representing whether or not each of the following four cyber-security measures are implemented iv) *two-factor* authentication, v) *bug-bounty* program[4], vi) *security-audit*, and vii) *cold-storage*[5]. Nevertheless, not all exchanges provide information regarding all four security programs on their website. In such cases of missing data, and for the purpose of maximising our sample size, we take a conservative approach and code missing samples as 0, implying that the exchange for which the data is missing does not implement the security measure in question. This is a reasonable assumption given that cryptocurrency investors worry about cyber risks and that digital exchanges compete on the basis of implemented security features. Finally, our dataset is completed with a variable capturing the extent of financial regulation for each exchange's base country. This remaining predictor is viii) *aml/cft* – the anti-money laundering and combating the financing of terrorism index of Verdugo Yepes (2011). Where an exchange operates in multiple countries we take a conservative approach and classify it as operating in the country with the lowest *aml/cft*.

Table 1 provides some summary statistics for our dataset.

[4] Bug bounty is a program offered by websites and software developers by which individuals can receive recognition and compensation for reporting bugs, especially those pertaining to security vulnerabilities.

[5] Cold storage (cold storage wallet) is a hardware device used to store cryptocurrency that is kept offline, thus protecting the funds from unauthorized access, cyber attacks and other vulnerabilities to which a system that is connected to the internet is susceptible.

Table 1: Descriptive Statistics

	mean	std	min	25%	50%	75%	max
active	0.55	0.50	0.00	0.00	1.00	1.00	1.00
breached	0.26	0.44	0.00	0.00	0.00	1.00	1.00
two-factor	0.90	0.30	0.00	1.00	1.00	1.00	1.00
bug-bounty	0.31	0.46	0.00	0.00	0.00	1.00	1.00
security-audit	0.30	0.46	0.00	0.00	0.00	1.00	1.00
cold-storage	0.80	0.40	0.00	1.00	1.00	1.00	1.00
aml/cft	27.23	6.82	11.90	22.84	28.33	33.67	35.33
volume	311.18	870.00	0.00	0.09	16.84	187.47	7344.85
lifetime	1303.03	842.00	19.00	730.25	1035.00	1894.25	3502.00

Notes: Computations are based on a dataset comprising 238 cryptocurrency exchanges. Volume is measured in millions of USD.

As indicated by the first column of the table about 55 percent of the 238 cryptocurrency exchanges contained in our database remain active, while 26 percent of the exchanges have suffered some form of security breach. A majority of the exchanges implement two-factor authentication (90 percent of all exchanges) as well as cold storage facilities (80 percent). On the other hand, bug bounty and security audits are less commonly implemented by digital exchanges, with respective frequencies of 31 and 30 percent. The anti-money laundering and combating of financing of terrorism (*aml/cft*) index varies substantially and exhibits a mean value of 27.23 out of 49.

Lastly we discuss the properties of the *lifetime* and *volume* predictors. The average lifetime for our sample of cryptocurrency exchanges appears to be about 1303.03 days, with a minimum of 19 days and the maximum of 3502 days. Thus, some exchanges have been quite short lived. In addition, the standard deviation of *lifetime* is 842.00 days which is high relative to its mean. The mean daily *volume* is USD 311.18 million, and also varies substantially from USD 300.616 (displayed as 0.00 in USD millions) to USD 7344.85 million. Thus our dataset is highly heterogeneous in terms of exchange properties.

We next turn our attention to pairwise correlations presented in Table 2, which contribute to prediction accuracy discussed in the next section.

Table 2: Correlation Matrix

	active	breached	two-	bug-	security-	cold-	aml/cft	volume	lifetime
			factor	bounty	audit	storage			
active	1.00	-0.13	0.33	0.35	0.40	0.41	-0.12	0.28	0.37
breached	-0.13	1.00	-0.26	-0.01	0.07	-0.20	0.05	0.02	0.04
two-factor	0.33	-0.26	1.00	0.22	0.18	0.58	-0.12	0.11	0.30
bug-bounty	0.35	-0.01	0.22	1.00	0.34	0.22	0.06	0.18	0.05
security-audit	0.40	0.07	0.18	0.34	1.00	0.26	0.02	0.16	0.29
cold-storage	0.41	-0.20	0.58	0.22	0.26	1.00	-0.04	0.15	0.17
aml/cft	-0.12	0.05	-0.12	0.06	0.02	-0.04	1.00	-0.05	0.01
volume	0.28	0.02	0.11	0.18	0.16	0.15	-0.05	1.00	0.02
lifetime	0.37	0.04	0.30	0.05	0.29	0.17	0.01	0.02	1.00

Notes: Computations based on a dataset comprising 238 cryptocurrency exchanges.

Considering correlations between the target variable and various predictors provided in the first row of the table we observe that *security-audit* and *cold-storage* exhibit relatively large and positive correlations with *active* which are respectively 0.40 and 0.41. These are followed in magnitude by the correlations between *active* and *lifetime*, *bug-bounty* and *two-factor* predictors, which are each estimated to be 0.37, 0.35 and 0.33. Additionally *volume* also has a positive and moderate correlation with *active* at 0.28, while *breached* and *aml/cft* variables are negatively correlated with the target. Thus, it would appear that implementing cyber-security features, a longer trading track record and greater transaction volume are positively associated with exchanges which remain active. On the other hand, experiencing a security breach and operating in countries with greater emphasis on anti-money laundering efforts is negatively related to the active markets, although these two correlations are rather small in magnitude.

In the second row of Table 2 we see that experiencing a security breach is negatively related to *two-factor* and *cold-storage*, as expected. These correlations are however not large in magnitude with the estimates of -0.26 and -0.20. Rows 3 – 6 suggest that the four cyber-security measures are all positively correlated, with correlations raging from 0.18 between *security-audit* and *two-factor*, to 0.58 which is found between *cold-storage* and *two-factor*. While some of these correlations may present a difficulty in disentangling the effects of individual features on the target, i.e. multicollinearity, as discussed in Section 4.2, they have no adverse effect on the classification performance. Given that we aim to maximize the forecasting ability we decide to leave all four security features in the dataset. Lastly, while *volume* exhibits relatively low correlations with other predictors, *lifetime* seems to be moderately and positively correlated with *two-factor* and *security-audit.*

4 Empirical Results

We start by discussing classification performance results, which we then follow with the analysis of feature importance and a visualisation of the predictions.

4.1 Measuring Classification Performance

Table 3 presents four measures of in-sample classification performance which are computed using the training dataset[6].

First, we note that all four algorithms achieve satisfactory performance across different measures of classification ability. The best performing algorithm according to classification accuracy is random forest, which reaches in-sample accuracy of 0.904. In the second and third places are decision tree and support vector classifiers with accuracy metrics of 0.880 and 0.843, respectively. Lastly logistic regression attains the accuracy of 0.783. Thus, the difference between the highest and the lowest classification accuracy is about 0.12, i.e. 12 percent, using the training dataset.

Table 3: In-sample Forecasting Performance (Training Dataset)

Algorithm	Accuracy	Precision	Recall	F1 Score
Random Forest	0.904	0.895	0.934	0.914
Decision Tree	0.880	0.838	0.967	0.898
Support Vector	0.843	0.865	0.846	0.856
Logistic Regression	0.783	0.816	0.780	0.798

Notes: Metrics are computed from the training dataset consisting of 166 samples (70 percent of all data).

Although Table 3 sorts values according to classification accuracy, all four performance measures are largely consistent in their rankings. As evident from the top row of the table random forest ranks first in terms of accuracy, precision and F1 score. The only metric which ranks random forest in the second place is recall where decision tree classifier achieve the highest value.

Having explored in-sample classification performance we now turn to out-of-sample metrics provided in Table 4, which are computed on the basis of the test dataset. These results provide a better representation of the true predictive ability since the test dataset has not been used for the purpose of training the algorithms and optimizing hyperparameters. While the out-of-sample performance rankings are similar to what we observed using in-sample data, they are less uniform. This is particularly notable for alternative measures of performance such as recall, where discrepancies between different classifiers can reach up to 20 percent.

According to out-of-sample measures of classification performance, random forest ranks first according to two out of four criteria – accuracy and F1 score. In contrast, random forest's precision is 0.812 which places it in the fourth place, while the top position is now taken by the support vector classifier with the precision of 0.892. These value are, nevertheless, similar in magnitude. Lastly, according to recall the decision tree classifier is best with the recall of 1.000 while random forest comes second with the recall of 0.975. Decision tree classifier ranks second according to accuracy (sharing the second place with the support vector classifier), recall and

[6] While we split our data according to the 70:30 percent ratio between the training and test datasets (as noted in Section 2.2), alternative 60:40 and 80:20 splits result in similar classification performances and relative rankings.

Table 4: Out-of-sample Forecasting Performance (Test Dataset)

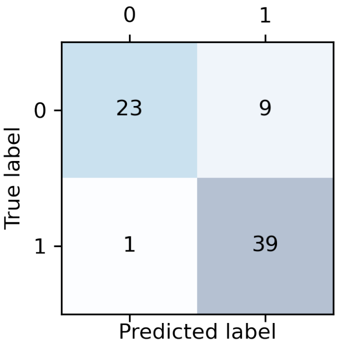
Algorithm	Accuracy	Precision	Recall	F1 Score
Random Forest	0.861	0.812	0.975	0.886
Decision Tree	0.847	0.784	1.000	0.879
Support Vector	0.847	0.892	0.825	0.857
Logistic Regression	0.819	0.865	0.800	0.831

Notes: Metrics are computed on test dataset consisting of 72 samples (30 percent of all data)

F1 score.

Next, we examine the results of our best classifier – random forest – in more detail using the confusion matrix produced in Figure 3.

Figure 3: Confusion Matrix – Random Forest Algorithm and Test Data



Out of the total of 72 samples contained in the test dataset, there are 40 exchanges which remain active (class 1) and 32 exchanges that have closed down (class 0). As can be seen from the second row, only 1 of the 40 active exchanges are misclassified as facing closure by the random forest classifier. This corresponds to the recall of 0.975 presented in Table 4. In contrast, of the 32 exchanges which went out of business we successfully predict 23 while random forest misclassifies 9 as remaining active. This results in a true negative rate (specificity) of 0.697. Thus, while we are able to separate the classes with high accuracy a certain amount of risk still prevails when predicting which exchanges will close down.

In order to gain further insight into the problem we consider which features contribute most to classification ability.

4.2 Feature Importance

Feature importance refers to the usefulness of predictors in forecasting the target variable. However, there is no single method to measuring feature importance for all algorithms. For instance our support vector classifier tackles potential nonlinearies via kernel methods such that even simple measures of feature importance, e.g. the magnitude of estimated weight coefficients[7], become unavailable. Nevertheless, three of the four classifiers we consider here provide measures of feature importance which we are able to easily compute. Should evidence from such multiple algorithms suggest that a certain feature is ”important” then we can have greater confidence in the impact of that predictor.

Figure 4: Feature Importance According to Random Forest Classifier

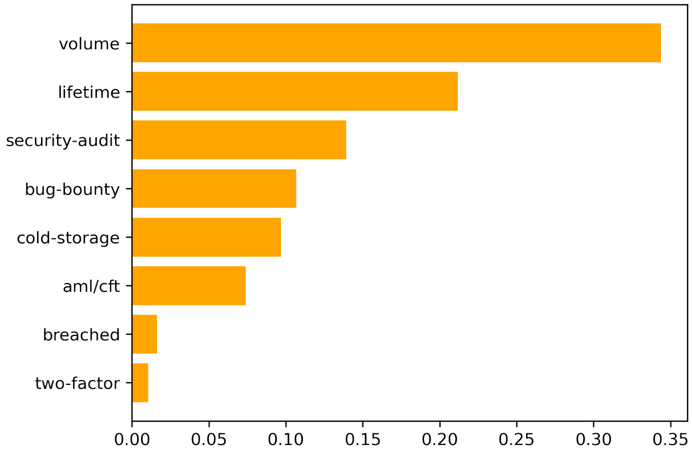


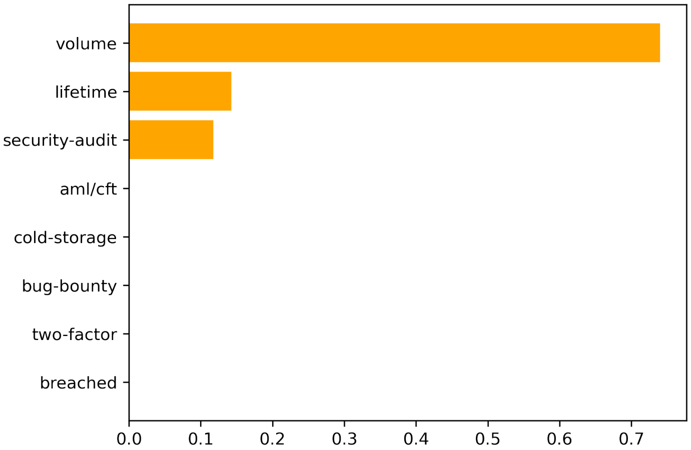
Figure 4 presents the estimates of Gini importance computed from the random forest classifier. These are calculated as normalised reductions in node impurity (Gini impurity) resulting from every feature and then averaged across all constituent trees. As evident from the figure, *volume* (in USD) appears to be the main predictor used in separating which exchanges will remain open and which markets will close. The second most important feature is *lifetime*. These two top predictors are followed by three cyber-security features, namely, *security-audit*, *bug-bounty* and *cold-storage*. The measure of anti-money laundering regulation in base countries *aml/cft* plays a smaller role, while *breached* and *two-factor* seem to have marginal impacts on the classification ability of random forest.

Considering feature importance according to the decision tree classifier in Figure 5 we confirm the importance of *volume*, *lifetime* and *security-audit* features. Here the importance of each feature is computed as the total reduction of Gini impurity resulting from that feature (similar to previosly discussed random forest feature importance). As depicted in the picture *volume* now plays a much more important role while *lifetime* and *security-audit* exhibit similar and smaller influences. The remaining five predictors play no role in the decision tree classifier.

Lastly we look at the logistic regression model. Table 5 reports marginal effects and their p-values which provide a different perspective on feature importance to what we discussed above.

[7] Considering the magnitude of estimated weight parameters provides information about feature importance when the features are measured on the same scale (standardised in some way).

Figure 5: Feature Importance According to Decision Tree Classifier



Marginal effects measure how the predicted probability of a binary outcome changes with a change in a risk factor. For instance, we can look at how the probability of remaining active changes with a 1-unit increase in (normalized) volume, or for an exchange with security audit versus an exchange without it. Using this approach, we can comment both on the magnitude of the impact of each predictor, i.e. the size of the marginal effect and their statistical significance.

Considering the p-values reported in the last column of the table, we see that *volume*, *cold-storage*, *bug-bounty*, *security-audit* and *lifetime* all exhibit statistically significant coefficients at the 5 percent level. All of these features also exhibit positive parameters, implying that they increase the probability of remaining active. For instance, increasing the (normalised) *volume* feature by one unit will increase the probability of remaining active by 0.414, while a 1-unit increase in (normalized) *lifetime* will result in a 0.117 change in the same probability. Of the binary variables, we see that implementing *cold-storage*, *bug-bounty* and *security-audit* respectively result in 0.211, 0.194 and 0.194 increases in the probability of remaining in business. We can see that these five variables also played an important role in the random forest classifier, while the decision tree classifier was stricter, identifying a three-variable subset in feature importance analysis.

The remaining features, i.e. *two-factor*, *aml/cft* and *breached*, are not statistically significant at any conventional level of significance. However, as we can see *two-factor* variable has a positive estimate coefficient while *aml/cft* and *breached* exhibit negative coefficients, as expected. The insignificance of *two-factor* could be due to the relatively high correlations between this variable and other security features making it difficult to disentangle individual effects (see Table 2). Another possible explanation, is that the lack of two-factor authentication may compromise user accounts of individual investors but not the viability of the exchange as a whole. The same explanation may be hypothesized for the large p-value on *breached*, which in fact identifies any type of security breach no matter how large or small the breach may be.

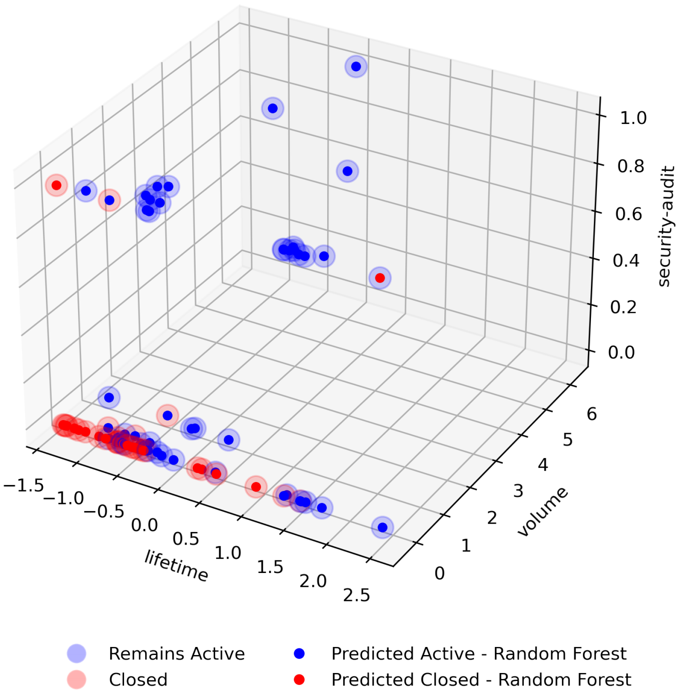
Table 5: Marginal Effects Estimated by Logistic Regression

	dy/dx	Std. Err.	z-stat.	p-value
volume	0.414	0.126	3.272	0.001
cold-storage	0.211	0.091	2.309	0.021
bug-bounty	0.194	0.058	3.324	0.001
security-audit	0.194	0.063	3.069	0.002
lifetime	0.117	0.030	3.942	0.000
two-factor	0.105	0.175	0.602	0.547
aml/cft	-0.039	0.029	-1.336	0.182
breached	-0.099	0.073	-1.351	0.177

Notes: The columns present i) marginal effects, ii) standard errors, iii) z-statistics and iv) *p*-values.

Lastly, we plot the predicted and realised samples from the test dataset against the three features which are designated as being important by multiple algorithms – namely *volume*, *lifetime* and *security-audit*. While the presented predictions are generated using all eight features, the visualisation illustrates the relationship between the forecasts and the plotted predictors in a 3-dimensional subspace. As can be seen from the graph most of the predictions (smaller solid circles) fall inside the realized data samples (larger transparent circles). More importantly, the colors of the larger and smaller circles mostly match indicating a high degree of classification accuracy.

Figure 6: Classification 3-dimensional Subspace



5 Conclusion

A large majority of investors conduct their cryptocurrency trades on organised digital exchanges despite having the possibility of peer-to-peer trading. Additionally, many investors also keep their cryptocurrencies in accounts with the exchanges, thus charging them with the safekeeping of their digital assets. These practices create exposure to the risk of digital exchange closures.

In this paper we compile a database containing eight publicly available characteristics on 238 cryptocurrency exchanges, 107 of which have closed since 2010. Using the collected data we build machine learning models to predict which digital markets will remain open and which will face closure. For the prediction task we employ four popular machine learning classifiers comprising i) decision tree, ii) random forest, iii) logistic regression and iv) support vector machine. Finally, we rank the alternative algorithms according to four different measures of classification performance, identify key predictor variables and visualize some predictions.

Our best algorithm is a random forest classifier which reaches in-sample classification accuracy of 0.904 (on training data) and out-of-sample accuracy of 0.861 (on independent test data). Nevertheless, all four classification methods accomplish relatively good performance with the minimum classification accuracy of 0.783 across all different algorithms. In-sample estimates of precision (0.895) and F1 score (0.914) also favour random forest, while the recall metric places decision tree classifier to the top (0.895) and random forest second. While in-sample estimates of accuracy show more variation across different classifiers than their out-of-sample counterparts, the estimates of recall are more disperse when measured out-of-sample. This highlights the importance of considering alternative measures of performance when comparing different classifiers.

From the list of eight exchange characteristics, average traded volume, exchange lifetime and security audit are found to be key predictors across multiple classifiers. Two additional cyber-security features, namely bug bounty and cold storage programs are also found to be important according to our random forest and logistic regression classifiers. On the other hand, experiencing a security breach, having two factor authentication and the extent of anti-money laundering regulation in the countries where the exchange operate does not seem to have a significant impact on the probability of remaining active.

While summarising risk factors inferred from several different models is not straightforward, we conclude that investors may be able to reduce their risk of digital exchange closures by trading on the markets which record relatively high transaction volumes, have a long track record of trading and implement multiple security features. Nevertheless, our results also show that a certain level of risk remains even after accounting for all the exchange characteristics considered in this paper. Traders should therefore aim to stay informed of any other pertinent information, as well as consider transferring their digital assets from organized exchanges to their own cryptocurrency wallets.

References

Hackernews (2019). A huge list of cryptocurrency thefts. [online]. Available: https://hackernoon.com/a-huge-list-of-cryptocurrency-thefts-16d6bf246389.

He, D. (2018). Monetary policy in the digital age. *Finance and Development*, 55(2):13–16.

Hileman, G. and Rauchs, M. (2017). Global cryptocurrency benchmarking study. *Cambridge Centre for Alternative Finance*, 33:33–113.

Johnson, B., Laszka, A., Grossklags, J., and Moore, T. (2018). Economic analyses of security investments on cryptocurrency exchanges. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 1253–1262. IEEE.

Moore, T. and Christin, N. (2013). Beware the middleman: Empirical analysis of bitcoin-exchange risk. In *International Conference on Financial Cryptography and Data Security*, pages 25–33. Springer.

Moore, T., Christin, N., and Szurdi, J. (2018). Revisiting the risks of bitcoin currency exchange closure. *ACM Transactions on Internet Technology (TOIT)*, 18(4):1–18.

Nakamoto, S. (2008). A peer-to-peer electronic cash system. Retreaved from https://bitcoin.org/bitcoin.pdf.

Newlands, C. (2018). Stiglitz, Roubini and Rogoff lead joint attack on bitcoin. *Barron's*.

Oosthoek, K. and Doerr, C. (2020). From hodl to heist: Analysis of cyber security threats to bitcoin exchanges. In *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–9. IEEE.

Roubini, N. (2018). Exploring the cryptocurrency and blockchain ecosystem. *Testimony for the Hearing of the US Senate Committee on Banking, Housing and Community Affairs*.

Selfkey (2019). Comprehensive list of cryptocurrency exchange hacks. [online]. Available: https://selfkey.org/list-of-cryptocurrency-exchange-hacks/.

Slowmist (2021). Exchange : 94 hack event(s). [online]. Available: https://hacked.slowmist.io/en/?c=Exchange.

Verdugo Yepes, C. (2011). Compliance with the aml/cft international standard: Lessons from a cross-country analysis. *IMF Working Papers*, pages 1–75.

Wilson, E. B. and Worcester, J. (1943). The determination of ld 50 and its sampling error in bio-assay. *Proceedings of the National Academy of Sciences of the United States of America*, 29(2):79.