Topic distribution modeling for categorization of market niches

Ivan Shadrin[1], Jihyun Kim[2] and Aleksei Pupyshev[3]

[1] Operations & Project Manager at Ventuary, gasp-gasp@mail.ru
[2] Data Scientist at Ventuary
[3] Founder, Research & Strategy at Ventuary

Abstract. Ventuary is building the analytical platform for ICO-organizations and their investors. Identifying ICO projects categories and finding a comprehensive way of scoring them using Python, Scikit-learn and NLTK was one of the first steps to absolute transparency and fairness in terms of data driven investment decisions. We used TF-IDF, Decomposition an Linear regression to identify ICO projects categories and to determine their score using geometric mean of ROI and ROI score. Finally, we were able to procedurally define ICO market and market performance using Machine learning and Natural language processing.

Keywords: Blockchain, Initial Coin Offering, Machine Learning.

1 Introduction

Ventuary is building the analytical platform for ICO-organizations and their investors. Our goal is absolute transparency and fairness in terms of data driven investment decisions. We aim to make it possible for every investor to make informed and educated decisions, at the right time providing access to quality and transparent data.
One of the first steps to our goal was identifying ICO projects categories and finding a comprehensive way of scoring them using Python, Scikit-learn and NLTK.

2 Methods

TF-IDF (term frequency–inverse document frequency) was a scoring method to calculate the frequency of a particular term in the target ICO project description with the consideration of the scarcity of that particular term in other projects description. We defined the TF-IDF as:

$$tf - idf_{d,t} = tf_t * idf_{d,t} = tf_t * \log \frac{n}{df_{d,t}}$$

(where n is the total number of descriptions and $df_{d,t}$ is the description frequency; the description frequency is the number of documents d that contain term t)

Then we used Decomposition to get 32 categories and their top terms from TF-IDF-weighted document-term matrix. Next step was finding competitors for each project. We implemented this computing cosine distance between projects. We defined cosine distance as:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

We also calculated ROI and ROI score for ended projects and then predicted ROI score for active and ongoing projects using Linear regression applied to calculated data and TF-IDF-weighted document-term matrix. As we predicted 3 values: ETH(ROI score), BTC(ROI score) and USD(ROI Score) we used geometric mean to calculate overall actual topic-based score and named it Ventuary rating.

3 Results

A total of 1712 ICO projects were added to our site. We defined top categories as: Payment, Gaming, Big Data, Content & Media, Energy, Real Estate, Trading, Betting, Industrial, Crowdfunding, Healthcare, Advertising, Startup Financing, Cryptocurrency, Cryptocurrency, VR, Security & Privacy, Mobile, Social Media, AI, Cryptocurrency, Travel, Banking, Cryptocurrency, Commerce, Betting, Ticketing, Consumer, Identity Verfication, Entertainment, Software, Banking.
Based on 1430 ended projects we calculated Ventuary rating for 282 ongoing and upcoming projects.

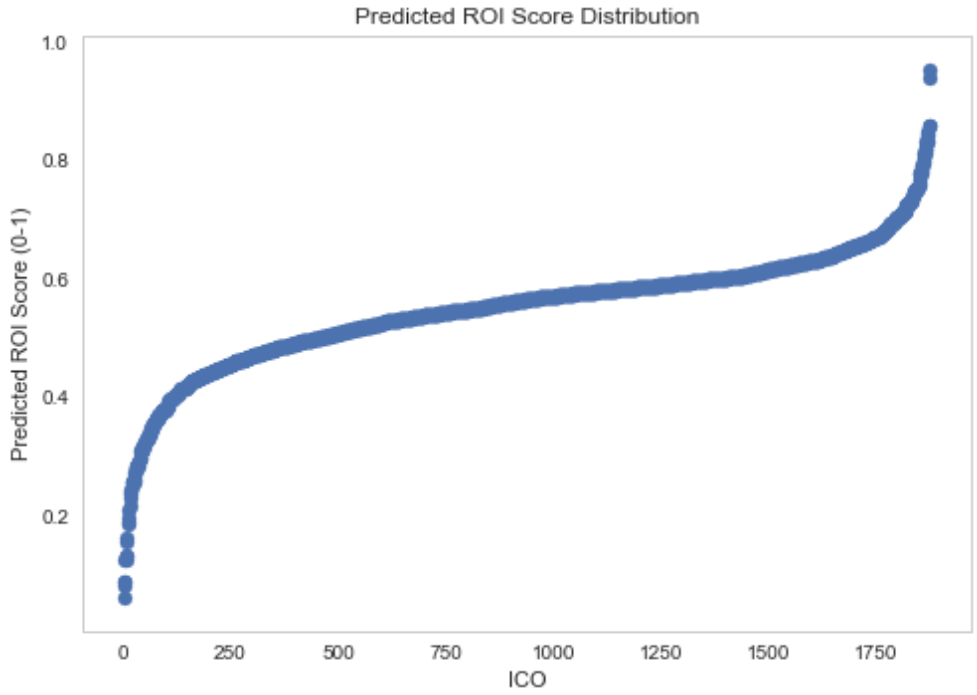

Fig. 1. Predicted ROI Score Distribution

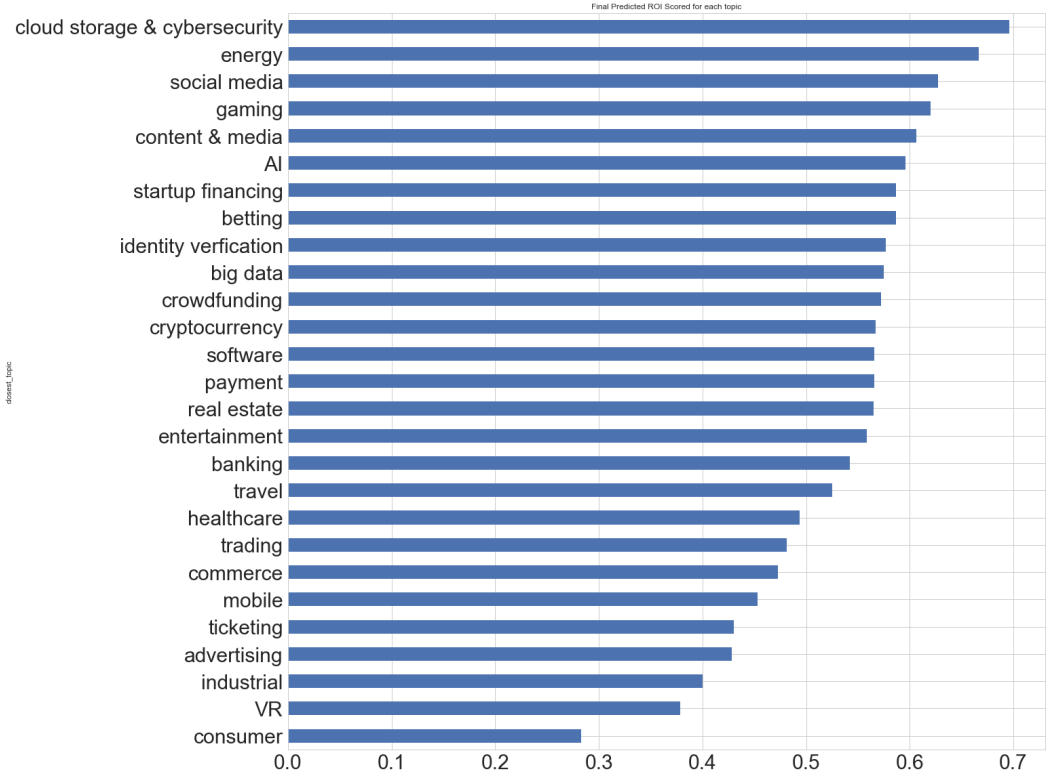

Fig. 2. Final Predicted ROI Scored for each topic

Finally, we were able to procedurally define ICO market and market performance using Machine learning and Natural language processing.

References

1. ICObench Homepage: https://icobench.com/, last accessed 2017/03/11.
2. ICOrating Homepage, https://icorating.com/, last accessed 2017/03/11.
3. Coinmarketcap Homepage, https://coinmarketcap.com/, last accessed 2017/03/11.
4. Scikit-Learn Homepage, http://scikit-learn.org/ last accessed 2017/03/11.