Machine learning approach to startup success prediction

Pavlov D.
MIPT
dima-pavlov@phystech.edu

Moiseev A.
MIPT

March 9, 2022

Abstract

This paper deals with a model for predicting the success of startups based on data about companies from CrunchBase and data about founders and investors associated with the company from LinkedIn. Unlike CrunchBase, where we have data with a clear structure, in a social network, data is filled in in the free form. In the last decade, natural language processing has shown good results when working with unstructured data. The approach for model construction using data about founders is new for the venture industry.

Keywords *Venture Capital · Unsupervised learning · Natural language processing*

I. Introduction

The number of startups born every day is growing from year to year. The growth leads to an increase in the load on the pipeline of a venture capital company. These companies are investing in the development of scoring models of startups to reduce the load on their pipelines. The scoring models summarize the experience of a venture analyst and automate routines such as reviewing reports and information about received investment rounds. However, in this industry, not only technical indicators are important, but also the network of people working on projects. The network consists of information that gets into social networks and news. At the moment, only human can process this information qualitatively. This is the difficulty that models for predicting the success of startups face.

The goal of the authors was to build a model for predicting the success of a startup using CrunchBase and LinkedIn data. The model uses unstructured data clustered according to their meaning.

II. Problem statement

1

III. Computational experiment

The task is to build a model for prediction the success of a startup. This task is decomposed into several subtasks:

1. Collect the data from CrunchBase and LinkedIn
2. Preprocess data
3. Create a model

The key issue in this workfow is data preprocesing. As mentioned earlier, the novelty of the model in the use of unstructured data from social networks.

<ToDo sample of data and show the problem>

Therefore, in the article we will focus on building an informational description of people's backgrounds from unstructured data. Our goal is to group people according to each of the available features:

- Education field of study
- Profession skills
- Academic degree
- Work experience

i. Data collection

Data from CrunchBase has this structure: ...<todo scheme describe>

Data from LinkedIn has Json structure: ...<todo scheme describe>

ii. Data preprocessing

Data from LinkedIn is processed using clustering and embeddings obtained using a neural network based on BERT: LaBSE (todo link to huggingface). The clustering is carried out using the kMeans algorithm.

iii. Model building

IV. Results

i. Preprocess data

<Show Word2Vec property of embeddings>

We can see that it turns out that the objects fall into clusters, but sometimes the clusters overlap. This is because TSNE transforms 754 dimensional space into 2 dimensional space.

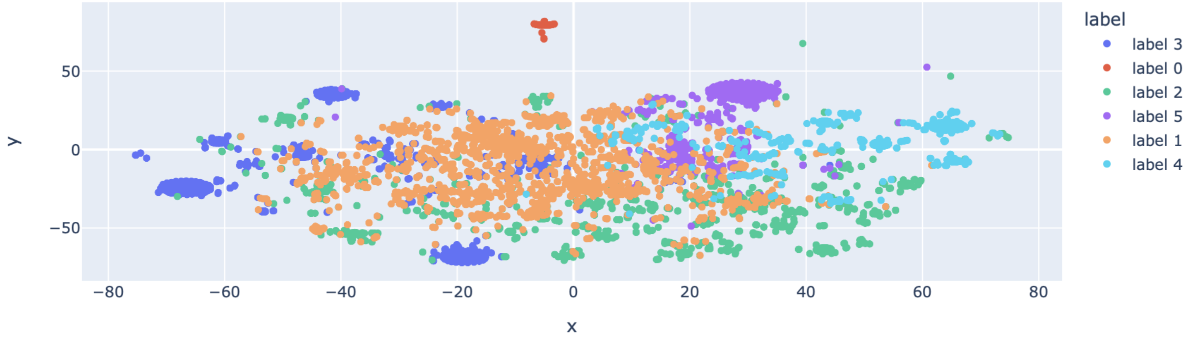

Figure 1: *Education field of study clustering*

The clustering coefficient – Silhouette metric is 0.26, which indicates that clusters are more likely to be allocated than not.