Cryptocoins market analysis

Lúcia Moreira, May13th 2019

Summary

The present report addresses the study of the crypto-coins market daily trend prediction. A binary classification problem was considered. The choice of the present data set is related to some interest by the author in market analysis and cryptocoins, moreover this database allows the practical study of some of the theoretical approaches considered in the discipline. Pre-processing of the data set allowed to conclude that the different variables created have a normal distribution. A random distribution of the different market trends is a quite common feature of stock markets information and due to this random nature is quite difficult to predict market trend. Several methodologies were considered: logistic linear regression, Support Vector Machines, Fischer discriminant analysis, Linear (LDA) and Quadratic (QDA) discriminant analysis, Mixture models with EM, k-near neighbors (knn) and Naïve Bayes. The best model is presented and discussed in terms of Bayes error. The ROC curves for the best fittings are also presented.

The dataset

The dataset was obtained from kaggle.com and contains the cryptocurrency market history from 2013 to 2018 for 887 different digital coins. The dataset englobes a total of 632218 entries for the open, higher, lower and close daily values as well as the 24h volume traded and the daily market capitalization (38 MB information) for each of the crypto coin considered. The dataset also includes not available information (“NA”) from 69712 entries (in majority in chronological order) that only miss volume and/or market capitalization information. Daily information from incomplete data rows were removed from the data set (ca. 11 % total) and this approach is considered to affect minimally the past information from a single day once usually sets of chronological information would be missing thus minimally affecting the information from the near previous days. With this removal the total number of different coins reduced from 887 to 765 coins. Further processing of the dataset reduced this number to 568 coins (see below) as the amount of coins introduced in this analysis.

Pre-processing of the data

Figure 1 shows the annual volume of total crypto-coins transacted and the logarithm of the annual High value per coin. Fig. 1 shows that the interest in crypto coins has indeed increased substantially in the more recent years and that their quoted value vary substantially between each coin.

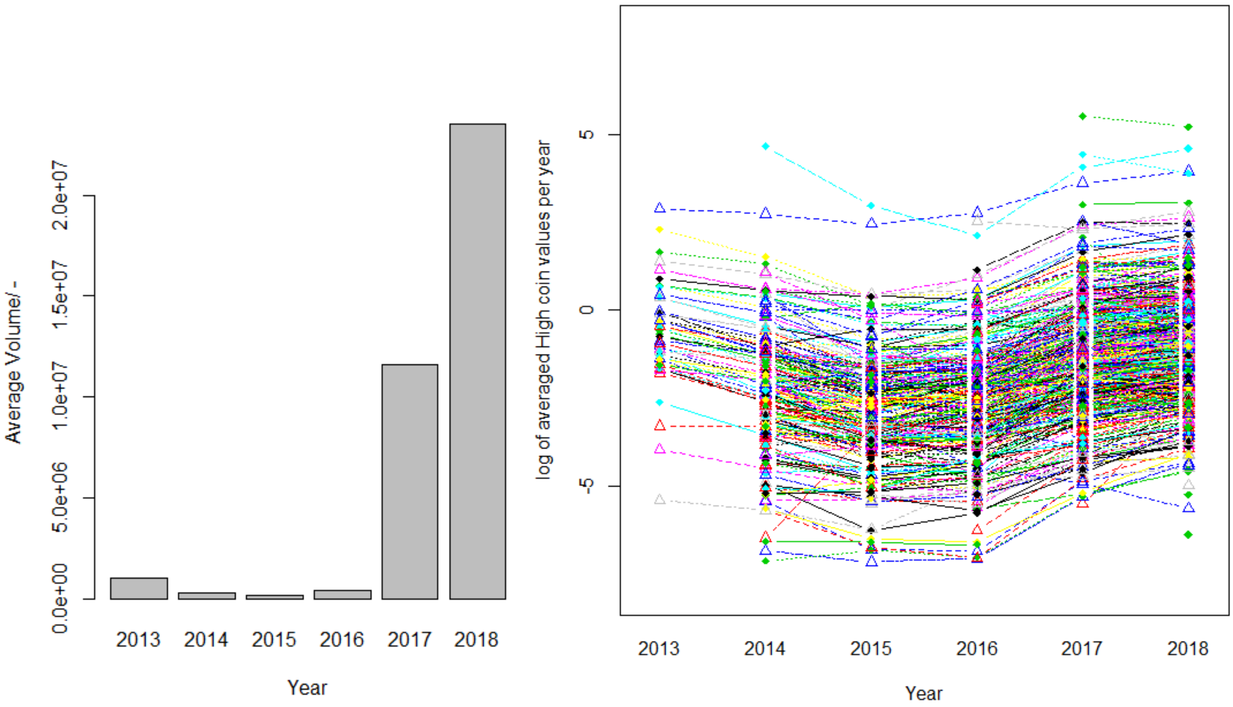

Fig. 1 – Averaged volume traded per year for all the 765 coins (left). Log version of the average High value per year for all the 765 coins (right).

Several additional variables were created taking into account the registered information in order to predict if the market trend of a particular day would increase or decrease (a binary classification problem) depending on the information available from the near previous days. For that, several percentage variations for each coin were created: the daily percentage variation for the value of the coin (calculated from the daily averaged coin values – the mean between open, high, low and close values), the daily percentage change of the volume traded and the daily percentage change for the market capitalization. These percentage calculations were also a way to normalize the values for the different coins. For predicting the increasing or decreasing class of the coin value trend, percentage variations from the previous day as well as up to 4 days before were used in different combinations in the considered models.

Bottom boxplots in Fig.2 show the average of the history of: 1-the daily percentage variation for 568 coins (left), 2- the daily percentage change of volume traded for 568 coins (middle) and 3- the daily percentage change for the market capitalization (right) for 568 coins. We can see (bottom left) that most common percentage variations for the daily value trend are around zero (some days increase while other days decrease a few percent) but some trends are substantially higher in the positive trend. Actually, coins with astronomic trends higher than a 1000% averaged history variation were removed from the data set to have a more uniform distribution of the values. The upper boxplots are the log version of the averaged values (then negative trends are not included) to give a better insight of the magnitude order of some percentage variations.

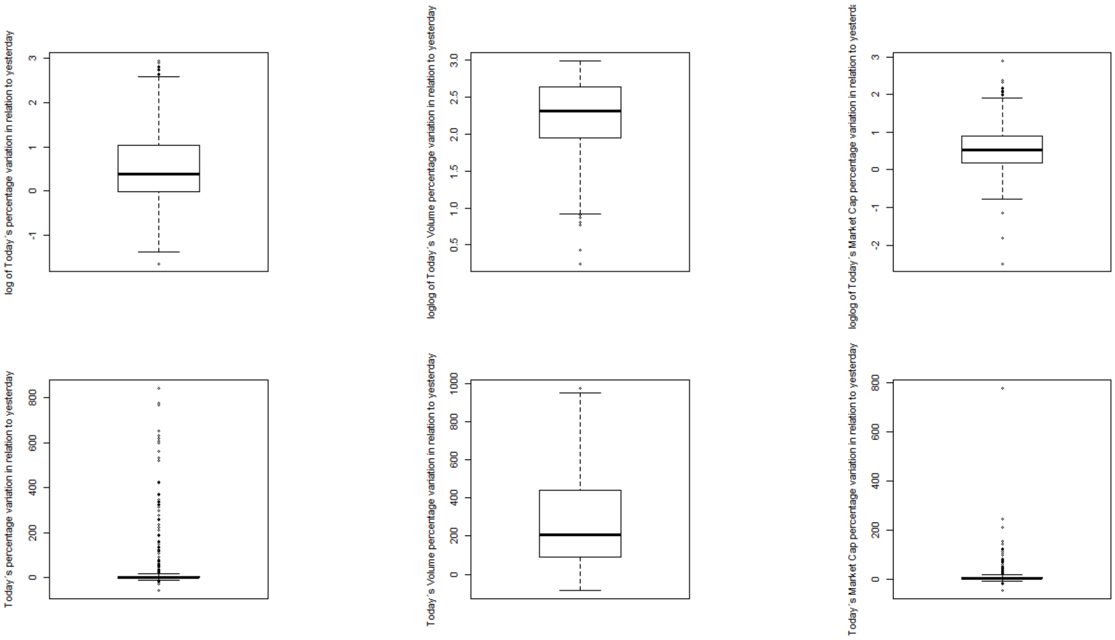

Fig. 2 - (Bottom) Boxplots with the averaged daily percentage variation for coin value (left), volume traded (middle) and market capitalization(right) for all the 568 different coins and during the history herein considered. (Top) Boxplots of the respective logarithmized versions (negative percentage variations thus not included. These are not highly representative of the presented values as this particular dataset contains only very high positive trends and not very high negative trends, probably related with the recent high increase in crypto coins mining).

Fig. 3 shows the histogram of some chosen variables. Variables have a normal distribution around mean with exception of the Volume based variables that may not be considered totally normal. As all the other variables are linear transformations of these variables so their distributions are also normal. These linear transformations (mainly related with calculating averages, differences and percentages) were performed in Excel and a new file was created in order to save multiple scans to the very big database (original file, excel file and text version of the file used are herein attached) and then being less time consuming to run the script.

A trend with mean around zero and a random distribution is a quite common feature of stock markets information and due to this random nature is quite difficult to predict market trend. Those who succeed have a powerful tool for valuing correctly their stock units and so substantially can make profit. This dataset was particularly chosen by the natural interest of the author in deeper knowledge about stock markets and crypto coins as well as its random nature that can be used to study some of the classification/regression methods learned in the discipline high based on probabilities information.

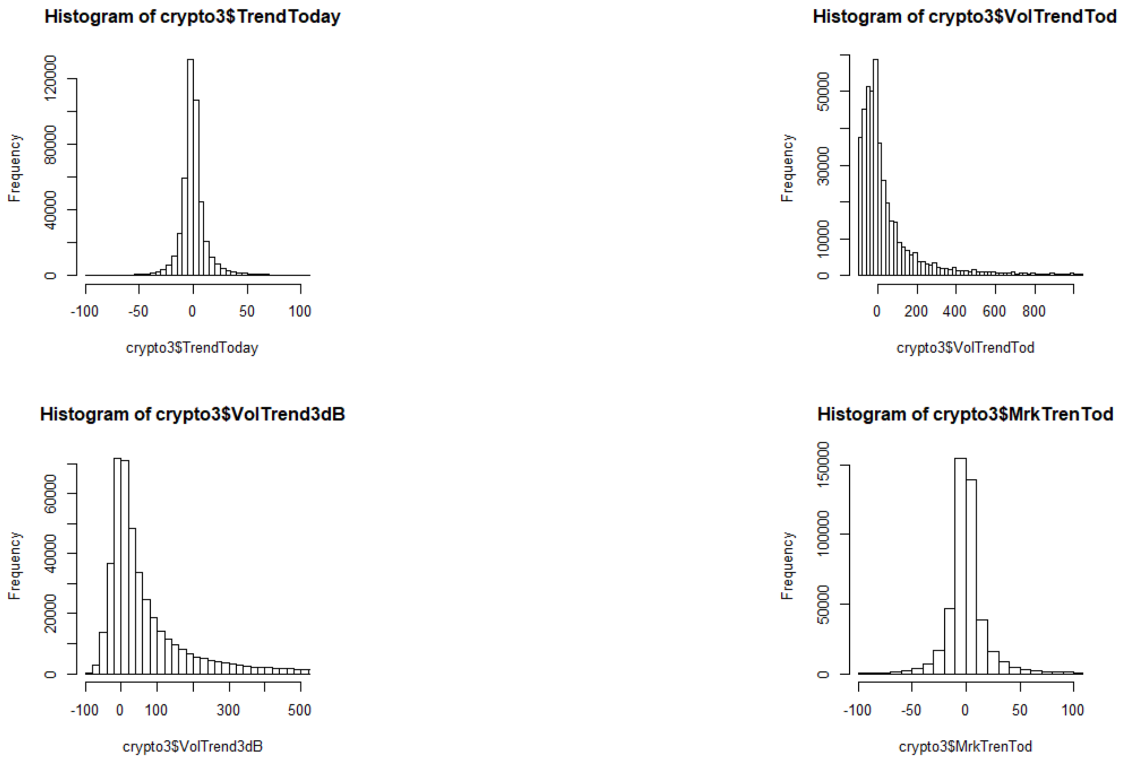

Fig. 3 – Histogram of some of the variables created showing a normal distribution around zero for most of the variables. (Top left) Variable *TrendToday*: represents the daily percentage change of the mean daily value of the coin for all the 568 coins. (Top right) Variable *VolTrendTod*: represents the daily percentage change of the volume traded for all the coins. (Bottom left) Variable *VolTrend3dB*: represents the mean of the percentage change of the volume traded during the last 3 days for all the coins. (Bottom right) Variable *MrkTrendTod*: represents the daily percentage change of the market capitalization value for all the 568 coins.

Classification approach

A binary value was created considering if the trend of a particular day represents the increase of the coin value in relation to the previous day, positive (+1), or the decrease of the coin value in relation to the previous day, negative (-1). The dataset thus contains ca. 54 % of negative trends and 46 % of positive trends. The classification problem was to predict correctly the daily trend of a particular day based on the information from the nearest previous days.

A regression approach was also considered taking into account predicting the value of the trend of a particular day taking into account previous days' information but fitting was extremely poor (not shown). A 3-classes approach was also considered using an invariant trend classified as 0 (stable), but classification results were also very poor (not shown).

The methods used to classify the binary trend of the daily coin value were: logistic linear regression, Support Vector Machines, Fischer discriminant analysis, Linear (LDA) and Quadratic (QDA) discriminant analysis, Mixture models with EM, k-near neighbors (knn) and Naïve Bayes. The dataset was divided in a 70% random split for training and the remaining for test.

Table 1 shows the correlation matrix between all the variables. It can be seen that all variables are mostly poorly correlated. However, from the original variables: Open, High, Low and Close variables are highly correlated to each other and Volume and Market Cap are also pretty correlated with each other and in a less extent with the coin values. It can also be seen that the created variables with percentage variations do not correlate with the original variables. In its turn, percentage change of the coin value correlate, although poorly, with the percentage variations from the previous days and with the moving percentage average from the close previous days being more intense the correlation when the same previous days are the same. For instance, the moving average for the two days before trend (Trend2dB) correlates well (0.71) with the 2 days before percentage variation value (Lag2d), once the moving average was calculated using that value. Also, volume trends for the different previous days also correlates with the nearest preceding volume trends. The same happens with market capitalization related variables (please see the Excel file for better clarification on how variables were calculated if necessary).

Discriminant Analysis

- Logistic linear regression

The summary of the linear regression using the value, volume and market cap trends from up to 3 days before is presented below:

```
Call:
glm(formula = dummiesTrend_3 ~ TrendYest + Trend2dB + Trend3dB +
    VolTrYest + +VolTrend2dB + VolTrend3dB + MrkTrenYest + MrkTren2dB
+ MrkTren3dB, data = trainCrypto2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8031  -0.9212  -0.9193   1.0788   2.9867

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.881e-02  1.778e-03 -44.328  < 2e-16 ***
TrendYest   -1.589e-06  7.126e-07  -2.230   0.0257 *
Trend2dB    -6.710e-07  1.360e-06  -0.493   0.6217
Trend3dB     3.130e-06  1.412e-06   2.216   0.0267 *
VolTrYest    3.418e-06  5.409e-07   6.319 2.63e-10 ***
VolTrend2dB  2.317e-07  1.087e-06   0.213   0.8312
VolTrend3dB -4.679e-06  1.177e-06  -3.977 6.99e-05 ***
MrkTrenYest -2.284e-06  2.643e-06  -0.864   0.3875
MrkTren2dB  -3.169e-06  6.231e-06  -0.509   0.6111
MrkTren3dB  -2.868e-06  7.064e-06  -0.406   0.6847
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.9935551)

    Null deviance: 316389  on 318384  degrees of freedom
Residual deviance: 316323  on 318375  degrees of freedom
AIC: 901491

Number of Fisher Scoring iterations: 2
```

Tabela 1 – Correlation matrix between all variables.

	Open	**High**	**Low**	**Close**	**Volume**	**MarketCap**	**TrendToday**	**TrendYest**	**Trend2dB**	**Trend3dB**	**Trend4dB**	**VolTrYest**	**VolTrendTod**	**VolTrend2dB**	**VolTrend3dB**	**MrkTrenTod**	**MrkTrenYest**	**MrkTren2dB**	**MrkTren3dB**	**Lag2d**	**Lag3d**	**LagVol2d**
Open	**1,00**	0,99	1,00	1,00	0,51	0,58	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
High	0,99	**1,00**	0,99	0,99	0,50	0,57	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Low	1,00	0,99	**1,00**	1,00	0,51	0,59	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Close	1,00	0,99	1,00	**1,00**	0,51	0,58	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Volume	0,51	0,50	0,51	0,51	**1,00**	0,87	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MarketCap	0,58	0,57	0,59	0,58	0,87	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TrendToday	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,33	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,00	0,00
TrendYest	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**1,00**	0,71	0,58	0,50	0,26	0,00	0,19	0,15	0,00	0,10	0,07	0,06	0,00	0,00	0,00
Trend2dB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,71	**1,00**	0,82	0,71	0,19	0,00	0,26	0,22	0,00	0,07	0,10	0,08	0,71	0,00	0,19
Trend3dB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,58	0,82	**1,00**	0,87	0,15	0,00	0,21	0,26	0,00	0,06	0,08	0,10	0,58	0,58	0,15
Trend4dB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,71	0,87	**1,00**	0,13	0,00	0,19	0,23	0,00	0,05	0,07	0,09	0,50	0,50	0,13
VolTrYest	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,26	0,19	0,15	0,13	**1,00**	0,00	0,71	0,58	0,01	0,15	0,11	0,09	0,00	0,00	0,00
VolTrendTod	0,00	0,00	0,00	0,00	0,00	0,00	0,33	0,00	0,00	0,00	0,00	0,00	**1,00**	0,00	0,00	0,11	0,00	0,00	0,00	0,00	0,00	0,00
VolTrend2dB	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,19	0,26	0,21	0,19	0,71	0,00	**1,00**	0,82	0,01	0,11	0,16	0,13	0,19	0,00	0,71
VolTrend3dB	-0,01	-0,01	-0,01	-0,01	0,00	0,00	0,00	0,15	0,22	0,26	0,23	0,58	0,00	0,82	**1,00**	0,00	0,09	0,13	0,16	0,15	0,15	0,57
MrkTrenTod	0,00	0,00	0,00	0,00	0,00	0,00	0,10	0,00	0,00	0,00	0,00	0,01	0,11	0,01	0,00	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00
MrkTrenYest	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,10	0,07	0,06	0,05	0,15	0,00	0,11	0,09	0,00	**1,00**	0,71	0,58	0,00	0,00	0,01
MrkTren2dB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,10	0,08	0,07	0,11	0,00	0,16	0,13	0,00	0,71	**1,00**	0,82	0,07	0,00	0,11
MrkTren3dB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,08	0,10	0,09	0,09	0,00	0,13	0,16	0,00	0,58	0,82	**1,00**	0,06	0,06	0,09
Lag2d	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,71	0,58	0,50	0,00	0,00	0,19	0,15	0,00	0,00	0,07	0,06	**1,00**	0,00	0,26
Lag3d	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,58	0,50	0,00	0,00	0,00	0,15	0,00	0,00	0,00	0,06	0,00	**1,00**	0,00
LagVol2d	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,19	0,15	0,13	0,00	0,00	0,71	0,57	0,00	0,01	0,11	0,09	0,26	0,00	**1,00**

This model produced a very poor fitting to the training data set and the same happened with other variable combinations (mostly predicted a class 0!, the middle). So a very poor classification algorithm for this problem.

- LDA, QDA and Fisher DA

LDA and QDA were also considered for different sets of variable combinations. QDA performed a little better than LDA. Despite that, results with discriminant analysis methodologies were quite poor for predicting the class in the train dataset (ca. 54 % accuracy and then only 4 percentage points above random guessing). In its turn, Fischer DA using the *lfda()* function from the "lfda" package was not possible to be evaluated because the system was not able to allocate GB-sized vectors.

Naïve Bayes

This algorithm was used in several sets of non- (or very poorly-) correlated variables for class prediction. Accuracy was in the order of 46% for the train set and worse than a simple random guessing!

k-NN

K-nearest neighbors algorithm gave the highest train accuracies using 5 neighbors: 72.4%. In its turn, test accuracies are ranging from 54.7% up to 56 % with 3 neighbors and different combinations of variables. A maximum test accuracy of 57.2 % was obtained when using 5 neighbors. This algorithm was very time consuming due to the very large dataset herein considered.

Mixture analysis with EM algorithm

The Gaussian finite mixture model for classification from the "mclust" package showed to be the overall best algorithm for this classification problem. Train accuracies varied from 48 % up to 60.7% for different sets of variables, being the set of variables presenting the highest train accuracy the combinations: A) the previous days (*TrendYest*), two(*Lag2d*) and three days(*Lag3d*) before value trend (60.7 % train accuracy); and B) the moving average from the value´s trend up to 3 days before (*TrendYest*, *Trend2B*, *Trend3dB*) (59.4 % train accuracy). In these cases, the test accuracy was 60.4% and 59%, respectively, meaning that this model could predict correctly the coin market trend by ca. 10% percentage points above simple random guessing. This way, a test error of 39.6% and 41%. The Bayes error using the set of variables B was calculated using the whole dataset. The Bayes error for this binary classification problem is 54 % assuming equal variances between classes. This gives us an idea of the classification error and because is indeed difficult to separate these two classes the Bayes error is high. So all the models tested will present quite high error rates.

Fig. 4 shows the ROC curves obtained for these two fittings with Set of variables A and B) using the mixture analysis algorithm. Set A) presents a higher area under curve thus being a better set of variables for predicting coin market daily trend.

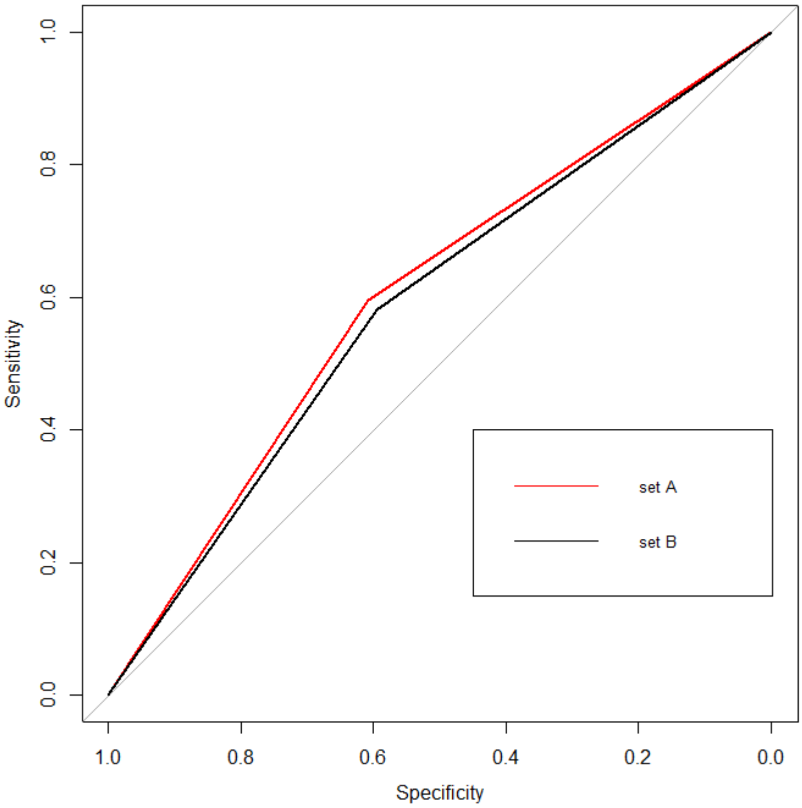

Fig. 4 - ROC curve for sets A and B of variables for the mixture analysis model. Set A) the previous days (*TrendYest*), two(*Lag2d*) and three days(*Lag3d*) before value trend; and set B) the moving average from the value´s trend up to 3 days before (*TrendYest*, *Trend2B*, *Trend3dB*).

Support Vector Machine

SVM was also considered however this classification approach is too time consuming due to the very big size of the dataset. One single simulation could be run for more than 3 hours and yet not be finished.

Conclusions

The present exercise analyzed the history of several crypto coins' market from 2013 to 2017 and address the prediction of the daily market trend in a binary classification approach considering different information from the previous nearest days. The methods used to classify the binary trend of the daily coin value were: logistic linear regression, Support Vector Machines, Fischer discriminant analysis, Linear (LDA) and Quadratic (QDA) discriminant analysis, Mixture models with EM, k-near neighbors (knn) and Naïve Bayes. The best method showed to be the Gaussian finite mixture model for classification fitted via EM algorithm. A test accuracy of 60% could be obtained, only 10% percentage points above random guessing. Indeed, calculation of the Bayes error shows that a quite high error can be present when considering this type of classification.